MODELING HOSPITAL LENGTH OF STAY AND COST WITH HETEROGENEITY

By

Xiaoqin Tang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Statistics

2010

ABSTRACT

MODELING HOSPITAL LENGTH OF STAY AND COST WITH HETEROGENEITY

By

Xiaoqin Tang

Hospital length of stay (LOS) is an important measure of healthcare utilization and is generally positively skewed and heterogeneous. We fit a Coxian phase-type distribution to LOS and identify the hidden states of the underlying latent homogeneous Markov model. We demonstrate that selecting an appropriate number of phases and a regression model for hazard rates can account for some heterogeneity in LOS. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) method enables us to dynamically uncover the hidden stochastic Markov structure. A classification method is used to assign patients to different latent LOS groups according to their mean LOS in hospitals.

Increasing availability of patient LOS and cost data permits joint analysis accounting for their possible correlations. A bivariate Coxian phase-type/log-normal (CPH-LN) model is proposed to assess the impacts of covariates simultaneously. Under marginal specification through parametric models for LOS and cost, shared random effects are introduced in the model as regressors and the model is easily estimated using SAS Proc NLMIXED.

We also propose an innovative method for the consideration of two-level correlations between LOS and cost. In our model, we are concerned with intra-hospital correlations, cross-equation correlations at both the hospital level and patient level. Full maximum likelihood (FML) is used to derive parameter estimates. A simulation study is conducted to illustrate our method.

The methodologies are illustrated with application to hospital admissions for acute my-

ocardial infarction (AMI) in the 2003 Nationwide Inpatient Sample (NIS) from the Healthcare Utilization Project (HCUP).

This thesis is dedicated to my parents for all their love and support!

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Joseph C. Gardiner for his valuable advice, patience, support and continuing encouragement during my graduate studies in the Department of Epidemiology, Michigan State University. He allowed me to investigate different areas of research. Working under his guidance on various research projects strengthened by training as a biostatistician. I also want to thank Dr. Zhehui Luo, for her devoted help on my research and her encouragement on my life. Appreciation is extended to other members in my committee, Dr. R.V. Ramamoorthi, Dr. Lijian Yang, and Dr. Connie Page, for their friendly support. My dissertation research was supported by the Agency for Healthcare Research Quality under grant 1R 01 HS 14206.

I give my special thanks to Dr. James Stapleton in the Department of Statistics and Probability. He is always ready to help students like a father. During the fives years of my graduate study at Michigan State University, he has given a lot of help and encouragement. I also thank all the other professors and my friends for their care and help.

I thank my parents for all the endless love and support they give me. They have always been supportive of my efforts for carrying out my career goals. I dedicate this dissertation work to both of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background

Studies of healthcare utilization from national and state administrative databases have been stymied by a lack of powerful methodological approaches to circumvent the lack of specificity that is often common in these databases. New statistical and econometric techniques are needed to account for unobserved heterogeneity and uncover hidden structures in healthcare utilization data. Predictive models are valuable in identifying factors associated with utilization so that adequate resources are properly allocated. These models are also used for risk-adjustment in payment for healthcare services and identifying high-risk patients for disease management programs. An important measure of healthcare resource use is hospital length of stay (LOS), measuring the number of days from admission to discharge. LOS can be terminated by cure, transfer to another care facility or death. In the past decade, health care providers and hospital administrators are interested in LOS predictions for both economic and organizational reasons. In addition to these aspects of quality control, there is also patient interest in anticipated dates of discharge. The development of LOS models has been useful to economists and statisticians. Some standard parametric models, for example, log-normal, Weibull, log-logistic and Generalized Gamma models have been widely applied to survival data. Another key measure in healthcare utilization analysis is the cost per individual patient, which is defined to be the total resource use from admission to discharge. Some analysts treat the hospital cost independent of LOS in standard linear mixed models after log transformation, or by standard survival analysis techniques, for example, proportional hazards models or accelerated failure time models which specify different error distributions, such as the extreme-value distribution, normal distribution and logistic distribution, etc. However, the independence assumption is generally not valid since an important feature of LOS and cost is the correlation between them, possibly in hierarchical levels, i.e. both at the hospital and individual levels. Hence, although separate univariate models may suffice to assess the correlates of LOS and hospital cost, the advantage of a multivariate model over independent models is that the correlation between the outcomes can be estimated resulting in inferences that are more efficient (Polverejan et al., 2003; Gardiner et al., 2002). The shared random-effects models and the copula-based models are two different ways to account for this correlation. For example, the correlation between the healthcare outcomes of individuals in a cluster is often assessed using measures of dependence such as the Pearson's correlation coefficient, the Spearman's rho and the Kendall's tau. In addition, to account for correlation in the healthcare outcomes within a cluster or equivalently heterogeneity between clusters, a random effect term is used, resulting in what are known as frailty models in survival analysis. In this thesis, however, we propose a joint bivariate copula random-effects model in a health economics study that includes both hospital-level and individual-level correlation.

This thesis is concerned with two closely related measures of healthcare resource utilization: LOS and cost. We consider models for LOS and joint models for LOS and cost. Methodological advances have been made for analysis of both measures during the past decade. In this chapter we first give a basic introduction to relevant models. We also briefly review related statistical literature. Models are described in detail in the subsequent chapters.

1.1.1 Length of Stay (LOS)

Length of stay (LOS) is a term commonly used to measure the duration of a single episode of hospitalization. Inpatient days are calculated by subtracting day of admission from day of discharge. People entering and leaving a hospital on the same day have a LOS of one. Hospital LOS is an important measure of healthcare resource use because hospital spending represents approximately one third of the national health spending which is projected to reach 4.4 trillion by 2018 (Konetzka et al., 2008; Sisko et al., 2009). Various statistical models have been used to assess the influence of clinical and demographic factors on LOS and hospital charge. The primary goal is to estimate the expected value of LOS (Faddy et al., 2009; Li, 1999; Wang et al., 2002). Analyses of LOS present several challenges to researchers as LOS data are often skewed and therefore standard regression methods cannot be applied directly. In addition, several other features of LOS data that need to be addressed include incomplete observations and heteroscedasticity. Validity of statistical inference on LOS will depend upon the strength of the methodology in addressing these characteristics of hospital LOS.

1.1.2 Cost

Hospital cost constitutes a significant proportion of overall expenditure in health care and analysis of cost data is another interesting area which has attracted considerable attention recently. Investigating appropriate models or assessing health outcomes and analysis of these data can help us identify the most cost-effective treatment and ascertain the determinants of cost (Liu, 2004).

Patient cost is collected as a final total cost in some national and state databases. For a treatment or intervention under study, let $C(t)$ be the accumulating cost over time t for an individual patient. Expenditures terminate at the discharge time from the hospital. In other situations we record cost $C(t_j)$ in a regular time interval $(t_{j-1}, t_j]$. Similar to LOS, analysis of hospital cost may have several technical problems, such as skewness, incompleteness, etc. Sometimes other components of cost such as inpatient or outpatient cost exhibit a two-part feature as a result of a significant proportion of zero cost. In addition, hospital cost is often correlated with LOS. With escalating cost, knowing the correlates (also called covariates) of LOS and in-hospital cost is important for decisions on allocating resources. Naive analysis treats the hospital cost independent of LOS by standard linear mixed models after log transformation (Sirbu, 2004), or by standard survival analysis techniques (such as Kaplan-Meier estimators) (Etzioni et al., 1999; Lin et al., 1997; Lin, 2000; Hallstrom and Sullivan, 1998). However the estimates might be invalid due to the correlation between LOS and cost we mentioned before. Therefore there is an interest in developing a joint modeling approach to analyze both LOS and hospital cost simultaneously.

1.1.3 Accelerated Failure Time (AFT) Model

The accelerated failure time (AFT) model is presented as an alternative to the proportional hazards model, which is widely used in medical research and time to event data analysis. The AFT models also have been studied extensively in the literature for analyzing right censored data (Buckley and James, 1979; Tsiatis, 1990; Wei et al., 1990; Jin et al., 2003). Recently, Tian and Cai (2006) studied a numerically efficient simple estimation procedure to the regression analysis of interval censored data using the AFT model.

Consider the following model,

$$\log T = \mathbf{x}'\boldsymbol{\beta} + \sigma\epsilon \tag{1.1}$$

AFT models have a number of advantages, in particular, they offer a wider variety of shapes of hazard functions than the parametric proportional hazards models that assume a particular distribution for survival times, since the family includes distributions with unimodal hazard functions, such as the log-normal, log-logistic distributions. Moreover, the log-linear formulation of such models emphasizes that the roles of the regression parameters and dispersion parameters are clearly separated (Keiding et al., 1997). The regression parameters in an AFT model are also robust with respect to neglected random effects (Hougaard, 1999), less affected by the choice of probability distribution, which is not the case for proportional hazards models (Hougaard et al., 1994). In addition, regression parameters in the proportional hazards model are more sensitive to the distribution of the random component.

The survivor function can be written as $S(t|\mathbf{x}) = S_0\left((t\exp(-\mathbf{x}'\boldsymbol{\beta}))^{\frac{1}{\sigma}}\right)$, where S_0 is the survivor function of the random variable $\exp(\epsilon)$. Clearly, the individual characteristics act on the duration distribution by transforming the time scale T to $T\exp(-\mathbf{x}'\boldsymbol{\beta})$. This may be an accurate description of the actual variation in the lifetime distribution of complex self-evolving organisms or mechanisms. Because of the one-to-one relation between a distribution and its hazard function, the AFT specification can be translated into a specification of the hazard function of T given $\mathbf{x}$. Assuming a normal or logistic distribution for ϵ provides a closed form expression for $S(t|\mathbf{x})$, which will be used in the later chapters when modeling cost.

1.1.4 Phase-type (PH) Model

Phase-type (PH) distributions are defined as distributions of absorption times T in Markov processes with $m < \infty$ transient states (the phases) and one absorbing state labeled $m + 1$. Since their introduction by Neuts in 1981 (Neuts, 1981), PH distributions have attracted a lot of attention in the past decades and have been widely used in queuing theory, reliability analysis, insurance risk, survival analysis, telecommunications, and healthcare utilization, such as hospital LOS. Aalen (2002) connects PH models to problems in survival analysis, for example the incubation time of acquired immune deficiency syndrome (AIDS) shown in Figure 1.1. Asmussen et al. (1996) propose EM algorithm to estimate PH distributions and exhibit four samples of the lengths of incoming telephone calls to the service center of one of Israels major television cable companies. Olsson (1996) extends EM algorithm to cover estimation from censored data including right-censored and interval-censored observations. Bitran and Dasu (1994) analyze a queue to which the arrival process is the superposition of separate arrival streams, each of whose interarrival time distributions is of phase type, and the service time distribution is also of phase type.

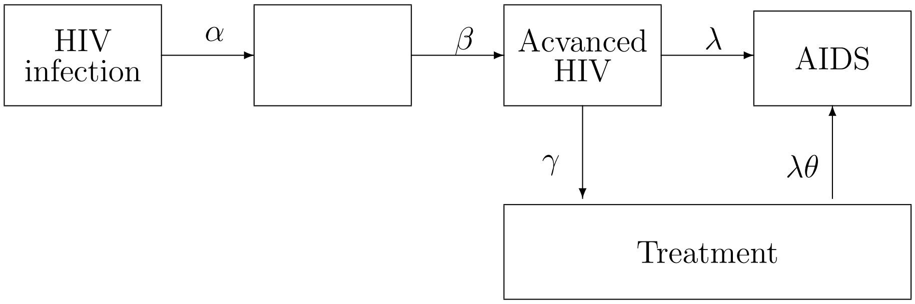


Figure 1.1: Phase-type model for incubation distribution of AIDS

PH distributions arise from a generalization of Erlang's method of stages in a form that is particularly well suited for numerical computation (Erlang, 1917). The simplest examples are mixtures and convolutions of exponential distributions (in particular Erlang distributions, defined as gamma distributions with an integer shape parameter). More generally, the class comprises all series/parallel arrangements of exponential distributions, possibly with feedback. PH distributions are a versatile class of distributions, which is dense in the class of distributions defined on the non-negative real line. Hence any distribution on $[0, \infty)$ can, at least, in principle be approximated by a PH distribution. Moreover, their formulation also allows the Markov structure of stochastic models to be retained when they replace the simple exponential distributions. Details will be provided in Chapter 2.

1.1.5 Random-Effects (RE) Model

The random-effects (RE) model arises in the context of analysis of survival data, event counts, jointly dependent continuous and discrete variables and so forth. Random effects have been used to model dependence, but they can also be viewed as a general approach of joint modeling. The RE modeling approach is built using regression models in which either common or correlated latent variables enter the models in the same manner as regressors. Such RE models have long history in statistics, for example, in the context of bivariate distributions they have appeared under a variety of names such as the shared frailty model in survival analysis (Hougaard, 2000), trivariate reduction model in multivariate count data analysis (Kocherlakota and Kocherlakota, 1992), and latent variable models for multilevel, longitudinal and structural equation analysis (Skrondal and Rabe-Hesketh, 2004). Generically they are all mixture models and can also be interpreted as random-effects models. Liu (2009) proposes a joint random effects model of longitudinal medical cost data and survival, taking into account the semi-continuous nature of medical costs. Liu and Huang (2009) propose a joint random-effects model for a repeated measures process and a recurrent events process, for which the correlation is modeled by shared random effects. RE models have been also established as an appealing approach to analyzing longitudinal and survival data (Vonesh et al., 2006; Ratcliffe et al., 2004; Wulfsohn and Tsiatis, 1997; Gruttola and Tu, 1994; Tsiatis et al., 1995; Henderson et al., 2000; Tsiatis and Davidian, 2004; Xu and Zeger, 2001; Guo and Carlin, 2004).

Let us consider a shared random-effects model for two possibly correlated outcomes Y_1 and Y_2 with joint density f:

$$f(y_1, y_2|x_1, x_2) = \int_0^\infty f_1(y_1|x_1, \nu) f_2(y_2|x_2, \nu) g(\nu) d\nu \tag{1.2}$$

where f_1, f_2, and g are univariate densities. For example, let $f_1(y_1|x_1, \nu)$ and $f_2(y_2|x_2, \nu)$ denote normal marginal distributions for continuous variables Y_1 and Y_2, with conditional mean $\mu_1 = x_1'\beta_1 + \nu$ and $\mu_2 = x_2'\beta_2 + \gamma\nu$, where γ is a scale parameter as Y_1 and Y_2 might be in different scales. This approach suggests a way of specifying correlated models based on a suitable choice of $g(.)$.

In fact, we can consider more flexible bivariate or multivariate parametric models by introducing correlated, rather than identical, unobserved heterogeneity components in marginal models. For example, suppose Y_1 and Y_2 are, respectively, $Normal(\mu_1|\nu_1)$ and $Normal(\mu_2|\nu_2)$, where ν_1 and ν_2 represent correlated unobserved heterogeneity. Dependence between Y_1 and Y_2 is induced if ν_1 and ν_2 are correlated. We refer to (ν_1, ν_2) as latent factors. For example, we can assume (ν_1, ν_2) to have bivariate normal distribution with correlation ρ. The joint density is of the form:

$$f(y_1, y_2|x_1, x_2) = \int\int f_1(y_1|x_1, \nu_1) f_2(y_2|x_2, \nu_2) g(\nu_1, \nu_2) d\nu_1 d\nu_2 \tag{1.3}$$

Similar to other generalized linear mixed models (GLMM's), a full likelihood analysis of the above joint model is hindered by the high-dimensional integration. To avoid these computational problems, several approaches have been proposed, for example, Gaussian quadrature techniques can be used as a practical estimation tool and computations carried out in available software such as freely available softwares aML (http://www.applied-ml.com/index.html) or SAS Proc NLMIXED.

1.1.6 Copula-based Regression Model

Copulas are another approach for modeling dependence or correlation in the context of linear or nonlinear regression models. Copulas, originally introduced by Sklar in 1959 (Sklar, 1959), have been suggested as a useful method for generating joint distributions from the given marginals. For example, a copula approach can generate a Gaussian distribution when the Gaussian copula is applied to the Gaussian marginal distributions. More importantly, a copula model can also generate many complicated non-Gaussian joint distributions. This approach is fruitful when the marginals can be specified with confidence, but the joint distribution is awkward to establish (Cameron et al., 2004). In the recent years, the copula models become popular for modeling dependencies between random variables, especially in such fields as biostatistics, finance and actuarial science (Genest and Rivest, 1993; Joe, 1997; Nelsen, 1999; Capéraà et al., 2000). Chen and Fan (2002) study the temporal dependence properties and the estimation of a class of semiparametric stationary Markov time series models, propose simple estimators of the unknown marginal distribution and the copula dependence parameter, and establish their large sample properties under easily verifiable conditions. Miller and Liu (2002) propose a minimum cross-entropy approach that recovers continuous joint distributions from the joint and marginal moments and the marginal densities. Smith (2003) models sample selection using Archimedean copulas with the specification of binary models that are designed to account for data selectivity.

A copula is a function that connects the marginal distributions to restore the joint distribution. In the bivariate case, a joint distribution $H(x, y)$ can be expressed in terms of its margins $F_X(x)$ and $F_Y(y)$ and a copula function $C(\cdot, \cdot)$ such that $C(F_X(x), F_Y(y)) = H(x, y)$. In this approach, C models the dependence structure. Copulas separate marginal distributions from the dependence structure, and the appropriate copula for a particular application is the one which best captures dependence features of the data. Details will be shown in Chapter 4.

1.2 Outline of the Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we will systematically address the principles of PH and Coxian PH regression models. We will also describe the estimation methods, including a context within the Bayesian framework. Later, we describe the classification and calculation of partial effects based on the Bayesian posterior samples. The methodology is illustrated with application to hospital admissions for acute myocardial infarction (AMI) in the 2003 Nationwide Inpatient Sample (NIS) from the Healthcare Utilization Project (HCUP).

In Chapter 3, we will describe a statistical approach based on shared random effects for joint modeling LOS and hospital cost. To begin with, conditional on a common latent factor ν, suppose hospital LOS (denote by T) has a Coxian PH distribution and cost (denote by $C(T)$) is log-normally distributed, they are independent. From this model, we derive the marginal means, variances, covariances and correlations. The shared random-effects methodology is applied to the same AMI data set described in Chapter 2, Section 2.9.

In Chapter 4, we address the problem in the correlation of LOS and cost with the bivariate copula random-effects model. We model the hospital-clustered unobserved heterogeneity through correlated random effects and individual correlation through the copula-based correlated measurement errors simultaneously. The full maximum likelihood (FML) is used for the estimation. We carry out a simulation study to assess model fit and performance of estimates. Similarly, the joint modeling approach is illustrated by the AMI data set again.

Finally, Chapter 5 summarizes our work and gives some suggestions for future research.

Chapter 2

Modeling Hospital Length of Stay with a Coxian PH Regression Model

In this chapter, we will give an overview of the phase-type (PH) distributions. To better understand this chapter, we need to know the definition of the PH distributions and its special subclasses, especially Coxian PH distributions, which we will use in our application with real data. We also provide some basic properties of the PH distributions in Section 2.4.

2.1 Introduction

Phase-type (PH) distributions, introduced by Neuts in 1981 (see Neuts, 1981), are defined as the distributions of the time to reach an absorbing state in a finite-state continuous-time homogeneous Markov process. PH distributions have been received a lot of attention in a wide range of stochastic modeling applications, such as telecommunication, teletraffic modeling, queuing theory, reliability theory, insurance risk, survival analysis and healthcare utilization, like hospital length of stay (LOS). PH distributions have enjoyed such popularity because they constitute a very versatile class of distributions defined on the non-negative real numbers that lead to algorithmically tractable models. Since the set of PH distributions is dense in the set of non-negative distributions with support on $[0, \infty)$, any non-negative distribution, at least in principle, can be approximated arbitrarily closely by a PH distribution. In addition, their formulation also allows the Markov structure of stochastic models to be retained when they replace the simple exponential distributions. A Coxian PH distribution is a special form of PH distributions with the transient states ordered so that transitions are forward flowing with exit from any transient state. Coxian PH distributions have considerably fewer parameters than general PH distributions (Cumani, 1982). In the past two decades, Aalen (2002), Marshall and his colleagues (Marshall et al., 2000, 2002; Marshall and McClean, 2003; Marshall et al., 2005) have applied PH models to analyze LOS data. Marshall et al. (2007) estimate total inpatient cost for a cohort of patients based on a Coxian PH distribution for their hospital LOS. Fackrell (2009) gives a general survey of PH models with applications to healthcare data. More recently, Ausín and colleagues (Ausín and Lopes, 2007; Ausín et al., 2008, 2009) proposed Bayesian methods for estimating a Coxian PH distribution, treated as "finite mixtures". McGrory et al. (2009) use a fully Bayesian approach for inference in Coxian PH models with covariate-dependent mean duration.

McGrory et al. (2009) incorporate covariates into Coxian PH distributions through a log-linear mean function with the intent of estimating covariate effects on the mean LOS. Ausín and collegues (Ausín and Lopes, 2007; Ausín et al., 2008, 2009) mention the importance of ordering the incidence rate parameters in identifying parameter estimates, which has consequences for classification. Their focus is on the estimation of ruin probabilities in certain risk reserve processes in insurance claims and on the busy period in queuing systems, but they do not consider the influence of covariates. Bayesian methods are used for estimation and inference.

In this chapter, we apply a similar Bayesian method to estimate Coxian PH regression models that combine identification of transition probabilities and coefficients of covariates. We describe a classification method to assign patients to different latent groups, and explain differences in these groups by covariates.

2.2 Continuous Phase-type Distribution

Continuous phase-type (PH) distributions describe the time to absorption T of an underlying finite-state continuous time Markov process $\{X(t); t \geq 0\}$. We denote by $E = \{1, 2, \ldots, m, m+1\}$ the state space where the single absorbing state is labeled '$m+1$' and the remaining states $1, 2, \ldots, m$ are transient. $X(t)$ is governed by the $(m+1) \times (m+1)$ transition intensity matrix, $\boldsymbol{A} = \{\alpha_{hj}\}, h, j \in E$, with the elements

$$\begin{aligned} \alpha_{hj} &= \lim_{\Delta t \downarrow 0} \frac{P\{X(t+\Delta t) = j | X(t) = h\}}{\Delta t} \\ \alpha_{hh} &= -\sum_{j \neq h} \alpha_{hj}, h, j \in E \end{aligned} \tag{2.1}$$

The probability density function, survival function and k-th non-central moment of T

have closed forms

$$f(t) = \boldsymbol{\pi} \exp(\mathbf{Q}t)(-\mathbf{Q}\mathbf{1}) \tag{2.2}$$

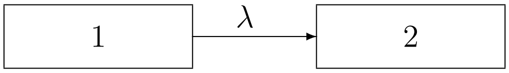
$$S(t) = \boldsymbol{\pi} \exp(\mathbf{Q}t)\mathbf{1} \tag{2.3}$$

$$m_k = (-1)^k k! \mathbf{Q}^{-k}\mathbf{1}, k = 1, 2, \ldots \tag{2.4}$$

Here, $\exp(\mathbf{A})$ denotes the matrix exponential of the square matrix $\mathbf{A}$ (Golub and Van Loan, 1996). The initial probability distribution over the transient states is $\boldsymbol{\pi}$ (row vector), $\mathbf{Q}$ is the $m \times m$ intensity matrix for the Markov chain, restricted to the transient states, and $\mathbf{1}$ is a column m-vector of 1's. Equation (2.2) is a representation of a PH distribution given $(\boldsymbol{\pi}, \mathbf{Q})$. The class of all PH distributions with generator $\mathbf{Q}$ is denoted by PH($\mathbf{Q}$), including the representation $(0, \mathbf{Q})$, which is the distribution with point mass at zero (or a process that is instantly absorbed in $m+1$). The class PH($\mathbf{Q}$) can be described as the convex hull of the points $(0, \mathbf{Q})$ and $(\boldsymbol{e}_k, \mathbf{Q})$, $k = 1, 2, \ldots, m$, where $\boldsymbol{e}_k$ is a $m \times 1$ vector with k-th component equals to 1, and all other components equal to 0.

We now give some special examples of PH distributions, assuming that there is no probability mass at zero.

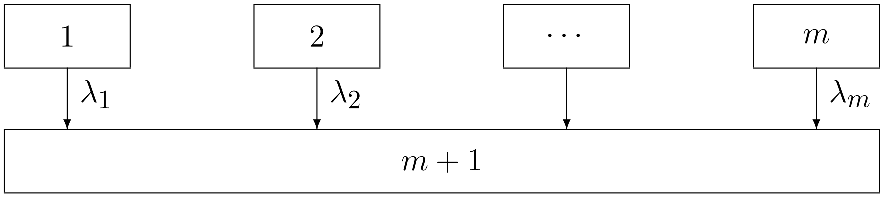
(1) Exponential distributions



The simplest non-trivial example of a PH distribution is the exponential distribution with the parameter λ , which has density function $f(x) = \lambda \exp(-\lambda x)$ with a representation

$$\boldsymbol{\pi} = 1, \mathbf{Q} = (\lambda)$$

(2) Hyper-Exponential (Mixture of exponential) distributions



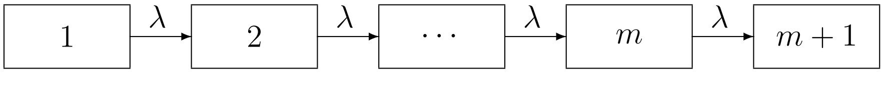
The hyper-exponential distribution with density function $f(x) = \sum_{i=1}^{m} \pi_i \lambda_i \exp(\lambda_i x)$ can be represented as a PH distribution with

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_m)$$

$$\mathbf{Q} = \begin{pmatrix} -\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_m \end{pmatrix}$$

where $\sum_{i=1}^{m} \pi_i = 1$.

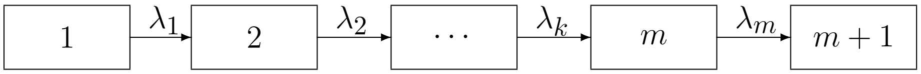
(3) Erlang distributions



The m-Erlang distribution has two parameters, the shape an integer $m > 0$ and the rate $\lambda > 0$. This is sometimes denoted $E(m, \lambda)$. The Erlang distribution with the density function $f(x) = \dfrac{\lambda^m x^{m-1}\exp(-\lambda x)}{(m-1)!}$ has the PH representation

$$\boldsymbol{\pi} = (1, 0, \cdots, 0)$$
$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\lambda & \lambda \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\lambda \end{pmatrix}_{m\times m}$$

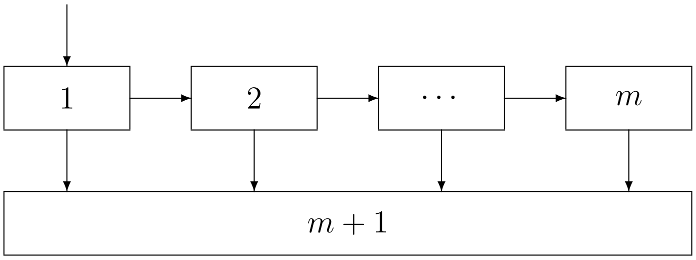
(4) Hypo-exponential distributions



The hypo-exponential distribution is a generalization of the Erlang distribution by having different rates for each transition (the non-homogeneous case), i.e.

$$\boldsymbol{\pi} = (1, 0, \cdots, 0)$$
$$\mathbf{Q} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_3 & \lambda_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\lambda_{m-1} & \lambda_{m-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\lambda_m \end{pmatrix}_{m\times m}$$

(5) Coxian PH distributions



The Coxian distribution is a generalization of the hypo-exponential distribution. The absorbing state can be reached from any phase. The PH representation is given by,

$$\boldsymbol{\pi} = (1, 0, \cdots, 0)$$

$$\mathbf{Q} = \begin{pmatrix} -\lambda_1 & p_{12}\lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & p_{23}\lambda_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_3 & p_{34}\lambda_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\lambda_{m-1} & p_{m-1,m}\lambda_{m-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\lambda_m \end{pmatrix}_{m\times m}$$

where $0 < p_{k,k+1} \leq 1, k = 1, 2, \cdots, m-1$. In the case when all $p_{k,k+1}$'s=1, it reduces to hypo-exponential distribution. The Coxian PH distribution is extremely important as any acyclic PH distribution has a generator that is upper triangular and has an equivalent Coxian representation. In addition, the process **generalised Coxian PH distribution** relaxes the condition that requires starting in the first phase.

(6) Mixture of Erlang distributons

The mixture of two Erlang distributions with parameters $E(m_1, \lambda_1)$, $E(m_2, \lambda_2)$ and

(ω_1, ω_2), such that $\omega_1 + \omega_2 = 1, \omega_i \geq 0$ for each i, has the PH representation

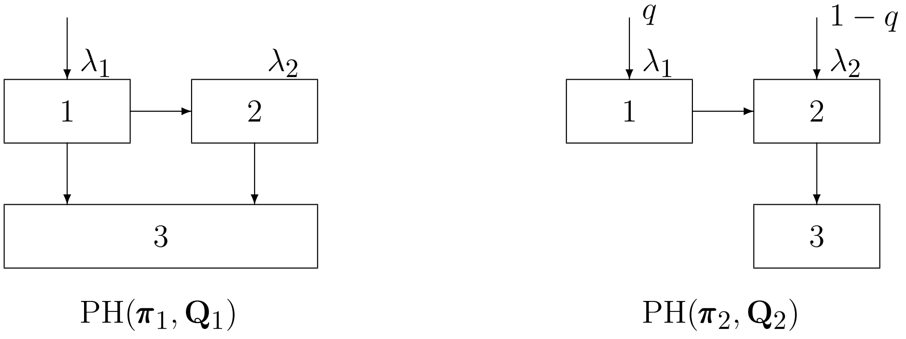
$$\boldsymbol{\pi} = (1, 0, \cdots, 0)$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}_{m \times m}$$

$$\text{where } \mathbf{Q}_k = \begin{pmatrix} -\lambda_k & \lambda_k & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_k & \lambda_k & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_k & \lambda_k & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\lambda_k & \lambda_k \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\lambda_k \end{pmatrix}_{m_k \times m_k}$$

As noted earlier the class of PH distributions is dense in the set of distributions on $[0, \infty)$. Unfortunately, this distribution family is over-parameterized and thus not identifiable. Every PH distribution has several alternative representations $(\boldsymbol{\pi}, \mathbf{Q})$. This makes parameter estimation difficult (Fackrell, 2009). For a simple example where one PH distribution can be represented by two structures, see Figure 2.1. Consider two structures with $m = 2$, PH$(\boldsymbol{\pi}_1, \mathbf{Q}_1)$ and PH$(\boldsymbol{\pi}_2, \mathbf{Q}_2)$, where $\boldsymbol{\pi}_1 = (1, 0)$, $\boldsymbol{\pi}_2 = (q, 1-q)$, $\mathbf{Q}_1 = \begin{pmatrix} -\lambda_1 & q\lambda_1 \\ 0 & -\lambda_2 \end{pmatrix}$, and $\mathbf{Q}_2 = \begin{pmatrix} -\lambda_2 & \lambda_2 \\ 0 & -\lambda_1 \end{pmatrix}$, $\lambda_1 \neq \lambda_2$. From (2.2), we get

$$f_1(t) = f_2(t) = (1-q)\lambda_1 \exp(-\lambda_1) + \frac{q\lambda_1\lambda_2}{\lambda_1 - \lambda_2}[\exp(-\lambda_2 t) - \exp(-\lambda_1 t)]$$



PH($\boldsymbol{\pi}_1, \mathbf{Q}_1$) PH($\boldsymbol{\pi}_2, \mathbf{Q}_2$)

Figure 2.1: Two different representations for one PH distribution

It is also apparent from the examples that representations for PH distributions do not necessarily have the same order, meaning that the number of transient phases need not be the same. In fact, there must be a representation that has the minimal order. A representation that has minimal order is called a minimal representation.

Imposing a special structure on $\mathbf{Q}$ can guarantee that every distribution PH($\mathbf{Q}$) has a unique representation in the form $(\boldsymbol{\pi}, \mathbf{Q})$. Although one could work with such a minimal representation, in this thesis we will use Coxian PH distributions because they present fewer problems in estimation and inference and also provide a simple interpretation of fit for LOS data.

2.3 Coxian Phase-type Models

A Coxian PH distribution results when the transient states have a natural order and only forward transitions between them may occur, beginning in state 1: $1 \to 2, 2 \to 3, \ldots, m{-}1 \to m$ and exiting from any transient state: $1 \to m+1, 2 \to m+1, \ldots, m \to m+1$. Transitions between the $m+1$ states is illustrated in Figure 2.2. The actual states of the Markov model are not observable, that is, we assume every patient enters the system from state 1, but do not know the state from which the patients exit.

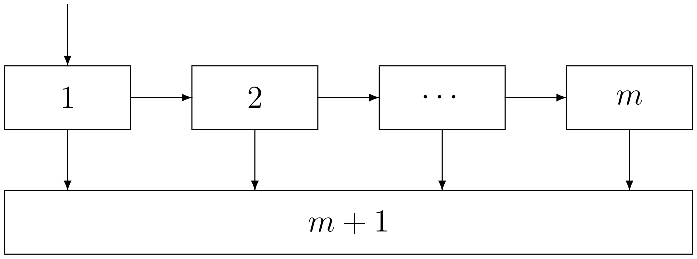


Figure 2.2: Representation of $(m+1)$-state Markov process with a Coxian PH distribution

A Coxian PH distribution is represented by $(\boldsymbol{\pi}, \mathbf{Q})$ where $\boldsymbol{\pi} = (1, 0, \ldots, 0)$, and

$$\mathbf{Q} = \begin{pmatrix} -(\alpha_{12}+\alpha_{1,m+1}) & \alpha_{12} & 0 & \cdots & 0 \\ 0 & -(\alpha_{23}+\alpha_{2,m+1}) & \alpha_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \alpha_{m,m+1} \end{pmatrix} \tag{2.5}$$

The number of parameters in a general PH distribution is $m^2 + m$, but the number of parameters in a Coxian PH distribution is reduced to $2m - 1$. The Coxian family is also dense in the class of all distributions on $[0, \infty)$ and is appropriate for estimating long-tailed distributions.

In the context of survival analysis we are interested in the hazard rate for sojourn time in each state and transition probabilities between states, thus it is constructive to reformulate this intensity rates $\boldsymbol{\alpha}$'s to hazard rate in transient state k as $\lambda_k = \alpha_{k,k+1} + \alpha_{k,m+1}, k = 1, 2, \ldots, m-1$, and in transient state m as $\lambda_m = \alpha_{m,m+1}$. Transition probabilities are $p_{k,k+1} = \alpha_{k,k+1} \big/ (\alpha_{k,k+1} + \alpha_{k,m+1})$ from $k \to k+1$, $k = 1, 2, \ldots, m-1$ with $p_{m,m+1} = 1$.

Suppose we observe duration times of n patients $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)'$ from a presumed Coxian PH with m transient states and covariate matrix $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where $\mathbf{x}_i =$

$(x_{1i}, x_{2i}, \ldots, x_{li})'$. We can incorporate covariates in the classical way comparable to a generalized linear model (Faddy et al., 2009). Assuming $\lambda_{ki} = \lambda_k(\mathbf{x}_i) = \lambda_{0k} \exp(-\mathbf{x}_i'\boldsymbol{\beta})$, the conditional mean distribution is log-linear in $\mathbf{x}_i$, i.e, $\log(E(T_i|\mathbf{x}_i)) = a_0 + \mathbf{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_l)'$ is the coefficient vector, and a_0 is a function of $p_{k,k+1}$'s and λ_{0k}'s.

From (2.2) the corresponding likelihood function can be written as

$$L(\boldsymbol{t}|\mathbf{X}, \lambda_0, \boldsymbol{\beta}, \boldsymbol{p}) = \prod_{i=1}^{n} \boldsymbol{\pi}[\exp(\boldsymbol{\Lambda}_i \mathbf{P} t_i)](-\boldsymbol{\Lambda}_i \mathbf{P}\mathbf{1}) = \prod_{i=1}^{n} \boldsymbol{\pi} \exp(\tilde{\mathbf{Q}}_i t_i)\tilde{\boldsymbol{q}}_i \tag{2.6}$$

where $\boldsymbol{\Lambda}_i = \begin{pmatrix} -\lambda_{1i} & 0 & \cdots & 0 \\ 0 & -\lambda_{2i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\lambda_{mi} \end{pmatrix}$, $\mathbf{P} = \begin{pmatrix} 1 & -p_{12} & 0 & \cdots & 0 \\ 0 & 1 & -p_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$,

$\boldsymbol{\pi} = (1, 0, \ldots, 0)$, $\tilde{\mathbf{Q}}_i t_i = \boldsymbol{\Lambda}_i \mathbf{P}$ and $\tilde{\boldsymbol{q}}_i = -\boldsymbol{\Lambda}_i \mathbf{P}\mathbf{1} = (p_{1,m+1}\lambda_{1i}, p_{2,m+1}\lambda_{2i}, \ldots, p_{m,m+1}\lambda_{mi})'$. The parameters of the model are $\boldsymbol{\lambda}_0 = (\lambda_{01}, \lambda_{02}, \ldots, \lambda_{0m})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_l)'$, and $\boldsymbol{p} = (p_{12}, p_{23}, \ldots, p_{m-1,m})'$.

The likelihood function (2.6) becomes

$$\prod_{i=1}^{n} \left\{ \boldsymbol{\pi} \exp\left(\exp(-\mathbf{x}_i'\boldsymbol{\beta}) \begin{pmatrix} -\lambda_{01} & p_{12}\lambda_{01} & \cdots & 0 \\ 0 & -\lambda_{02} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\lambda_{0m} \end{pmatrix} t_i \right) \left(\exp(-\mathbf{x}_i'\boldsymbol{\beta}) \begin{pmatrix} (1-p_{12})\lambda_{01} \\ (1-p_{23})\lambda_{02} \\ \vdots \\ \lambda_{0m} \end{pmatrix} \right) \right\} \tag{2.7}$$

We refer to (2.7) as the likelihood of our Coxian PH regression model. Covariates enter through a specific parameterization of the mean $E(T|\mathbf{x}) = \exp(a_0 + \mathbf{x}'\boldsymbol{\beta})$. This maintains the general form of the mean common to accelerated failure time models, $\log T = \mathbf{x}'\boldsymbol{\beta} + \sigma\epsilon$ when σ is a scale parameter and ϵ has a specified distribution on the real line. For example, if ϵ is extreme-value, then T has the Weibull distribution with $E(T|\mathbf{x}) = \exp\left(\log\Gamma(1+\sigma) + \mathbf{x}'\boldsymbol{\beta}\right)$ where Γ is the Gamma function. The log-normal, log-logistic and Generalized Gamma failure time distributions also share this structure.

At the expense of a formidable increase in the number of parameters, a general Coxian PH regression model could be described by specifying $\lambda_k(\mathbf{x}_i) = \lambda_{0k}\exp(-\mathbf{x}_i'\boldsymbol{\beta}_k)$ with phase-type specific parameters $\boldsymbol{\beta}_k$. In addition, a logit model $\log\left(p_{k,k+1}/(1-p_{k,k+1})\right) = \mathbf{z}_i'\boldsymbol{\gamma}_k$ could be specified for transition probabilities. However, stability of ML estimates is often in doubt unless some structural simplifications can be made and exclusion restrictions can be imposed on the covariates $\mathbf{x}, \mathbf{z}$ in the two parts of the regression model. Exploration of various strategies for model formulation and estimation is the subject of on-going research.

2.4 Properties of Phase-type Distributions

In this section, we discuss some of the basic properties of the PH distributions. First, the set of PH distributions is quite broad and, in theory, any non-negative distribution can be approximated arbitrarily closely by a PH distribution.

Proposition 2.1. *The distribution of PH($\boldsymbol{\pi}, \mathbf{Q}$) is given by*

$$F(t) = 1 - \boldsymbol{\pi}\exp\left(\mathbf{Q}t\right)\mathbf{1}$$

for $t \geq 0$, where the matrix exponential is defined by $\exp(\mathbf{A}) = \sum_{i=0}^{\infty}\dfrac{1}{i!}\mathbf{A}^i$. The density *function of PH($\boldsymbol{\pi}, \mathbf{Q}$) is given by*

$$f(t) = \boldsymbol{\pi} \exp\left(\mathbf{Q}t\right)(-\mathbf{Q}\mathbf{1})$$

for $t \geq 0$. Let T be a random variable with the PH($\boldsymbol{\pi}, \mathbf{Q}$) distribution, then

$$E\left[T^k\right] = (-1)^k k! \mathbf{Q}^{-k}\mathbf{1}, k = 1, 2, \ldots$$

The Laplace transformation of PH($\boldsymbol{\pi}, \mathbf{Q}$) is given by

$$\tilde{T}(s) = \boldsymbol{\pi}(s\mathbf{I} - \mathbf{Q})(-\mathbf{Q}\mathbf{1})$$

where $\mathbf{I}$ is an identity matrix and s is a complex number.

Theorem 2.2. *Suppose that F and G are PH distributions with representations $(\boldsymbol{\pi}_{\mathbf{1}}, \mathbf{Q}_1)$ of order m_1, and $(\boldsymbol{\pi}_{\mathbf{2}}, \mathbf{Q}_2)$ of order m_2, respectively. Then we have the following.*

1. *The convolution $F_1 {}^*F_2$ is a PH distribution with a representation $(\boldsymbol{\pi}, \mathbf{Q})$ of order $m_1 + m_2$ where*

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}_1 & \pi_{10}\boldsymbol{\pi}_2 \end{pmatrix}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & -\mathbf{Q}_1 \boldsymbol{e} \boldsymbol{\pi}_1 \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}$$

where $\mathbf{0}$ is a $m_1 \times m_2$ matrix of zeros and π_{10} is known as the point mass at zero.

2. *The mixture $\omega F_1 + (1 - \omega)F_2$, where $0 \leq \omega \leq 1$, is a PH distribution with a represen-*

tation $(\boldsymbol{\pi}, \mathbf{Q})$ *of order* $m_1 + m_2$ *where*

$$\boldsymbol{\pi} = \begin{pmatrix} \omega\boldsymbol{\pi}_1 & (1-\omega)\boldsymbol{\pi}_2 \end{pmatrix}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}$$

Proof. See Neuts (1981). □

2.5 Bayesian Estimation Methods

Several approaches have been proposed for the estimation of PH/Coxian PH models. Asmussen et al. (1996) develop an expectation-maximization (EM) algorithm to calculate maximum likelihood (ML) estimators for general PH distributions. Olsson (1996) extends this algorithm to right censored and interval censored data. Using Matlab or R routines, Faddy and McClean (1999, 2005) use ML methods to estimate the LOS of geriatric patients with Coxian PH distributions. In Faddy (2002), a penalized ML method is carried out to fit Coxian PH distributions with high orders. The method of moments has also been used to fit PH distributions. Johnson (1993) develops an algorithm that matched the first three moments of a mixture of Erlang distributions to the empirical moments. Horváth and Telek (2000) propose a method that separately approximates the main and tail parts of a PH distribution. More recently, McGrory et al. (2009) propose an innovative fully Bayesian approach for inference where the number of phases is unknown and the mean duration depends on covariates. Unfortunately, the Coxian model is only identifiable up to permutation of the intensity rates (Cumani, 1982). This identifiability issue often affects the convergence of the Markov

Chain Monte Carlo (MCMC) algorithm and the interpretation of the estimated parameters. Ausín and colleagues (Ausín and Lopes, 2007; Ausín et al., 2008, 2009) use Bayesian methods for estimating PH distributions and place ordering constraints on the intensity rates to address this identifiability problem. We incorporate covariates into the model and extend their Bayesian method to fit the Coxian PH regression model.

Moment Matching Algorithm

The moment matching algorithms are broadly used in computer science, engineering, operation research, etc., where the performance evaluation or optimization in stochastic environment is needed. The idea is to map a general probability distribution G, into a PH distribution P, such that some moments of P and G agree. Matching the first moment of any non-negative distribution is possible by a single exponential distribution, but unfortunately it is often not sufficient, as ignoring the higher moments can result in misleading conclusions. Generally in cases where matching only two moments suffices, it is possible to achieve solutions which perform very well. Sauer and Chandy (1975) provide a closed-form solution for matching two moments of a general distribution with squared coefficient of variation (CV). They use a two-phase hyper-exponential distribution, for matching distributions with squared coefficient of variability $\text{CV}^2 > 1$, and a generalized Erlang distribution for matching distributions with $\text{CV}^2 < 1$. Marie (1980) provides a closed-form solution for matching two moments of a general distribution with $\text{CV}^2 > 0$. He uses a two-phase Coxian PH distribution for distributions with $\text{CV}^2 > 1$, and a generalized Erlang distribution for distributions with $\text{CV}^2 < 1$. Osogami (2005) improves existing estimates with respect to methods computational efficiency and provides the closed-form solutions for matching distributions with Erlang-Coxian (EC) distributions. For example, if G has sufficiently high second and third moments, then a two-phase Coxian PH distribution alone suffices. If the variability of G is lower, however, we might append several exponential distributions to the two-phase Coxian PH distribution in order to get the variability of P to be low enough.

Maximum likelihood estimate (EM algorithm)

The Expectation-Maximization (EM) algorithm is an iterative method that aims to maximize the log-likelihood function (Dempster. et al., 1977). The EM algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete data problems. On each iteration of the EM algorithm, there are two steps: the *Expectation step* (E-step) and the *Maximization step* (M-step).

Asmussen et al. (1996) propose a novel EM algorithm fitting PH Distributions. Later Olsson (1996) extends the EM algorithm to censored data. EMpht is a C program for fitting PH distributions (http://home.imf.au.dk/asmus/pspapers.html). It can be used either to fit a PH distribution to a sample (which may contain censored observations), or to make a PH approximation of another continuous distribution. In addition, it is complemented by a Matlab program, PHplot, for graphical display of the fitted PH distribution.

2.5.1 Reversible Jump Markov Chain Monte Carlo (RJMCMC)

Green's (see Green, 1995) Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm is basically of Metropolis-Hastings type with specific trans-dimensional proposals carefully designed to move between different models in a way that is consistent with the desired stationary distribution of the MCMC algorithm. The idea has been widely applied in the finite normal mixture models (Richardson and Green, 1997). The key in RJMCMC is to allow moves between parameter subspaces of different dimensionality by permitting a series of different 'move types'. Details can be found in Denison et al. (1998); Dellaportas and Forster (1999).

A typical way to estimate parameters in finite mixture models is the EM algorithm we mentioned above. However, the basic EM algorithm has two main drawbacks, slow convergence and lack of an in-built procedure to compute the covariance matrix of parameter estimates. Moreover, some complex problems lead to intractable E-steps, for which Monte Carlo methods have been shown to provide efficient solutions. Though in the finite mixture problems, the E-step is easy to implement, the M-step is more difficult using Newton-Raphson method, Quasi-Newton method or other optimization methods. Diebolt and Robert (1994) provide the standard Bayesian formulation of the finite mixure models with a known number of components and its implementation via Markov Chain Monte Carlo (MCMC). For PH distributions, we first need to do some transformation to make the density function more like "finite mixture models".

2.5.2 Transformation Strategies

We can assume, without loss of generality, $\lambda_{01} \geq \lambda_{02} \geq \ldots \geq \lambda_{0m}$. One way to incorporate the ordering restriction is to represent the hazard rates of the model as follows (Ausín and Lopes, 2007; Ausín et al., 2008, 2009):

$$\lambda_{0k} = \lambda_{01}\nu_2\nu_3 \cdots \nu_k, 0 < \nu_j \leq 1, j, k = 2, 3, \cdots, m.$$

In addition, we can easily recover the transition probabilities from $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_m)$.

$$\begin{aligned}
\eta_1 &= 1 - p_{12} \\
\eta_2 &= p_{12}(1 - p_{23}) \\
&\vdots \\
\eta_{m-1} &= p_{12}p_{23} \cdots p_{m-2,m-1}(1 - p_{m-1,m}) \\
\eta_m &= p_{12}p_{23} \cdots p_{m-2,m-1}p_{m-1,m}
\end{aligned}$$

where η_k is the total probability of existing from transient state k. Instead of the parameters $(m, \boldsymbol{\lambda}_0, \boldsymbol{\beta}, \boldsymbol{p})$, we can work with the equivalent set of parameters $(m, \boldsymbol{\eta}, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta})$, which is relatively stable and easy to sample in the MCMC framework.

Based on the new parameter set, the likelihood function (2.6) can be written as

$$\begin{aligned}
L(\mathbf{t}|\mathbf{X}, m, \lambda_{01}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{m} \eta_k f_k(t_i|\mathbf{x}_i, k, \boldsymbol{\nu}, \boldsymbol{\beta}) \right\} \\
&= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{m} \eta_k \boldsymbol{\pi} \exp(\mathbf{Q}^{\star}_{ik} t_i)(-\mathbf{Q}^{\star}_{ik}\mathbf{1}) \right\}
\end{aligned} \tag{2.8}$$

where $f_k(t_i|\mathbf{x}_i, k, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta}) = \boldsymbol{\pi} \exp(\mathbf{Q}^{\star}_{ik} t_i)(-\mathbf{Q}^{\star}_{ik}\mathbf{1})$ is a k-phase generalized Erlang distribution and

$$\mathbf{Q}^{\star}_k = \lambda_{01} \exp(-\mathbf{x}'\boldsymbol{\beta}) \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -\nu_2 & \nu_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\nu_2\nu_3\cdots\nu_k \end{pmatrix}$$

If all the rates are unequal, we can obtain an explicit form of (2.8) without complicated matrix exponential computation.

Let

$$C_{j,k} = \prod_{r=1, r\neq j}^{k} \frac{\prod_{i=2}^{r} \nu_i}{\prod_{i=2}^{r} \nu_i - \prod_{i=2}^{j} \nu_i} \text{ for } j = 1, 2, \cdots, k$$

Then

$$f_k(t_i|\mathbf{x}_i, k, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta}) = \sum_{j=1}^{k} C_{j,k} \lambda_{01} \exp(-\mathbf{x}_i'\boldsymbol{\beta}) \left(\prod_{r=2}^{j} \nu_j \right) \exp\left(-\lambda_{01} \exp(-\mathbf{x}_i'\boldsymbol{\beta}) t_i \prod_{r=2}^{j} \nu_j \right)$$

Due to this re-parameterization, we can borrow ideas from finite mixture models and it is relatively easy to implement the imputation steps in MCMC.

2.5.3 RJMCMC for Coxian PH Distribution

We now define the prior distributions for the parameters $(m, \lambda_{01}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\beta})$ as follows:

First if the number of phases m is known, we assume the following prior distributions:

$$\begin{aligned}
\boldsymbol{\eta} &\sim Dirichlet(\xi_1, \xi_2, \cdots, \xi_m) \\
\lambda_{01} &\sim \text{Improper, i.e. density} \propto \frac{1}{\lambda_{01}} \\
\nu_k &\sim Beta(a, b), \text{ for } k = 2, \cdots, m \\
\boldsymbol{\beta} &\sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_l)
\end{aligned}$$

We specify the priors by setting $\xi_1 = \xi_2 = \cdots \xi_m = 1$, $a = 1.1$, $b = 1$ and $\sigma = 1000$ (Ausín et al., 2008). The RJMCMC algorithm for our model consists of the following steps.

(1) Update the total exiting weights $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_m)'$;

(2) Update the intensity rate λ_{01};

(3) Update the vector of $\boldsymbol{\nu} = (\nu_2, \nu_3, \cdots, \nu_m)'$;

(4) Update the coefficient parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_l)'$;

(5) Update the current number of phases.

Steps (1)-(4) do not alter the dimension of the current model and they are performed using the Gibbs Sampler with Metropolis-Hastings step. But step (5) accounts for the jumps between adjacent values of m, which is done using the two types of moves discussed in Richardson and Green (1997), i.e. the split-combine and birth-death moves suitably modified for specific needs.

We construct a MCMC algorithm using a data augmentation step, introducing a missing indicator variable Z_i such that $p(Z_i = k|\boldsymbol{\eta}, m) = \eta_k$, $f(t_i|Z_i = k) = f_k(t_i|k, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta}), i =$

$1, 2, \cdots, n$.

Then the posterior distribution is

$$\begin{aligned} p(Z_i = k|t_i, \boldsymbol{\eta}, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta}) &= \frac{\eta_k f_k(t_i|k, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta})}{\sum_{j=1}^{m} \eta_j f_j(t_i|j, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta})} \\ &\propto \eta_k f_k(t_i|k, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta}), k = 1, 2, \cdots, m \end{aligned}$$

With this missing data approach, the complete data set now is $(t_i, \mathbf{x}_i, z_i), i = 1, 2, \cdots, n$. Given the missing data $\mathbf{z}$, the complete likelihood is calculated by

$$L(\mathbf{t}|\mathbf{z}, \mathbf{X}, m, \boldsymbol{\eta}, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_{z_i}(t_i|\mathbf{x}_i, k, \lambda_{01}, \nu_2, \cdots, \nu_{Z_i}, \boldsymbol{\beta})$$

Therefore, step (1) can be achieved as

$$\boldsymbol{\eta}|\mathbf{t}, \mathbf{z} \sim Dirichlet(1 + n_1, 1 + n_2, \cdots, 1 + n_m)$$

where n_j is the number of observations exited from phase j, $j = 1, 2, \cdots, m$. One advantage of this reparameterization is that the total exiting probability $\boldsymbol{\eta}$ can be obtained directly from Gibbs sampling.

Next we obtain the conditional posterior distributions of $\lambda_{01}, \boldsymbol{\nu}$ and $\boldsymbol{\beta}$, which do not have explicit forms, and require the Metropolis-Hastings algorithm in order to obtain the new updates through iterations.

$$f(\lambda_{01}|\mathbf{t},\mathbf{z},\mathbf{X},\boldsymbol{\nu},\boldsymbol{\beta}) \propto \prod_{i=1}^{n} f_{z_i}(t_i|\mathbf{x}_i,k,\lambda_{01},\nu_2,\cdots,\nu_{Z_i},\boldsymbol{\beta})\pi(\lambda_{01})$$

$$f(\nu_k|\mathbf{t},\mathbf{z},\mathbf{X},\lambda_{01},\boldsymbol{\nu}_{-k},\boldsymbol{\beta}) \propto \prod_{i=1,z_i\geq k}^{n} f_{z_i}(t_i|\mathbf{x}_i,k,\lambda_{01},\nu_2,\cdots,\nu_{Z_i},\boldsymbol{\beta})\pi(\nu_k), k=2,3,\cdots,m$$

$$f(\boldsymbol{\beta}|\mathbf{t},\mathbf{z},\mathbf{X},\lambda_{01},\boldsymbol{\nu}) \propto f_{z_i}(t_i|\mathbf{x}_i,k,\lambda_{01},\nu_2,\cdots,\nu_{Z_i},\boldsymbol{\beta})\pi(\boldsymbol{\beta})$$

We use a Gamma proposal distribution $\tilde{\lambda} \sim G(2, 2/\lambda^{(s)})$ to update λ_{01} (step (2)) at step s. To sample $\nu_k^{(s+1)}$ (step (3)), we propose a Beta mixture distribution as a proposal distribution in order to preserve the mode of $\nu_k^{(s)}$ and avoid clustering around 1 (Ausín et al., 2008).

$$\tilde{\nu}_k \sim \frac{1}{2}Beta\left(\frac{1}{1-\nu_k^{(s)}}, 2\right) + \frac{1}{2}Beta\left(2, \frac{1}{\nu_k^{(s)}}\right), k=2,3,\cdots,m$$

The coefficient $\boldsymbol{\beta}$ are generated with the aid of a Metropolis-Hastings step as well, using a symmetric random walk. The log density of $\boldsymbol{\beta}$, ignoring terms that do not depend on $\boldsymbol{\beta}$, is

$$T(\boldsymbol{\beta}) = \log\left(f_{z_i}(t_i|\mathbf{x}_i,k,\lambda_{01},\nu_2,\cdots,\nu_{Z_i},\boldsymbol{\beta})\right) + \log(\pi(\boldsymbol{\beta}))$$

Assume that the maximum of $T(\boldsymbol{\beta})$ exists and that the Hessian matrix $\mathbf{H}$ is negative definite in a neighborhood of $\boldsymbol{\beta}^{(s)}$, define $\mathbf{V}^{(s)} = -\mathbf{H}(\boldsymbol{\beta}^{(s)})^{-1}$, and propose $\tilde{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}^{(s)}, \mathbf{V}^{(s)})$.

We now describe the relevant maps and constraints that must be satisfied to make these moves (in step (5)) reversible, so called “detailed balance” (Green, 1995). Assume a discrete uniform prior defined on $[1, M_{\max}]$ for m, where $M_{\max}$ represents an assumed upper limit of m. If the current number of phases is M, then the proposed number is $M^\star$, where $M^\star = M \pm 1$. We will need to consider split and combine moves, i.e. split one phase into two adjacent phases, or combine two adjacent phases into one. We take a mapping approach to combine or split phases as follows.

In the combine move, parameters are updated according to $\tilde{\eta}_r = \eta_{r1} + \eta_{r2}, \tilde{\nu}_r = \nu_{r1}\nu_{r2}$, where we can obtain $\tilde{\lambda}_{0r} = \lambda_{0,r2}$. For the case $r = 1, \tilde{\lambda}_{01} = \lambda_{01}\nu_2$.

In the split move, we generate u_1 and u_2 from $U(0, 1)$. Let

$$\begin{aligned}
\tilde{\eta}_{r1} &= u_1\eta_r \\
\tilde{\eta}_{r2} &= (1 - u_1)\eta_r \\
\tilde{\nu}_{r1} &= u_2 + \nu_r(1 - u_2) \\
\tilde{\nu}_{r2} &= \frac{\nu_r}{u_2 + \nu_r(1 - u_2)}
\end{aligned}$$

For the case $r = 1$, we generate $\tilde{\nu}_{r2} = u_2$, and $\tilde{\lambda}_{01} = \lambda_{01}/u_2$. The parameters in the remaining phases are not modified.

2.6 Model Checking

To assess model fitting, Faddy et al. (2009) define a generalized residual as observed outcome divided by the estimated mean from the fitted model, and produce a quantile-quantile plot for the residuals. However, the residual distribution of a Coxian PH model is unknown. An alternative approach, which also allows for censored data, compares the empirical Kaplan-Meier estimate of the survival function with the model based estimates (Lambert et al., 2004). The model-based estimated survival for each patient is

$$\begin{aligned}
\hat{S}_i(t_i) &= \boldsymbol{\pi} \exp\left(\hat{\mathbf{Q}}_i t_i\right) \mathbf{1} \\
&= \boldsymbol{\pi} \exp\left(\exp(-\mathbf{x}_i' \hat{\boldsymbol{\beta}}) \begin{pmatrix} -\hat{\lambda}_{01} & \hat{p}_{12}\hat{\lambda}_{01} & \cdots & 0 \\ 0 & -\hat{\lambda}_{02} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\hat{\lambda}_{0m} \end{pmatrix} t_i \right) \mathbf{1}
\end{aligned} \tag{2.9}$$

Average predicted survival is obtained from point-wise averages of the survival curves over patients with a specified covariate profile. This procedure is an informal assessment of model adequacy. Johnson (2004) extends the classical Pearson χ^2-test for goodness of fit to the Bayesian context. A formal diagnostic to departures of the assumed model is beyond the scope of the present thesis. A unique aspect of PH model specification is the choice of the number of phases of the hidden Markov process, which is described in the previous section.

2.7 Classification

One of the goals in analyzing healthcare utilization is to predict resource use by different groups of patients who exhibit a semblance of within-group homogeneity while having between-group heterogeneity. Such a classification would be valuable in healthcare resource management and efficient allocation. For inpatient stays, Marshall and McClean (2004) investigate the potential of Coxian PH distributions to identify common characteristics in groups of patients according to their observed LOS. We focus here on the expected mean LOS and exploit this measure to group patients in the study. Specifically, we use the following algorithm to group LOS in m classes. Recall the exit probabilities $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_m)'$

from the transient states which can be obtained directly from MCMC posterior samples after the reparameterization.

Patients are assigned into different clusters according to their LOS in the ratio $\eta_1 : \eta_2 : \cdots \eta_m$. The k-th LOS group G_k is determined by:

$$G_k = \left\{ \hat{t}_{(i)} : n \sum_{k=1}^{K-1} \eta_k < i \leq n \sum_{k=1}^{K} \eta_k \right\}, K = 1, 2, \cdots, m. \tag{2.10}$$

where $\hat{t}_{(i)}, i = 1, 2, \cdots, n$ denotes the ordered expected mean LOS and n is the number of patients. Patient characteristics within each latent LOS group may then be explored to determine if they have any common characteristics and across LOS groups is also be explored to see if they differ.

2.8 Computation of Partial Effects

Social scientists are sometimes interested in estimating quantities other than survival in Eq (2.9) or classifying individuals into relatively more homogeneous groups using (2.10). One estimand favored by social scientists is the marginal effect (or predictive margin, recycled prediction, partial effect) of covariates with respect to the mean of the dependent variable, which is, in our case, the mean survival time (Basu and Rathouz, 2005). Under the parameterization in (2.6) and (2.7), the conditional mean duration is $E(T|\mathbf{x}) = \exp(a_0 + \mathbf{x}'\boldsymbol{\beta})$ (Faddy et al., 2009). The marginal effects are defined below for continuous and discrete covariates.

2.8.1 Continuous Covariates

For continuous covariates, partial effects (PE) are usually computed as the derivative of the conditional mean with respect to the covariate. In order to assess the overall impact of a continuous covariate x_b on the mean LOS, it is reasonable to consider the following PE:

$$\begin{aligned}\mathbf{PE} &= \frac{d}{d\mathbf{x}}\left(E(T|\mathbf{x})\right) \\ &= E(T|\mathbf{x})\boldsymbol{\beta} \\ &= \exp(a_0 + \mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}\end{aligned} \tag{2.11}$$

The **PE** is estimated from the posterior samples $\left\{\boldsymbol{\beta}^{(s)}; 1 \leq s \leq S\right\}$ by

$$\widehat{\mathbf{PE}}^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \exp(a_0^{(s)} + \mathbf{x}_i'\boldsymbol{\beta}^{(s)})\boldsymbol{\beta}^{(s)} \tag{2.12}$$

$$\widehat{\mathbf{PE}} = \frac{1}{S}\sum_{s=1}^{S} \widehat{\mathbf{PE}}^{(s)} \tag{2.13}$$

The standard deviation (SD) of PE_b is given by

$$\text{SD}(\text{PE}_b) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\widehat{\text{PE}}_b^{(s)} - \overline{\widehat{\text{PE}}}_b\right)^2} \tag{2.14}$$

where the subscript 'b' denotes the b-th component of $\widehat{\mathbf{PE}}$. Here, Eq (2.11) measures the local rate of change in $E(T|\mathbf{X})$ considered as a function in $\mathbf{x}$, relative to $E(T|\mathbf{X})$, and can be viewed as an approximation to the percentage change in $E(T|\mathbf{X})$ for one unit increase in $\mathbf{x}$. Eq (2.13) and (2.14) compute the estimated average PE and SD derived directly from the posterior samples, instead of bootstrapping by resampling methods.

2.8.2 Discrete Covariates

For a discrete covariate, the formula (2.11) is no longer valid. The partial effects are more appropriately derived by partial difference rather than partial derivatives. Without loss of generality, suppose a covariate has B categories. We can generate B-1 dummy variables for each level of the covariate other than the reference level. We will estimate the conditional mean twice. First, we set x_b to 1 (e.g., b-th category indicator) in the entire sample; and then second, we set x_b to 0. In both occasions, the values of the other B-2 dummy variables are set to 0 and the values of other covariates are left untouched. Then the proper analogs of (2.11) and (2.12) are

$$\mathrm{PE}_b = E(Y|x_b = 1, \mathbf{x}_{(-b)}) - E(Y|x_b = 0, \mathbf{x}_{(-b)}) \tag{2.15}$$

$$\widehat{\mathrm{PE}}_b^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(a_0^{(s)} + \beta_b^{(s)} + \mathbf{x}_i'\boldsymbol{\beta}_{(-b)}^{(s)}\right) - \frac{1}{n}\sum_{i=1}^{n} \exp\left(a_0^{(s)} + \mathbf{x}_i'\boldsymbol{\beta}_{(-b)}^{(s)}\right) \tag{2.16}$$

The estimated average PE and SD are hence calculated in the same way as formula (2.13) and (2.14). The partial effect of the binary variable x_b is the change in the mean LOS when x_b changes from 0 to 1 with all other variables kept constant at their observed values.

2.9 Modeling Hospital LOS for AMI Patients

2.9.1 Patient Sample

We apply the above models to the analysis of length of stay (LOS) for patients with acute myocardial infarction (AMI) in the 2003 Nationwide Inpatient Sample (NIS). The NIS is a database of all hospital inpatient stays from a stratified sample of approximately 1,000 community hospitals in the US. In 2003, the NIS contains nearly 8 million discharges from 37 states. Sixty strata are defined by a combination of geographic region (Northeast, South, Midwest and West), location (urban, rural), ownership (public, private), teaching status, and bed size (small, medium and large). Summary measures derived from the sample can be extrapolated to national estimates using sampling weights. Patient demographics in the NIS include age at admission, gender, race, and primary payer. Patient clinical characteristics include treatment procedures undergone and comorbidities assessed during the stay. In addition to the stratification variables in the NIS, we used the following covariates in our Coxian regression model: age at admission in years, gender, procedure, comorbidity and insurance status (Pompei et al., 1991). The primary procedure that the patient underwent was based on the ICD-9 CM Procedure codes: CABG=Coronary Artery Bypass Graft, PTCA=Percutaneous Transluminal Coronary Angiography, CATH=Cardiac Catheterization, Other, or None=no procedure performed. We constructed the Charlson Comorbidity Index (CCI) to capture comorbidity based on ICD-9 CM diagnoses during the stay (Charlson et al., 1987; Matsui et al., 1996). The CCI is a weighted sum of 19 conditions such as diabetes, with and without complications, congestive heart failure, peripheral vascular disease, pulmonary disease, renal disease, etc. We categorized the CCI score into 4 subgroups: 1, 2, 3 and ≥ 4. Insurance status is defined as the expected primary payer: Medicare, Medicaid, Self-pay or Other (which includes private insurance 33.6%, no charge $< 1\%$).

The LOS of AMI patients could vary widely depending on severity of illness, comorbidity and type of procedure. Most patients are discharged after a relatively short period of time. However, some patients may remain in hospital over a long time period receiving continuous care. This suggests that the distribution of LOS reflects different hidden structures. Figure 2.3 shows the histogram (truncated at 50 days) of the distribution of LOS for 11,749 AMI patients in the 2003 NIS. The data are highly skewed to the right with a large number of outliers, because of variation in patient characteristics. The LOS ranges between 1 and 142 days with mean of 5.51 days and standard deviation of 5.90 days. The mean LOS of female patients (mean=6.01) was higher than that of male patients (mean=5.21). Patients who underwent CABG had the highest LOS, followed by patients who underwent OTHER, CATH, PTCA and NONE. Mean LOS was different according to CCI scores. Patients with CCI$\geq$4 had the highest mean LOS (mean=7.89), followed by CCI=3, CCI=2, and CCI=1 with means 6.70, 5.61 and 3.87, respectively.

Approximately 63% of our patient sample is male. Mean age at admission was 64.6 years (SD=12.7). The most prevalent procedure performed on these patients was PTCA (40.3%) followed by CATH (19.4%). By definition all patients had CCI$\geq$1. Approximately 37.8% of patients had no other comorbidity (CCI=1), followed by CCI=2 (29.1%), CCI=3 (16.9%) and CCI$\geq$4 (16.2%). As expected in our sample, Medicare was the primary payer for 52.7% of the patients. We dropped 23 patients from the sample who had invalid or missing insurance information. See Table 2.1 for additional information.

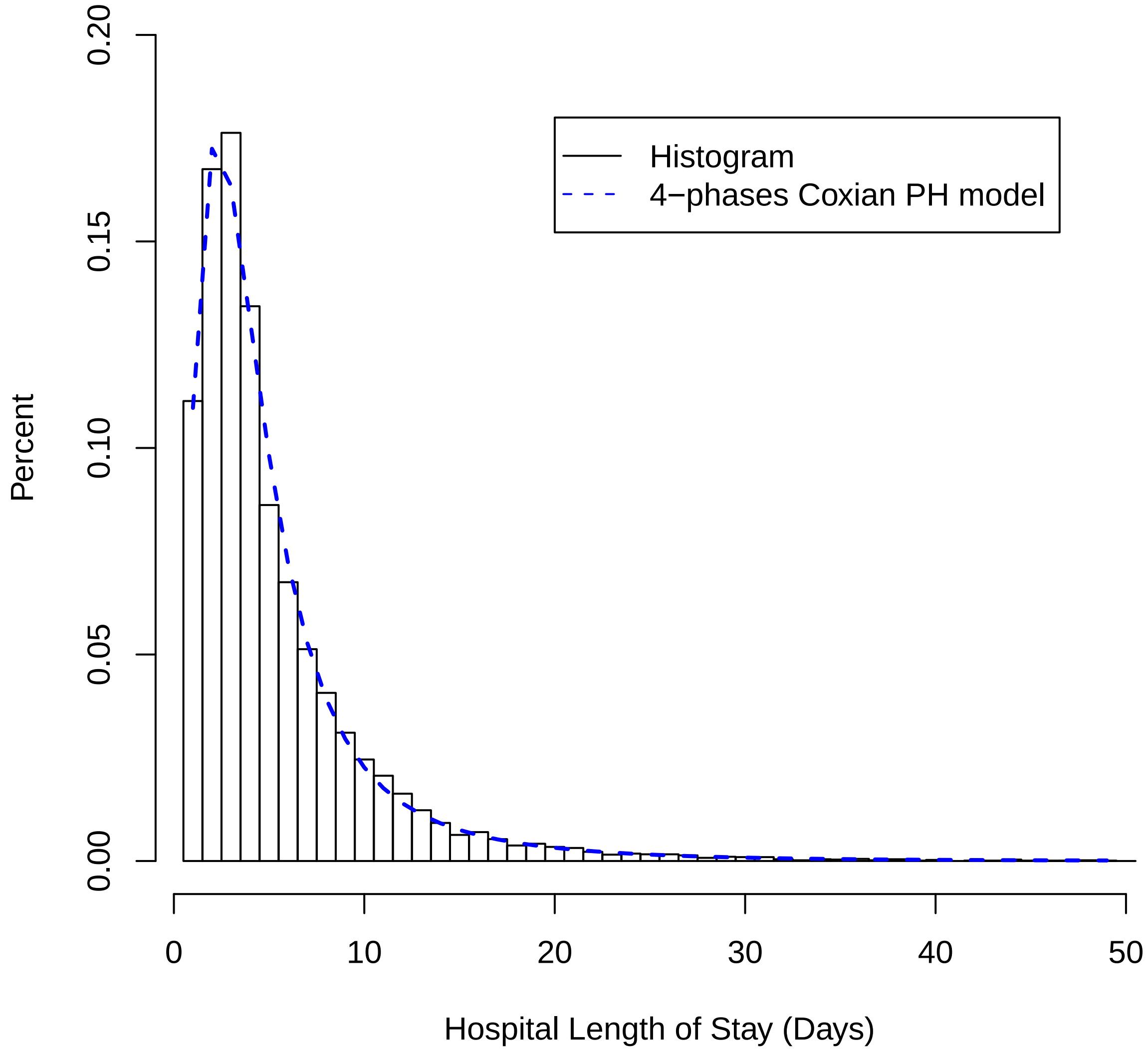


Figure 2.3: Histogram of hospital length of stay (LOS) and fitted 4-phases Coxian PH model. The solid line (—) is the histogram of LOS and the dotted line (- - -) is the model-based density function estimate (truncated at 50 days).

Table 2.1: Descriptive Statistics for AMI patients in NIS in 2003 (N=11,749)

Characteristics	Group	N(%)	LOS (days) Mean (SD)
All Patients		11,749 (100)	5.51 (5.90)
Gender	Male	7378 (62.8)	5.21 (5.54)
	Female	4371 (37.2)	6.01 (6.43)
Procedure	CABG&PTCA	107 (0.9)	11.73 (8.78)
	CABG	1301 (11.1)	11.59 (8.81)
	PTCA	4736 (40.3)	4.04 (3.38)
	CATH	2275 (19.4)	5.08 (5.12)
	Other	1490 (12.7)	7.06 (8.16)
	None	1840 (15.7)	3.88 (2.89)
Comorbidity, CCI	1	4438 (37.8)	3.87 (3.79)
	2	3423 (29.1)	5.61 (6.22)
	3	1981 (16.9)	6.70 (6.54)
	4^+	1907 (16.2)	7.89 (7.33)
Disposition of Patients	Routine	7635 (65.0)	4.58 (3.51)
	Transfer: Short-term Hospital	1451 (12.3)	2.96 (2.96)
	Transfer: Other Type of Facility	1119 (9.5)	11.33 (10.3)
	Home Health Care	883 (7.5)	9.25 (10.8)
	Died in Hospital	661 (5.6)	6.96 (6.65)
Hospital Region	Northeast	2524 (24.5)	5.76 (6.60)
	Midwest	2633 (22.4)	5.42 (5.43)
	South	4986 (42.4)	5.34 (5.72)
	West	1606 (13.7)	5.16 (6.00)
Hospital Location/ Teaching Status	Rural	1041 (8.9)	4.64 (4.20)
	Urban non-teaching	5301 (45.1)	5.22 (5.16)
	Urban teaching	5407 (46.0)	5.95 (6.76)
Hospital Bedsize	Small	950 (8.1)	4.79 (5.65)
	Medium	2658 (22.6)	5.30 (6.24)
	Large	8141 (69.3)	5.66 (5.81)
Primary Payer	Medicare	6186 (52.7)	6.25 (6.36)
	Medicaid	690 (5.9)	6.17 (7.17)
	Self-pay	562 (4.8)	4.32 (3.64)
	Other	4311 (36.7)	4.48 (4.97)

2.9.2 Estimation of Coxian phase-type Regression Model

For the Bayesian estimation of the Coxian regression model described in Section 2.5, we performed 100,000 MCMC iterations, discarded the first 50,000 of these iterations as burn-in period and retained 10,000 thinned posterior samples of $(\boldsymbol{\eta}, \lambda_{01}, \boldsymbol{\nu}, \boldsymbol{\beta})$. The selection of the best estimate for the number of phases m, using the RJMCMC algorithm, is based on the highest percent of times out of the 10,000 posterior samples where a particular number of phases is reached. In our data, the most likely number of phases of the underlying Coxian PH distribution is 4, with posterior probability of 0.313, followed by a 3-phase model with posterior probability of 0.273 and the 5-phase model with posterior probability of 0.255 (see Figure 2.4). The posterior means of the intercept parameters and covariate coefficients are given in Column 1, Table 2.2. We also use these posterior means as the starting values and apply the fully Bayesian approach (McGrory et al., 2009) to generate the posterior samples for measures of interest. Similar results are presented in Column 2, Table 2.2, except for the first two intensity rates without ordering.

Based on the estimates of the 4-phase model, all patients start in state 1, almost all of them transfer from state 1 to state 2, then from state 2 to state 3. The total probabilities of exiting from states 1, 2, 3, and 4 were 0.0002, 0.001, 0.923 and 0.076, respectively. Among correlates, LOS was positively associated with age, female gender, CCI and procedure type (Table 2.2). As expected, older patients had longer LOS. A possible explanation is that older patients have a longer recovery period from their procedures that lengthen their hospital stay. The effect of female gender on LOS was an estimated 0.076 (β-coefficient), corresponding to an expected increase in LOS of 1.079 times that of males. Higher comorbidity would indicate a more severe condition possibly leading to a longer stay in hospital. For example, CCI comorbidity scores 2, 3 and $\geq$4, led to an increase in LOS by 1.23, 1.46, and 1.78 times the LOS of patients with a CCI score of 1. The expected LOS for patients who underwent CABG, PTCA, CATH and 'other' procedures were, respectively, 3.24, 1.25, 1.31, 1.43 times the LOS of patients without any procedure. In the present study, we found that Medicare and Medicaid insured individuals had increased mean hospital LOS, compared to self-pay and 'other' (primarily private insurance). LOS variation by region and hospital size has been explored previously (Xiao et al., 1997). In our study, the highest mean LOS was in the North East region whereas hospitals in the West region had the smallest mean LOS.

To further justify the proposed 4-phase Coxian model, an alternative model using a heavy tailed distribution, like the log-normal distribution, was fitted for comparison with the same structure for the mean LOS. Some similarities are found in the covariate coefficients between the two different models. However, the adequacy of the specified log-normal is inferior to the proposed Coxian model. It is discussed later and visually illustrated in Figure 2.5.

Mean LOS can be estimated from the model using a specific covariate profile or by averaging over the predicted values for each observation. The partial effects, defined in Section 2.8 for continuous and discrete covariates are reported in Table 2.3. For example, the mean LOS for patients with CCI equal to 3 was 1.97 days longer than the mean LOS for patients with CCI equal to 1. Compared to those having no procedures, patients who underwent CABG had much longer mean LOS, the difference was 8.30 days.

Table 2.4 highlights the sample average of estimates of the fitted mean for different classes. Combining the fairly small total exit probabilities in the first two phases, an estimated 92.4% of the sample exited from first three phases. This is regarded as short-stay group, with predicted average LOS of 4.8 days. The remaining 7.6% of the sample are long-stay patients, with predicted average LOS of 13.8 days. For the whole sample, mean LOS estimated from the model is 5.48 days, with a minimum of 1.75 days and a maximum of 22.46 days. There is large variation in the mean LOS between these two classes.

To assess model fit, we compared the empirical Kaplan-Meier survival curves to the model-based average of subject-specific fitted curves from the 4-phase Coxian distribution. We can obtain values of estimated discharge probability $1 - S(t)$ for the individual, for LOS ranging from 1 to 142 days. The overall averaged fitted curves and empirical estimates are shown in the top left plot in Figure 2.5. The step functions, solid lines (—), are the empirical Kaplan-Meier estimates of $\hat{S}(t)$ the dash lines (- - -), the Coxian PH model-based estimates, are visually superimposed which indicate an adequate level of fitting. Within each stratum defined by procedure type, the model also fitted fairly well. Patients who had CABG surgery had the highest discharge probability. For comparison, the dotted lines $(\cdots)$ in Figure 2.5 are the log-normal model-based estimates. Relative to the better fitting Coxian PH model, the log-normal model gives similar estimates in covariates coefficents, but inferior fit, especially for patients with CABG surgery. These patients tended to have longer LOS.

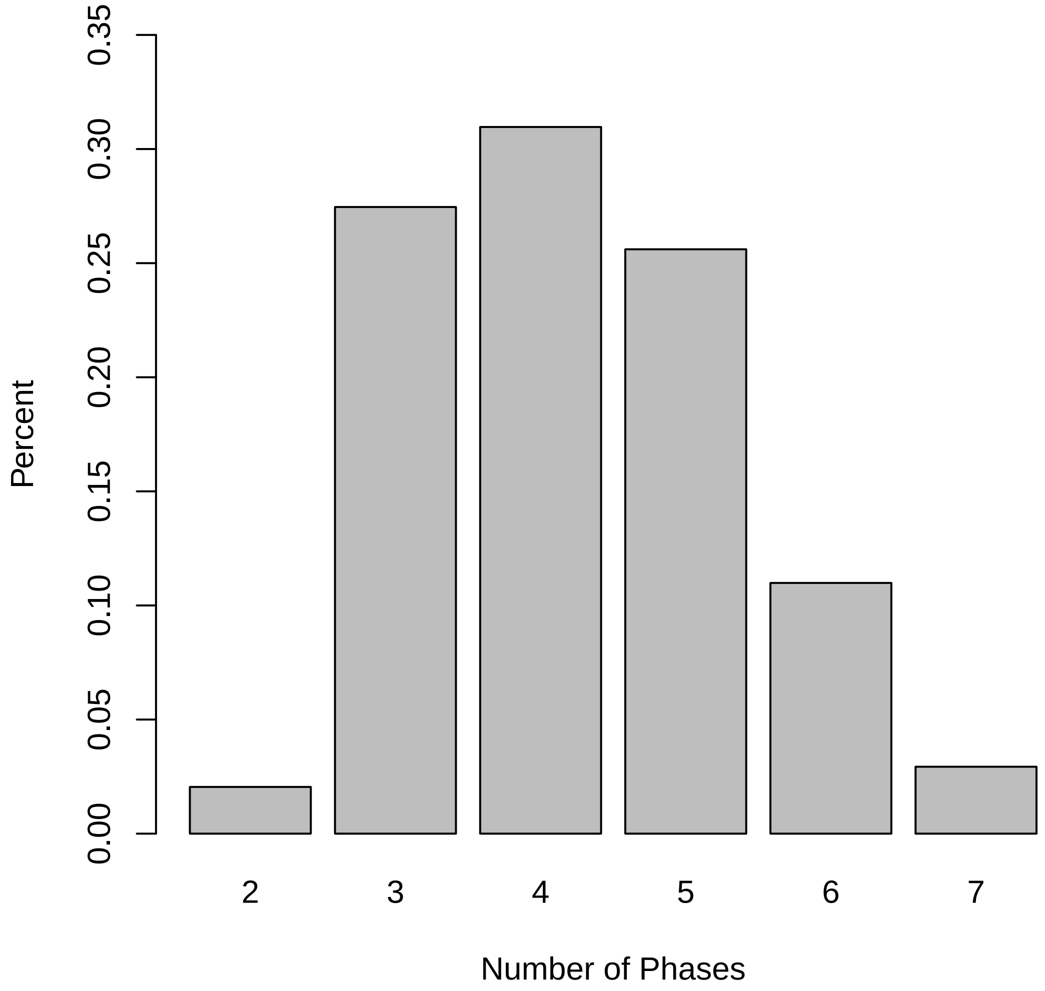


Figure 2.4: Number of phases dynamically selected by RJMCMC

Table 2.2: Estimates of posterior means (SD) based on 4-phase Coxian PH model and log-normal model

Covariate		RJMCMC	RJMCMC	log-normal
			(McGrory)	
		Posterior Mean	Posterior Mean	Posterior Mean
		(SD)	(SD)	(SD)
λ_{01}		1.297 (0.052)	1.202 (0.103)	N.A.
λ_{02}		1.230 (0.046)	1.203 (0.104)	
λ_{03}		1.131 (0.065)	1.231 (0.104)	
λ_{04}		0.209 (0.011)	0.206 (0.013)	
η_1		0.0002 (0.0002)	0.0003 (0.0003)	
η_2		0.001 (0.001)	0.002 (0.002)	
η_3		0.923 (0.005)	0.923 (0.007)	
η_4		0.076 (0.005)	0.075 (0.007)	
Intercept		1.042 (0.026)	1.050 (0.024)	0.365 (0.055)
Age		0.008 (0.0007)	0.008 (0.0007)	0.007 (0.0007)
Gender	Female	0.076 (0.012)	0.072 (0.012)	0.071 (0.012)
Procedure	CABG	1.177 (0.021)	1.171 (0.022)	1.237 (0.025)
	PTCA	0.222 (0.018)	0.220 (0.017)	0.237 (0.022)
	CATH	0.272 (0.020)	0.268 (0.018)	0.285 (0.021)
	Other	0.360 (0.025)	0.361 (0.023)	0.339 (0.022)
CCI	2	0.205 (0.015)	0.205 (0.015)	0.194 (0.024)
	3	0.378 (0.019)	0.377 (0.019)	0.357 (0.019)
	4^+	0.576 (0.019)	0.574 (0.019)	0.565 (0.020)
Region	Northeast	0.094 (0.021)	0.092 (0.021)	0.084 (0.022)
	Midwest	0.055 (0.020)	0.056 (0.019)	0.055 (0.021)
	South	0.085 (0.019)	0.086 (0.016)	0.084 (0.019)
Location/	Rural	-0.135 (0.022)	-0.134 (0.023)	-0.167 (0.024)
Teaching status	Urban non-teaching	-0.069 (0.013)	-0.072 (0.014)	-0.075 (0.014)
Bedsize	Small	-0.155 (0.026)	-0.156 (0.024)	-0.176 (0.025)
	Medium	-0.067 (0.015)	-0.066 (0.012)	-0.085 (0.015)
Primary payer	Medicaid	0.072 (0.029)	0.072 (0.028)	0.056 (0.029)
	Self-Pay	-0.031 (0.029)	-0.031 (0.031)	-0.036 (0.030)
	Other	-0.063 (0.018)	-0.069 (0.019)	-0.064 (0.018)
Scale				0.433 (0.006)

Note: CCI=Charlson Comorbidity Index.
Reference group: Gender=male, Procedure=none, CCI=1, Region=West,
Bed size=large, Location/teaching status=urban teaching, Primary payer=Medicare.

Table 2.3: Estimated average partial effects on mean LOS (SD) associated with different covariates

Covariate		Estiamte (SD)	Covariate		Estimate (SD)
Age*		0.043 (0.004)	Location/teaching status		
Procedure				Rural	-0.719 (0.111)
	CABG	8.296 (0.205)		Urban non-teaching	-0.379 (0.071)
	PTCA	0.918 (0.072)			
	CATH	1.154 (0.087)	Bedsize		
	Other	1.602 (0.117)		Small	-0.809 (0.125)
				Medium	-0.365 (0.081)
CCI					
	2	0.974 (0.072)	Primary Payer		
	3	1.971 (0.106)		Mediciad	0.424 (0.172)
	≥4	3.343 (0.127)		Self-Pay	-0.167 (0.159)
				Other	-0.340 (0.097)
Region					
	Northeast	0.507 (0.111)			
	Midwest	0.288 (0.102)			
	South	0.455 (0.097)			

Note: CCI=Charlson Comorbidity Index.

* The partial effect for age was estimated at the mean age.

Reference group:Procedure=none, CCI=1, Region=West,
Bed size=large, Location/teaching status=urban teaching, Primary payer=Medicare.

Table 2.4: Patient characteristics by classification of LOS

Covariate		Total	Short LOS[a]	Long LOS[a]
		N=11,749	N=10,856	N=893
Mean LOS (SD)		5.50 (5.90)	4.8	12.8
Mean Age (SD)		64.6 (12.7)	64.26	68.22
		N(%)	N(%)	N(%)
Gender	Female	4371 (37.0)	4019 (37.0)	352 (39.4)
	Male	7378 (63.0)	6837 (63.0)	541 (60.6)
Procedure	CABG	1408 (12.0)	516 (0.5)	892 (99.9)
	PTCA	4736 (40.3)	4736 (43.6)	0 (0.0)
	CATH	2275 (19.4)	2275 (21.0)	0 (0.0)
	Other	1490 (12.7)	1489 (3.7)	1 (0.1)
	None	1840 (15.6)	1840 (16.9)	0 (0.0)
CCI	1	4438 (37.8)	4363 (40.2)	75 (8.4)
	2	3423 (29.1)	3054 (28.1)	369 (41.3)
	3	1981 (16.9)	1722 (15.9)	252 (29.0)
	≥4	1907 (16.2)	1717 (15.8)	190 (21.3)
Insurance	Medicare	6186 (52.7)	5628 (51.8)	558 (62.5)
	Medicaid	690 (5.9)	622 (5.7)	68 (7.6)
	Self-Pay	562 (4.8)	535 (4.9)	27 (3.0)
	Other	4311 (36.7)	4071 (37.5)	240 (26.9)

Note: CCI=Charlson Comorbidity Index.

Reference group: Gender=male, Procedure=none, CCI=1, Region=West,
Bed size=large, Location/teaching status=urban teaching, Primary payer=Medicare.

[a] Sample averages for LOS and age within the short-stay and long-stay groups

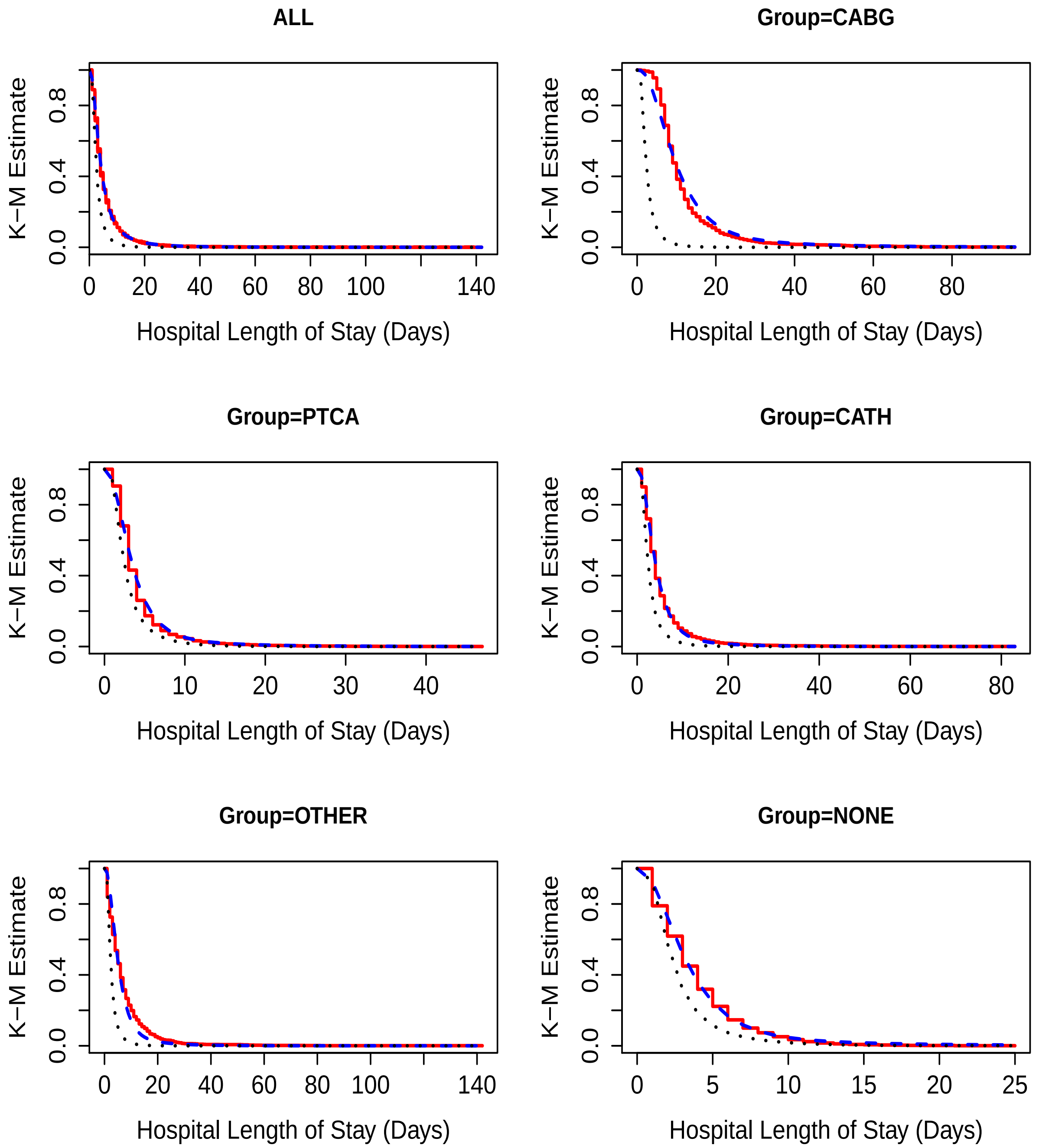


Figure 2.5: Goodness-of-fit curves by procedure type. The solid lines (—) are the Empirical estimates of discharge probability, the dash lines (- - -) are the Coxian PH model-based estimates, the dotted lines (...) are the log-normal model-based estimates.

2.10 Discussion

This chapter demonstrates the application of Coxian PH stochastic regression models to hospital LOS to account for the heavy skewness and heterogeneity in the data. A Bayesian method based on RJMCMC was applied to dynamically select the number of phases. This method avoids arbitrary trimming and transformation of the data. In addition, the approach allows us to obtain estimates of mean LOS, median, other percentiles, and class characteristics. While a complete description of the underlying hidden states of the Markov process would depend on specific applications, we found in this example a meaningful interpretation based on short-stay, and long-stay subgroups for AMI patients. The two groups differed primarily on comorbidity and procedure type. Short-stay patients had the lowest CCI; long-stay patients comprised largely of those who underwent CABG surgery. Procedure type and CCI were significant correlates of LOS.

The strength of Coxian PH regression models for LOS lies in their flexibility in accommodating extreme values, while revealing hidden features such as short and long stays in hospitals. The management of hospital LOS has become an important issue in cost containment and control. Determination of relevant factors and hidden structures could inform discharge planning and allocation of resources (Lanzarone et al., 2010; McDermott and Stock, 2007; Ramiarina et al., 2008). We believe this work contributes to the development of statistical models for analysis of LOS distributions and other consumption variables in healthcare resource analysis. For example, our proposed Coxian PH model is applicable to other skewed data such as inpatient cost as well as censored duration data. It can be extended to analyze longitudinal data exhibiting unobserved heterogeneity.

Chapter 3

Joint Modeling of Hospital Length of Stay and Cost with Shared Random Effects

In this chapter, we consider hospital length of stay (LOS) and cost jointly and construct a bivariate model from a common constellation of covariates that might influence their joint distribution. Shared random-effects modeling approach is then developed to conduct the joint analysis which permits simultaneous assessment of the correlates of LOS and in-hospital cost. Our model also provides important information for decisions on resource allocation.

3.1 Introduction

Random-effects (RE) models are of interests and widely used in capturing statistical dependence for a variety of data types, and allow for prediction, imputation, and hypothesis testing within a general regression context. The shared random effect approach has a very intuitive appeal to a lot of researchers who generally believe that there might be some latent quantity underlying an individual's susceptibility to both in-hospital LOS and cost. This latent quantity may represent environmental risk factors or hospital site effects yet to be identified. The latent process also induces dependence between the two explicitly observed processes. In this chapter, we illustrate the use of shared random effects to account for the correlation, build a joint Coxian phase-type /Log-normal (CPH-LN) model and describe the details of the marginal measurement. The idea of introducing random effects to incorporate correlation is not new, but a lot of previous work has focused on joint modeling the multivariate survival times through frailty models, with the same underlying model and β -coefficient for covariates. Here, we develop a flexible joint modeling approach with different distribution types and scales. We propose a modeling framework using a finite-state continuous time Markov process with a single absorbing state (discharge) to describe the hospital LOS, introduced in Chapter 2. Each phase represents an unknown health status and absorbing state is interpreted as the healthcare is terminated and discharge from the hospital or death. In addition, we assume cost is log-normally distributed. The correlations between outcomes are introduced by shared random effects, which can also affect the marginal means, variances and covariances. This approach adjusts for patient level characteristics while allowing for the inherent correlation structure.

The rest of the chapter is organized as follows. Section 3.2 introduces the joint modeling approach of LOS and cost via shared random effects. We provide a detailed description for marginal mean, variance and covariance structure. Then Section 3.3 describes the estimation methods using likelihood reformulation (LR). Application of the proposed model is given in Section 3.4. Finally in Section 3.5, we conclude and outline further research direction.

3.2 Shared Random-effects Model

Let Y_{i1} $(i = 1, 2, \dots, n)$ denote in-hospital LOS and $\mathbf{x}_{1i}$ be a vector of l_1 explanatory variables that might have impact on the distribution of Y_{i1}, and similarly, Y_{i2} $(i = 1, 2, \dots, n)$ denote cost, $\mathbf{x}_{2i}$ are the l_2 variables that may affect Y_{i2}. In practice, covariates $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ may represent the same covariate constellation, but not necessarily so. The observed data for patient i are $(Y_{i1}, Y_{i2}, \mathbf{x}_{1i}, \mathbf{x}_{2i}, 1 \leq i \leq n)$. The underlying relationship between LOS and its risk factors is modeled by a Coxian phase-type regression introduced in Chapter 2, where the mean structure is assumed to be a log-linear function of covariates $\mathbf{x}_1$, i.e. $\log{(\mu_i^{CPH})} = \beta_0^{CPH} + \mathbf{x}_{1i}'\boldsymbol{\beta}^{CPH}$. Similarly, the relationship between cost and risk factors is modeled by a log-normal regression $\log Y_{i2} = \beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN} + \sigma\epsilon = \mu_i + \sigma\epsilon_i$, where μ_i is the location parameter linear in $\mathbf{x}_{2i}$, σ is the scale parameter and $\epsilon_i \sim N(0, 1)$. Corresponding mean function of cost is given by $\log{(\mu_i^{LN})} = \mu_i + \frac{1}{2}\sigma^2 = \beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN} + \dfrac{1}{2}\sigma^2$. That is

$$\begin{aligned}
Y_{i1} &\sim \text{Coxian PH}(\lambda_0, \boldsymbol{p}, \boldsymbol{\beta}) \\
\mu_i^{CPH} &= \exp{(\beta_0^{CPH} + \mathbf{x}_{1i}'\boldsymbol{\beta}^{CPH})} \\
Y_{i2} &\sim \text{log-normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN}
\end{aligned} \tag{3.1}$$

The separate regression models described above are not suitable for dependent data because the two healthcare utilization measures, LOS and cost, are most of the time, correlated. To accommodate the inherent dependence between LOS and cost at the subject (individual, patient) level, shared random effects $b_i, 1 \leq i \leq n$ are incorporated into the linear prediction

(Liu et al., 2007). Then (3.1) can be rewritten as:

$$
\begin{aligned}
Y_{i1}|b_i &\sim \text{Coxian PH}(\lambda_0, \boldsymbol{p}, \boldsymbol{\beta}) \\
\mu_i^{CPH} &= \exp\left(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} + b_i\right) \\
Y_{i2}|b_i &\sim \text{log-normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \gamma b_i \\
b_i &\sim \text{i.i.d.}F(.)
\end{aligned}
\tag{3.2}
$$

where $F(.)$ is the distribution function for i.i.d. random effects $b_i (1 \leq i \leq n)$. Condition on b_i, Y_{i1} and Y_{i2} are independent, i.e. *conditional independence.* γ measures the different scale for the variance components. In (3.2), b_i characterizes the correlation between Y_{i1} and Y_{i2} within the same subject. When the correlation is high, fitting separate models might lead to biased covariate effect estimates. Economists are interested in marginal means, variances, covariances or correlations, which can be derived from conditional measurements by the following basic formulas:

$$
E[Y_{ij}] = E[E(Y_{ij}|b_i)], i = 1, 2, \dots n \quad j = 1, 2 \tag{3.3}
$$

$$
\begin{aligned}
&Cov(Y_{ij}, Y_{kl}) = E[Cov(Y_{ij}, Y_{kl}|b_i)] + Cov(E[Y_{ij}|b_i], E[Y_{kl}|b_i]), \\
&i, k = 1, 2, \dots, n \quad j, l = 1, 2
\end{aligned}
\tag{3.4}
$$

In addition, under the assumption of *conditional independence*, we have $Cov(Y_{i1}, Y_{i2}|b_i) = 0$.

Marginal Means

The marginal means for LOS and cost, averaging over the distribution of random effects b, are easy to calculate from (3.4):

$$
\begin{aligned}
E[Y_{i1}] &= E[\exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} + b_i)] \\
&= \exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH})E[\exp{(b_i)}]
\end{aligned}
\tag{3.5}
$$

$$
\begin{aligned}
E[Y_{i2}] &= E[\exp(\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \gamma b_i + \frac{1}{2}\sigma^2)] \\
&= \exp(\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2)E[\exp{(\gamma b_i)}]
\end{aligned}
\tag{3.6}
$$

Without loss of generality (WLOG), we can assume $E[\exp{(b_i)}] = 1$, under which the marginal means keep the same in both separate models and joint model for LOS and only a scale changed for cost. For example, assume $W_i = \exp{(b_i)}$, and $W_i \sim$ Gamma$(\frac{1}{\theta}, \theta)$, i.e.

$$
p(w_i) = \frac{w_i^{(\frac{1}{\theta}-1)} \exp{(-\frac{w_i}{\theta})}}{\Gamma(\theta^{-1})\theta^{\frac{1}{\theta}}}
$$

Then $E[W_i] = 1$, and $Var(W_i) = \theta$. Large values of θ signify a closer positive relationship between LOS and cost for the same subject and greater heterogeneity among subjects. Accordingly, $b_i \sim$ log-Gamma$(\frac{1}{\theta}, \theta)$ with density function

$$
p(b_i) = \frac{(\exp{(b_i)})^{\frac{1}{\theta}} \exp{(-\frac{\exp{(b_i)}}{\theta})}}{\Gamma(\theta^{-1})\theta^{\frac{1}{\theta}}}
$$

For some identifiability reasons, we restrict γ=1.

$$
\begin{aligned}
E[Y_{i1}] &= \exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH}) \\
E[Y_{i2}] &= \exp(\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2)
\end{aligned}
$$

which retains the unchaged marginal means.

Marginal Variances

From (3.4), we have

$$Cov(Y_{ij}, Y_{ij}) = E[Cov(Y_{ij}, Y_{ij}|b_i)] + Cov(E[Y_{ij}|b_i], E[Y_{ij}|b_i]), i = 1, 2, \ldots, n \quad j = 1, 2 \tag{3.7}$$

Therefore under the log-Gamma assumption for b_i, the marginal variances can be derived:

$$\begin{aligned} Cov(Y_{i1}, Y_{i1}) &= E[Var(Y_{i1}|b_i)] + Cov(E[Y_{i1}|b_i], E[Y_{i1}|b_i]) \\ &= E[Var(Y_{i1}|b_i)] \\ &+ Cov(\exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH})\exp{(b_i)}, \exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH})\exp{(b_i)}) \\ &= E[Var(Y_{i1}|b_i)] + \exp(2(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH}))Var(\exp(b_i)) \\ &= E[Var(Y_{i1}|b_i)] + \exp(2(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH}))\theta \end{aligned} \tag{3.8}$$

From Coxian PH properties in Chapter 2, we can obtain

$$\begin{aligned} Var(Y_{i1}|b_i) &= E(Y_{i1}^2|b_i) - (E(Y_{i1}|b_i))^2 \\ &= 2(\boldsymbol{\pi}\tilde{\mathbf{Q}}^{-2}\mathbf{1}) - (\boldsymbol{\pi}\tilde{\mathbf{Q}}^{-1}\mathbf{1})^2 \\ &= E[\exp{(2(\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} + b_i))}]\left(2\boldsymbol{\pi}(\boldsymbol{\Lambda}_0\mathbf{P})^{-2}\mathbf{1} - (\boldsymbol{\pi}(\boldsymbol{\Lambda}_0\mathbf{P})^{-1}\mathbf{1})^2\right) \\ &= \exp(2(\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH}))ME[\exp(2b_i)] \end{aligned}$$

where $M = 2\boldsymbol{\pi}(\boldsymbol{\Lambda}_0\mathbf{P})^{-2}\mathbf{1} - (\boldsymbol{\pi}(\boldsymbol{\Lambda}_0\mathbf{P})^{-1}\mathbf{1})^2$, $\boldsymbol{\Lambda}_0, \mathbf{P}$ can be found in Chapter 2, Section 2.2.

$$\begin{aligned}
E[\exp(2b_i)] &= E[W_i^2] \\
&= \int w_i^2 \frac{w_i^{\frac{1}{\theta}-1} \exp{(-\frac{w_i}{\theta})}}{\Gamma(\theta^{-1})\theta^{\frac{1}{\theta}}} dw_i \\
&= \int \frac{w_i^{\frac{1}{\theta}+1} \exp{(-\frac{w_i}{\theta})}}{\Gamma(\theta^{-1})\theta^{\frac{1}{\theta}}} dw_i \\
&= \frac{\Gamma(\theta^{-1}+2)\theta^{\theta^{-1}+2}}{\Gamma(\theta^{-1})\theta^{\theta^{-1}}} \\
&= \theta + 1
\end{aligned}$$

With above results, marginal variance of Y_{i1} in (3.8) can be rewritten as:

$$\begin{aligned}
Cov(Y_{i1}, Y_{i1}) &= \exp(2(\mathbf{x}_{1i}'\boldsymbol{\beta}^{CPH})M(\theta+1) + \exp(2(\beta_0^{CPH} + \mathbf{x}_{1i}'\boldsymbol{\beta}^{CPH}))\theta \\
&= \exp(2(\mathbf{x}_{1i}'\boldsymbol{\beta}^{CPH})(M(\theta+1) + \exp(2(\beta_0^{CPH}))\theta)
\end{aligned} \tag{3.9}$$

$$\begin{aligned}
Cov(Y_{i2}, Y_{i2}) &= E[Var(Y_{i2}|b_i)] + Cov(E[Y_{i2}|b_i], E[Y_{i2}|b_i]) \\
&= E[(\exp(\sigma^2) - 1)\exp(2\mu_i + \sigma^2)] \\
&+ Cov(\exp(\beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2)e^{b_i}, \exp(\beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2)e^{b_i}) \\
&= (\exp(\sigma^2) - 1)\exp(2(\beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN}) + \sigma^2)E[e^{2b_i}] \\
&+ \exp(2(\beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2))Var(e^{b_i}) \\
&= \exp(2(\beta_0^{LN} + \mathbf{x}_{2i}'\boldsymbol{\beta}^{LN}) + \sigma^2)[(\exp(\sigma^2) - 1)(\theta+1) + \theta]
\end{aligned} \tag{3.10}$$

Marginal Covariances

Under the assumption of *conditional independecne*,

$$
\begin{aligned}
Cov(Y_{i1}, Y_{i2}) &= E[Cov(Y_{i1}, Y_{i2}|b_i)] + Cov(E[Y_{i1}|b_i], E[Y_{i2}|b_i]) \\
&= 0 + Cov(E[Y_{i1}|b_i], E[Y_{i2}|b_i]) \\
&= [\exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} + b_i), \exp(\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2 + b_i)] \\
&= \exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} + \beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2)Var(\exp(b_i)) \\
&= \exp(\beta_0^{CPH} + \mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} + \beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + \frac{1}{2}\sigma^2)\theta
\end{aligned}
\tag{3.11}
$$

Marginal Correlations

$$
\begin{aligned}
Corr(Y_{i1}, Y_{i2}) &= \frac{Cov(Y_{i1}, Y_{i2})}{\sqrt{Var(Y_{i1})}\sqrt{Var(Y_{i2})}} \\
&= \frac{\exp(\beta_0^{CPH})\theta}{\sqrt{(M(\theta+1) + \exp(2\beta_0^{CPH})\theta)}\sqrt{(\exp(\sigma^2) - 1)(\theta+1) + \theta)}}
\end{aligned}
\tag{3.12}
$$

Obviously,

$\theta \rightarrow 0, Corr(Y_{i1}, Y_{i2}) \rightarrow 0$

$\theta \rightarrow \infty, Corr(Y_{i1}, Y_{i2}) \rightarrow \dfrac{\exp(\beta_0^{CPH})}{\sqrt{M + \exp(2\beta_0^{CPH})}\sqrt{\exp(\sigma^2)}}$

3.3 Estimation

Let $\mathbf{O}_i$ denote the observed data of a particular subject i, i.e. $\mathbf{O}_i = (Y_{i1}, Y_{i2}, \mathbf{x}_{1i}, \mathbf{x}_{2i})$ are i.i.d. for subjects, $i = 1, 2, \ldots, n$. The likelihood for $\mathbf{O}_i$ is

$$\begin{aligned} L(\mathbf{O}_i) &= \int f(y_{i1}|b_i) f(y_{i2}|b_i) p(b_i|\theta) db_i \\ &= \int \Big\{ \boldsymbol{\pi} [\exp(\exp(-\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} - b_i)\boldsymbol{\Lambda}_0 \mathbf{P} y_{i1})](-\exp(\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} - b_i)\boldsymbol{\Lambda}_0 \mathbf{P1}) \Big\} \qquad (3.13) \\ & \frac{1}{y_{i2}\sqrt{2\pi}\sigma} \exp(-\frac{(\log y_{i2} - (\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + b_i))^2}{2\sigma^2}) p(b_i|\theta) db_i \end{aligned}$$

A variety of numerical methods have been used to assess the above integral, for example, Laplace approximation, partial quasilikelihood, Gauss-hermite quadrature, adaptive Gaussian quadrature, and various Monte Carlo techniques. We will use SAS Proc NLMIXED (SAS Institute, Cary, NC, USA) for estimation in our joint modeling with shared random effects, based on the likelihood reformulation (LR) method proposed by Liu and Yu (2008). The basic idea is to reformulate the conditional likelihood on non-normal random effects to standard normal random effects and improve its computational efficiency. Apply LR method to reformulate (3.13) with respect to a log-Gamma density to a standard normal density $\phi(a_i)$, which gives,

$$\begin{aligned}
L(\mathbf{O}_i) &= \int f(y_{i1}|a_i)f(y_{i2}|a_i)\frac{p(a_i|\theta)}{\phi(a_i)}\phi(a_i)da_i \\
&= \int \Big\{\boldsymbol{\pi}[\exp(\exp(-\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} - a_i)\boldsymbol{\Lambda}_0\mathbf{P}y_{i1})](-\exp(\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} - a_i)\boldsymbol{\Lambda}_0\mathbf{P1})\Big\} \\
&\frac{1}{y_{i2}\sqrt{2\pi}\sigma}\exp(-\frac{(\log y_{i2} - (\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + a_i))^2}{2\sigma^2})\frac{p(a_i|\theta)}{\phi(a_i)}\phi(a_i)da_i \\
&= \int \exp(l_i^{A_1} + l_i^{A_2} + l_i^B - l_i^C)
\end{aligned} \tag{3.14}$$

where $l_i^{A_j}$ is the conditional log-likelihood of Y_{ij}, $j = 1, 2$, i.e.

$$\begin{aligned}
l_i^{A_1} &= \log \Big\{\boldsymbol{\pi}[\exp(\exp(-\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} - a_i)\boldsymbol{\Lambda}_0\mathbf{P}y_{i1})](-\exp(\mathbf{x}'_{1i}\boldsymbol{\beta}^{CPH} - a_i)\boldsymbol{\Lambda}_0\mathbf{P1})\Big\} \\
l_i^{A_2} &= -\log y_{i2} - \log\sigma - 0.5\log 2\pi - \frac{(\log y_{i2} - (\beta_0^{LN} + \mathbf{x}'_{2i}\boldsymbol{\beta}^{LN} + a_i))^2}{2\sigma^2}
\end{aligned}$$

l_i^B is log of log-Gamma density function, i.e.

$$l_i^B = \log p(a_i\theta) = -\theta^{-1}\log\theta - \log\Gamma(\theta^{-1}) + \frac{a_i}{\theta} - \frac{\exp(a_i)}{\theta}$$

and l_i^C is log of standard normal density function, i.e.

$$l_i^C = \log\phi(a_i) = -\frac{1}{2}a_i^2 + \text{Constant}$$

After above reformulation, likelihood function in (3.13) can be easily constructed by regular SAS programming statement and maximized by 'general' option in the 'Model' statement in Proc NLMIXED with adaptive Gaussian quadrature method.

3.4 Application to 2003 NIS AMI Hospitalized Patients

In this section, we apply the proposed joint modeling approach to the analysis of LOS and cost for acute myocardial infarction (AMI) patients in the 2003 Nationwide Inpatient Sample (NIS), which is introduced in Chapter 2, Section 2.9. The LOS of AMI patients could vary widely depending on severity of illness, comorbidity and type of procedure. Most patients are discharged after a relatively short period time. However, some patients may remain in hospital over a long time period receiving continuous care. The LOS ranged between 1 and 142 days with mean 5.50 days and SD 5.90 days. Cost had a range from \$90 to \$962,611, with the median \$31,704. Preliminary descriptive statistics is shown in Table 3.1. The mean LOS of female patients (mean=6.01) was higher than that of male patients (mean=5.21). The mean cost of female patients was \$45,166, approximately \$2,000 lower than that of males. Patients who underwent CABG had the highest LOS, followed by patients who underwent OTHER, CATH, PTCA and NONE. The same conclusion was obtained for CABG patients having highest mean cost of \$109,188. Mean LOS and median cost were different according to CCI scores. Patients with CCI≥4 had the highest mean LOS (=7.89) and mean cost (=\$52,282), followed by CCI=3, CCI=2 and CCI=1, respectively. However when concerning with median cost, the order could be changed. The plot of log-total charge versus LOS in Figure 3.1 indicates the positive correlation between the two outcomes.

Table 3.1: Characteristic of patients for LOS and Cost (N=1,1749)

Variable	Subgroup	N(%)	Mean LOS (SD)	Mean Cost (SD)	Median (Cost)
Gender	Female	4371 (37.2)	6.01 (6.43)	45,166 (54,755)	29,381
	Male	7378 (62.8)	5.21 (5.54)	47,054 (54,394)	33,278
CCI	CCI=1	4438 (37.8)	3.87 (3.79)	39,782 (39,497)	30,724
	CCI=2	3423 (29.1)	5.61 (6.22)	48,883 (57,602)	33,489
	CCI=3	1981 (16.9)	6.71 (6.54)	50,985 (61,542)	32,802
	CCI≥4	1907 (16.2)	7.89 (7.33)	52,282 (68,103)	30,561
Procedure	CABG	1408 (12.0)	11.60 (8.81)	109,188 (89,822)	81,701
	PTCA	4736 (40.3)	4.04 (3.38)	46,555 (33,949)	375,71
	CATH	2275 (19.4)	5.08 (5.12)	34,825 (40,781)	23,087
	Other	1490 (12.7)	7.06 (8.16)	41,722 (65,561)	21,710
	None	1840 (15.7)	3.88 (2.90)	15,744 (13,404)	11,946

Because healthcare utilization outcomes LOS and cost may be correlated, we apply our joint modeling approach proposed in Section 3.2 to fit the 2003 AMI patients data, linking the LOS and cost at the individual level with shared random effects. Potential correlates of LOS and cost included demographic and clinical variables that could be identified at admission, and the use of any cardiac procedures. In many applications, a two-phase Coxian model can provide sufficient flexibility to describe the evolution of the process over time. Therefor for simplicity we assume LOS has a 2-phase Coxian distribution. States typically represent unknown healthcare status. Gardiner et al. (2002) shows it is suitable to assume that in-hospital cost is log-normally distributed. SAS Proc NLMIXED with 10 quadrature points is used for estimation. Results are summarized in Table 3.2. Both LOS and cost are positively associated with age, female gender, CCI and procedure types. As expected older patients had longer LOS, higher cost, and the effect of female gender on LOS was an estimated 0.073 (β-coefficient) corresponding to an expected relative increase in LOS of 1.08, while female gender effect on cost is not significant. Higher comorbidity led to an increased stay by 1.25, 1.46, 1.78, and increased cost by 1.18, 1.29, 1.44 for CCI scores 2, 3, and ≥ 4

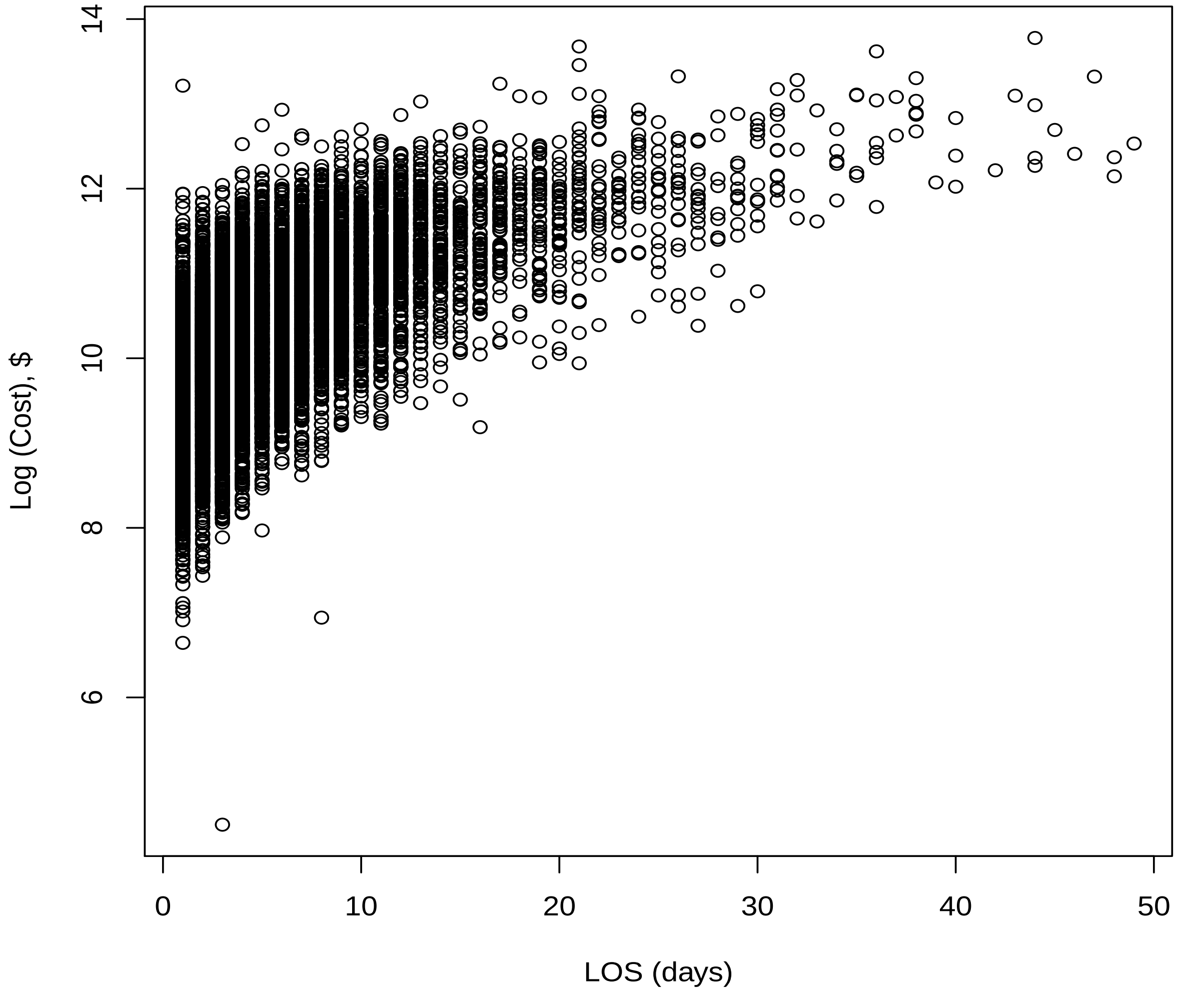


Figure 3.1: Plot of LOS vs. Log (Cost), truncated at 50 days

relative to a CCI score of 1. The effects of CABG, PTCA, CATH and 'other' procedures, compared to no procedure was an estimated relative increase by 3.78, 1.42, 1.49, 1.63 for LOS and 8.24, 3.82, 2.33, 2.17 for cost, respectively. These results were consistent with the raw data. In addition, from Table 3.2 we observe that random effects exist among the subject level. The significant random effect variance component ($\hat{\theta} = 0.290, p -$ value < 0.0001) indicates that a patient with longer length of stay tends to have higher cost.

For comparison, we also fit the separate models in (3.1) without random effects assuming 2-phase Coxian for LOS, and log-normal for cost. Results are provided in Table 3.3. We found that estimated coefficient for patients with other procedure types is larger than our model. There also exists some noticeable difference in the scale parameter estimate for σ.

Table 3.2: Estimates of coefficients and random effects ($\gamma = 1$)

Variable		CPH-Lognormal	
		LOS (CPH)	Cost (LN)
		Estimate (SE)	Estimate (SE)
Intercept		1.005 (0.026)	9.312 (0.021)
AGE		0.008 (0.0007)	0.003 (0.0006)
Gender	Female	0.073 (0.018)	0.005 (0.010)
CCI	CCI=2	0.221 (0.021)	0.162 (0.017)
	CCI=3	0.378 (0.026)	0.257 (0.021)
	CCI≥4	0.577 (0.027)	0.362 (0.021)
Procedure	CABG	1.330 (0.032)	2.109 (0.026)
	PTCA	0.353 (0.027)	1.341 (0.021)
	CATH	0.401 (0.029)	0.846 (0.023)
	Other	0.489 (0.032)	0.774 (0.025)
Scale σ			0.450 (0.005)
Random Effect θ		0.290 (0.004)	

Note:CCI=Charlson Comorbidity Index.
Reference group: Gender=male, Procedure=none, CCI=1.

Table 3.3: Estimates of coefficients without random effects

Variable		CPH-Lognormal	
		LOS (CPH)	Cost (LN)
		Estimate (SE)	Estimate (SE)
Intercept		0.926 (0.021)	9.190 (0.020)
AGE		0.008 (0.0005)	0.003 (0.0005)
Gender	Female	0.092 (0.014)	0.012 (0.014)
CCI	CCI=2	0.261 (0.017)	0.155 (0.016)
	CCI=3	0.429 (0.021)	0.258 (0.019)
	CCI≥4	0.618 (0.021)	0.370 (0.020)
Procedure	CABG	1.228 (0.003)	2.084 (0.025)
	PTCA	0.265 (0.021)	1.285 (0.020)
	CATH	0.346 (0.023)	0.772 (0.022)
	Other	0.518 (0.026)	0.663 (0.022)
Scale σ			0.698 (0.004)

Note:CCI=Charlson Comorbidity Index.
Reference group: Gender=male, Procedure=none, CCI=1.

3.5 Discussion

In this chapter we propose a novel joint model for correlated LOS and cost data. Our study demonstrates the application of joint modeling of correlated LOS and total cost for hospitalized patients. The underlying distributions, with *conditional independence* assumption are Coxian PH for LOS and log-normal for cost. Gaussian quadrature technique implemented in SAS Proc NLMIXED is used for ML estimation.

Our model is comprehensive yet easy to fit. These advantages make it valuable in practical data analysis. The strength of Coxian PH regression models for LOS lies in their flexibility in accommodating extreme values, while revealing hidden status possibly due to the presence of other comorbid conditions. We find a Coxian PH model serves well for LOS. Cost easily fitted through a log-normal regression model. The correlation between them can be accommodated by shared random effects. Joint modeling approach is primarily necessary when researchers are more interested in the association of different two outcomes.

The management of hospital LOS and cost has become an important issue, since the determination of relevant factors and hidden structures could inform discharge planning and allocation of resources. In the application, we showed that there exists significant association between LOS and cost at the subject level. Moreover, this approach can be extended to analyze longitudinal data or clustered data exhibiting unobserved heterogeneity. If the data are incomplete, for example, when patients who die in hospital, this joint model can be generalized to deal with censored observations by modifying the log-likelihood function (Liu et al., 2007).

Chapter 4

Bivariate Copula Random-effects (BCRE) Model for Length of Stay and Cost

This chapter is concerned with regression models for correlated outcomes constructed using copula functions and correlated random effects. Our approach entails specifying conditional marginal regression models with random effects for the outcomes and combining them to form a joint model via a specified copula.

4.1 Introduction

As we mentioned in Chapter 3, mixed outcomes have attracted more attention in health and medicine, and joint analysis of such outcomes entails specification of models flexible enough to accommodate them. Such joint models are potentially advantageous in several statistical and practical respects. For example, a multivariate model enables analysts to account for relationships between outcomes and assess at the same time the joint influence of predictors/covariates have on them.

Multilevel models (also called hierarchical linear models, nested models, mixed models, random-effects models, random-coefficient models, or split-plot designs) are statistical models for addressing variation at more than one level. They can be viewed as generalizations of linear models.

The multivariate normal distribution is by far the most commonly used to model multivariate outcomes. However, multivariate normality may not be a proper assumption in many situations. For example, if two outcomes are positive and skewed, the log-normal or log-logistic regression models might be more appropriate. This is especially true for LOS and cost where positive right skewness and correlation are present. In the recent years, some attempts have been made to relate those two outcomes that permit consideration of the correlation between LOS and medical charges. Gardiner et al. (2002) propose a two-equation model for total cost and duration of treatment with the endogeneity of the later accounted for in the model for cost. In their model, correlation is assessed either under the assumption of bivariate normal for measurement error or by the "seemly unrelated regression". Often different survival distributions give better fit to LOS and cost, for example, a log-logistic LOS and log-normal cost. The new approach that we consider here provides a more general and flexible model for the related variables.

Our approach is based on copula dependence modeling. Copula functions are useful tools to model dependence for multivariate outcomes. There is an increasing use of copulas in several scientific fields, such as economics (Cameron et al., 2004), survival analysis (Lambert and Vandenhende, 2002), finance (Bee, 2004; Breymann et al., 2003) and insurance (Klugman and Parsa, 1999). A copula is a function that connects the marginal distributions to restore the joint distribution with a dependence parameter. For example, multivariate Gaussian distribution can be generated by a Gaussian copula applied to the Gaussian marginal distributions. More importantly, a copula can provide many flexible, complicated non-Gaussian joint distributions.

The context of our application of copula models is LOS and cost. They are likely to have different marginal distributions being correlated. In addition to the correlation for LOS and cost for each patient, another potential correlation exists at the cluster (hospital) level. For example, unmeasured latent variables (hospital efficiency, provider characteristics, etc.) induce a random effect shared among patients within the same cluster that affects the outcomes and results in a within-cluster correlation. We can integrate out this latent random effect to derive a new marginal distribution for the outcomes, but the forms are typically not closed and complicated. In fact, the random effect (RE) approach is intuitive to researchers who believe that there may be some latent quantity underlying a cluster's LOS and cost. The random-effects models have been used extensively in clustered and longitudinal data analysis, see Allison (2005); Chen and Dunson (2003); Wooldridge (2002). For example, Chen and Dunson (2003) propose an approach for random-effects model selection and apply it to study the relationship between prenatal exposure to polychlorinated biphenyls and motor development in young children with possible heterogeneity among the 12 centers.

In this chapter, we develop a new flexible joint model based on correlated measurement errors modeled by copulas and incorporate a cluster level random effect to account for individual and within-cluster correlations simultaneously. The proposed approach tries to capture the various dependence structures of LOS and cost (symmetric or asymmetric) in the copula function, and takes advantage of the relative ease in specifying the marginal distributions and introduction of within-cluster correlation based on the cluster level random effects.

4.2 Copulas and Dependence: a Brief Overview

The study of copula functions was initiated in the 1940s by Hoeffding, and further developed by Fréchet. Model theories about copulas were introduced by Sklar in 1959 (Sklar, 1959). Nelsen (1999); Joe (1997); Trivedi and Zimmer (2005) provide a comprehensive discussion of copulas and their applications. Copulas are parametrically specified joint distributions generated from given marginal distributions. Therefore, properties of copulas are analogous to properties of joint distributions. In this section we provide some fundamental properties of copulas. The copula approach is a modeling strategy whereby a joint distribution is induced by specifying marginal distributions and a copula function with a dependence parameter. Sklar's theorem states that there exists a copula function which acts to represent the joint cumulative distribution functions (CDF) of random variables in terms of its underlying one dimensional margins. An m-copula can be defined as an m-dimensional CDF whose support is contained in $[0,1]^m$ and whose one-dimensional margins are uniform on $[0,1]$. Consider a continuous m-variate distribution $F(y_1, y_2, \ldots, y_m)$ with univariate marginal distributions $F_1, F_2, \ldots, F_m$ and inverse functions $F_1^{-1}, F_2^{-1}, \ldots, F_m^{-1}$. Let $u_1, u_2, \ldots, u_m \in [0,1]$, $y_1 =$

$F_1^{-1}(u_1), y_2 = F_2^{-1}(u_2), \ldots, y_m = F_m^{-1}(u_m)$. Then

$$\begin{aligned} F(y_1, y_2, \ldots, y_m) &= F\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_m^{-1}(u_m)\right) \\ &= \Pr\left(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_m \leq u_m\right) \\ &= C(u_1, u_2, \ldots, u_m) \end{aligned} \tag{4.1}$$

where $(U_1, U_2, \ldots, U_m)$ are each marginally uniform on $[0,1]$. C is the unique copula associated with the distribution function F. Here for simplicity, we discuss the bivariate copulas to illustrate copula modeling approach. Consider two random variables X and Y with continuous distributions F and G, respectively, and joint distribution function H.

Theorem 4.1 (Sklar's theorem)**.** *Let H be a joint distribution function with margins F and G, then there exists a 2-dimensional copula function $C_\theta : [0,1]^2 \rightarrow [0,1]$ such that*

$$H(x, y) = C_\theta(F(x), G(y))$$

where θ is the dependence parameter, which measures dependence between the margins.

Assume C_θ, F, G are differentiable, we can write down the joint density function of X and Y in the form:

$$h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y} = \frac{\partial^2 C_\theta(F(x), G(y))}{\partial x \partial y} = c_\theta(u_1, u_2) f(x) g(y) \tag{4.2}$$

where $u_1 = F(x), u_2 = G(y)$, $c_\theta(u_1, u_2) = \dfrac{\partial^2 C_\theta(u_1, u_2)}{\partial u_1 \partial u_2}$ and f, g are the marginal univariate densities.

Three bivariate copulas of importance are

1. Independent (Product) Copula : $\Pi(u, v) = uv$

2. Maximum Copula: $W(u, v) = \max(u + v - 1, 0)$

3. Minimum Copula: $M(u, v) = \min(u, v)$

where $(u, v) \in [0, 1]^2$. W and M are called the Fréchet lower and upper bounds and have the property that all bivariate copulas C satisfy $W \leq C \leq M$ (Smith, 2003).

Family of copulas C_θ are indexed by the association parameter θ which captures dependence between the random variables interested since copula parameters have different ranges and interpretations, they are not comparable to each other. Here, we present three dependence concepts.

1. Pearson correlation. The Pearson correlation coefficient $\rho(X, Y)$ for two random variables is a measure of linear dependence, given by

   $$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

   The Pearson correlation coefficient is the most popular measure of the linear association. For elliptical copulas (such as Gaussian copula, ρ appears naturally as the . An alternative measure of dependence is rank correlation, including Kendall's tau and Spearman's rho, which are measures of concordance.

2. Kendall's tau and Spearman's rho. Kendall's tau and Spearman's rho are associations between rankings instead of the actual values of the observations. Hence Kendall's tau

is an alternative measure of association for non-elliptical distributions.

$$\begin{aligned}\tau(X,Y) &= \Pr[\text{concordance}] - \Pr[\text{discordance}] \\ &= \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]\end{aligned}$$

where (X_1, Y_1) and (X_2, Y_2) are two independent pairs of random variables (X, Y) from H.

Spearman's rho is the linear correlation between $F(X)$ and $G(Y)$ defined as:

$$\rho_S(X,Y) = \rho(F(X), G(Y))$$

Both $\tau(X,Y)$ and $\rho_S(X,Y)$ can be expressed in term of copulas:

$$\begin{aligned}\tau(X,Y) &= 4E\left[C(U_1,U_2)\right] - 1 = 4\int_0^1\int_0^1 C(u_1,u_2)dC(u_1,u_2) - 1 \\ \rho_S(X,Y) &= 12E[U_1U_2] - 3 = 12\int_0^1\int_0^1 u_1u_2dC(u_1u_2) - 3\end{aligned}$$

where $(U_1, U_2) \sim C$ with uniform margins.

For continuous variables the above measures take value [-1,1]. For the independent copula Π, both measures are 0. For the Fréchet lower bound W, both measures are -1. For the Fréchet upper bound M, both measures are 1. From the above expressions for $\tau(X,Y), \rho_S(X,Y)$ both measures depend on the copula of the joint distribution and not on the margins. They are also invariant with respect to strictly increasing transformation of (X,Y). Recall that the Pearson correlation is not a measure of independence: for example, $\rho(X,Y) = 0$ does not imply independence of the two variables. Table **??** below gives the functional form of some selected copulas.

For simplicity, we here write down copulas in terms of random variables U_1 and U_2 that have standard uniform marginal distributions. Table **??** summarizes several bivariate copula functions in terms of U_1 and U_2.

Farlie-Gumbel-Morgenstern (FGM) copula

The FGM copula is defined as

$$C_\theta(u_1, u_2) = u_1 u_2 \left(1 + \theta(1 - u_1)(1 - u_2)\right), \theta \in [-1, 1].$$

The joint distribution of (X, Y) is

$$H(x, y) = F(x)G(y)\left(1 + \theta(1 - F(x))(1 - G(y))\right) \tag{4.3}$$

X and Y are independent as $\theta = 0$, and the FGM copula collapses to independent copula. The FGM copula does not caputre the Fréchet lower & upper bounds as special cases. The joint density of the FGM coula is

$$h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y} = \left(1 + \theta(1 - 2F(x))(1 - 2G(y))\right) f(x)g(y) \tag{4.4}$$

The FGM copula is attractive due to its simplicity but it can only be useful when dependence between the two margins is modest.

Gaussian Copula

The Gaussian copula is defined by

$$C_\theta(u_1, u_2) = \Phi_2\left\{\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\right\}, \theta \in [-1, 1].$$

The joint distribution of (X, Y) is

$$H(x, y) = \Phi_2\left\{\Phi^{-1}[F(x)], \Phi^{-1}[G(y)]; \theta\right\} \tag{4.5}$$

where $\Phi(.)$ is the CDF of the standard normal distribution, and $\Phi_2(.)$ is the standard bivariate normal CDF with correlation parameter θ. X and Y are independent if $\theta = 0$. The Gaussian copula contains the Fréchet bounds: for $\theta = -1$ we get W, and for $\theta = 1$ we get M. The density of H is

$$h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y} = \phi_2\left\{\Phi^{-1}[F(x)], \Phi^{-1}[G(y)]; \theta\right\} \frac{f(x)}{\phi(\Phi^{-1}[F(x)])} \frac{g(x)}{\phi(\Phi^{-1}[G(x)])} \tag{4.6}$$

where ϕ_2 is the density of bivariate standard normal with correlation θ, i.e.

$$\phi_\theta(x, y) = \frac{1}{2\pi\sqrt{1 - \theta^2}} \exp\left\{-\frac{x^2 - 2\theta xy + y^2}{2(1 - \theta^2)}\right\}$$

$\phi(.)$ is the standard normal density, $f(x)$ and $g(y)$ are densities of X and Y, respectively. The higher the association parameter θ , the stronger the dependence. We can see the dependence structure is symmetric, see left top in Figure 4.1.

Clayton Copula

Clayton copula has the form

$$C_\theta(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}, \theta \in (0, \infty).$$

The corresponding joint distribution function and density function are

$$H(x, y) = \left(F(x)^{-\theta} + G(y)^{-\theta} - 1\right)^{-1/\theta} \tag{4.7}$$

$$h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y} = (1+\theta)[F(x)G(y)]^{-\theta-1}\left(F(x)^{-\theta} + G(y)^{-\theta} - 1\right)^{-1/\theta-2} f(x)g(y) \tag{4.8}$$

X and Y are independent as $\theta \to 0$. When correlation between two variables is strongest in the left tail of the joint distribution, Clayton copula is an appropriate choice.

Frank Copula

The Frank copula takes the form:

$$C_\theta(u_1, u_2) = -\theta^{-1} \log\left\{1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right\}$$

The corresponding distribution function is

$$H(x, y) = -\theta^{-1} \log\left\{1 + \frac{\exp(-\theta F(x)) - 1)(\exp(-\theta G(y)) - 1)}{\exp(-\theta) - 1}\right\} \tag{4.9}$$

where dependence parameter $\theta \in (-\infty, \infty) \setminus \{0\}$. Values $-\infty$, and ∞ correspond (limit) to the Fréchet lower bound and upper bound, respectively. We also have $H(x, y) = F(x)G(y)$

if $\theta \rightarrow 0$. The left bottom panel of Figure 4.1 shows the strongest dependence is centered in the middle of distribution.

Gumbel Copula

The Gumbel copula has the form:

$$C_\theta(u_1, u_2) = \exp\left\{-[(-\log u_1)^\theta + (-\log u_2)^\theta]^{1/\theta}\right\}, \theta \in [1, \infty].$$

The joint distribution function is

$$H(x, y) = \exp\left\{-[(-\log F(x))^\theta + (-\log G(y))^\theta]^{1/\theta}\right\} \tag{4.10}$$

X and Y are independent as $\theta = 1$. Gumbel copula exhibits strong right tail dependence, while does not allow negative dependence.

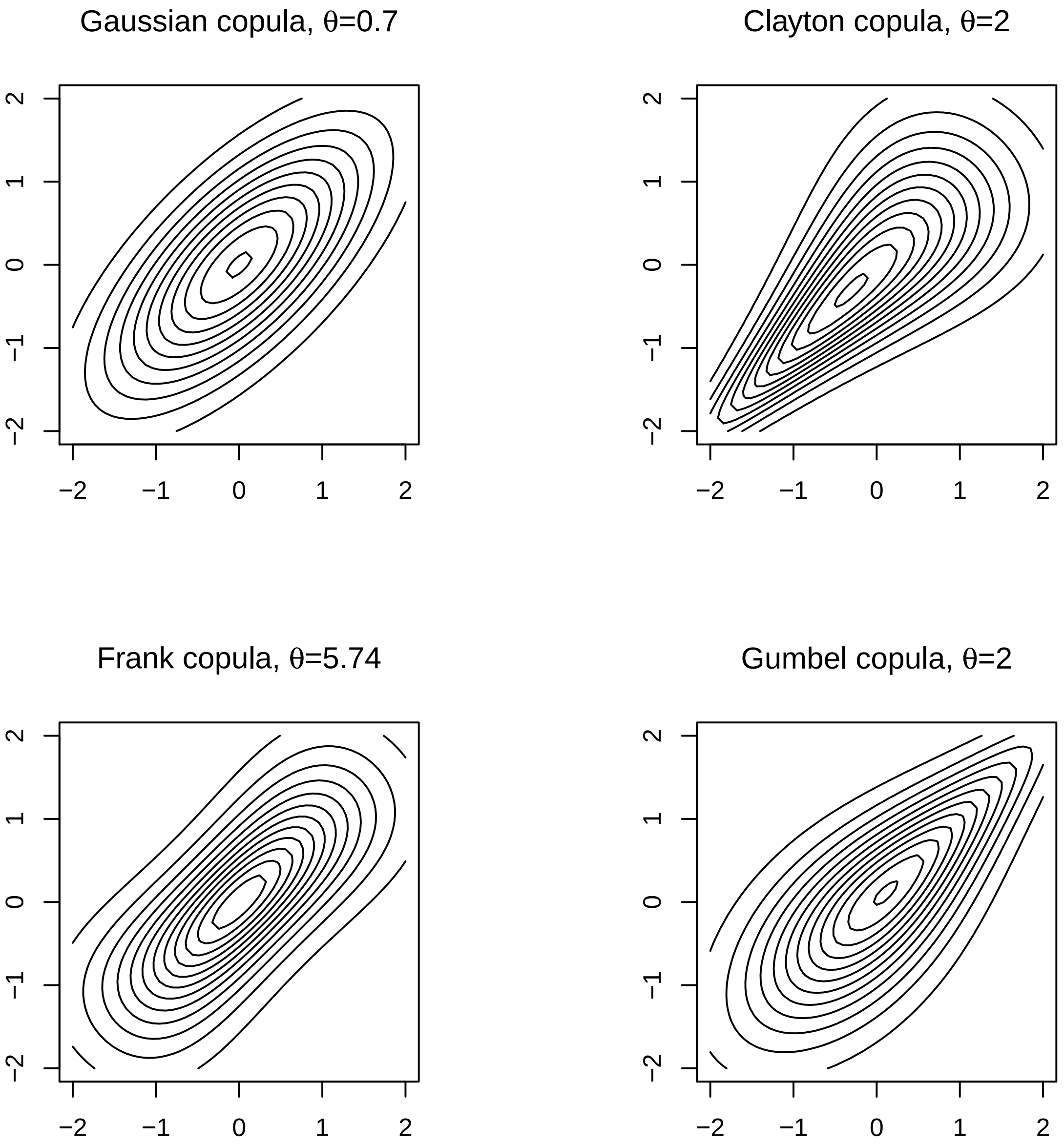


Figure 4.1: Bivariate pdf contour plots induced by copula, $N(0,1)$ margins (Kendall's tau=0.5)

Table 1: Selected examples of copulas

Copula	Copula function $C_\theta(u_1,u_2)$	Density $c_\theta(u_1,u_2)$	θ-domain	Kendall's τ
FGM[a]	$u_1u_2(1+\theta(1-u_1)(1-u_2))$	$1+\theta(1-2u_1)(1-2u_2)$	$-1\leq\theta\leq 1$	$\frac{2}{9}\theta$
Gaussian	$\Phi_\theta\left\{\Phi^{-1}(u_1),\Phi^{-1}(u_2)\right\}$	$\dfrac{\phi_2\left\{\Phi^{-1}(u_1),\Phi^{-1}(u_2);\theta\right\}}{\phi(\Phi^{-1}(u_1))\phi(\Phi^{-1}(u_2))}$	$-1\leq\theta\leq 1$	$\frac{2}{\pi}\arcsin(\theta)$
Clayton	$\left(u_1^{-\theta}+u_2^{-\theta}\right)^{-1/\theta}$	$(1+\theta)[u_1u_2]^{-\theta-1}\left(u_1+u_2-1\right)^{-1/\theta-2}$	$\theta\in(0,\infty)$	$\frac{\theta}{\theta+2}$
Frank	$-\theta^{-1}\log\left\{1+\dfrac{(\exp(-\theta u_1)-1)(\exp(-\theta u_2)-1)}{\exp(-\theta)-1}\right\}$	$\dfrac{-\theta(e^{-\theta}-1)e^{-\theta(u_1+u_2)}}{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)+(e^{\theta}-1)^2}$	$\theta\in(-\infty,\infty)$ $\backslash\{0\}$	$1-\frac{4}{\theta}[1-D_1(\theta)]$[b]
Gumbel	$\exp\left\{-\left[(-\log u_1)^{\theta}+(-\log u_2)^{\theta}\right]^{1/\theta}\right\}$	$C_\theta(u_1,u_2)(u_1u_2)^{-1}\dfrac{(\tilde{u}_1\tilde{u}_2)^{\theta-1}}{(\tilde{u}_1^{\theta}+\tilde{u}_2^{\theta})^{2-1/\theta}}\left[(\tilde{u}_1^{\theta}+\tilde{u}_2^{\theta})^{1/\theta}+\theta-1\right]$ where $\tilde{u}_1=-\log u_1$, and $\tilde{u}_2=-\log u_2$	$\theta\in[1,\infty)$	$1-\frac{1}{\theta}$

[a] FGM denotes Farlie-Gumbel-Morgenstern.

[b] the density function $D_k(z)=kz^{-k}\int_0^z t_k(e^t-1)^{-1}dt$ for k any positive integer.

4.3 Bivariate Copula Random-effects (BCRE) Model

4.3.1 BCRE Model and Likelihood

In this section, we present our model for the hospital length of stay (LOS) and cost at two levels. We define notation as follows. Suppose that we observe LOS T_{ij} and cumulative cost $C_{ij} = C(T_{ij})$ for the jth subject in ith hospital, where $i = 1, 2, \cdots, n, j = 1, 2, \cdots, n_i$. Denote by $\mathbf{x}_{1,ij}$, $\mathbf{x}_{2,ij}$ the covariate vectors for the fixed effect, specific to the type of outcome. In practice, they may represent the same common covariate constellation. Let u_i and v_i be the correlated random effects at the hospital level with joint density $\phi(u_i, v_i)$, that is $\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N(\mathbf{0}, \Sigma)$, with $\Sigma = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}$ being a positive definite matrix, $\rho \in [-1, 1]$.

Denote by $\epsilon_{1,ij}$ and $\epsilon_{2,ij}$ the correlated measurement error terms for the positive values of T_{ij} and C_{ij}. We assume that measurement errors $\epsilon_{1,ij}$ and $\epsilon_{2,ij}$ are independent of random effects u_i and v_i, but jointly distributed with some specified marginal distributions (normal, logistic, etc.) from a particular copula with dependence parameter θ, for example, assume $\epsilon_{1,ij} \sim N(0, 1)$ and $\epsilon_{2,ij} \sim N(0, 1)$. Our Bivariate copula random-effects (BCRE) model is defined as

$$\begin{aligned} \log T_{ij} &= \mathbf{x}'_{1,ij}\boldsymbol{\beta}_1 + u_i + \sigma_1\epsilon_{1,ij} \\ \log C_{ij} &= \mathbf{x}'_{2,ij}\boldsymbol{\beta}_2 + v_i + \sigma_2\epsilon_{2,ij} \end{aligned} \tag{4.11}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are regression coefficient vectors, respectively, σ_1 and σ_2 are two scale parameters. The joint distribution of random effects defines the correlation (between, and cross equation) among clustered measures (T_i, C_i). In one aspect, u_i and v_i model intra-hospital correlation, which induces the "within-hospital correlation", i.e. they generate dependence between those individuals in the same cluster (hospital), whereas conditional on the random effects those individuals are independent. In another aspect, the correlation between u_i and v_i describes the "cross-equation correlation" at the hospital level. There might also be "cross-equation correlation" at the individual level between the two equations, which is modeled by copula dependence parameter θ. For example, patients with longer hospital length of stay tend to incur more cost. Also, patients with higher cost are more likely to be hospitalized.

Under the assumptions that $\{(\epsilon_{2,ij}, \epsilon_{2,ij}), 1 \leq i \leq n, 1 \leq j \leq n_i\}$ are correlated and $\{(\epsilon_{m,ij}, \epsilon_{m,ik}), m = 1, 2, j \neq k\}$ are uncorrelated, the variances and covariances are easily calculated

$$\begin{aligned}
Var(\log T_{ij}) &= Var(u_i) + \sigma_1^2 Var(\epsilon_{1,ij}) = \tau_1^2 + \sigma_1^2 \\
Var(\log C_{ij}) &= Var(v_i) + \sigma_2^2 Var(\epsilon_{2,ij}) = \tau_2^2 + \sigma_2^2 \\
Cov(\log T_{ij}, \log T_{ik}) &= Var(u_i) + \sigma_1^2 Cov(\epsilon_{1,ij}, \epsilon_{1,ik}) = \tau_1^2 \\
Cov(\log C_{ij}, \log C_{ik}) &= Var(v_i) + \sigma_2^2 Cov(\epsilon_{2,ij}, \epsilon_{2,ik}) = \tau_2^2 \\
Cov(\log T_{ij}, \log C_{ij}) &= Cov(u_i, v_i) + \sigma_1 \sigma_2 Cov(\epsilon_{1,ij}, \epsilon_{2,ij}) \\
&= \rho \tau_1 \tau_2 + \sigma_1 \sigma_2 Cov(\epsilon_{1,ij}, \epsilon_{2,ij}) \\
Cov(\log T_{ij}, \log C_{ik}) &= Cov(u_i, v_i) + \sigma_1 \sigma_2 Cov(\epsilon_{1,ij}, \epsilon_{2,ik}) = \rho \tau_1 \tau_2
\end{aligned}$$

where $i = 1, 2, \ldots, n, j, k = 1, 2, \ldots, n_i, j \neq k$. Such equations are often necessary by identifying all the variance/covariance components.

Likelihood

Denote by $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})$ and $\mathbf{c}_i = (c_{i1}, c_{i2}, \ldots, c_{in_i})$ the vectors of response variables, LOS and cost, for the i-th unit (cluster), $i = 1, 2, \ldots, n$. The likelihood for $(\mathbf{t}_i, \mathbf{c}_i)$ is

$$L_i = \int \int \prod_j C_\theta \left\{ F(\frac{\log t_{ij} - \mathbf{x}'_{1,ij}\boldsymbol{\beta}_1 - u_i}{\sigma_1}), G(\frac{\log c_{ij} - \mathbf{x}'_{2,ij}\boldsymbol{\beta}_2 - v_i}{\sigma_2}) \right\}$$

$$\frac{1}{\sigma_1 t_{ij}} f(\frac{\log t_{ij} - \mathbf{x}'_{1,ij}\boldsymbol{\beta}_1 - u_i}{\sigma_1}) \frac{1}{\sigma_2 c_{ij}} g(\frac{\log c_{ij} - \mathbf{x}'_{2,ij}\boldsymbol{\beta}_2 - v_i}{\sigma_2}) p(u_i, v_i | \tau_1, \tau_2, \rho) du_i dv_i \tag{4.12}$$

where F and G are CDF of ϵ_1 and ϵ_2, similarly, f and g are PDF of ϵ_1 and ϵ_2, separately. The whole estimation process can be conveniently implemented in SAS.

4.3.2 Estimation

The EM algorithm (Dempster. et al., 1977) is commonly used for estimation in joint random-effects models, with random effects treated as missing data. However, for the above models, the conditional expectations of random effects given observed data do not have a closed form. Monte Carlo methods (e.g. the Metropolis–Hastings algorithm) are often needed in the E-step to approximate these terms, making the estimation highly computationally intensive. Furthermore, the implementation (programming) of the Monte Carlo EM method is quite difficult and must be treated case by case. Therefore, joint models are not yet widely adopted in practical data analysis.

Likelihood (4.12) involves an integral with respect to random effects. Numerical integration techniques, e.g. Gaussian quadrature, thus can be adopted for estimation. The resulting parametric likelihood can be maximized conveniently by Gaussian quadrature tools in standard statistical packages such as Proc NLMIXED in SAS. An introduction to the Gaussian quadrature technique and implementation in SAS is given in Appendix A.

Generally, we have two estimation methods when using copulas. Full maximum likelihood (FML) approach is the most direct estimation method. To obtain FML estimates, one maximizes the loglikelihood function $l(\Omega_{\text{FML}}) = \sum_{i=1}^{n} L_i$ where L_i is obtained from Eq (4.12) and Ω denotes all parameters. By standard likelihood theory under regularity conditions, the ML estimates $\widehat{\Omega}_{\text{FML}}$ is consistent for the true parameter vector Ω_0 and retains its asymptotic normality (Trivedi and Zimmer, 2005). Joe and Xu (1996) propose a two-stage estimation method called inference function for margins (IFM). The set of parameters of the model are estimated through a (nonlinear) system of estimating equations, with each estimating equation being a score function (partial derivative of a loglikelihood) from some marginal distribution of the multivariate model. However, the standard error of $\widehat{\Omega}_{\text{IFM}}$ is not appropriately taken care of, which is typically solved by bootstrapping method. In this chapter, we apply the FML with Gaussian quadrature for the model estimation. By simulations, we found that such method yields satisfactory estimates.

4.4 Simulation

In this section, we conduct a simulations study to evaluate the performance of the proposed estimation method. Data are simulated under a regression model with random effects. In each simulated data set, $n = 30$ clusters and $n_i = 20$ observations are fixed within each cluster, which gives 600 observations per simulated data set. The covariate vectors $\mathbf{x}_{1,ij}$ and $\mathbf{x}_{2,ij}$ consist of the constant 1 (representing the intercept term) and values randomly generated from the uniform $(-0.5,\, 0.5)$ distribution, with associated regression coefficients

(1, -0.5) for the LOS measure, and (4, 0.5) for the cost measure. The correlation of the random effects ρ is chosen to be 0.5. We assume that measurement errors $\epsilon_{1,ij}$ and $\epsilon_{2,ij}$ have normal marginal distributions. The true value of the association parameter θ for each copula is chosen so that Kendall's tau is approximately equal to 0.3, except for the FGM copula, which cannot accommodate dependence of this magnitude. Alternatively in the FGM case, we set $\theta = 0.5$, with corresponding Kendall's tau value of 0.11. According to this simulation design, 30 realizations of the random cluster effects from a normal distribution are generated for each outcome. The simulation study is designed to evaluate the performance of the estimators. The number of replications is 500 for each setting considered. For each replication, new realizations of $\epsilon_{1,ij}$ and $\epsilon_{2,ij}$ are randomly drawn from each type of copula, which results in the new observations of LOS and cost.

Results of the simulation study are presented in Table 4.2, reporting the average estimate and mean standard errors of the parameter estimates over the 500 replicates. It is evident that the FML estimators of the regression coefficients have negligible biases and relatively small MSE. For variance component parameters, the FML estimator also performs reasonably well in each setting considered. The simulation results thus confirm the applicability of the FML for parameter estimation in the correlated regression model with random effects.

Table 2: Simulation results for bivariate copulas with normal margins

		$\rho = 0.5$														
		Gaussian Copula			Clayton Copula			Frank Copula			Gumbel Copula			FGM Copula		
		$(\theta = 0.45)$			$(\theta = 0.86)$			$(\theta = 5.40)$			$(\theta = 1.43)$			$(\theta = 0.50)$		
	DGP	Est	SE	SEM	Est	SE	SEM	Est	SE	SEM	Est	SE	SEM	Est	SE	SEM
β_{01}	1.00	0.999	0.152	0.149	1.001	0.151	0.148	1.000	0.149	0.147	1.001	0.147	0.151	1.000	0.149	0.147
β_{11}	-0.50	-0.503	0.120	0.120	-0.502	0.124	0.121	-0.499	0.118	0.127	-0.498	0.121	0.124	-0.499	0.162	0.157
β_{02}	4.00	3.997	0.247	0.246	3.991	0.227	0.228	3.990	0.239	0.224	3.988	0.235	0.227	3.989	0.231	0.222
β_{12}	0.50	0.500	0.267	0.270	0.500	0.260	0.241	0.502	0.258	0.256	0.502	0.271	0.260	0.501	0.262	0.257
σ_1	1.00	1.001	0.030	0.031	0.999	0.031	0.029	0.996	0.030	0.029	0.996	0.031	0.031	1.002	0.030	0.028
σ_2	2.00	1.999	0.059	0.059	1.998	0.059	0.057	1.997	0.057	0.058	1.995	0.059	0.059	2.004	0.061	0.058
τ_1	0.80	0.775	0.113	0.113	0.778	0.116	0.108	0.777	0.111	0.110	0.778	0.113	0.116	0.776	0.111	0.112
τ_2	1.20	1.166	0.180	0.176	1.167	0.175	0.170	1.164	0.180	0.175	1.160	0.176	0.175	1.162	0.183	0.179
ρ	0.50	0.502	0.151	0.147	0.504	0.149	0.147	0.505	0.159	0.154	0.503	0.147	0.149	0.502	0.147	0.149
θ		0.452	0.154	0.156	0.863	0.223	0.231	5.409	0.337	0.340	1.434	0.156	0.162	0.504	0.503	0.041

Est is the mean of the parameter estimates (based on 500 replicates); SE is the sampling standard error of the parameter estimates; SEM is the sampling mean of the standard error estimates.

4.5 Application to 2003 NIS AMI Hospitalized Patients

We apply our method discussed above to the 2003 NIS for LOS and charges for patients hospitalized for AMI. There is a high correlation between LOS and cost (rank correlation $r = 0.576$ and Kendall's tau $\tau = 0.431$, $N = 11,749$). The range of the LOS was 1 to 142 days, and cost ranged from \$90 and \$962,611. Figure 4.2 show a plot of log(LOS) and log(cost).

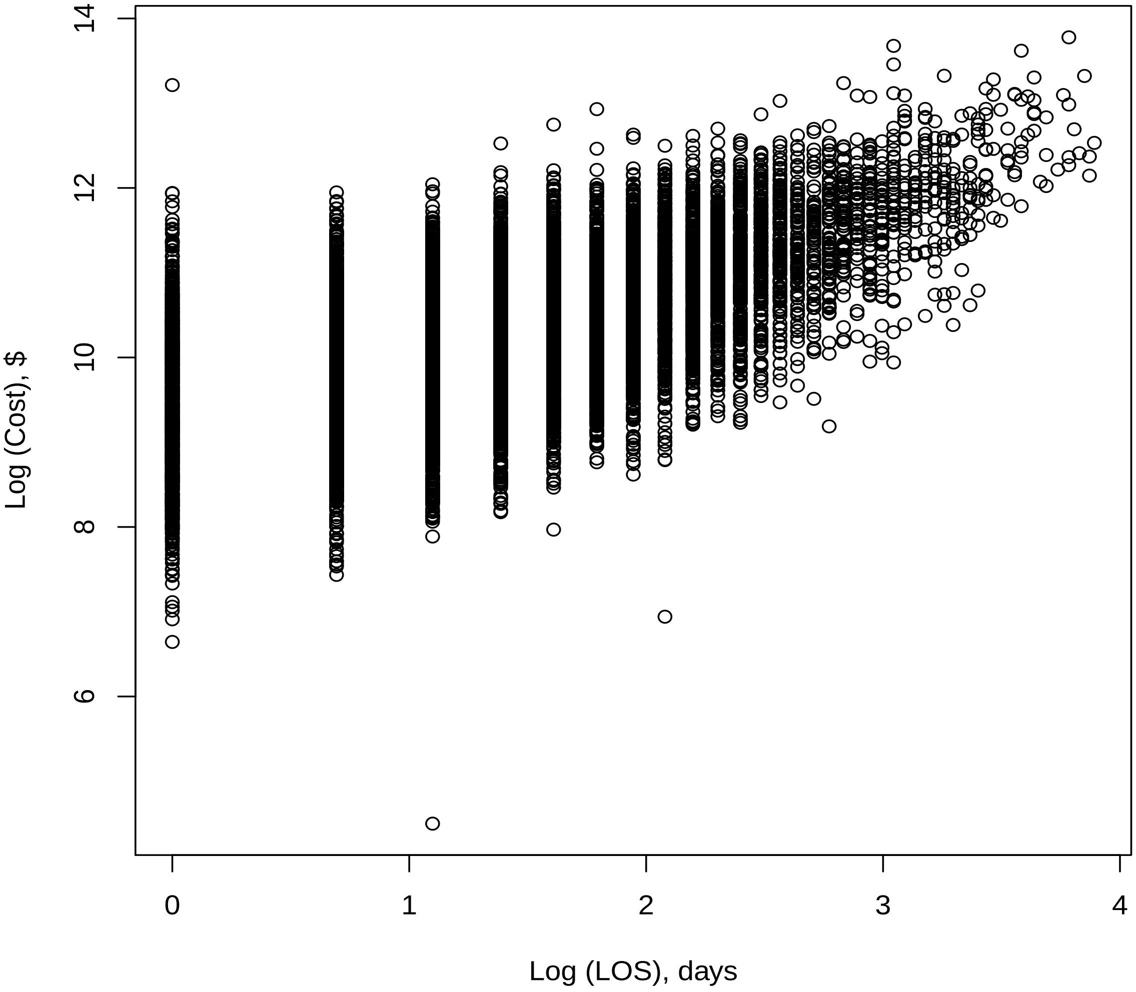


Figure 4.2: Plot of Log (LOS) vs. Log (Cost), truncated at 50 days

The bivariate model we use for LOS and cost allow inference about regression parameters simultaneously for these two outcomes. Our method accounts for the existence of both the non-zero intra-hospital and individual level correlations. For comparison, we also fit two reduced models as special cases. A reduced model A is fit without random effects and a reduced model B is fit with independent measurement errors. The covariates of interest include age, gender, comorbidity and procedure type. To correct for skewness to the right of LOS and cost, we take the logarithm of the outcomes. Level 1 is the subject level, Level 2 is the hospital level. We use the adaptive Gaussian quadrature estimation method with five quadrature points in Proc NLMIXED. Different starting values are used in the estimation. We fit with three different copulas (Gaussian, Clayton and Gumbel) and chose the best one based on BIC. We conclude that in our application, the Gumbel copula performs the best. We include the results for covariates with Gumbel copula in Table 4.3.

Table 4.3 shows that comorbidities and procedure types are highly significant in both the LOS and cost of the model. CABG patients have significant longer LOS and higher cost. Subjects with high comorbidity (CCI$\geq$4) are more likely to stay longer and incur more cost at a rate of 1.72 and 1.41 respectively, compared to those with low comorbidity CCI=1 ($p < 0.0001$). Patient age has a significant effect in both equations of the model, but with very small effect size ($\hat{\beta}_{1,\text{age}} = 0.0008, \hat{\beta}_{2,\text{age}} = 0.0003, p < 0.0001$). Gender has a significant effect in LOS, but there is no difference between males and females in cost.

Bottom of Table 4.3 shows the estimates of the covariance matrix and copula association parameter. We note that both the random effects are present for LOS and cost, as τ_1 and τ_2 are highly significant. A significant cross-equation correlation is seen ($\hat{\rho}_{uv} = 0.615, p < 0.0001$), suggesting that a hospital with higher LOS also has higher cost. At the same time, there is an association between LOS and cost at the patient level, $\hat{\theta} = 2.051 (p < 0.0001)$, resulting in the estimates of Kendall's tau=0.512. Results for the Gaussian and Clayton copulas are shown in Table 4.6 and Table 4.7.

For comparison, we fit two additional models. Reduced model A assumes that there are no random effects at the hospital level. In this model, we only consider association between LOS and cost at the patient level via measurement errors connected with a copula function. Results are shown in Table 4.4. We observe that this model yields the estimates of age, gender, comorbidities and procedure type close to those in our model. However some notable difference are present. We note a larger variance σ^2 for both LOS and cost, which might be due to ignoring the heterogeneity at the hospital level. Reduced model B only involves the random effects, assuming independent measurement errors. Table 4.5 shows the estimates for model B. We see that the estimate $\hat{\rho}_{uv}$ is higher than that in our model possibly transferring some correlation to the hospital level.

Table 4.3: Parameter estimates (SE) for log-normal margins with Gumbel copula

		LOS		**Cost**	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Intercept		0.753 (0.019)	< 0.0001	9.131 (0.018)	< 0.0001
Age		0.008(0.0005)	< 0.0001	0.003(0.0004)	< 0.0001
Gender	Female	0.071 (0.012)	< 0.0001	0.001 (0.011)	0.922
Procedure	CABG	1.236 (0.023)	< 0.0001	1.928 (0.021)	< 0.0001
	PTCA	0.270 (0.019)	< 0.0001	1.189 (0.018)	< 0.0001
	CATH	0.314 (0.021)	< 0.0001	0.704 (0.019)	< 0.0001
	Other	0.366 (0.022)	< 0.0001	0.672 (0.020)	< 0.0001
CCI	2	0.096 (0.018)	< 0.0001	0.131 (0.013)	< 0.0001
	3	0.337 (0.019)	< 0.0001	0.222 (0.016)	< 0.0001
	≥ 4	0.542 (0.018)	< 0.0001	0.343 (0.016)	< 0.0001
Scale (σ)		0.649 (0.004)	< 0.0001	0.564 (0.004)	< 0.0001
Random Effects (τ)		0.129 (0.007)	< 0.0001	0.424 (0.007)	< 0.0001
Correlation (ρ)		0.615 (0.011)		< 0.0001	
θ		2.051 (0.020)		< 0.0001	
BIC			310631		

Note:CCI=Charlson Comorbidity Index.

Reference group: Gender=male, Procedure=none, CCI=1.

Table 4.4: Parameter estimates (SE) for log-normal margins with Gumbel copula, reduced model A without random effects

		LOS		**Cost**	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Intercept		0.758 (0.019)	< 0.0001	9.203 (0.018)	< 0.0001
Age		0.008(0.0005)	< 0.0001	0.003(0.0005)	< 0.0001
Gender	Female	0.078 (0.013)	< 0.0001	0.004 (0.011)	0.787
Procedure	CABG	1.287 (0.027)	< 0.0001	2.052(0.026)	< 0.0001
	PTCA	0.303 (0.021)	< 0.0001	1.283 (0.022)	< 0.0001
	CATH	0.318 (0.026)	< 0.0001	0.774 (0.025)	< 0.0001
	Other	0.330 (0.033)	< 0.0001	0.656 (0.023)	< 0.0001
CCI	2	0.201 (0.014)	< 0.0001	0.142 (0.017)	< 0.0001
	3	0.362 (0.016)	< 0.0001	0.243 (0.016)	< 0.0001
	≥ 4	0.565 (0.017)	< 0.0001	0.362 (0.019)	< 0.0001
Scale (σ)		0.662 (0.004)	< 0.0001	0.702 (0.005)	< 0.0001
θ		1.640 (0.015)		< 0.0001	

Note:CCI=Charlson Comorbidity Index.

Reference group: Gender=male, Procedure=none, CCI=1.

Table 4.5: Parameter estimates (SE) for log-normal margins with Gumbel copula, reduced model B with independent errors

		LOS		**Cost**	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Intercept		0.758 (0.019)	< 0.0001	9.236 (0.021)	< 0.0001
Age		0.008(0.0005)	< 0.0001	0.003(0.0004)	< 0.0001
Gender	Female	0.077 (0.013)	< 0.0001	0.008 (0.011)	0.469
Procedure	CABG	1.254 (0.024)	< 0.0001	1.922 (0.022)	< 0.0001
	PTCA	0.257 (0.019)	< 0.0001	1.181 (0.019)	< 0.0001
	CATH	0.294 (0.021)	< 0.0001	0.682 (0.020)	< 0.0001
	Other	0.351 (0.023)	< 0.0001	0.657 (0.021)	< 0.0001
CCI	2	0.197 (0.015)	< 0.0001	0.134 (0.013)	< 0.0001
	3	0.358 (0.018)	< 0.0001	0.228 (0.016)	< 0.0001
	≥ 4	0.579 (0.019)	< 0.0001	0.348 (0.016)	< 0.0001
Scale (σ)		0.649 (0.004)	< 0.0001	0.543 (0.004)	< 0.0001
Random Effects (τ)		0.145 (0.008)	< 0.0001	0.461 (0.011)	< 0.0001
Correlation (ρ)		0.840 (0.008)		< 0.0001	

Note:CCI=Charlson Comorbidity Index.

Reference group: Gender=male, Procedure=none, CCI=1.

Table 4.6: Parameter estimates (SE) for log-normal margins with Gaussian copula

		LOS		**Cost**	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Intercept		0.732 (0.019)	< 0.0001	9.171 (0.019)	< 0.0001
Age		0.008(0.0005)	< 0.0001	0.003(0.0004)	< 0.0001
Gender	Female	0.080 (0.013)	< 0.0001	0.014 (0.011)	0.205
Procedure	CABG	1.257 (0.024)	< 0.0001	1.966 (0.022)	< 0.0001
	PTCA	0.257 (0.020)	< 0.0001	1.224 (0.018)	< 0.0001
	CATH	0.297 (0.022)	< 0.0001	0.715 (0.019)	< 0.0001
	Other	0.354 (0.023)	< 0.0001	0.666 (0.020)	< 0.0001
CCI	2	0.202 (0.015)	< 0.0001	0.144 (0.013)	< 0.0001
	3	0.361 (0.019)	< 0.0001	0.240 (0.016)	< 0.0001
	≥ 4	0.579 (0.019)	< 0.0001	0.365 (0.016)	< 0.0001
Scale (σ)		0.651 (0.004)	< 0.0001	0.549 (0.004)	< 0.0001
Random Effects (τ)		0.122 (0.007)	< 0.0001	0.467 (0.008)	< 0.0001
Correlation (ρ)		0.599 (0.010)		< 0.0001	
θ		0.715 (0.005)		< 0.0001	
BIC			310979		

Note:CCI=Charlson Comorbidity Index.

Reference group: Gender=male, Procedure=none, CCI=1.

Table 4.7: Parameter estimates (SE) for log-normal margins with Clayton copula

		LOS		**Cost**	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Intercept		0.768 (0.020)	< 0.0001	9.098 (0.019)	< 0.0001
Age		0.008(0.0005)	< 0.0001	0.003(0.0004)	< 0.0001
Gender	Female	0.076 (0.013)	< 0.0001	0.018 (0.011)	0.085
Procedure	CABG	1.250 (0.026)	< 0.0001	1.978 (0.022)	< 0.0001
	PTCA	0.180 (0.021)	< 0.0001	1.240 (0.018)	< 0.0001
	CATH	0.305 (0.022)	< 0.0001	0.734 (0.018)	< 0.0001
	Other	0.376 (0.023)	< 0.0001	0.682 (0.019)	< 0.0001
CCI	2	0.228 (0.016)	< 0.0001	0.171 (0.013)	< 0.0001
	3	0.383 (0.019)	< 0.0001	0.270 (0.015)	< 0.0001
	≥ 4	0.628 (0.020)	< 0.0001	0.395 (0.016)	< 0.0001
Scale (σ)		0.696 (0.005)	< 0.0001	0.565 (0.004)	< 0.0001
Random Effects (τ)		0.160 (0.008)	< 0.0001	0.459 (0.010)	< 0.0001
Correlation (ρ)		0.560 (0.012)		< 0.0001	
θ		1.387 (0.035)		< 0.0001	
BIC			314048		

Note:CCI=Charlson Comorbidity Index.
Reference group: Gender=male, Procedure=none, CCI=1.

4.6 Discussion

In this chapter, we have introduced a joint bivariate copula random-effects model to describe the interplay of LOS and cost. The basic idea is that we start out with a specific latent response process (conditional on the random effects), connect those two equations through a copula with the association parameter. This approach flexibly models two-level correlations and at the same time account for hospital heterogeneity.

- u_i and v_i, model intra-hospital correlation, which induces the "within-hospital correlation", i.e. they generate dependence between those individuals in the same cluster, whereas conditional on the random effects those individuals are independent.

$$\begin{aligned}
&Cov(\log T_{ij}, \log T_{ik}) = Var(u_i)\\
&Cov(\log C_{ij}, \log C_{ik}) = Var(v_i)\\
&i = 1, 2, \ldots, n; j, k = 1, 2, \ldots, n_i, j \neq k.
\end{aligned}$$

- Correlation between u_i and v_i ρ describes the "cross-equation correlation" at the hospital level.

$$\begin{aligned}
&Cov(\log T_{ij}, \log C_{ij}) = Cov(u_i, v_i) + \sigma_1\sigma_2 Cov(\epsilon_{1,ij}, \epsilon_{2,ij})\\
&Cov(\log T_{ij}, \log C_{ik}) = Cov(u_i, v_i), j \neq k
\end{aligned}$$

- Association parameter in copula model θ describes "cross-equation correlation" at the individual level due to $Cov(\epsilon_{1,ij}, \epsilon_{2,ij})$. For example, patients with longer hospital length of stay tend to incur more cost.

Our proposed BCRE model overcomes some of the problems in existing models for mixed outcomes. Full maximum likelihood method is used to estimate parameters simultaneously, which is easily obtain in SAS Proc NLMIXED. Simulation results indicate that FML performs well. To illustrate our approach, we apply our model to 2003 NIS AMI patient data. Results obtained from our analysis are similar to previous results in Chapter 2 and 3.

To accommodate model misspecification, our model can be generalized in several aspects.

(a) In this chapter we consider only bivariate normal random effects. It can be easily extended to other bivariate distributions, eg. Gamma. Liu and Yu (2008) proposed a likelihood reformulation method which is much faster and can handle more complicated non-normal random-effects cases.

(b) ϵ's are assumed to have normal margins, with some pre-specified copula. We may also apply logistic distribution for measurement errors, which might be more proper for LOS (Gardiner et al., 2002) and cost (Luo et al., 2007).

With the easy implementation and satisfactory estimate results, our method greatly facilitates the application of joint random-effects models. We expect our estimation method to gain more popularity in practical data analysis of multiple measures of healthcare utilization.

Chapter 5

Conclusion and Future Research

5.1 Conclusion

In this thesis, we consider the two closely related measures of healthcare resource utilization: hospital length of stay (LOS) and cost. The management of hospital LOS has become an important issue in cost containment and control. Determination of relevant factors and hidden structures could inform discharge planning and allocation of resources (Lanzarone et al., 2010; McDermott and Stock, 2007; Ramiarina et al., 2008). In Chapter 2, a Coxian phase-type (PH) stochastic regression is proposed to model LOS and account for the heavy skew and heterogeneity in the data. Coxian PH distributions describe the time to absorption T of an underlying finite-state continuous time Markov process with only forward transitions. Transitions between states are governed by the Markov assumptions. The actual states of the Markov process are not observable, i.e, we do not know the state from which a patient enters the system, or the state from which the patient exits. The Coxian PH model presented in this thesis retains the general form of the common mean structure $E(T|\mathbf{x}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta}$.

A Bayesian method based on RJMCMC was applied to dynamically select the number of phases. This method avoids arbitrary trimming and transformation of the data. In addition, the approach allows us to obtain the estimates of mean LOS, median and other percentiles, and class memberships. While a complete description of the underlying hidden states of the Markov process would depend on specific applications, it is easy for us to classify patients into different groups, for example, short, medium, and long LOS groups. Partial effects are derived directly from the posterior samples, instead of using any bootstrapping method.

Shared random-effects models are introduced to jointly analyze LOS and cost in Chapter 3, which simultaneously assess the correlates of LOS and in-hospital cost, and provide important information for decisions on resource allocation. This model helps us ascertain the degree of correlation between LOS and cost, and to estimate the underlying mechanism for both processes.

When there is a complex random-effects structure in the model, the corresponding likelihood may be computationally prohibitive. Liu and Yu (2008) propose a likelihood reformulation method for non-normal random-effects models, which substantially reduces computational time, while yielding similar estimates to the probability integral transformation (PIT) method (Nelson et al., 2006).

In this model, we assume that the hospital cost may be correlated with LOS. In the application, we showed that there exists significant association between LOS and cost at the subject level, therefore separate models which disregard the correlations are not appropriate.

In Chapter 4, we develop a novel bivariate copula random-effects (BCRE) model for the analysis of LOS and cost with two-level correlations.

$$\begin{aligned}\log T_{ij} &= \mathbf{x}'_{1,ij}\boldsymbol{\beta}_1 + u_i + \sigma_1\epsilon_{1,ij} \\ \log C_{ij} &= \mathbf{x}'_{2,ij}\boldsymbol{\beta}_2 + v_i + \sigma_2\epsilon_{2,ij}\end{aligned} \tag{5.1}$$

- u_i and v_i, $1 \leq i \leq n$ model intra-hospital correlation, which induces the "within-hospital correlation" among patients for LOS and cost, respectively.

- Correlation between u_i and v_i, $1 \leq i \leq n$ describes the "cross-equation correlation" at the hospital level.

- Association parameter θ in measurement errors $(\epsilon_{1,ij}, \epsilon_{2,ij})$, $1 \leq i \leq n, 1 \leq j \leq n_i$, describes "cross-equation correlation" at the individual level.

Full maximum likelihood (FML) is implemented using SAS Proc NLMIXED to derive the parameter estimates simultaneously. The estimation method yields satisfactory results as demonstrated by the simulation study. Our model is very comprehensive yet easy to fit. These advantages make it valuable in real data analysis.

In the application to the 2003 NIS AMI patients, we showed that there are significant associations between LOS and cost at both the hospital level and the individual level. For comparison, we also fit the two measures with two reduced models, assuming that there are no random effects, and that the measurement errors are independent. Those reduced models are not appropriate due to ignoring either the hospital or the individual level correlations. Therefore our more comprehensive model should be preferred in such situations.

5.2 Future Research

The strength of Coxian PH regression models for LOS lies in their flexibility in accommodating extreme values, while revealing hidden features such as short, long stays in hospitals. A natural extension of the proposed Coxian PH method is the modeling of censored duration data, as well as multiple discharge destinations. Olsson (1996) extends the EM algorithm for estimation of PH distributions for censored data. We further extend the models by incorporating covariates in the mean specification.

In Chapter 3, an equivalent joint model can be proposed as follows:

$$\begin{aligned} Y_{i1}|a_i &\sim \text{Coxian PH}(\lambda_0, \boldsymbol{p}, \boldsymbol{\beta}) \\ \log E(Y_{i1}|a_i) &= \beta_0^{CPH} + \mathbf{x}_{1i}' \boldsymbol{\beta}^{CPH} + a_i \\ \log(Y_{i2}|b_i) &= \mathbf{x}_{2i}' \boldsymbol{\beta}^{LN} + b_i \end{aligned} \tag{5.2}$$

where a_i and b_i are assumed correlated. $\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N(\mathbf{0}, \Sigma)$, with Σ being a positive definite matrix. This equivalent form is more flexible in modeling random effects for two outcomes with different scales.

In Chapter 4, we assumed that (1) measurement errors $(\epsilon_{1,ij}, \epsilon_{2,ij})$, $1 \leq i \leq n$, $1 \leq j \leq n_i$ have normal margins with some specified copula. (2) the hospital level random effects (u_i, v_i), $1 \leq i \leq n$ are bivariate normally distributed. To accommodate possible model misspecification, our model can be generalized in several aspects.

(a) Bivariate normal random effects can be easily extended to other bivariate distributions, eg. Gamma. Liu and Yu (2008) proposed a likelihood reformulation method which is much faster in comparison to PIT method, and can handle more complicated non-normal random-effects cases.

(b) We may also apply logistic distribution for measurement errors, which might be more proper for LOS (Gardiner et al., 2002) and cost (Luo et al., 2007).

Another attractive extension of the proposed methods in Chapter 3 and 4 is the joint modeling of LOS and repeated measures via shared random-effects models as wells as copula-based models. This model can be used to model hospital LOS and daily (monthly) cost jointly in heathcare utilization studies. More research on identifiability, estimation and inference of this model is needed.

We apply our methods to hospital LOS and cost for patients with acute myocardial infarction (AMI) in the 2003 Nationwide Inpatient Sample (NIS). More applications can be found for other heathcare measures. Generalization to other related areas is of further interest.

APPENDICES

Appendix A

Gaussian Quadrature

Gaussian quadrature is often used to approximate the likelihood with no closed form. It approximates the integration by a weighted average of the integrand assessed at Q predetermined quadrature points $c_q(q = 1, 2, \ldots, Q)$ over the normal random effects u_i with density $p(u_i)$. The likelihood

$$L_i = \int \prod_{j=1}^{n_i} f(c_{ij}, t_{ij} | \mathbf{x}_{ij}, u_i) p(u_i) du_i$$

can be approximated by

$$L_i \approx \sum_{q=1}^{Q} \prod_{j=1}^{n_i} f(c_{ij}, t_{ij} | \mathbf{x}_{ij}, c_q) p(c_q) \omega_q$$

with $c_q = \sqrt{2} z_q$ and $\omega_q = \sqrt{2} \eta_q \exp(z_q^2)$, where η_q and z_q can be can be obtained from tables (Abramowitz and Stegun, 2002). In the adaptive Gaussian quadrature, the integral is centered at the empirical Bayes estimate of u_i , while it is centered at 0 in the non-adaptive Gaussian quadrature. Thus, the adaptive Gaussian quadrature provides a better approximation for badly behaved integrands. Refer Littell et al. (2006) for details of Gaussian quadrature in SAS.

Appendix B

Example SAS code for copula estimates

The data format for a sample data set is given below.

```
data gaussian;

input groupid id x1 x2 los;

cards;
 1     1    −0.36473     0.43007     8.833     31.83
 1     2    −0.37871    −0.47626    14.563    573.40
 1     3     0.02560    −0.19365    34.126    190.54
 2     1    −0.08726    −0.10627    18.973      7.47
 2     2     0.22927     0.05269    16.582      4.40
 2     3    −0.23027    −0.00650    33.323    140.89
 3     1     0.26273    −0.48568     2.724      8.55
 . . .  . . .       . . .        . . .
;
```

B.1 SAS Code for Gaussian Copula

```
proc nlmixed data=gaussian gconv=0 qpoints=5;
parms b01 0.994 b11 -0.46 sigma1 1.308 b02 3.98 b12 0.35 sigma2 0.938 alpha 0.48
tau1 0.5 tau2 1 rho 0.5;
bounds sigma1>0, sigma2>0, -1≤alpha≤1,tau1>0,tau2>0,-1≤rho≤1;
mu1=b01+b11*x1+u1;
mu2=b02+b12*x2+u2;
p1=cdf(’lognormal’,y1,mu1,sigma1);
if p1>.999999 then p1=.999999;
F1=quantile(’normal’,p1);
p2=cdf(’lognormal’,y2,mu2,sigma2);
if p2>.999999 then p2=.999999;
F2=quantile(’normal’,p2);
logGaussian=-.5*log(1-alpha**2)-.5*(F1**2+F2**2-2*alpha*F1*F2)/(1-alpha**2)
+logpdf(’lognormal’,y1,mu1,sigma1)+logpdf(’lognormal’,y2,mu2,sigma2)-(-F1**2/2-F2**2/2);
model y1∼general(logGaussian);
random u1 u2∼normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=group;
run;
```

B.2 SAS Code for Clayton Copula

```
proc nlmixed data=clayton gconv=0 qpoints=5;
parms b01 1.05 b11 -0.48 sigma1 1 b02 4.07 b12 0.6 sigma2 2.1 alpha 0.86 tau1 0.8
tau2 1.15 rho 0.5;
bounds sigma1>0, sigma2>0,tau1>0,tau2>0,-1≤rho≤1;
mu1=b01+b11*x1+u1;
mu2=b02+b12*x2+u2;
p1=cdf(’lognormal’,y1,mu1,sigma1);
if p1>.999999 then p1=.999999;
p2=cdf(’lognormal’,y2,mu2,sigma2);
if p2>.999999 then p2=.999999;
logClayton=log(1+alpha)-(alpha+1)*(log(p1)+log(p2))-(1/alpha+2)*log(p1**(-alpha)
+p2**(-alpha)-1)+logpdf(’lognormal’,y1,mu1,sigma1)+logpdf(’lognormal’,y2,mu2,sigma2);
model y1∼general(logClayton);
random u1 u2∼normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=group;
run;
```

B.3 SAS Code for Gumbel Copula

```
proc nlmixed data=gumbel gconv=0 qpoints=5;
parms b01 1.1 b11 -0.46 sigma1 1 b02 4.07 b12 0.6 sigma2 2.1 alpha 1.2 tau1 0.8
tau2 1.15 rho 0.5;
bounds sigma1>0, sigma2>0,tau1>0,tau2>0,-1≤rho≤1;
```

```
mu1=b01+b11*x1+u1;
mu2=b02+b12*x2+u2;
p1=cdf(’lognormal’,y1,mu1,sigma1);
if p1>.999999 then p1=.999999;
logp1=-log(p1);
p2=cdf(’lognormal’,y2,mu2,sigma2);
if p2>.999999 then p2=.999999;
logp2=-log(p2);
logGumbel=-(logp1**alpha+logp2**alpha)**(1/alpha)-log(p1*p2)+(alpha-1)*log(logp1*logp2)
-(2-1/alpha)*log(logp1**alpha+logp2**alpha)
+log((logp1**alpha+logp2**alpha)**(1/alpha)+alpha-1)
+logpdf(’lognormal’,y1,mu1,sigma1)+logpdf(’lognormal’,y2,mu2,sigma2);
model y1∼general(logGumbel);
random u1 u2∼normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=group;
run;
```

BIBLIOGRAPHY

BIBLIOGRAPHY

Aalen, O. O., 2002. Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, 22:447–463.

Abramowitz, M. and I. A. Stegun, editors, 2002. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables.* Dover Publications.

Allison, P. D., 2005. *Fixed Effects Regression Methods for Longitudinal Data Using SAS.* SAS Publishing.

Asmussen, S., O. Nerman, and M. Olsson, 1996. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441.

Ausín, M. C. and H. F. Lopes, 2007. Bayesian estimation of ruin probabilities with a heterogeneous and heavy-tailed insurance claim-size distribution. *Australian & New Zealand Journal of Statistics*, 49(4):415–434.

Ausín, M. C., M. P. Wiper, and R. E. Lillo, 2008. Bayesian prediction of the transient behaviour and busy period in short- and long-tailed GI/G/1 queueing systems. *Comput. Stat. Data Anal.*, 52(3):1615–1635.

Ausín, M. C., M. P. Wiper, and R. E. Lillo, 2009. Bayesian estimation of finite time ruin probabilities. *Applied Stochastic Models in Business and Industry*, 25(6):787–805.

Basu, A. and P. J. Rathouz, 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*, 6(1):93–109.

Bee, M., 2004. Modelling credit default swap spreads by means of normal mixtures and copulas. *Applied Mathematical Finance*, 11(2):125–146.

Bitran, G. R. and S. Dasu, 1994. Analysis of $\sum PH_i/PH/1$ queues. *Operations Research*, 42:158–174.

Breymann, W., A. Dias, and P. Embrechts, 2003. Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3(1):1469–7688.

Buckley, J. and I. James, 1979. Linear regression with censored data. *Biometrika*, 66(3):429–436.

Cameron, A. C., T. Li, P. K. Trivedi, and D. M. Zimmer, 2004. Modeling differences in counted outcomes using bivariate copula models: With application to mismeasured counts. *Econometrics Journal*, 7(2):566–584.

Capéraà, P., A. L. Fougères, and C. Genest, 2000. Bivariate distributions with given extreme value attractor. *Journal of Multivariate Analysis*, 72(1):30–49.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie, 1987. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373 – 383.

Chen, X. and Y. Fan, 2002. Estimation of copula-based semiparametric time series models. Working Papers 0226, Department of Economics, Vanderbilt University.

Chen, Z. and D. B. Dunson, 2003. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Cumani, A., 1982. On the canonical representation of homogeneous markov processes modelling failure-time distributions. *Microelectron. Reliab.*, 22:583–602.

Dellaportas, P. and J. J. Forster, 1999. Markov chain monte carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633.

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

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith, 1998. Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):333–350.

Diebolt, J. and C. P. Robert, 1994. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375.

Erlang, A. K., 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13:5.

Etzioni, R. D., E. J. Feuer, S. D. Sullivan, D. Lin, C. Hu, and S. D. Ramsey, 1999. On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics*, 18(3):365–380.

Fackrell, M., 2009. Modelling healthcare systems with phase-type distributions. *Health Care Management Science*, 12(1):11–26.

Faddy, M., N. Graves, and A. Pettitt, 2009. Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314.

Faddy, M. J., 2002. Penalized maximum likelihood estimation of the parameters in a coxian phase-type distribution. In Latouche, G. and P. Taylor, editors, *Matrix-analytic methods: theory and applications*, pages 107–114. World scientific.

Faddy, M. J. and S. I. McClean, 1999. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15(4):311–317.

Faddy, M. J. and S. I. McClean, 2005. Markov chain modelling for geriatric patient care. *Methods of Information in Medicine*, 44(3):369–373.

Gardiner, J. C., Z. Luo, C. J. Bradley, E. Polverejan, M. Holmes-Rovner, and D. Rovner, 2002. Longitudinal assessment of cost in health care interventions. *Health Services and Outcomes Research Methodology*, 3:149–168.

Genest, C. and L.-P. Rivest, 1993. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043.

Golub, G. and C. Van Loan, 1996. *Matrix computations.* The Johns Hopkins University Press, Baltimore, MD, 3rd edition.

Green, P. J., 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Gruttola, V. D. and X. M. Tu, 1994. Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4):1003–1014.

Guo, X. and B. P. Carlin, 2004. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58:16–24.

Hallstrom, A. P. and S. D. Sullivan, 1998. On estimating costs for economic evaluation in failure time studies. *Medical Care*, 36(3):433–436.

Henderson, R., P. Diggle, and A. Dobson, 2000. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Horváth, A. and M. Telek, 2000. Approximating heavy tailed behaviour with phase type distributions. Leuven, Belgium.

Hougaard, P., 1999. Fundamentals of survival data. *Biometrics*, 55(1):13–22.

Hougaard, P., 2000. *Analysis of multivariate survival data.* Springer, New York.

Hougaard, P., P. Myglegaard, and K. Borch-Johnsen, 1994. Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. *Biometrics*, 50(4):1178–1188.

Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying, 2003. Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

Joe, H., 1997. *Multivariate Models and Multivariate Dependence Concepts.* Chapman and Hall, London, 1 edition.

Joe, H. and J. J. Xu, 1996. The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia.

Johnson, M. A., 1993. Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and erlang distributions. *ORSA Journal on Computing*, 5(1):69–83.

Johnson, V. E., 2004. A bayesian chi-squared test for goodness of fit. *The Annals of Statistics*, 32(6):2361–2384.

Keiding, N., P. K. Andersen, and J. P. Klein, 1997. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, 16(2):215–224.

Klugman, S. A. and R. Parsa, 1999. Fitting bivariate loss distributions with copulas. *Insurance: Mathematics and Economics*, 24(1-2):139 – 148.

Kocherlakota, S. and K. Kocherlakota, 1992. *Bivariate discrete distributions.*

Konetzka, R. T., W. Spector, and M. R. Limcangco, 2008. Reducing hospitalizations from long-term care settings. *Medical Care Research and Review*, 65(1):40–66.

Lambert, P., D. Collett, A. Kimber, and R. Johnson, 2004. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*, 23(20):3177–3192.

Lambert, P. and F. Vandenhende, 2002. A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, 21(21):3197–3217.

Lanzarone, E., A. Matta, and G. Scaccabarozzi, 2010. A patient stochastic model to support human resource planning in home care. *Production Planning & Control*, 21(1):3–25.

Li, J., 1999. An application of lifetime models in estimation of expected length of stay of patients in hospital with complexity and age adjustment. *Statistics in Medicine*, 18(23):3337–3344.

Lin, D. Y., 2000. Proportional Means Regression for Censored Medical Costs. *Biometrics*, 56(3):775–778.

Lin, D. Y., E. J. Feuer, R. Etzioni, and Y. Wax, 1997. Estimating medical costs from incomplete follow-up data. *Biometrics*, 53(2):419–434.

Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberber, 2006. *SAS for Mixed Models.* SAS Publishing, 2 edition.

Liu, L., 2004. *Modeling Recurrent Events and Medical Cost Data in the Presence of a Correlated Terminating Event*. PhD thesis, University of Michigan.

Liu, L., 2009. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*, 28(6):972–986.

Liu, L. and X. Huang, 2009. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):65–81.

Liu, L., R. A. Wolfe, and J. D. Kalbfleisch, 2007. A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*, 26(1):139–155.

Liu, L. and Z. Yu, 2008. A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine*, 27(16):3105–3124.

Luo, Z., J. Goddeeris, J. C. Gardiner, and R. C. Smith, 2007. Costs of an intervention for primary care patients with medically unexplained symptoms: A randomized controlled trial. *Psychiatr Serv.*, 58(8):1079–1086.

Marie, R., 1980. Calculating equilibrium probabilities for $\lambda(n)/c_k/1/n$ queues. In *In Proceedings of the Performance*, pages 117–125.

Marshall, A. and S. McClean, 2004. Using coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Science*, 7:285–289.

Marshall, A., S. McClean, C. Shapcott, and P. Millard, 2002. Modelling patient duration of stay to facilitate resource management of geriatric hospitals. *Health Care Management Science*, 5:313–319.

Marshall, A., S. McClean, M. Shapcott1, and P. Millard, 2000. Learning dynamic bayesian belief networks using conditional phase-type distributions. In *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *Lecture Notes in Computer Science*, pages 1–11. Springer Berlin / Heidelberg.

Marshall, A., C. Vasilakis, and E. El-Darzi, 2005. Length of stay-based patient flow models: Recent developments and future directions. *Health Care Management Science*, 8(3):213–220.

Marshall, A. H. and S. I. McClean, 2003. Conditional phase-type distributions for modelling patient length of stay in hospital. *International Transactions in Operational Research*, 10(6):565–576.

Marshall, A. H., B. Shaw, and S. I. McClean, 2007. Estimating the costs for a group of geriatric patients using the coxian phase-type distribution. *Statistics in Medicine*, 26(13):2716–2729.

Matsui, K., L. Goldman, P. Johnson, K. Kuntz, E. Cook, and T. Lee, 1996. Comorbidity as a correlate of length of stay for hospitalized patients with acute chest pain. *Journal of General Internal Medicine*, 11(5):262–268.

McDermott, C. and G. N. Stock, 2007. Hospital operations and length of stay performance. *International Journal of Operations & Production Management*, 27(9):1020–1042.

McGrory, C., A. Pettitt, and M. Faddy, 2009. A fully bayesian approach to inference for coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis*, 53(12):4311 – 4321.

Miller, D. J. and W. Liu, 2002. On the recovery of joint distributions from limited information. *Journal of Econometrics*, 107(1-2):259–274.

Nelsen, R. B., 1999. *An Introduction to Copulas*. Springer, New York.

Nelson, K. P., S. R. Lipsitz, G. M. Fitzmaurice, J. Ibrahim, M. Parzen, and R. Strawderman, 2006. Use of the probability integral transformation to fit nonlinear mixed-effects modelswith nonnormal random effects. *Journal of Computational and Graphical Statistics*, 15(1):39–57.

Neuts, M. F., 1981. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Johns Hopkins University Press, Baltimore, MD.

Olsson, M., 1996. Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, 23(4):443–460.

Osogami, T., 2005. *Analysis of Multi-server Systems via Dimensionality Reduction of Markov Chains*. PhD thesis, Carnegie Mellon University.

Polverejan, E., J. C. Gardiner, C. J. Bradley, M. H. Rovner, and D. Rovner, 2003. Estimating mean hospital cost as a function of length of stay and patient characteristics. *Health Economics*, 12(1):935–947.

Pompei, P., M. E. Charlson, K. Ales, C. R. Mackenzie, and M. Norton, 1991. Relating patient characteristics at the time of admission to outcomes of hospitalization,. *Journal of Clinical Epidemiology*, 44(10):1063 – 1069.

Ramiarina, R., R. Almeida, and W. A. Pereira, 2008. Hospital costs estimation and prediction as a function of patient and admission characteristics. *The International Journal of Health Planning and Management*, 23(4):345–355.

Ratcliffe, S. J., W. Guo, and T. R. T. Have, 2004. Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, 60(4):892–899.

Richardson, S. and P. J. Green, 1997. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59(4):731–792.

Sauer, C. H. and K. M. Chandy, 1975. Approximate analysis of central server models. *IBM Journal of Research and Development*, 19:31–3.

Sirbu, C. M., 2004. *Assessing Medical Costs from a Longitudinal Model.* PhD thesis, Michigan State University.

Sisko, A., C. Truffer, S. Smith, S. Keehan, J. Cylus, J. A. Poisal, M. K. Clemens, and J. Lizonitz, 2009. Health spending projections through 2018: recession effects add uncertainty to the outlook. *Health Affairs (Millwood)*, 28(2):w346–w357.

Sklar, A., 1959. Fonctions de repartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, pages 229–231.

Skrondal, A. and S. Rabe-Hesketh, 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman & Hall/CRC, Boca Raton, FL.

Smith, M. D., 2003. Modelling sample selection using archimedean copulas. *Econometrics Journal*, 6:99–123.

Tian, L. and T. Cai, 2006. On the accelerated failure time model for current status and interval censored data. *Biometrika*, 93(2):329–342.

Trivedi, P. K. and D. M. Zimmer, 2005. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111.

Tsiatis, A. A., 1990. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18(1):354–372.

Tsiatis, A. A. and M. Davidian, 2004. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–834.

Tsiatis, A. A., V. DeGruttola, and M. S. Wulfsohn, 1995. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37.

Vonesh, E. F., tom Greene, and M. D. Schluchter, 2006. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, 25:143–163.

Wang, K., K. K. W. Yau, and A. H. Lee, 2002. A hierarchical poisson mixture regression model to analyse maternity length of hospital stay. *Statistics in Medicine*, 21(23):3639–3654.

Wei, L. J., Z. Ying, and D. Y. Lin, 1990. Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77(4):845–851.

Wooldridge, J. M., 2002. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, Cambridge, Massachusetts.

Wulfsohn, M. S. and A. A. Tsiatis, 1997. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Xiao, J., D. Douglas, A. H. Lee, and S. R. Vemuri, 1997. A delphi evaluation of the factors influencing length of stay in australian hospitals. *The International Journal of Health Planning and Management*, 12(3):207–218.

Xu, J. and S. L. Zeger, 2001. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50(3):375–387.