2016

Parametric Reversed Hazards Model for Left Censored Data with Application to HIV

Farahnaz Islam
University of South Carolina

Parametric Reversed Hazards Model for Left Censored Data with Application to HIV

by

Farahnaz Islam

Bachelor of Medical Radiation Science
University of Newcastle, 2012

Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Public Health in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2016

Accepted by:

Hrishikesh Chakraborty, Director of Thesis

Alexander C. McLain, Reader

James Hussey, Reader

Cheryl L. Addy, Vice Provost and Dean of Graduate Studies

Abstract

Left censoring is generally a rare type of censoring in time-to-event data, however there are some fields such as HIV related studies where it commonly occurs. Currently, there is no clear recommendation in the literature on the optimal model and distribution to analyze left-censored data. Recommendations can help researchers apply more accurate models for this type of censoring. This study derives the Parametric Reversed Hazards (PRH) Model for a variety of distributions which may be appropriate for left censored data. The performance of these derived PRH models to analyze HIV viral load data are compared using extensive simulations and a guideline is established for which distribution/s are most appropriate. Each simulation setup is varied by sample size and proportion of censoring to find a consistently high performance distribution. The best distribution is determined using the information criteria: AIC, AICC, HQIC, and CAIC. The South Carolina Enhanced HIV/AIDS Reporting Surveillance System (SC eHARS) data were utilized and a bootstrap study provided further insights towards appropriateness of the distributions in analyzing HIV viral load data. Results from simulation studies point to the Generalized Inverse Weibull distribution to outperform all others across censoring rates and sample sizes. The bootstrap study, however, contradicts this and suggests the Marshal-Olkin distribution to be the superior performer. This disagreement may have resulted from the special heavy tail nature of viral load data that demands further attention. Application of the best performing models on the SC eHARS database revealed important effects explaining trends of viral load over time.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Analysis of survival data is used in various fields such as social science (event history), economics (duration analysis), engineering (reliability analysis), and medicine (survival analysis). For consistency, we will refer to this type of analysis as survival analysis, although techniques mentioned here can be applied in any field. Survival analysis focuses on measuring time to event data, for example, time to death or time to recovery. One of the issues with survival analysis is censoring. This occurs when the time to the event of interest is not available for all subjects in the study due to loss of follow-up, the event does not occur within the study period, or death occurs from reasons not related to the research. There are three types of censoring; right censoring, interval censoring, and left censoring. Right censoring is when the event of interest is not observed in the study period, for example, if the subject drops out before the end of the study. Interval censoring is when we know the event of interest occurs in a certain time interval but the exact time of occurrence is unknown. This commonly occurs in medical research and epidemiological studies with periodic monitoring. Left censoring is when the event of interest has occurred before enrollment into the study, but it is not known exactly when. For example, consider that the event is the age at which children are able to learn the alphabet at school. There may be some children who are able to recite the alphabet before starting school, these subjects are left censored. Compared to the other types of censoring, left censoring

less frequently occurs. As a result, it is often overlooked and understudied. In this study, we focus on survival data which have left censored observations.

It is important to differentiate between censoring and truncation. When data is truncated, there is a cut point beyond which observations cannot occur. For example, age is naturally truncated since it cannot take on values less than zero. When data is censored, the censored observations take on a range of values as they are only known to be equal to or more extreme than a certain point. With respect to modeling, censoring requires making probability calculations on a wide range of values whereas truncation requires making probability calculations after rescaling the distribution to reflect the truncated data.

If censoring is ignored when analyzing data, it can lead to underestimation of the survival probability or mean, and inconsistent covariate effects.[1] Additionally, the impact of ignoring censoring increases as the proportion of censored observations increases.[2] The most frequently used approaches to deal with censoring is to replace the censored value with an arbitrary value such as the detection limit value or half of the limit value.[3,4] This arbitrary replacement method usually results in overestimation because the predicted values based on the arbitrary value would be higher than the predicted values based on the unknown true values. These approaches also underestimate the variability in the data because the same value is imputed several times. Another potential approach developed by Paxton et al.[5] is a two-stage imputation procedure which is used to predict the censored values by first substituting half of the lower detection limit and then refitting the model by imputing the new estimated values. This method is a slight improvement from the arbitrary replacement methods since it removes much of the bias in the parameter estimates, but the effect on the variability is less predictable.[5] There is one major disadvantage to utilizing these convenient techniques: they ignore the correlated structure of longitudinal data and do not adjust for the variability of the parameter estimates due to the loss

of information from censoring. Ganser and Hewett[6] developed a more sophisticated substitution method they termed the β-substitution method which calculates a β factor depending on the uncensored data and replaces the limit of detection with the β factor multiplied by the limit of detection. This approach has been shown to be less biased than the simpler substitution methods.[6]

An alternative to crude imputation methods is a maximum likelihood (ML) approach in which the censored data is incorporated into the log-likelihood functions of the observed data.[7,8] Although censored data lack information on the event of interest, incorporating them in this way can provide valuable information to the model. Hughes[7] modified the usual mixed effects model by using a likelihood-based Monte Carlo Expectation-Maximization (MCEM) algorithm to account for censoring. This method removes the bias in the parameter estimates and the within-person variability but there is some bias in the between-person variability which is mostly due to the variability in the ML estimates of the uncensored data.[7] Jacqmin-Gadda et al.[8] used a general likelihood with cumulative distribution function (CDF) to account for left-censored observations. The formulation of the likelihood is conditioned on observed measures and the marginal likelihood is used to make inferences about the unknown parameters. The approach by Lyles, Lyles, and Taylor[9] is based on a hierarchical formulation of the likelihood where the estimation is carried out by direct maximization of the likelihood. These likelihood approaches correct for the bias obtained when an arbitrary value is assigned to the censored data.[7–9] One disadvantage is it makes stringent Gaussian distribution assumptions and is not easy to implement in standard software.[10]

Traditional regression models are not able to handle censored data directly, and as a result a wide class of non-parametric, semi-parametric and parametric survival models have been developed to handle data with censored observations. These models all explore the relationship between the hazard rate of a subject and several independent

variables. The commonly used form of these models can be written as

$$\lambda(t) = \lambda_0(t)e^{x_i\beta}$$

where $\lambda_0(t)$ is the baseline hazard function, x_i is the set of covariates, and β are parameters estimating covariate effects on hazard. This proportional hazards (PH) model assumes that the survival curves for any two subjects have hazard functions which are proportional over time, i.e., they have a relative constant hazard. This assumption can be checked by confirming that the complementary log-log survival curves for the two subjects are parallel.

In semi-parametric models, the regression coefficients are estimated leaving the baseline hazard unspecified. For example, the Cox Proportional Hazards model[11] introduced the use of the partial-likelihood function to estimate the coefficients without needing to characterize the baseline hazard rate. There are several studies which use non-parametric methods to correct for left-censoring.[12–15] An advantage of using this type of method is that distributional assumptions about the baseline hazard do not have to be made. However, this can also be disadvantageous.

In a parametric regression model, a particular shape or distribution is specified for the baseline hazard rate. These models let the parameters of the assumed distribution depend on the covariates. An advantage of using parametric regression models is that they naturally smooth the data by assuming an underlying distribution so that censoring has less effect on parameter estimates than for semi-parametric methods. If the characterization of the underlying time-dependency is accurate, i.e., if someone chooses the correct distribution, then parameter estimates are generally more precise than estimates from semi-parametric models where the underlying time-dependency is left unspecified. However, problems can arise if the incorrect parametric form is selected.

Determining which distribution to assume in the presence of left censoring is difficult as the existing literature on this is scarce and inconsistent due to the lack of

guidelines. Many studies assume specific underlying distributions based on guidelines for right censored data or the shape of the data being analyzed. Thompson, Voit, and Scott[16] compared different distributions recommended for right censored data using probability plots to find which one best fits their left censored data, prior to running survival models. Annan, Liu, and Zhang[17] compared various estimators for left censored data using simulation studies in which they assume the underlying distribution is exponential. They selected the exponential distribution reasoning that common distributions usually associated with left/right censored data such as the normal, lognormal and gamma distribution all belong to the exponential family. Pajek et al.[18] simulated data from a log-normal distribution to compare various estimators for left censored data on trace element concentrations since this distribution was experimentally validated in a prior study.[19] Another study by Luczynska et al.[20] assumed a Normal distribution with no validation as to why this distribution was used in the analyses. Some survival studies[16,21–23] assumed a Weibull distribution for the left censored data simply because it is commonly used in survival analysis, especially in the field of medicine, and approximately fits the data. Gupta and Kundu[24] proposed a new family of distributions, the generalized exponential distribution, which is very similar to the corresponding shape of a gamma or Weibull distribution. The probability density function (pdf) is of the form:

$$f(x;\alpha,\lambda) = \alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x}$$

where α and λ are the shape and scale parameters, respectively. Since these distributions are often used for censored data in survival analysis, the generalized exponential distribution is a possible alternative to use in survival models. Expanding on this, Mitra and Kundu[25] derived the maximum likelihood estimator for data with left censored observations from a generalized exponential distribution:

$$\hat{\alpha}(\lambda) = -\frac{n-r}{rln(1-e^{-\lambda x_{(r+1)}} + \sum_{i=r+1}^{n} ln(1-e^{-\lambda x_{(i)}})}$$

1.2 Motivating Example

Human immunodeficiency virus (HIV) is a chronic disease which weakens the immune system, leading to increased susceptibility to a wide range of infections and some types of cancer. HIV RNA or viral load (VL) measures the number of actively replicating HIV virus in a subject and is an important biomarker for HIV disease progression.[26] There is no cure for HIV, but the success of highly active antiretroviral therapy (ART) to suppress VL to undetectable levels for prolonged periods of time has transformed HIV into a manageable chronic disease.[26] Suppression of VL to undetectable levels improves physical functioning, reduces opportunistic infections, reduces HIV related mortality, and is associated with a substantial decrease in the probability of transmitting HIV to others.[27–29] By CDC guideline, VL is detectable if > 200 copies/mL and undetectable if ≤ 200 copies/mL. Not only is suppressing VL important on an individual level, it also has the potential to decrease HIV incidence rates in a community because of reduced infectivity.[29,30] Consequently, the focus of care has shifted from survival to improving health outcomes among people with HIV.[26]

The HIV endemic disproportionately impacts the Southern states in the US in terms of the overall number of people living with HIV/AIDS (PLWHA), and survival rates after HIV/AIDS diagnosis.[31] South Carolina (SC), like many Southern states, ranks high for poverty, unemployment, and low educational completion which are all characteristics that may promote disease transmission. The number of PLWHA in SC has increased from 12,089 in 2004 to 16,311 in 2014.[32] Recent studies on retention in HIV care found that a large proportion of PLWHA in SC failed to remain in care on a regular basis.[33,34] Given the HIV burden in SC and the need to focus on retention in HIV care within the context of the National HIV/AIDS Strategy goals, it is important to identify factors which suppress VL. Identifying these factors will assist in developing targeted strategies to reduce the HIV burden in SC.

To provide insight into the HIV endemic in SC, survival models of time to undetectable levels of VL should be analyzed. These models can also be used to assess the effect of various drugs on the VL. Some subjects may have undetectable VL at the beginning of the study, in which case the first point at which they reach detectable level will be the start of their observations. Thus, these subjects are not censored. However, this dataset is complicated with many left censored subjects where the time of infection, i.e., the exact time VL reaches detectable levels is unknown. Furthermore, the distribution of VL patterns from detectable to undetectable levels varies from person to person so there is a need to establish a baseline distribution which can be used when analyzing this data.

In the literature on HIV datasets with left censoring, the most common approach is to use a PH model because of its relative simplicity.[35–39] One study showed the consistency and asymptotic normality for the maximum likelihood estimator of the PH model for doubly censored HIV data, i.e., data with both left and right censoring present.[39] An advantage of using the PH model is the ability to fit survival models without knowing (or assuming) the underlying distribution. However, as explained above in section 1.1, if the distribution is known or an appropriate distribution can be assumed, then the maximum likelihood estimates from a parametric model are more accurate than this simple approach. One study used a log logistic Accelerated Failure Time (AFT) model to estimate the effect of age on time to VL suppression.[40] However, the use of AFT models is very rare in cases where the data are left censored. Parametric regression models are more commonly applied to HIV data with left censoring present.[23,41–43] Studies by Zaba et al.[41] and Isingo et al.[23] used a parametric regression model based on the Weibull distribution to assess survival after HIV infection. The authors from these studies selected a Weibull distribution as it more closely fit the data, but no results of this comparison was provided in either article.

More specifically considering studies measuring time to HIV VL suppression,

Thiébaut et al.[42] applied a lognormal survival model using a full parametric approach to take into account the left censored HIV VL and CD4+ counts. The lognormal distribution was utilized in their study as suggested by Henderson, Diggle, and Dobson[44] because the estimated lognormal survival distribution function was contained within the 95% confidence interval of non-parametric Kaplan Meier estimate. The authors did further sensitivity analysis by comparing the lognormal survival model to a univariate mixed model and a Cox PH model. However, other survival distributions were not considered in their sensitivity analysis.

A study by Cole et al.[43] applied a parametric likelihood-based approach to handle left-censoring of HIV VL measurements assuming it follows a standard Normal distribution. They defined the marginal likelihood for participant i and visit j as:

$$L_{ij} = \left[\phi\left(\frac{Y_{ij}^{\star} - \mu_{ij}}{\sigma}\right)\right]^{w_{ij}(1-d_{ij})} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_{ij} - \mu_{ij})^2}{2\sigma^2}\right)\right]^{w_{ij}d_{ij}}$$

where ϕ is the cumulative distribution function (CDF) of a standard normal random variable. Detectable VL measurements contribute the second term, while undetected (censored) VL measurements contribute to the first term in the likelihood.

While there have been studies comparing the fit of various distributions to right-censored and interval-censored data,[45–47] there are no recommendations in the literature on optimal distributions to use for left-censored data. Recommendations can help researchers apply more accurate models for this type of censoring, specifically in HIV related studies where it is a common occurrence.

1.3 Objectives

The objectives of this study are to:

1. review the non-parametric, semi-parametric, and parametric statistical methods for analyzing survival data in the presence of left censored data, outlined in Chapter 2.

2. derive the Proportional Reverse Hazards (PRH) model for a variety of distributions which may be appropriate for left censored data. These include the Exponential, Log-normal, Inverse Gaussian, Log-logistic, Gompertz-Makeham, Gamma, Generalized Gamma, Inverse Gamma, Generalized Inverse Gamma, Weibull, Inverse Weibull, Generalized Inverse Weibull, Modified Weibull, Flexible Weibull, Power Generalized Weibull, and the Marshal-Olkin distributions. These derivations are outlined in Chapter 3.

3. conduct simulation studies to assess performance of the derived PRH models and compare these to establish a guideline for which distribution/s would "best" fit left censored HIV viral load data. Sample sizes and the proportion of censored observations will be varied for each distribution to simulate different data conditions. Details of the simulation setup are provided in Chapter 4. Then, using a bootstrapping technique, determine which distribution under the PRH model is best suited for analyzing the VL of HIV infected individuals using the SC eHARS database.

4. apply the selected best performing models to the SC eHARS database to explain effects of different demographic, social, and treatment factors on patients' viral load transition from detectable to undetectable levels.

Chapter 2

Methods for Analyzing Left Censored Data

In this chapter, we will review the statistical methods which have been developed to analyze time to event data with a focus on methods applied to left censored data.

2.1 Non-Parametric Methods

Non-parametric methods use related data to estimate survival rate instead of assuming a distributional shape for the data. The well-known Kaplan-Meier (KM)[48] estimator is a non-parametric approach originally developed for handling right-censored data, which estimates the survivor function, or $1 - F(t)$, where $F(t)$ is the CDF. It is defined as

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

where n_i is the number of survivors (persons at risk prior to time t_i minus the number of censored observations) and d_i is the number of deaths at time t_i.

The KM method can be used for left censored data in two ways. The first way involves converting the left censored data to right censored, calculating the survival probabilities using the KM method, then flipping it back to the original scale.[49] The second way deals directly with the left censored data and has been termed the Reverse Kaplan-Meier (RKM) estimator[50], or equivalently, Turnbull's method,[51] which generalizes the KM estimator to include both left and right censoring. The RKM estimator is calculated similarly to the above formula but with the censoring indicator reversed, i.e., it estimates the CDF, $F(t)$, rather than $1 - F(t)$. It can be

denoted as

$$\hat{S}(t) = \prod_{t>t_i} \frac{n_i - d_i}{n_i}$$

These KM and RKM estimators are mainly used to describe the survivorship patterns of a population or compare the survivorship patterns of two populations. While it may be advantageous not to assume a distributional shape, especially in cases where the data do not follow a standard distribution, there are many disadvantages. Note that these methods are descriptive in nature and do not have the ability to control for time-invariant or time-dependent covariates.

2.2 Semi-Parametric Methods

Semi-parametric methods are termed as such as they have parametric and non-parametric components. The most common semi-parametric method used in survival analysis which can account for covariates, is the Cox Proportional Hazards Model,[11] denoted as

$$\lambda(t) = \lambda_0(t)e^{x_i\beta}$$

where $\lambda_0(t)$ is the baseline hazard function, x_i is the set of covariates, and β are parameters estimating covariate effects on hazard. This model is classed as semi-parametric since no assumptions are made on the baseline hazard function (non-parametric component) but the effect of the covariates on the hazard rate assumes a parametric form. This is advantageous in settings where the distribution of the underlying hazard is not known or it is not of interest to know the distribution of the baseline hazard rate for the research question. In these cases, the risk of incorrectly specifying the baseline hazard is more detrimental than not knowing the shape of the hazard function.

Covariate effects are estimated by maximizing the partial likelihood as opposed

to the likelihood. The partial likelihood function can be written concisely as

$$PL = \prod_{i=1}^{n} \left[\frac{e^{x_i \beta}}{\sum_{j=1}^{n} Y_{ij}\ e^{x_j \beta}} \right]^{\delta_i}$$

where $Y_{ij} = 1$ if $t_j \geq t_i; 0$ otherwise, and δ_i is the censoring indicator.

2.3 Parametric Methods

Parametric models involve assuming a specific distribution for the baseline Hazard Rate (HR). Let T be a non-negative random variable denoting time to some event. Then the HR of T is the instantaneous rate of the event occurring in the interval $[t, t + \Delta t)$ given that the event has not yet occurred. It is defined in notation as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

2.3.1 Parametric Reversed Hazards Model

For left censored data, the Parametric Reversed Hazard (PRH) model,[52] which is a fully parametric model based on the Reversed Hazard Rate (RHR) has been developed. In the case of analyzing survival data in the presence of left censoring, reversed hazard rates are more appropriate to use since estimators of hazard rates tend to be unstable.[52] The RHR[53] of T is the instantaneous rate of the event occurring in an infinitesimal time width, Δt, preceding t, given that the event occurred before time t. It is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t - \Delta t \leq T | T \leq t)}{\Delta t}$$

In terms of the distribution function, $F(t)$, and probability density function, $f(t)$, this can be written as

$$\lambda(t) = \frac{f(t)}{F(t)}$$

Let X be a $p \times 1$ vector of covariates. We can now define the PRH model by

$$\lambda(t|X) = \lambda_0(t) g(\beta; X)$$

where $\lambda_0(t)$ is the baseline RHR, $g(.)$ is a nonnegative function of X and β (a $p \times 1$ vector of regression parameters). $\lambda(t|X)$ is the RHR of T given the covariates X.

The PRH model can be expressed in terms of the distribution function as

$$F(t|X) = F_0(t)^{g(\beta;X)}$$

where $F(t|X)$ is the distribution function of T given X and $F_0(t)$ is the baseline distribution function in the absence of covariates.

Suppose that the lifetime random variable T is randomly left censored by Z. In practice, we may observe the vectors (Y, δ, X), where $Y = \max(T, Z)$ and $\delta = I(T = Y)$ with $I(.)$ being the indicator function. The likelihood function can then be written as

$$L(\beta, y) = \prod_{i=1}^{n} f(y_i|x_i)^{\delta_i} F(y_i|x_i)^{1-\delta_i}$$

Using the method of maximum likelihood, we can then derive estimates for the parameters in this model. This general notation can be applied to any distribution where the specifications of the PRH model is derived for the distributions used in this study (shown in Chapter 4).

2.4 Bayesian Methods

Bayesian inference starts with the likelihood distribution of the data given the model parameters, $p(Y|\theta)$, and the prior information on the distribution of the model parameters, $p(\theta)$. Then, using Bayes' Theorem, inference is made based on the posterior distribution:

$$\begin{aligned} p(\theta|Y) &= \frac{p(Y|\theta)p(\theta)}{p(Y)} \\ &= \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta)d\theta} \\ &\propto p(Y|\theta)p(\theta) \end{aligned}$$

Using this approach, Huynh et al.[54] developed a Bayesian model for analyzing left censored data. In this model, the censored observations, $Y_{i,cen}$, are treated as missing values. The posterior distribution of these censored values, in addition to the model parameters, θ, are obtained based on the observed data, $Y_{i,obs}$.

$$p(\theta, Y_{cen}|Y_{obs}) \propto p(\theta) \times \prod_{observed} [p(Y_{i,obs}|\theta) I(Y_{i,obs} > LOD_i)] \times \prod_{censored} [p(Y_{i,cen}|\theta) I(Y_{i,cen} \leq LOD_i)]$$

where $p(\theta)$ is the prior distribution, LOD_i is the limit of detection for each observation, and $I(.)$ denotes an indicator function which will ensure that each imputed censored value is not greater than its' respective limit of detection. This model can use either the PDF or CDF for the censored values since theoretically they would be equivalent.[54]

Chapter 3

Derivation of Parametric Reversed Hazards Model

In this chapter, we extend the work done by Variyath and Sankaran[52] on developing a PRH model using an Inverse Weibull distribution. Using the same technique, we will derive the PRH model for the Exponential, Generalized Exponential, Log-normal, Inverse Gaussian, Log-logistic, Gompertz-Makeham, Gamma, Generalized Gamma, Inverse Gamma, Generalized Inverse Gamma, Weibull, Generalized Inverse Weibull, Modified Weibull, Flexible Weibull, Power Generalized Weibull, and the Marshal-Olkin distribution.

3.1 Inverse Weibull Distribution

When the lifetime random variable follows an inverted Weibull distribution, the baseline distribution function is given by

$$F_0(t) = e^{-\gamma/t^\alpha}, \quad t > 0; \alpha, \gamma > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\gamma\alpha}{t^{\alpha+1}}$$

Note that the baseline Reversed Hazard Rate is decreasing as t increases. In the presence of the covariates X and assuming that $g(\beta; X) = \exp(x_i\beta)$ (see Section 2.3), we have the following

$$\lambda(t|X) = \frac{\gamma\alpha}{t^{\alpha+1}} \exp(x_i\beta)$$

$$F(t|X) = e^{-(\gamma/t^{\alpha})\exp(x_i\beta)}$$

$$f(t|X) = \frac{\gamma\alpha\exp(x_i\beta)}{t^{\alpha+1}} e^{-(\gamma/t^{\alpha})\exp(x_i\beta)}$$

From these, the likelihood function for the inverted Weibull is obtained as

$$L(\beta,\alpha,\gamma,y) = \prod_{i=1}^{n}\left[\frac{\gamma\alpha\exp(x_i\beta)}{y_i^{\alpha+1}} e^{-(\gamma/y_i^{\alpha})\exp(x_i\beta)}\right]^{\delta_i}\left[e^{-(\gamma/y_i^{\alpha})\exp(x_i\beta)}\right]^{1-\delta_i}$$

so that the log likelihood function is

$$l(\beta,\alpha,\gamma,y) = \sum_{i=1}^{n}\delta_i x_i\beta + \sum_{i=1}^{n}\delta_i(\ln\gamma + \ln\alpha) - (\alpha+1)\sum_{i=1}^{n}\delta_i \ln y_i - \gamma\sum_{i=1}^{n}\frac{\exp(x_i\beta)}{y_i^{\alpha}}$$

3.2 Exponential Distribution

When the lifetime random variable follows an Exponential distribution, the baseline distribution function is given by

$$F_0(t) = 1 - e^{-t/\gamma}, \quad t > 0; \gamma > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{e^{-t/\gamma}}{\gamma(1-e^{-t/\gamma})}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{e^{-t/\gamma}}{\gamma(1-e^{-t/\gamma})}\exp(x_i\beta)$$

$$F(t|X) = (1-e^{-t/\gamma})^{\exp(x_i\beta)}$$

$$f(t|X) = \frac{e^{-t/\gamma}}{\gamma}\exp(x_i\beta)(1-e^{-t/\gamma})^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Exponential distribution is obtained as

$$L(\beta,\gamma,t) = \prod_{i=1}^{n}\left[\frac{e^{-t_i/\gamma}}{\gamma}\exp(x_i\beta)(1-e^{-t_i/\gamma})^{\exp(x_i\beta)-1}\right]^{\delta_i}\left[(1-e^{-t_i/\gamma})^{\exp(x_i\beta)}\right]^{1-\delta_i}$$

so that the log likelihood function is

$$l(\beta,\gamma,t) = \sum_{i=1}^{n}\delta_i x_i\beta - \frac{1}{\gamma}\sum_{i=1}^{n}\delta_i t_i - \sum_{i=1}^{n}\delta_i \ln\gamma + \sum_{i=1}^{n}(e^{x_i\beta}-\delta_i)\ln(1-e^{-t_i/\gamma})$$

3.3 Log-normal Distribution

When the lifetime random variable follows a Log-normal distribution, the baseline distribution function is given by

$$F_0(t) = \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right), \quad t > 0; \mu, \sigma > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\frac{1}{\sqrt{2\pi}\sigma t}\exp\left(-\frac{[\ln(t)-\mu]^2}{2\sigma^2}\right)}{\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\exp\left(-\frac{[\ln(t)-\mu]^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma t\ \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}\exp(x_i\beta)$$

$$F(t|X) = \left[\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \frac{1}{\sqrt{2\pi}\sigma t}\exp\left(-\frac{[\ln(t)-\mu]^2}{2\sigma^2} + x_i\beta\right)\left[\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)\right]^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Log-normal distribution is obtained as

$$L(\mu,\sigma,t) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma t}\exp\left(-\frac{[\ln(t)-\mu]^2}{2\sigma^2} + x_i\beta\right)\left[\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)\right]^{\exp(x_i\beta)-1}\right]^{\delta_i}$$

$$\times\left[\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)\right]^{(1-\delta_i)\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\mu,\sigma,t) = \sum_{i=1}^{n}\delta_i x_i\beta - \sum_{i=1}^{n}\delta_i\ln(\sqrt{2\pi}\sigma t_i) + \sum_{i=1}^{n}\delta_i\frac{[\ln(t)-\mu]^2}{2\sigma^2}$$

$$+\sum_{i=1}^{n}(e^{x_i\beta}-\delta_i)\ln\left[\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)\right]$$

3.4 Inverse Gaussian Distribution

When the lifetime random variable follows a Inverse Gaussian distribution, the baseline distribution function is given by

$$F_0(t) = \Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}-1\right)\right] - exp\left(\frac{2\gamma}{\alpha}\right)\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}+1\right)\right], \quad t>0; \alpha,\gamma>0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\sqrt{\frac{\gamma}{2\pi t^3}}\exp\left[\frac{-\gamma(t-\alpha)^2}{2\alpha^2 t}\right]}{\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}-1\right)\right] - \exp\left(\frac{2\gamma}{\alpha}\right)\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}+1\right)\right]}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\sqrt{\frac{\gamma}{2\pi t^3}}\exp\left[\frac{-\gamma(t-\alpha)^2}{2\alpha^2 t} + x_i\beta\right]}{\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}-1\right)\right] - \exp\left(\frac{2\gamma}{\alpha}\right)\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}+1\right)\right]}$$

$$F(t|X) = \left\{\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}-1\right)\right] - \exp\left(\frac{2\gamma}{\alpha}\right)\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}+1\right)\right]\right\}^{\exp(x_i\beta)}$$

$$f(t|X) = \sqrt{\frac{\gamma}{2\pi t^3}}\exp\left[\frac{-\gamma(t-\alpha)^2}{2\alpha^2 t} + x_i\beta\right]$$

$$\times\left\{\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}-1\right)\right] - \exp\left(\frac{2\gamma}{\alpha}\right)\Phi\left[-\sqrt{\frac{\gamma}{t}}\left(\frac{t}{\alpha}+1\right)\right]\right\}^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Inverse Gaussian distribution is obtained as

$$L(\alpha,\gamma,t) = \prod_{i=1}^{n}\left\{\sqrt{\frac{\gamma}{2\pi t_i^3}}\exp\left[\frac{-\gamma(t_i-\alpha)^2}{2\alpha^2 t_i} + x_i\beta\right]\right\}^{\delta_i}$$

$$\times\left\{\Phi\left[-\sqrt{\frac{\gamma}{t_i}}\left(\frac{t_i}{\alpha}-1\right)\right] - \exp\left(\frac{2\gamma}{\alpha}\right)\Phi\left[-\sqrt{\frac{\gamma}{t_i}}\left(\frac{t_i}{\alpha}+1\right)\right]\right\}^{\exp(x_i\beta)-\delta_i}$$

so that the log likelihood function is

$$l(\alpha,\gamma,t) = \sum_{i=1}^{n}\delta_i x_i\beta + \frac{1}{2}\sum_{i=1}^{n}\delta_i\ln\left(\frac{\gamma}{2\pi t_i^3}\right) - \sum_{i=1}^{n}\left[\frac{\delta_i\gamma(t_i-\alpha)^2}{2\alpha^2 t_i}\right]$$

$$+\sum_{i=1}^{n}\left[e^{x_i\beta}-\delta_i\right]\ln\left\{\Phi\left[-\sqrt{\frac{\gamma}{t_i}}\left(\frac{t_i}{\alpha}+1\right)\right]\right\} + \sum_{i=1}^{n}\left[e^{x_i\beta}-\delta_i\right]\ln\left[1-\exp\left(\frac{2\gamma}{\alpha}\right)\right]$$

3.5 Log-logistic Distribution

When the lifetime random variable follows a Log-logistic distribution, the baseline distribution function is given by

$$F_0(t) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{-\omega}}, \quad t > 0; \alpha, \omega > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\frac{\omega}{t}}{1+\left(\frac{t}{\alpha}\right)^{\omega}}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\frac{\omega}{t}\exp(x_i\beta)}{1+\left(\frac{t}{\alpha}\right)^{\omega}}$$

$$F(t|X) = \left[\frac{1}{1+\left(\frac{t}{\alpha}\right)^{-\omega}}\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \left[\frac{\frac{\omega}{t}\exp(x_i\beta)}{1+\left(\frac{t}{\alpha}\right)^{\omega}}\right]\left[\frac{1}{1+\left(\frac{t}{\alpha}\right)^{-\omega}}\right]^{\exp(x_i\beta)}$$

From these, the likelihood function for the Log-logistic distribution is obtained as

$$L(\alpha, \omega, t) = \prod_{i=1}^{n}\left[\frac{\frac{\omega}{t_i}\exp(x_i\beta)}{1+\left(\frac{t_i}{\alpha}\right)^{\omega}}\right]^{\delta_i}\left[\frac{1}{1+\left(\frac{t_i}{\alpha}\right)^{-\omega}}\right]^{\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\alpha, \omega, t) = \sum_{i=1}^{n}\delta_i x_i\beta + \sum_{i=1}^{n}\delta_i \ln\left(\frac{\omega}{t_i}\right) - \sum_{i=1}^{n}\delta_i \ln\left[1+\left(\frac{t_i}{\alpha}\right)^{\omega}\right] + \sum_{i=1}^{n}e^{x_i\beta}\ln\left[1+\left(\frac{t_i}{\alpha}\right)^{-\omega}\right]$$

3.6 Gompertz-Makeham Distribution

When the lifetime random variable follows a Gompertz-Makeham distribution, the baseline distribution function is given by

$$F_0(t) = 1 - \exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right], \quad t > 0; \alpha, \gamma > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\alpha e^{\gamma t} \exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right]}{1-\exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right]}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\alpha e^{\gamma t} \exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right] \exp(x_i\beta)}{1-\exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right]}$$

$$F(t|X) = \left\{1-\exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right]\right\}^{\exp(x_i\beta)}$$

$$f(t|X) = \alpha e^{\gamma t} \exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right] e^{x_i\beta} \left\{1-\exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t}-1\right)\right]\right\}^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Gompertz-Makeham distribution is obtained as

$$L(\alpha,\gamma,t) = \prod_{i=1}^{n} \left\{\alpha e^{\gamma t_i} \exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t_i}-1\right)\right] e^{x_i\beta}\right\}^{\delta_i} \times \left\{1-\exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t_i}-1\right)\right]\right\}^{\exp(x_i\beta)-\delta_i}$$

so that the log likelihood function is

$$l(\alpha,\gamma,t) = \sum_{i=1}^{n}\delta_i x_i\beta + \sum_{i=1}^{n}\delta_i \ln\alpha + \sum_{i=1}^{n}\delta_i\gamma t_i - \sum_{i=1}^{n}\frac{\delta_i\alpha}{\gamma}\left(e^{\gamma t_i}-1\right) + \sum_{i=1}^{n}\left(e^{x_i\beta}-\delta_i\right)\ln\left\{1-\exp\left[-\frac{\alpha}{\gamma}\left(e^{\gamma t_i}-1\right)\right]\right\}$$

3.7 Gamma Distribution

When the lifetime random variable follows a Gamma distribution, the baseline distribution function is given by

$$F_0(t) = \frac{\gamma(\alpha,\omega t)}{\Gamma(\alpha)}, \quad t>0; \alpha,\omega>0$$

where $\gamma(\alpha,t)$ is the incomplete Gamma function and $\Gamma(\alpha)$ is the complete Gamma function. The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\omega^\alpha t^{\alpha-1} e^{-\omega t}}{\gamma(\alpha,\omega t)}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\omega^{\alpha} t^{\alpha-1} \exp(-\omega t + x_i\beta)}{\gamma(\alpha, \omega t)}$$

$$F(t|X) = \left[\frac{\gamma(\alpha, \omega t)}{\Gamma(\alpha)}\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \frac{\omega^{\alpha} t^{\alpha-1} \exp(-\omega t + x_i\beta)}{\gamma(\alpha, \omega t)} \left[\frac{\gamma(\alpha, \omega t)}{\Gamma(\alpha)}\right]^{\exp(x_i\beta)}$$

From these, the likelihood function for the Gamma distribution is obtained as

$$L(\alpha, \omega, t) = \prod_{i=1}^{n} \left[\frac{\omega^{\alpha} t_i^{\alpha-1} \exp(-\omega t_i + x_i\beta)}{\gamma(\alpha, \omega t_i)}\right]^{\delta_i} \left[\frac{\gamma(\alpha, \omega t_i)}{\Gamma(\alpha)}\right]^{\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\alpha, \omega, t) = \sum_{i=1}^{n} \delta_i x_i\beta + \sum_{i=1}^{n} \delta_i \alpha \ln \omega + (\alpha - 1) \sum_{i=1}^{n} \delta_i \ln t_i - \sum_{i=1}^{n} \delta_i \omega t_i$$

$$+ \sum_{i=1}^{n} \left(e^{x_i\beta} - \delta_i\right) \ln\left[\gamma(\alpha, \omega t_i)\right] - \sum_{i=1}^{n} e^{x_i\beta} \ln\left[\Gamma(\alpha)\right]$$

Note that the exponential distribution is a special case of this result.

3.8 Generalized Gamma Distribution

When the lifetime random variable follows a Generalized Gamma distribution, the baseline distribution function is given by

$$F_0(t) = \frac{\gamma\left[\frac{\alpha}{\omega}, (\lambda t)^{\omega}\right]}{\Gamma\left(\frac{\alpha}{\omega}\right)}, \quad t > 0; \alpha, \omega, \lambda > 0$$

where $\gamma(\alpha, t)$ is the incomplete Gamma function and $\Gamma(\alpha)$ is the complete Gamma function. The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\omega \lambda^{\alpha} t^{\alpha-1} e^{-(\lambda t)^{\omega}}}{\gamma\left[\frac{\alpha}{\omega}, (\lambda t)^{\omega}\right]}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\omega \lambda^{\alpha} t^{\alpha-1} e^{-(\lambda t)^{\omega} + x_i\beta}}{\gamma\left[\frac{\alpha}{\omega}, (\lambda t)^{\omega}\right]}$$

$$F(t|X) = \left[\frac{\gamma\left[\frac{\alpha}{\omega}, (\lambda t)^{\omega}\right]}{\Gamma\left(\frac{\alpha}{\omega}\right)}\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \frac{\omega\lambda^{\alpha}t^{\alpha-1}e^{-(\lambda t)^{\omega}+x_i\beta}}{\gamma\left[\frac{\alpha}{\omega}, (\lambda t)^{\omega}\right]}\left[\frac{\gamma\left[\frac{\alpha}{\omega}, (\lambda t)^{\omega}\right]}{\Gamma\left(\frac{\alpha}{\omega}\right)}\right]^{\exp(x_i\beta)}$$

From these, the likelihood function for the Generalized Gamma distribution is obtained as

$$L(\alpha, \omega, \lambda, t) = \prod_{i=1}^{n}\left[\frac{\omega\lambda^{\alpha}t_i^{\alpha-1}e^{-(\lambda t_i)^{\omega}+x_i\beta}}{\gamma\left[\frac{\alpha}{\omega}, (\lambda t_i)^{\omega}\right]}\right]^{\delta_i}\left[\frac{\gamma\left[\frac{\alpha}{\omega}, (\lambda t_i)^{\omega}\right]}{\Gamma\left(\frac{\alpha}{\omega}\right)}\right]^{\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\alpha, \omega, \lambda, t) = \sum_{i=1}^{n}\delta_i x_i\beta + \sum_{i=1}^{n}\delta_i\alpha\ln(\omega\lambda) + (\alpha-1)\sum_{i=1}^{n}\delta_i\ln t_i - \sum_{i=1}^{n}\delta_i(\lambda t_i)^{\omega}$$

$$+\sum_{i=1}^{n}(e^{x_i\beta}-\delta_i)\ln\left[\gamma\left[\frac{\alpha}{\omega}, (\lambda t_i)^{\omega}\right]\right] - \sum_{i=1}^{n}e^{x_i\beta}\ln\left[\Gamma\left(\frac{\alpha}{\omega}\right)\right]$$

3.9 Inverse Gamma Distribution

When the lifetime random variable follows a Inverse Gamma distribution, the baseline distribution function is given by

$$F_0(t) = \frac{\gamma(\alpha, t)}{\Gamma(\alpha)}, \quad t > 0; \alpha, \omega > 0$$

where $\gamma(\alpha, t)$ is the incomplete Gamma function and $\Gamma(\alpha)$ is the complete Gamma function. The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\omega^{\alpha}t^{-\alpha-1}e^{-\omega/t}}{\gamma(\alpha, t)}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\omega^{\alpha}t^{-\alpha-1}e^{-\omega/t}}{\gamma(\alpha, t)}\exp(x_i\beta)$$

$$F(t|X) = \left[\frac{\gamma(\alpha, t)}{\Gamma(\alpha)}\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \frac{\omega^{\alpha} t^{-\alpha-1} e^{-\omega/t}}{\gamma(\alpha,t)} \exp(x_i\beta) \left[\frac{\gamma(\alpha,t)}{\Gamma(\alpha)}\right]^{\exp(x_i\beta)}$$

From these, the likelihood function for the Inverse Gamma distribution is obtained as

$$L(\alpha,\omega,t) = \prod_{i=1}^{n} \left[\frac{\omega^{\alpha} t^{-\alpha-1} e^{-\omega/t+x_i\beta}}{\gamma(\alpha,t)}\right]^{\delta_i} \left[\frac{\gamma(\alpha,t)}{\Gamma(\alpha)}\right]^{\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\alpha,\omega,t) = \sum_{i=1}^{n} \delta_i x_i\beta + \sum_{i=1}^{n} \delta_i \alpha \ln\omega - (\alpha+1)\sum_{i=1}^{n} \delta_i \ln t_i - \sum_{i=1}^{n} \frac{\delta_i \omega}{t_i}$$
$$+ \sum_{i=1}^{n} \left(e^{x_i\beta} - \delta_i\right) \ln\left[\gamma(\alpha,t_i)\right] - \sum_{i=1}^{n} e^{x_i\beta} \ln\left[\Gamma(\alpha)\right]$$

3.10 Weibull Distribution

When the lifetime random variable follows a Weibull distribution, the baseline distribution function is given by

$$F_0(t) = 1 - e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}, \quad t > 0; \alpha, \gamma > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\left(\frac{\alpha}{\gamma}\right)\left(\frac{t}{\gamma}\right)^{\alpha-1} e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}}{1 - e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\left(\frac{\alpha}{\gamma}\right)\left(\frac{t}{\gamma}\right)^{\alpha-1} e^{-\left(\frac{t}{\gamma}\right)^{\alpha}+x_i\beta}}{1 - e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}}$$

$$F(t|X) = \left[1 - e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \left[\left(\frac{\alpha}{\gamma}\right)\left(\frac{t}{\gamma}\right)^{\alpha-1} e^{-\left(\frac{t}{\gamma}\right)^{\alpha}+x_i\beta}\right]\left[1 - e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}\right]^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Weibull distribution is obtained as

$$L(\alpha,\gamma,t) = \prod_{i=1}^{n} \left[\left(\frac{\alpha}{\gamma}\right)\left(\frac{t_i}{\gamma}\right)^{\alpha-1} e^{-\left(\frac{t_i}{\gamma}\right)^{\alpha}+x_i\beta}\right]^{\delta_i} \left[1 - e^{-\left(\frac{t_i}{\gamma}\right)^{\alpha}}\right]^{\exp(x_i\beta)-\delta_i}$$

so that the log likelihood function is

$$l(\alpha,\gamma,t)=\sum_{i=1}^{n}\delta_i x_i\beta+\sum_{i=1}^{n}\delta_i\ln\alpha-\sum_{i=1}^{n}\delta_i\alpha\ln\gamma+(\alpha-1)\sum_{i=1}^{n}\delta_i\ln t_i$$
$$-\sum_{i=1}^{n}\delta_i\left(\frac{t_i}{\gamma}\right)^{\alpha}+\sum_{i=1}^{n}\left(e^{x_i\beta}-\delta_i\right)\ln\left[1-e^{-\left(\frac{t}{\gamma}\right)^{\alpha}}\right]$$

3.11 Generalized Inverse Weibull Distribution

When the lifetime random variable follows a Generalized Inverse Weibull distribution, the baseline distribution function is given by

$$F_0(t)=e^{-\gamma\left(\frac{\lambda}{t}\right)^{\alpha}},\quad t>0;\alpha,\gamma,\lambda>0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t)=\alpha\gamma\lambda^{\alpha}t^{-(\alpha-1)}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X)=\alpha\gamma\lambda^{\alpha}t^{-(\alpha-1)}e^{x_i\beta}$$

$$F(t|X)=\left[e^{-\gamma\left(\frac{\lambda}{t}\right)^{\alpha}}\right]^{\exp(x_i\beta)}$$

$$f(t|X)=\alpha\gamma\lambda^{\alpha}t^{-(\alpha-1)}e^{x_i\beta}\left[e^{-\gamma\left(\frac{\lambda}{t}\right)^{\alpha}}\right]^{\exp(x_i\beta)}$$

From these, the likelihood function for the Generalized Inverse Weibull distribution is obtained as

$$L(\alpha,\gamma,\lambda,t)=\prod_{i=1}^{n}\left[\alpha\gamma\lambda^{\alpha}t_i^{-(\alpha-1)}e^{x_i\beta}\right]^{\delta_i}\left[e^{-\gamma\left(\frac{\lambda}{t_i}\right)^{\alpha}}\right]^{\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\alpha,\gamma,\lambda,t)=\sum_{i=1}^{n}\delta_i x_i\beta+\sum_{i=1}^{n}\delta_i\ln\alpha+\sum_{i=1}^{n}\delta_i\ln\gamma+\sum_{i=1}^{n}\delta_i\alpha\ln\lambda$$
$$-(\alpha-1)\sum_{i=1}^{n}\delta_i\ln t_i-\sum_{i=1}^{n}\gamma\left(\frac{\lambda}{t_i}\right)^{\alpha}e^{x_i\beta}$$

3.12 Modified Weibull Distribution

When the lifetime random variable follows a Modified Weibull distribution[55], the baseline distribution function is given by

$$F_0(t) = 1 - \exp(-\gamma t^{\alpha} e^{\lambda t}), \quad t > 0; \alpha, \gamma, \lambda > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\gamma(\alpha + \lambda t)t^{\alpha-1} \exp(\lambda t) \exp(-\gamma t^{\alpha} e^{\lambda t})}{1 - \exp(-\gamma t^{\alpha} e^{\lambda t})}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\gamma(\alpha + \lambda t)t^{\alpha-1} \exp(\lambda t) \exp(-\gamma t^{\alpha} e^{\lambda t}) \exp(x_i\beta)}{1 - \exp(-\gamma t^{\alpha} e^{\lambda t})}$$

$$F(t|X) = \left[1 - \exp\left(-\gamma t^{\alpha} e^{\lambda t}\right)\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \gamma(\alpha + \lambda t)t^{\alpha-1} \exp(\lambda t) \exp(-\gamma t^{\alpha} e^{\lambda t}) \exp(x_i\beta) \left[1 - \exp(-\gamma t^{\alpha} e^{\lambda t})\right]^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Modified Weibull distribution is obtained as

$$L(\alpha, \gamma, \lambda, t) = \prod_{i=1}^{n} \left[\gamma(\alpha + \lambda t_i)t_i^{\alpha-1} \exp(\lambda t_i) \exp(-\gamma t_i^{\alpha} e^{\lambda t_i}) \exp(x_i\beta)\right]^{\delta_i}$$

$$\times \left[1 - \exp(-\gamma t_i^{\alpha} e^{\lambda t_i})\right]^{\exp(x_i\beta)-\delta_i}$$

so that the log likelihood function is

$$l(\alpha, \gamma, \lambda, t) = \sum_{i=1}^{n} \delta_i x_i\beta + \sum_{i=1}^{n} \delta_i \ln\gamma + \sum_{i=1}^{n} \delta_i \ln(\alpha + \lambda t_i) + (\alpha - 1)\sum_{i=1}^{n} \delta_i \ln t_i + \sum_{i=1}^{n} \delta_i \lambda t_i$$

$$- \sum_{i=1}^{n} \delta_i \gamma t_i^{\alpha} e^{\lambda t_i} + \sum_{i=1}^{n} \left(e^{x_i\beta} - \delta_i\right) \ln\left[1 - \exp(-\gamma t_i^{\alpha} e^{\lambda t_i})\right]$$

3.13 Flexible Weibull Distribution

When the lifetime random variable follows a Flexible Weibull distribution[56], the baseline distribution function is given by

$$F_0(t) = 1 - \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right), \quad t > 0; \alpha, \gamma > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\left(\alpha + \frac{\gamma}{t^2}\right) \exp\left(\alpha t - \frac{\gamma}{t}\right) \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right)}{1 - \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right)}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\left(\alpha + \frac{\gamma}{t^2}\right) \exp\left(\alpha t - \frac{\gamma}{t}\right) \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right) \exp(x_i\beta)}{1 - \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right)}$$

$$F(t|X) = \left[1 - \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right)\right]^{\exp(x_i\beta)}$$

$$f(t|X) = \left(\alpha + \frac{\gamma}{t^2}\right) \exp\left(\alpha t - \frac{\gamma}{t}\right) \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right) e^{x_i\beta} \left[1 - \exp\left(-e^{\alpha t - \frac{\gamma}{t}}\right)\right]^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Flexible Weibull distribution is obtained as

$$L(\alpha, \gamma, t) = \prod_{i=1}^{n} \left[\left(\alpha + \frac{\gamma}{t_i^2}\right) \exp\left(\alpha t_i - \frac{\gamma}{t_i}\right) \exp\left(-e^{\alpha t_i - \frac{\gamma}{t_i}}\right) \exp(x_i\beta)\right]^{\delta_i}$$

$$\times \left[1 - \exp\left(-e^{\alpha t_i - \frac{\gamma}{t_i}}\right)\right]^{\exp(x_i\beta)-\delta_i}$$

so that the log likelihood function is

$$l(\alpha, \gamma, t) = \sum_{i=1}^{n} \delta_i x_i\beta + \sum_{i=1}^{n} \delta_i \ln\left(\alpha + \frac{\gamma}{t_i^2}\right) + \sum_{i=1}^{n} \delta_i \left(\alpha t_i - \frac{\gamma}{t_i}\right) - \sum_{i=1}^{n} \delta_i e^{\alpha t_i - \frac{\gamma}{t_i}}$$

$$+ \sum_{i=1}^{n} \left(e^{x_i\beta} - \delta_i\right) \ln\left[1 - \exp\left(-e^{\alpha t_i - \frac{\gamma}{t_i}}\right)\right]$$

3.14 Power Generalized Weibull Distribution

When the lifetime random variable follows a Power Generalized Weibull distribution[57], the baseline distribution function is given by

$$F_0(t) = 1 - \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right], \quad t > 0; \alpha, \gamma, \lambda > 0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t) = \frac{\frac{\alpha}{\gamma\lambda^{\alpha}} t^{\alpha-1}\left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]^{\frac{1}{\gamma}-1} \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]}{1 - \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X) = \frac{\frac{\alpha}{\gamma\lambda^{\alpha}} t^{\alpha-1}\left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]^{\frac{1}{\gamma}-1} \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right] \exp(x_i\beta)}{1 - \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]}$$

$$F(t|X) = \left\{1 - \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]\right\}^{\exp(x_i\beta)}$$

$$f(t|X) = \frac{\alpha}{\gamma\lambda^{\alpha}} t^{\alpha-1}\left[1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right]^{\frac{1}{\gamma}-1} \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right] \exp(x_i\beta)$$
$$\times \left\{1 - \exp\left[1 - \left(1 + \left(\frac{t}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]\right\}^{\exp(x_i\beta)-1}$$

From these, the likelihood function for the Power Generalized Weibull distribution is obtained as

$$L(\alpha, \gamma, \lambda, t) = \prod_{i=1}^{n} \left\{\frac{\alpha}{\gamma\lambda^{\alpha}} t_i^{\alpha-1}\left[1 + \left(\frac{t_i}{\lambda}\right)^{\alpha}\right]^{\frac{1}{\gamma}-1} \exp\left[1 - \left(1 + \left(\frac{t_i}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right] \exp(x_i\beta)\right\}^{\delta_i}$$
$$\times \left\{1 - \exp\left[1 - \left(1 + \left(\frac{t_i}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]\right\}^{\exp(x_i\beta)-\delta_i}$$

so that the log likelihood function is

$$l(\alpha, \gamma, \lambda, t) = \sum_{i=1}^{n} \delta_i x_i\beta + \sum_{i=1}^{n} \delta_i \ln\alpha - \sum_{i=1}^{n} \delta_i \ln\gamma - \sum_{i=1}^{n} \delta_i\alpha \ln\lambda + (\alpha - 1)\sum_{i=1}^{n} \delta_i \ln t_i$$

$$+\left(\frac{1}{\gamma}-1\right)\sum_{i=1}^{n}\delta_i \ln\left[1+\left(\frac{t_i}{\lambda}\right)^{\alpha}\right]+\sum_{i=1}^{n}\delta_i\left[1-\left(1+\left(\frac{t_i}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]$$
$$+\sum_{i=1}^{n}\left(e^{x_i\beta}-\delta_i\right)\ln\left\{1-\exp\left[1-\left(1+\left(\frac{t_i}{\lambda}\right)^{\alpha}\right)^{\frac{1}{\gamma}}\right]\right\}$$

3.15 Marshal-Olkin Distribution

When the lifetime random variable follows a Marshal-Olkin distribution[58], the baseline distribution function is given by

$$F_0(t)=\frac{1-e^{-(\lambda t)^{\alpha}}}{1-(1-\gamma)e^{-(\lambda t)^{\alpha}}},\quad t>0;\alpha,\gamma,\lambda>0$$

The baseline Reversed Hazard Rate of T is then obtained as

$$\lambda_0(t)=\frac{\gamma\alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}}{\left[1-(1-\gamma)e^{-(\lambda t)^{\alpha}}\right]\left[1-e^{-(\lambda t)^{\alpha}}\right]}$$

In the presence of the covariates X, we have the following

$$\lambda(t|X)=\frac{\gamma\alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}e^{x_i\beta}}{\left[1-(1-\gamma)e^{-(\lambda t)^{\alpha}}\right]\left[1-e^{-(\lambda t)^{\alpha}}\right]}$$

$$F(t|X)=\left[\frac{1-e^{-(\lambda t)^{\alpha}}}{1-(1-\gamma)e^{-(\lambda t)^{\alpha}}}\right]^{\exp(x_i\beta)}$$

$$f(t|X)=\frac{\gamma\alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}e^{x_i\beta}\left[1-e^{-(\lambda t)^{\alpha}}\right]^{\exp(x_i\beta)-1}}{\left[1-(1-\gamma)e^{-(\lambda t)^{\alpha}}\right]^{\exp(x_i\beta)+1}}$$

From these, the likelihood function for the Marshal-Olkin distribution is obtained as

$$L(\alpha,\gamma,\lambda,t)=\prod_{i=1}^{n}\left\{\frac{\gamma\alpha\lambda(\lambda t_i)^{\alpha-1}e^{-(\lambda t_i)^{\alpha}}e^{x_i\beta}\left[1-e^{-(\lambda t_i)^{\alpha}}\right]^{\exp(x_i\beta)-1}}{\left[1-(1-\gamma)e^{-(\lambda t_i)^{\alpha}}\right]^{\exp(x_i\beta)+1}}\right\}^{\delta_i}$$
$$\times\left\{\frac{1-e^{-(\lambda t_i)^{\alpha}}}{1-(1-\gamma)e^{-(\lambda t_i)^{\alpha}}}\right\}^{(1-\delta_i)\exp(x_i\beta)}$$

so that the log likelihood function is

$$l(\alpha,\gamma,\lambda,t)=\sum_{i=1}^{n}\delta_i x_i\beta+\sum_{i=1}^{n}\delta_i\ln\gamma+\sum_{i=1}^{n}\delta_i\ln\alpha+\sum_{i=1}^{n}\delta_i\alpha\ln\lambda+(\alpha-1)\sum_{i=1}^{n}\delta_i\ln t_i$$

$$-\sum_{i=1}^{n}\delta_i\left(\lambda t_i\right)^{\alpha}+\sum_{i=1}^{n}\left(e^{x_i\beta}-\delta_i\right)\ln\left[1-e^{-(\lambda t_i)^{\alpha}}\right]$$
$$-\sum_{i=1}^{n}\left(e^{x_i\beta}+\delta_i\right)\ln\left[1-(1-\gamma)e^{-(\lambda t_i)^{\alpha}}\right]$$

Chapter 4

Simulation Study

In this chapter, we will perform a simulation study to determine if the derived distributions from the previous chapter are adaptable for use in a PRH model and compare these distributions to establish a guideline for which distribution/s would best fit left censored HIV VL data.

4.1 Simulation Setup

We generated the simulated data from a Skewed Normal distribution using the *sn* package in R as we expected that this would most closely match the left censored HIV VL data. Different parameters were tested under the Skewed Normal distribution until the closest matching simulated data could be generated. The final parameters chosen were 5 (location), 30 (scale), and 50 (shape) with 100000 observations randomly generated. The distribution of the simulated data can be seen in Figure 4.1.

From the simulated data, we randomly generated samples of size 1000, 2000, and 3000. The percentage of censored observations was 20, 30, and 40 percent. The censoring rate was ensured by creating a censoring indicator where 0 represents a censored observation, 1 otherwise. The indicator was then randomly assigned to the corresponding proportion of observations. Each of the simulation setups was repeated 5000 times to ensure reliability. To assess which distribution model fits best, we used 4 information criteria:

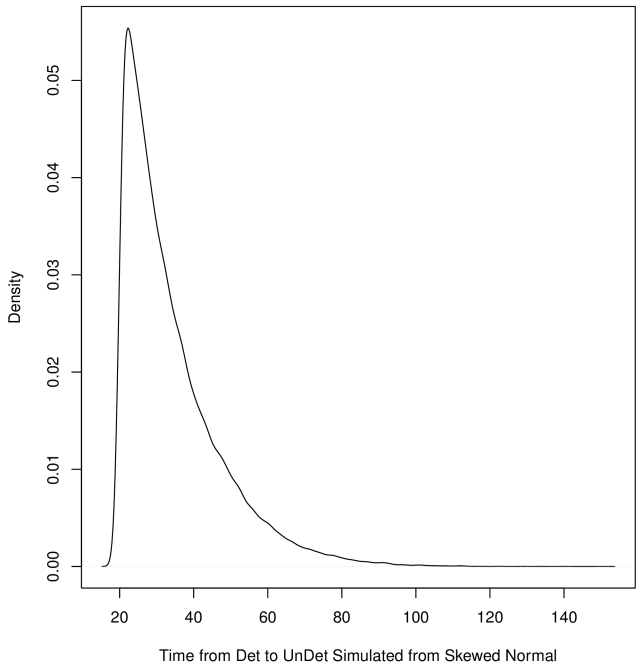

Figure 4.1 Distribution of Simulated Data

- Akaike Information Criterion (AIC) rewards goodness of fit but penalizes the model for increasing the number of estimated parameters.

$$AIC = 2k - 2\ln(L)$$

- Corrected Akaike Information Criterion (AICC) corrects the AIC for overfitting of the data in cases where the sample size is relatively small compared to the number of parameters in the model.

$$AICC = AIC + \frac{2k(k+1)}{n-k-1}$$

- Hannan-Quinn Information Criterion (HQIC) is often cited in the literature but, unlike AIC, it is not asymptotically efficient.

$$HQIC = 2k\ln(\ln(n)) - 2\ln(L)$$

- Bozdogan's Consistent Akaike Information Criterion (CAIC) is another adjusted form of AIC which is consistent.

$$CAIC = k(\ln(n) + 1) - 2\ln(L)$$

where k is the number of parameters to be estimated, L is the maximum value of the likelihood function, and n is the number of observations. The model with the smallest average AIC, AICC, HQIC, and CAIC value was determined to be the model with the best fit. This simulation study was conducted using the Statistical Computing Software, R version 3.2.5[59] with summary results presented in the following section.

4.2 Simulation Results

Results of the simulation study are summarized in Tables 4.1-4.3. Table 4.1 summarizes the results for the simulated data with a censoring rate of 20%, Table 4.2 for data with censoring rate of 30%, and Table 4.3 for data with censoring rate 40%. From these tables, it is clear that the Generalized Inverse Weibull distribution performs the best, having the lowest average AIC, AICC, HQIC, and CAIC values. Following closely behind in performance are the Log-Logistic, Log-Normal, Inverse Gaussian, and Gamma distributions, respectively. This is consistent across all censoring rates and sample sizes. The consistently worst performing distributions are the Modified Weibull, Inverse Weibull, Inverse Gamma, Power Generalized Weibull, and Exponential distributions, respectively.

Table 4.1 Average summary measures across 5000 simulations from simulation study with censoring rate 20%

Distribution	Sample Sizes											
	1000				2000				3000			
	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC
Exponential	8753.05	8753.06	8754.92	8758.96	17504.19	17504.20	17506.25	17510.79	26255.36	26255.36	26257.52	26262.37
Log-Normal	6162.11	6162.12	6165.84	6173.92	12323.86	12323.87	12327.97	12337.06	18485.63	18485.64	18489.96	18499.65
Inverse Gaussian	6163.50	6163.51	6167.23	6175.31	12326.67	12326.68	12330.79	12339.88	18489.88	18489.89	18494.20	18503.90
Gamma	6199.30	6199.31	6203.03	6211.12	12398.14	12398.14	12402.25	12411.34	18597.16	18597.16	18601.48	18611.17
Generalized Gamma	6206.54	6206.57	6212.14	6224.27	12410.79	12410.80	12416.96	12430.60	18615.33	18615.34	18621.82	18636.35
Inverse Gamma	10622.02	10622.03	10625.75	10633.83	21240.12	21240.13	21244.24	21253.32	31858.19	31858.19	31862.51	31872.20
Log-Logistic	6115.87	6115.88	6119.60	6127.68	12230.91	12230.92	12235.03	12244.12	18346.09	18346.09	18350.41	18360.10
Weibull	6502.31	6502.32	6506.04	6514.13	13007.93	13007.94	13012.05	13021.14	19515.02	19515.03	19519.34	19529.04
Inverse Weibull	13992.74	13992.75	13996.47	14004.55	27981.55	27981.56	27985.66	27994.75	41970.35	41970.35	41974.67	41984.36
Generalized Inverse Weibull	5959.85	5959.87	5965.44	5977.57	11916.55	11916.56	11922.72	11936.35	17873.49	17873.50	17879.97	17894.51
Flexible Weibull	7768.04	7768.05	7771.77	7779.86	15488.38	15488.39	15492.50	15501.59	23311.80	23311.80	23316.12	23325.81
Marshal-Olkin	7532.91	7532.94	7538.51	7550.63	15059.88	15059.90	15066.05	15079.69	22586.86	22586.87	22593.34	22607.88
Power Generalized Weibull	10492.37	10492.38	10496.10	10504.19	20980.81	20980.82	20984.92	20994.01	31469.28	31469.28	31473.60	31483.29
Modified Weibull	2.97e63	2.97e63	2.97e63	2.97e63	1.27e68	1.27e68	1.27e68	1.27e68	1.00e68	1.00e68	1.00e68	1.00e68
Gompertz	8177.79	8177.80	8181.52	8189.60	16177.93	16177.94	16182.05	16191.13	24099.00	24099.01	24103.32	24113.01

Table 4.2 Average summary measures across 5000 simulations from simulation study with censoring rate 30%

Distribution	Sample Sizes											
	1000				2000				3000			
	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC
Exponential	7752.73	7752.74	7754.60	7758.64	15503.37	15503.37	15505.43	15509.97	23254.14	23254.14	23256.30	23261.15
Log-Normal	5571.09	5571.10	5574.82	5582.91	11141.41	11141.41	11145.52	11154.61	16711.43	16711.43	16715.75	16725.44
Inverse Gaussian	5572.56	5572.57	5576.29	5584.38	11144.38	11144.39	11148.49	11157.58	16715.90	16715.90	16720.22	16729.91
Gamma	5603.77	5603.78	5607.50	5615.58	11206.68	11206.68	11210.79	11219.88	16809.04	16809.04	16813.36	16823.05
Generalized Gamma	5611.47	5611.50	5617.07	5629.20	11219.82	11219.84	11225.99	11239.63	16828.00	16828.01	16834.48	16849.02
Inverse Gamma	9731.22	9731.23	9734.95	9743.04	19458.34	19458.34	19462.45	19471.54	29185.65	29185.66	29189.97	29199.67
Log-Logistic	5521.91	5521.92	5525.64	5533.73	11042.28	11042.28	11046.39	11055.48	16562.88	16562.88	16567.20	16576.89
Weibull	5848.74	5848.75	5852.47	5860.55	11700.72	11700.73	11704.83	11713.92	17551.99	17552.00	17556.31	17566.01
Inverse Weibull	12430.99	12431.01	12434.72	12442.81	24857.86	24857.87	24861.98	24871.07	37284.93	37284.94	37289.25	37298.95
Generalized Inverse Weibull	5401.80	5401.83	5407.40	5419.53	10799.21	10799.22	10805.38	10819.01	16197.20	16197.20	16203.68	16218.22
Flexible Weibull	6843.80	6843.81	6847.53	6855.61	13628.37	13628.37	13632.48	13641.57	20402.93	20402.93	20407.25	20416.94
Marshal-Olkin	7118.74	7118.77	7124.34	7136.47	14231.37	14231.38	14237.54	14251.17	21344.14	21344.15	21350.62	21365.16
Power Generalized Weibull	9251.89	9251.90	9255.62	9263.70	18499.64	18499.64	18503.75	18512.84	27747.57	27747.58	27751.89	27761.59
Modified Weibull	6.15e63	6.15e63	6.15e63	6.15e63	1.95e63	1.95e63	1.95e63	1.95e63	5.20e64	5.20e64	5.20e64	5.20e64
Gompertz	7117.94	7117.95	7121.67	7129.76	14424.68	14424.68	14428.79	14437.88	21422.25	21422.25	21426.57	21436.26

Table 4.3 Average summary measures across 5000 simulations from simulation study with censoring rate 40%

Distribution	Sample Sizes											
	1000				2000				3000			
	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC
Exponential	6741.30	6741.31	6743.17	6747.21	13480.48	13480.48	13482.54	13487.08	20220.29	20220.29	20222.45	20227.30
Log-Normal	4953.90	4953.91	4957.63	4965.71	9905.84	9905.85	9909.95	9919.04	14861.39	14861.39	14865.71	14875.40
Inverse Gaussian	4955.39	4955.40	4959.12	4967.20	9908.88	9908.88	9912.99	9922.08	14865.96	14865.97	14870.29	14879.98
Gamma	4982.80	4982.82	4986.53	4994.62	9964.45	9964.46	9968.56	9977.65	14948.26	14948.26	14952.58	14962.27
Generalized Gamma	4986.95	4986.98	4992.55	5004.68	9972.46	9972.47	9978.63	9992.26	14961.84	14961.85	14968.33	14982.86
Inverse Gamma	8650.96	8650.97	8654.69	8662.77	17297.85	17297.85	17301.96	17311.05	25945.08	25945.09	25949.41	25959.10
Log-Logistic	4905.78	4905.79	4909.51	4917.59	9809.98	9809.99	9814.09	9823.18	14716.61	14716.62	14720.93	14730.62
Weibull	5174.16	5174.17	5177.89	5185.97	10349.18	10349.19	10353.29	10362.38	15527.55	15527.55	15531.87	15541.56
Inverse Weibull	10840.67	10840.68	10844.40	10852.48	21677.25	21677.26	21681.37	21690.46	32514.24	32514.24	32518.56	32528.25
Generalized Inverse Weibull	4815.45	4815.48	4821.05	4833.18	9626.12	9626.13	9632.29	9645.92	14440.87	14440.88	14447.35	14461.89
Flexible Weibull	5926.41	5926.42	5930.14	5938.23	11890.95	11890.96	11895.07	11904.15	17877.73	17877.73	17882.05	17891.74
Marshal-Olkin	6704.64	6704.66	6710.23	6722.36	13403.17	13403.18	13409.34	13422.97	20102.14	20102.15	20108.62	20123.16
Power Generalized Weibull	8003.90	8003.91	8007.63	8015.72	16003.69	16003.69	16007.80	16016.89	24003.98	24003.98	24008.31	24018.00
Modified Weibull	4.60e63	4.60e63	4.60e63	4.60e63	2.09e62	2.09e62	2.09e62	2.09e62	9.34e65	9.34e65	9.34e65	9.34e65
Gompertz	6218.98	6218.99	6222.71	6230.79	12435.16	12435.17	12439.27	12448.36	18726.64	18726.65	18730.97	18740.66

Chapter 5

Real Data Application: SC eHARS Database

In this chapter, we apply the PRH model to the South Carolina Enhanced HIV/AIDS Reporting Surveillance System (SC eHARS) database using the distribution which was found to be the best fit from our simulation study.

5.1 Background

Since January 2004, all health care providers, hospitals, and laboratories are legally mandated to report all CD4 count and VL measurements to the SC Department of Health and Environmental Control (DHEC).[60] This data is stored in the SC eHARS database along with the patient's socio-demographic characteristics. The quality rating of the SC eHARS database exceeds the CDC minimum standards of reporting timeliness with 95% of new cases being reported within 6 months of HIV diagnosis and 98% of all HIV cases reported.[61] Our sample was reduced based on the following selection criteria (summarized in Figure 5.1):

- aged ≥ 13 years or older
- diagnosed or living with HIV infection between January 1, 2005 and December 31, 2013
- having detectable VL at the start of the study period
- had at least two reported VL values during the study period

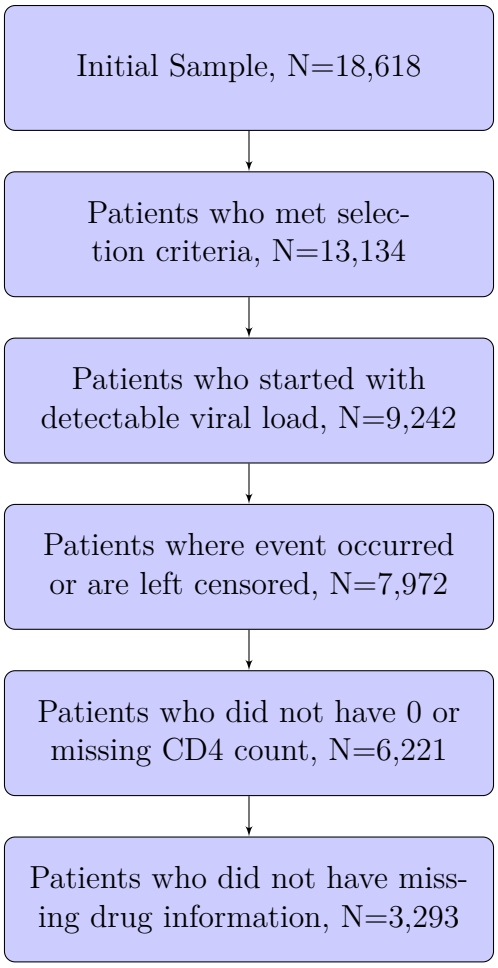

Figure 5.1 Flowchart of analytic sample selection procedure and exclusion criteria

The aim of applying the PRH model to this dataset is to explain the risk behavior of transitioning from detectable VL to undetectable VL. Patients with undetectable VL at the beginning of the study were defined as being left censored. Covariates that were assessed include gender (male or female), race (White, Black, or other), HIV risk exposure group (heterosexual, men who have sex with men, or other), place of residence (rural or urban), age at baseline, initial treatment regimen (single tablet regimen, multiple tablet regimen, or neither), and baseline log CD4 count. Note that HIV risk exposure group refers to how the patient was first exposed to HIV with options including heterosexual HIV infected partner, men who have sex with other men, injecting drug user, no identifiable risk, and no risk reported. Results from the

Table 5.1 Characteristics of persons living with HIV in South Carolina, 2005-2013

Characteristics	Frequency (%)/Summary statistics
Gender	
Female	2564 (41.22%)
Male	3657 (58.78%)
Race	
Black	4966 (79.83%)
White	1086 (17.46%)
Others	169 (2.72%)
Risk of Exposure	
Heterosexual	2295 (36.89%)
Men who have Sex with Men (MSM)	1911 (30.72%)
Others	2015 (32.39%)
Place of Residence	
Urban	4208 (67.64%)
Rural	2013 (32.36%)
Starting Treatment Regimen	
Single Tablet Regimen (STR)	1056 (16.97%)
Multiple Tablet Regimen (MTR)	2237 (35.96%)
N/A	2928 (47.07%)
Baseline Age (in years)	Range = 14.84-81.58; Mean = 39.99; SD = 11.46
Log Baseline CD4 Count (cell/mm^3)	Range = 0.00-3.56; Mean = 2.34; SD = 0.57
Outcome	
Event (Det VL to Undet VL)	4518 (72.62%)
Left censored	1703 (27.38%)

Abbreviations: SD = standard deviation; Det = detectable; VL = viral load; Undet = undetectable.

PRH model are presented and discussed in the next section.

Of the individuals in our sample, 1703 (27%) were classified as being left censored (Table 5.1). Mean age of the sample at baseline was 40.0 years (range = 14.8 - 81.6). The majority of subjects were male (n=3657, 58.8%), Black (n=4966, 79.8%), and lived in an urban county when diagnosed with HIV (n=4208, 67.6%). The mean log CD4 count at the beginning of the study was 2.34 cells/mm^3 (range = 0.00 - 3.56 cells/mm^3). Almost half of the sample had missing drug information (n=2928, 47.1%).

Comparing the SC eHARS database and the simulated data, you can see that they match up well on the left tail (Figure 5.2). However, the right tail of the simulated data is heavier than the right tail of the observed data. This may be a reason for concern when selecting the best model from the simulation study which led to us to conduct a bootstrap study. We used a bootstrapping sampling technique to generate

samples of size 1000, 2000, and 3000 from the observed data i.e. the SC eHARS database. We used 5000 bootstraps for each setup. Results of this are provided in the following section.

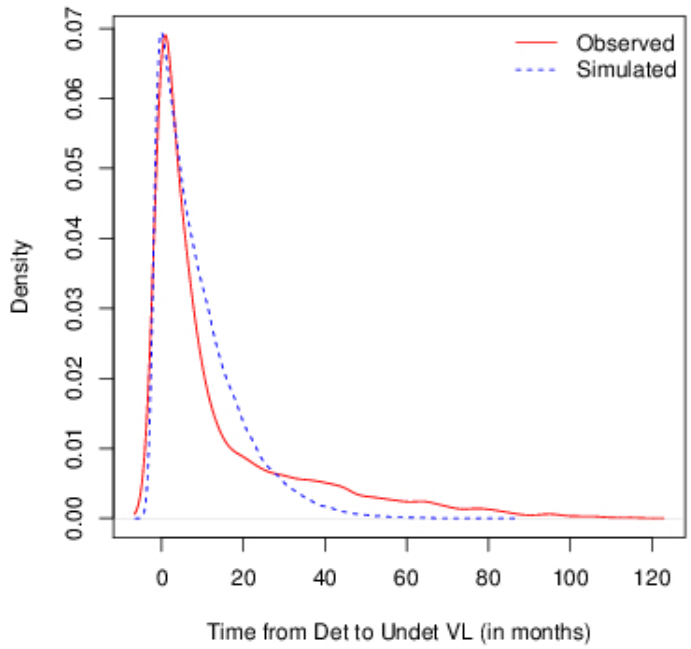

Figure 5.2 Observed vs Simulated Data

5.2 Results and Discussion

Table 5.2 presents the results from the bootstrap study. From this study, the best performing distributions were found to be the Marshal-Olkin, Modified Weibull, Generalized Gamma, Gamma, and Flexible Weibull distribution, respectively. This conflicts with the results from the simulation study, and it's most likely that this discrepancy is due to the heavy tail nature of the simulation study data as compared to the observed data from the SC eHARS database.

Both the Generalized Inverse Weibull and Marshal-Olkin distributions are applied

to the SC eHARS database. Table 5.3 shows the results of the estimated reverse hazard model using a Generalized Inverse Weibull distribution for persons living with HIV and Table 5.4 shows the results using the Marshal-Olkin distribution. Information on treatment regimen is a very important variable to use in our model. However, this information is missing in almost 50% of the subjects in our sample. Thus we fit the model without this variable first (Model 1) and then we fit a second model with reduced sample size after including the treatment variable in model (Model 2). It should be noted that if there was not such a large proportion of missing values in the treatment variable, we would fit only one model, Model 2.

While several covariates have been shown to have an effect on the time from detectable to undetectable VL level, the significant change in behavior of some of these covariates comparing the model incorporating the treatment variable compared to the model without this important factor suggests that an interaction may be present. Note that Models 1 and 2 cannot be compared using AIC, AICC, HQIC, and CAIC due to the large difference in sample size. Additional models were run testing interactions between drug regimen and each of the other covariates. The only significant interaction found was between drug regimen and age, the results of which are shown in Table 5.5.

Comparing the two models in Table 5.5, we can see that the Generalized Inverse Weibull based PRH model fits the data better than the Marshal-Olkin based PRH model in terms of the information criteria. Hence, the best model applied to the SC eHARS database is the Generalized Inverse Weibull based PRH model with all covariates including an interaction of drug regimen and age. Note here how estimates of the reversed hazard rate and significance differs in these two models, particularly the interaction between drug and age which changes in direction and maintains significance leading to opposing conclusions. This reflects the importance of selecting the appropriate distribution in a parametric model to analyze left censored data.

Table 5.2 Average summary measures from bootstrap study

Distribution	Sample Sizes											
	1000				2000				3000			
	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC	AIC	AICC	HQIC	CAIC
Exponential	9873.81	9873.82	9875.68	9879.72	19744.94	19744.94	19747.00	19751.54	29608.05	29608.05	29610.21	29615.06
Log-Normal	8309.58	8309.59	8313.31	8321.39	16613.97	16613.98	16618.09	16627.17	24922.77	24922.77	24927.09	24936.78
Inverse Gaussian	10452.75	10452.76	10456.48	10464.56	20899.89	20899.90	20904.01	20913.10	31353.79	31353.79	31358.11	31367.80
Gamma	7592.29	7592.30	7596.02	7604.11	15179.62	15179.63	15183.73	15192.82	22770.88	22770.88	22775.20	22784.89
Generalized Gamma	7475.02	7475.04	7480.62	7492.74	14944.71	14944.73	14950.88	14964.52	22418.40	22418.40	22424.88	22439.41
Log-Logistic	8309.85	8309.87	8313.59	8321.67	16614.53	16614.54	16618.64	16627.73	24922.87	24955.87	24927.19	24936.88
Weibull	7919.79	7919.80	7923.52	7931.60	15834.46	15834.46	15838.57	15847.66	23752.54	23752.54	23756.86	23766.55
Inverse Weibull	8857.13	8857.14	8860.86	8868.94	17708.94	17708.94	17713.05	17722.14	26570.04	26570.04	26574.36	26584.05
Generalized Inverse Weibull	8657.62	8657.64	8663.22	8675.34	17308.05	17308.06	17314.22	17327.85	25963.86	25963.87	25970.34	25984.88
Flexible Weibull	7772.01	7772.02	7775.74	7783.83	15541.72	15541.72	15545.83	15554.92	23319.71	23319.71	23324.03	23333.72
Marshal-Olkin	5029.27	5029.29	5034.87	5046.99	10051.39	10051.40	10057.56	10071.19	15068.73	15068.73	15075.21	15089.74
Power Generalized Weibull	7998.86	7998.87	8002.59	8010.67	15992.53	15992.53	15996.64	16005.73	23990.21	23990.21	23994.53	24004.22
Modified Weibull	7372.54	7372.56	7378.13	7390.26	14741.53	14741.54	14747.69	14761.33	22115.79	22115.80	22122.27	22136.81
Gompertz	9869.09	9869.10	9872.82	9880.90	19742.41	19742.42	19746.53	19755.62	29601.23	29601.24	29605.56	29615.25

Table 5.3 Estimated Reverse Hazard Rates (HR) using Generalized Inverse Weibull Reverse Hazard model of SC adult HIV patients

Characteristics	Model 1 Reverse HR (95% CI)	Model 2 Reverse HR (95% CI)
Drug Regimen		
Single Tablet Regimen (STR)	—	Ref
Multiple Tablet Regimen (MTR)	—	1.56 (1.37, 1.77)
Gender		
Female	Ref	Ref
Male	0.81 (0.75, 0.87)	1.11 (0.99, 1.25)
Race		
Black	Ref	Ref
White	0.60 (0.55, 0.66)	1.44 (1.30, 1.59)
Others	3.60 (3.26, 4.00)	1.64 (1.29, 2.08)
Risk of Exposure		
Heterosexual	Ref	Ref
Men who have Sex with Men (MSM)	2.08 (1.88, 2.32)	1.73 (1.52, 1.97)
Others	1.96 (1.81, 2.14)	1.20 (1.07, 1.34)
Place of Residence		
Rural	Ref	Ref
Urban	2.35 (2.16, 2.56)	0.47 (0.43, 0.51)
Baseline Age (in years)	1.02 (1.01, 1.03)	1.03 (1.03, 1.04)
Log Baseline CD4 Count(cell/mm^3)	0.30 (0.29, 0.31)	0.37 (0.35, 0.39)
AIC	56332.53	31807.20
AICC	56332.57	31807.29
HQIC	56358.21	31833.40
CAIC	56417.62	31892.39

Table 5.4 Estimated Reverse Hazard Rates (HR) using Marshal-Olkin Reverse Hazard model of SC adult HIV patients

Characteristics	Model 1 Reverse HR (95% CI)	Model 2 Reverse HR (95% CI)
Drug Regimen		
Single Tablet Regimen (STR)	—	Ref
Multiple Tablet Regimen (MTR)	—	1.00 (0.91, 1.10)
Gender		
Female	Ref	Ref
Male	0.72 (0.67, 0.78)	1.17 (1.05, 1.30)
Race		
Black	Ref	Ref
White	2.32 (2.16, 2.49)	1.21 (1.09, 1.34)
Others	1.23 (1.03, 1.48)	0.94 (0.74, 1.21)
Risk of Exposure		
Heterosexual	Ref	Ref
Men who have Sex with Men (MSM)	0.75 (0.67, 0.84)	0.98 (0.86, 1.11)
Others	1.67 (1.55, 1.80)	1.32 (1.19, 1.46)
Place of Residence		
Rural	Ref	Ref
Urban	1.22 (1.14, 1.31)	1.01 (0.93, 1.11)
Baseline Age (in years)	1.04 (1.04, 1.04)	1.02 (1.02, 1.03)
Log Baseline CD4 Count(cell/mm^3)	0.81 (0.79, 0.84)	0.93 (0.88, 0.97)
AIC	46170.84	25552.72
AICC	46170.88	25552.82
HQIC	46196.52	25578.92
CAIC	46255.93	25637.92

Table 5.5 Estimated Reverse Hazard Rates using Generalized Inverse Weibull and Marshal-Olkin Reverse Hazard model of SC adult HIV patients

Characteristics	Generalized Inverse Weibull Reverse HR (95% CI)	Marshal-Olkin Reverse HR (95% CI)
Drug Regimen		
Single Tablet Regimen (STR)	Ref	Ref
Multiple Tablet Regimen (MTR)	0.88 (0.70, 1.10)	0.89 (0.62, 1.29)
Gender		
Female	Ref	Ref
Male	1.35 (1.23, 1.47)	0.86 (0.73, 1.01)
Race		
Black	Ref	Ref
White	1.25 (1.16, 1.34)	1.04 (0.87, 1.24)
Others	0.40 (0.29, 0.55)	1.01 (0.70, 1.44)
Risk of Exposure		
Heterosexual	Ref	Ref
Men who have Sex with Men (MSM)	1.28 (1.16, 1.42)	0.90 (0.72, 1.12)
Others	1.37 (1.26, 1.49)	1.39 (1.19, 1.63)
Place of Residence		
Rural	Ref	Ref
Urban	1.11 (1.04, 1.19)	1.20 (1.04, 1.39)
Baseline Age (in years)	0.99 (0.99, 1.00)	0.91 (0.88, 0.94)
Log Baseline CD4 Count(cell/mm^3)	0.88 (0.81, 0.95)	0.96 (0.88, 1.06)
Interaction of drug by age	0.97 (0.97, 0.98)	1.14 (1.10, 1.18)
AIC	35038.08	37409.31
AICC	35038.19	37409.43
HQIC	35066.47	37437.70
CAIC	35130.38	37501.61

From this final model, we can make the following conclusions on the behavior of transitioning from detectable VL to undetectable VL level. Males are more likely to reach undetectable levels faster than females (RHR:1.35; CI:1.23, 1.47). This trend is also evident in several recent studies[62,63]. A possible reason for this disparity may be higher rates of treatment adherence among males compared to females. Though some studies did not find an association between gender and treatment adherence, a meta-analysis[64] of 207 studies concluded that males adhere more to ART than females.

Individuals who classify as White are more likely to reach undetectable levels faster than Black individuals (RHR:1.25; CI:1.16, 1.34). This is supported by previous studies which highlight that Black individuals are disproportionately affected by HIV/AIDS as they tend to have poorer access to health care, are less likely to receive treatment, less likely to adhere to treatment, and less likely to survive HIV/AIDS.[62,65–67]

People with high risk of exposure such as men who have sex with men (RHR:1.28; CI:1.16, 1.42) and other high risk groups (RHR:1.37; CI:1.26, 1.49) are more likely to reach undetectable levels faster than heterosexual men. It is unclear why this trend is evident in higher risk groups but another study has shown similar results.[60]

People who live in urban areas are more likely to reach undetectable levels faster than those who live in rural areas at the time of diagnosis (RHR:1.11; CI:1.04, 1.19). A possible reason for this effect may be due to the typically increased access to health care and higher range of specialists available to people living with HIV/AIDS in urban areas. This is supported by a study which does an in depth analysis on the effect of place of residence on the timing of diagnosis and stage of disease at diagnosis.[61]

Individuals with higher CD4 count at baseline are less likely to reach undetectable levels faster than those with lower levels of log CD4 count (RHR:0.88; CI:0.81, 0.95). The literature on the association between changes in viral load and CD4 is inconclusive. Some studies[68] support our finding, while others highlight an opposing trend[69]. It has been suggested that those with higher CD4 count may be less adherent due to the absence of symptoms and hence patients do not complete the treatment regimen as they feel better.

Finally, the interaction between drug regimen and age highlights that older people who are on a multiple treatment regimen are likely to reach undetectable levels slower than their younger counterparts (RHR:0.97; CI:0.97, 0.98). There are mixed findings on this is the literature. Young people with HIV tend to have delayed diagnosis and thus higher VL at baseline. One study [60] suggests that this along with underutilization of health care due to HIV-related stigma explains their finding that younger people with HIV reach undetectable levels slower than their older counterparts. A possible explanation of our result may be that older people are not as adherent to treatment [64] or perhaps they have a co-existing morbidity which effects the rate at which they reach undetectable levels.

Chapter 6

Conclusions

The current study derived several extensions of the PRH model and conducted extensive simulation studies to evaluate the usefulness of parametric regression models based on the reversed hazard rate for analyzing left censored HIV viral load data. Simulation studies suggests the best distribution to use under the PRH model is the Generalized Inverse Weibull distribution with the Log-Logistic, Log-Normal, Inverse Gaussian, and Gamma distribution following next in ranking of performance, respectively. The bootstrap analysis suggested the Marshal-Olkin distribution to be the superior performer with the Modified Weibull, Generalized Gamma, Gamma, and Flexible Weibull distributions following behind. Although the bootstrap study was conducted to support the guidelines established in the simulation study, our results are inconsistent. This disagreement may be a result of the characteristic heavy tail nature of VL data that requires further attention and more research. However, when both the top performers of the simulation study and the boostrap study are applied to the SC eHARS database, the Generalized Inverse Weibull based PRH model outperforms the Marshal-Olkin based PRH model. Application of this best performing model on the SC eHARS data revealed important factors on the time to transition from detectable to undetectable viral load levels.

There are several limitations of the SC eHARS database. One limitation is that almost 50% of drug related information is missing which creates complications in estimating hazard rates. Future research using this data should attempt to account for this missingness to make meaningful conclusions on the population. Data on VL

and CD4 count measurements were not available for those who dropped out of medical care after initial diagnosis - this includes those who passed away, moved to a different state, etc. Additionally, persons living with HIV/AIDS who have not been diagnosed were not captured in this database. The database also does not include information on morbidities which may be co-existing with HIV/AIDS which can impact the effect of drug regimens, especially in older people. Since the interaction between age and drug regimen is significant, co-existing conditions warrant further exploration. Regardless of these limitations, the application to the SC eHARS database provides important information on the trajectories of viral load in SC over time.

In conclusion, we recommend that the Generalized Inverse Weibull PRH model be used for analyses involving skewed, heavy tailed left censored HIV VL data.

Bibliography

[1] Elisa T Lee and John Wang. *Statistical methods for survival data analysis*. Vol. 476. John Wiley & Sons, 2003.

[2] Yan Jin et al. "Analysis of lognormally distributed exposure data with repeated measures and values below the limit of detection using SAS". In: *Annals of Occupational Hygiene* 55.1 (2011), pp. 97–112.

[3] IPM Keet et al. "Longitudinal analysis of CD4 T cell counts, T cell reactivity, and human immunodeficiency virus type 1 RNA levels in persons remaining AIDS-free despite CD4 cell counts< 200 for> 5 years". In: *Journal of Infectious Diseases* 176.3 (1997), pp. 665–671.

[4] Thomas R O'Brien et al. "Longitudinal HIV-1 RNA levels in a cohort of homosexual men." In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 18.2 (1998), pp. 155–161.

[5] William B Paxton et al. "Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with $\geq$ 400 CD4 lymphocytes: Implications for applying measurements to individual patients". In: *Journal of Infectious Diseases* 175.2 (1997), pp. 247–254.

[6] Gary H Ganser and Paul Hewett. "An accurate substitution method for analyzing censored data". In: *Journal of occupational and environmental hygiene* 7.4 (2010), pp. 233–244.

[7] James P Hughes. "Mixed effects models with censored data with application to HIV RNA levels". In: *Biometrics* 55.2 (1999), pp. 625–629.

[8] Hélène Jacqmin-Gadda et al. “Analysis of left-censored longitudinal data with application to viral load in HIV infection”. In: *Biostatistics* 1.4 (2000), pp. 355–368.

[9] Robert H Lyles, Cynthia M Lyles, and Douglas J Taylor. “Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49.4 (2000), pp. 485–497.

[10] Rodolphe Thiébaut and Hélène Jacqmin-Gadda. “Mixed models for longitudinal left-censored repeated measures”. In: *Computer methods and programs in biomedicine* 74.3 (2004), pp. 255–260.

[11] D. R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246. URL: `http://www.jstor.org/stable/2985181`.

[12] Anna Christina D’Addio, Michael Rosholm, et al. *Left-censoring in duration data: Theory and applications*. Tech. rep. 2002.

[13] Francesco Passamonti et al. “Prognostic factors for thrombosis, myelofibrosis, and leukemia in essential thrombocythemia: a study of 605 patients”. In: *haematologica* 93.11 (2008), pp. 1645–1651.

[14] Immo Kleinschmidt et al. “Marked increase in child survival after four years of intensive malaria control”. In: *The American journal of tropical medicine and hygiene* 80.6 (2009), pp. 882–888.

[15] Michael Rosholm. “An analysis of the processes of labor market exclusion and (re-) inclusion”. In: *IZA Disussion Paper* 332 (2001).

[16] RE Thompson, EO Voit, and GI Scott. “Statistical modeling of sediment and oyster PAH contamination data collected at a South Carolina estuary (complete and left-censored samples)”. In: *Environmetrics* 11.1 (2000), pp. 99–119.

[17] Samuel Young Annan, Piaomu Liu, and Yuan Zhang. "Comparison of the Kaplan-Meier, Maxilllum Likelihood, and ROS Estimators for Left-Censored Data Using Simulation Studies". In: (2009).

[18] M Pajek et al. "Random left-censoring: a statistical approach accounting for detection limits in x-ray fluorescence analysis". In: *X-Ray Spectrometry* 33.4 (2004), pp. 306–311.

[19] A Kubala-Kukus et al. "Determination of concentration distribution of trace elements near the detection limit". In: *Spectrochimica Acta Part B: Atomic Spectroscopy* 56.11 (2001), pp. 2037–2044.

[20] C Luczynska et al. "Indoor factors associated with concentrations of house dust mite allergen, Der p 1, in a random sample of houses in Norwich, UK". In: *Clinical and Experimental Allergy* 28 (1998), pp. 1201–1209.

[21] Patricia M Odell, Keaven M Anderson, and Ralph B D'Agostino. "Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model". In: *Biometrics* (1992), pp. 951–959.

[22] Raymond P Brettle et al. "Progression of HIV: follow-up of Edinburgh injecting drug users with narrow seroconversion intervals in 1983-1985." In: *Aids* 10.4 (1996), pp. 419–430.

[23] Raphael Isingo et al. "Survival after HIV infection in the pre-antiretroviral therapy era in a rural Tanzanian cohort". In: *Aids* 21 (2007), S5–S13.

[24] Rameshwar D Gupta and Debasis Kundu. "Theory & methods: Generalized exponential distributions". In: *Australian & New Zealand Journal of Statistics* 41.2 (1999), pp. 173–188.

[25] Sharmishtha Mitra and Debasis Kundu. "Analysis of left censored data from the generalized exponential distribution". In: *Journal of Statistical Computation and Simulation* 78.7 (2008), pp. 669–679.

[26] U.S. Department of Health & Human Services. *Viral Load*. 2015. URL: `https://www.aids.gov/hiv-aids-basics/index.html` (visited on 03/25/2016).

[27] Hrishikesh Chakraborty et al. "Viral burden in genital secretions determines male-to-female sexual transmission of HIV-1: a probabilistic empiric model". In: *Aids* 15.5 (2001), pp. 621–627.

[28] Christopher J Gill et al. "Relationship of HIV viral loads, CD4 counts and HAART use to health-related quality of life". In: *Journal of acquired immune deficiency syndromes (1999)* 30.5 (2002), pp. 485–492.

[29] Myron S Cohen et al. "Prevention of HIV-1 infection with early antiretroviral therapy". In: *New England Journal of medicine* 365.6 (2011), pp. 493–505.

[30] Carl W Dieffenbach. "Preventing HIV transmission through antiretroviral treatment mediated virologic suppression: aspects of an emerging scientific agenda". In: *Current Opinion in HIV and AIDS* 7.2 (2012), pp. 106–110.

[31] Susan Reif et al. "HIV/AIDS epidemic in the South reaches crisis proportions in last decade". In: *Duke Center for Health Policy and Inequalities Research* (2011).

[32] South Carolina Department of Health and Environmental Control. *An Epidemiologic Profile of HIV and AIDS in South Carolina 2015*. Report. 2015.

[33] Bankole A Olatosi et al. "Patterns of engagement in care by HIV-infected adults: South Carolina, 2004–2006". In: *Aids* 23.6 (2009), pp. 725–730.

[34] Avnish Tripathi et al. "The impact of retention in early HIV medical care on viro-immunological parameters and survival: a statewide study". In: *AIDS research and human retroviruses* 27.7 (2011), pp. 751–758.

[35] Veronica Miller et al. "Relations among CD4 lymphocyte count nadir, antiretroviral therapy, and HIV-1 disease progression: results from the EuroSIDA study". In: *Annals of internal medicine* 130.7 (1999), pp. 570–577.

[36] Jianguo Sun, Qiming Liao, and Marcello Pagano. “Regression analysis of doubly censored failure time data with applications to AIDS studies”. In: *Biometrics* 55.3 (1999), pp. 909–914.

[37] Scott D Holmberg et al. “Protease inhibitors and cardiovascular outcomes in patients with HIV-1”. In: *The Lancet* 360.9347 (2002), pp. 1747–1748.

[38] T Cai and S Cheng. “Semiparametric regression analysis for doubly censored data”. In: *Biometrika* 91.2 (2004), pp. 277–290.

[39] Yongdai Kim, Bumsoo Kim, and Woncheol Jang. “Asymptotic properties of the maximum likelihood estimator for the proportional hazards model with doubly censored data”. In: *Journal of Multivariate Analysis* 101.6 (2010), pp. 1339–1351.

[40] Leah Szadkowski et al. “Short communication: effects of age on virologic suppression and CD4 cell response in HIV-positive patients initiating combination antiretroviral therapy”. In: *AIDS research and human retroviruses* 28.12 (2012), pp. 1579–1583.

[41] Basia Zaba et al. “Age-specific mortality patterns in HIV-infected individuals: a comparative analysis of African community study data”. In: *Aids* 21 (2007), S87–S96.

[42] Rodolphe Thiébaut et al. “Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection”. In: *Statistics in medicine* 24.1 (2005), pp. 65–82.

[43] Stephen R Cole et al. “Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model”. In: *American journal of epidemiology* 166.2 (2007), pp. 219–227.

[44] Robin Henderson, Peter Diggle, and Angela Dobson. “Joint modelling of longitudinal measurements and event time data”. In: *Biostatistics* 1.4 (2000), pp. 465–480.

[45] Kwan-Moon Leung, Robert M Elashoff, and Abdelmonem A Afifi. “Censoring issues in survival analysis”. In: *Annual review of public health* 18.1 (1997), pp. 83–104.

[46] JK Lindsey. “A study of interval censoring in parametric regression models”. In: *Lifetime Data Analysis* 4.4 (1998), pp. 329–354.

[47] Yvonne H Sparling et al. “Parametric survival models for interval-censored data with time-dependent covariates”. In: *Biostatistics* 7.4 (2006), pp. 599–614.

[48] Edward L Kaplan and Paul Meier. “Nonparametric estimation from incomplete observations”. In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[49] Mustafa Agah Tekindal, Beyza Doğanay Erdoğan, and Yasemin Yavuz. “Evaluating Left-Censored Data Through Substitution, Parametric, Semi-parametric, and Nonparametric Methods: A Simulation Study”. In: *Interdisciplinary Sciences: Computational Life Sciences* (2015), pp. 1–20.

[50] Brenda W Gillespie et al. “Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator”. In: *Epidemiology* 21.4 (2010), S64–S70.

[51] Bruce W Turnbull. “Nonparametric estimation of a survivorship function with doubly censored data”. In: *Journal of the American Statistical Association* 69.345 (1974), pp. 169–173.

[52] AM Variyath and PG Sankaran. “Parametric Regression Models Using Reversed Hazard Rates”. In: *Journal of Probability and Statistics* 2014 (2014).

[53] PK Anderson et al. *Statistical methods based on counting processes*. 1993.

[54] Tran Huynh et al. "A Comparison of the β-Substitution Method and a Bayesian Method for Analyzing Left-Censored Data". In: *Annals of Occupational Hygiene* 60.1 (2016), pp. 56–73.

[55] CD Lai, Min Xie, and DNP Murthy. "A modified Weibull distribution". In: *Reliability, IEEE Transactions on* 52.1 (2003), pp. 33–37.

[56] Mark Bebbington, Chin-Diew Lai, and Ričardas Zitikis. "A flexible Weibull extension". In: *Reliability Engineering & System Safety* 92.6 (2007), pp. 719–726.

[57] Vilijandas Bagdonavicius and Mikhail Nikulin. *Accelerated life models: modeling and statistical analysis*. CRC Press, 2001.

[58] Albert W Marshall and Ingram Olkin. "A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families". In: *Biometrika* 84.3 (1997), pp. 641–652.

[59] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2016. URL: `https://www.R-project.org/`.

[60] Hrishikesh Chakraborty et al. "Disparities in Viral Load and CD4 Count Trends Among HIV-Infected Adults in South Carolina". In: *AIDS patient care and STDs* 29.1 (2015), pp. 26–32.

[61] Kristina E Weis et al. "Associations of Rural Residence With Timing of HIV Diagnosis and Stage of Disease at Diagnosis, South Carolina 2001-2005". In: *The Journal of Rural Health* 26.2 (2010), pp. 105–112.

[62] Linda Beer et al. "Understanding Cross-Sectional Racial, Ethnic, and Gender Disparities in Antiretroviral Use and Viral Suppression Among HIV Patients in the United States". In: *Medicine* 95.13 (2016), e3171.

[63] Loredana Manolescu and Paul Marinescu. “Sex differences in HIV-1 viral load and absolute CD4 cell count in long term survivors HIV-1 infected patients from Giurgiu, Romania”. In: *Romanian Review of Laboratory Medicine* 21.2 (2013), pp. 217–224.

[64] Nienke Langebeek et al. “Predictors and correlates of adherence to combination antiretroviral therapy (ART) for chronic HIV infection: a meta-analysis”. In: *BMC medicine* 12.1 (2014), p. 142.

[65] Meg C Kong et al. “Association between race, depression, and antiretroviral therapy adherence in a low-income population with HIV infection”. In: *Journal of general internal medicine* 27.9 (2012), pp. 1159–1164.

[66] Susan Reif, Kristin Lowe Geonnotti, and Kathryn Whetten. “HIV infection and AIDS in the Deep South”. In: *American Journal of Public Health* 96.6 (2006), pp. 970–973.

[67] Dharushana Muthulingam et al. “Disparities in engagement in care and viral suppression among persons with HIV”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 63.1 (2013), pp. 112–119.

[68] Jan CM Hendriks et al. “Use of immunological markers and continuous-time Markov models to estimate progression of HIV infection in homosexual men.” In: *AIDS* 10.6 (1996), pp. 649–656.

[69] Ayele Taye Goshu and Zelalem Getahun Dessie. “Modelling progression of HIV/AIDS disease stages using semi-Markov processes”. In: *Journal of Data Science* 11.2 (2013), pp. 269–280.