Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Extensões dos Modelos de Sobrevivência referente a Distribuição Weibull

Valdemiro Piedade Vigas

Orientador: Prof. Dr. Francisco Louzada Neto

São Carlos

Março de 2014

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Extensões dos Modelos de Sobrevivência referente a Distribuição Weibull

Valdemiro Piedade Vigas

Orientador: Prof. Dr. Francisco Louzada Neto

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos

Março de 2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

V672em

Vigas, Valdemiro Piedade.
Extensões dos modelos de sobrevivência referente a distribuição Weibull / Valdemiro Piedade Vigas. -- São Carlos : UFSCar, 2014.
74 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Análise de sobrevivência. 2. Dados censurados. 3. Distribuição Weibull. 4. Modelos de regressão. 5. Fração de cura. I. Título.

CDD: 519.9 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Valdemiro Piedade Vigas

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 07/03/2014
PELA COMISSÃO JULGADORA:

Presidente _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP/Orientador)

1º Examinador _____
Profa. Dra. Estela Maris Pereira Bereta (DEs-UFSCar)

2º Examinador _____
Profa. Dra. Giovana Oliveira Silva (UFBA)

Agradecimentos

Agradeço primeiramente a Deus, o qual sua representatividade em minha luta independeu de religião, fé ou coisas semelhantes.

À minha família, em especial aos meus pais Waldemar Carvalho Vigas e à guerreira Marina Piedade Vigas, pelo amor e carinho. Aos meus irmãos Valdemir e Vlademir, além das minhas irmãs Nina e Val que seguraram a barra nos momentos mais difíceis desse período. Ao Vanderlei (meu padrinho, cunhado, irmão de consideração e amigo) pelas palavras de incentivo.

Ao orientador Prof. Dr. Francisco Louzada Neto pela orientação e conhecimento compartilhado, tornando possível a realização deste trabalho.

Aos professores do curso de graduação/pós-graduação em Estatística da UFSCar e aos membros da banca. Em especial à Profa. Dra. Maria Aparecida de Paiva, à Profa. Dra. Estela Maris Bereta, ao Prof. Dr. Adriano Polpo e ao Prof. Dr. Josmar Mazucheli pelas sugestões para a melhoria da dissertação.

Aos meus colegas da pós-graduação em Estatística, por compartilhar o conhecimento, pela troca de experiências acadêmicas e pelos momentos descontraídos fora da sala de aula.

À CAPES pelo auxílio financeiro que permitiu a realização deste trabalho.

Aos funcionários do Departamento de Estatística da UFSCar. Em especial, a funcionária Isabel, pelo trabalho na pós-graduação.

Aos professores do Departamento de Estatística da UFBA, em especial à Profa. Dra. Giovana Oliveira Silva pela amizade, incentivo e confiança em mim depositada.

Aos amigos (as): Robert, Joel, Marcus, Gerson, Tiago, Brandão, Léo, Nei, Lucília, Khalla, Letícia, Verônica, Lorena, Taci, Xandy, Rogério, Otávio, Jonatas, Jairo, Cacau, Shirley, Ana, Urbano (e ainda outros que não citei) por me proporcionarem momentos

importantes nessa caminhada.

Ao meu irmão de República Jurandir e à Stela, os quais foram tão importantes para o meu crescimento pessoal. Sou grato pela paciência, pelas conversas, alegrias, tristezas, palavras de conforto e discussões.

Enfim, agradeço a todos que de forma direta ou indireta contribuíram para o meu crescimento como ser humano, pois esta é uma luta diária.

Minha eterna gratidão!

Resumo

Nesta dissertação são revistos dois modelos de distribuições de probabilidade para os tempos de vida até a ocorrência do evento provocado por uma causa específica para elementos em uma população. O primeiro modelo revisto é o denominado Weibull-Poisson (WP) que foi proposto por Louzada *et al.* (2011a), esse modelo generaliza as distribuições exponencial-Poisson proposta por Kus (2007) e Weibull. O segundo, denominado modelo de longa duração, foi proposto por vários autores e considera que a população não é homogênea em relação ao risco de ocorrência do evento pela causa em estudo. A população possui uma sub-população constituída de elementos que não estão sujeitos ao evento pela causa específica em estudo, sendo considerados como imunes ou curados. Em relação a parcela dos elementos que estão em risco observa-se o valor mínimo dos tempos da ocorrência do evento. Na revisão sobre a WP são detalhadas as expressões da função de sobrevivência, da função quantil, da função densidade de probabilidade e da função de risco, bem como a expressão dos momentos não centrais de ordem k e a distribuição de estatísticas de ordem. A partir desta revisão, é proposta de forma original, estudos de simulação com o objetivo de analisar as propriedades frequentistas dos estimadores de máxima verossimilhança dos parâmetros desta distribuição. E apresenta-se resultados relativos à inferência sobre os parâmetros desta distribuição, tanto no caso em que o conjunto de dados consta de observações completas de tempos de vida, como no caso em que ele possa conter observações censuradas. Além disso, apresentamos de forma original neste trabalho um modelo de regressão na forma de locação e escala quando T tem distribuição WP. Outra contribuição original dessa dissertação é propor a distribuição de longa duração Weibull-Poisson (LWP), além de estudar a LWP na situação em que as covariáveis são incluídas na análise. Realizou-se também a descrição das funções que caracterizam essa distribuição (função distribuição, função quantil, função densidade de

probabilidade e função de risco). Assim como a descrição da expressão do momento de ordem k e da função densidade da estatística de ordem. É feito um estudo por simulação desta distribuição via máxima verossimilhança. Aplicações à conjuntos de dados reais ilustram a utilidade dos dois modelos considerados.

Palavras-Chave: Análise de Sobrevivência, Dados censurados, Distribuição Weibull-Poisson, Regressão Log Weibull-Poisson, Longa duração.

Abstract

In this dissertation, two models of probability distributions for the lifetimes until the occurrence of the event produced by a specific cause for elements in a population are reviewed. The first revised model is called the Weibull-Poisson (WP) which has been proposed by Louzada *et al.* (2011a). This model generalizes the exponential-Poisson distributions proposed by Kus (2007) and Weibull. The second, called long-term model, has been proposed by several authors and it considers that the population is not homogeneous in relation to the risk of event occurence by the cause studied. The population has a sub-population that consists of elements who are not liable do die by the specific cause in study. These elements are considered as immune or cured. In relation to the elements who are at risk the minimum value of time of the event accurance is observed. In the review of WP the expressions of the survival function, quantile function, probability density function, and of the hazard function, as well the expression of the non-central moments of order k and the distribution of order statistics are detailed. From this review we propose, in an original way, studies of the simulation to analyze the paramenters of frequentist properties of maximum likelihood estimators for this distribution. And also we also present results related to the inference about the parameters of this distribution, both in the case in which the data set consists of complete observations of lifetimes, and also in the case in which it may contain censored observations. Furthermore, we present in this paper, in an original way a regression model in a form of location and scale when T has WP distribution. Another original contribution of this dissertation is to propose the distribution of long-term Weibull-Poisson (LWP). Besides studying the LWP in the situation in which the covariates are included in the analysis. We also described the functions that characterize this distribution (distribution function, quantile function, probability density function and the hazard function). Moreover we describe the expression of the moment of order k, and the density function of a statistical order. A study by simulation

of this distribution is made through maximum likelihood estimators. Applications to real data set illustrate the applicability of the two considered models.

Keywords: Survival analisys, Censored data, Weibull-Poisson distribution, Log Weibull-Poisson regression, Long-term.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Em diversas aplicações da área de Estatística a variável resposta consiste do tempo até a ocorrência de um evento e este tempo é geralmente denominado tempo de falha ou tempo de sobrevivência. Alguns exemplos são: o tempo de vida de equipamentos industriais, tempo de pessoas na situação de desemprego, tempo até surgimento de uma doença qualquer, entre outros. Em estudos sobre este tipo de variável é muito comum a ocorrência de censura em algumas das observações e isto consiste na observação incompleta do tempo de interesse, devido ao fato de que só é possível observar o menor entre dois tempos aleatórios e identificar se a observação corresponde ao tempo até a falha ou o tempo até a ocorrência de um outro evento, que será chamado de tempo até a censura. Existem várias distribuições de probabilidade que podem modelar a distribuição do tempo de sobrevivência. Mas apesar da existência na literatura de várias destas distribuições, a busca por distribuições novas é justificada pelo fato de que os modelos usuais tais como as distribuições exponencial e Weibull, muitas vezes não se ajustam bem ao conjunto de dados reais sob a análise. Neste sentido, surgem distribuições que são criadas a partir das distribuições exponencial e Weibull.

Também é comum que nem sempre seja possível observar o valor exato do tempo de sobrevivência e sim o valor mínimo ou máximo dos tempos. Isso ocorre, por exemplo, quando o interesse é observar o tempo de vida de um sistema em série ou em paralelo cuja duração depende da duração de um conjunto de componentes. Especificamente, considera-se o caso em que não há possibilidade de identificar o fator responsável pela falha do sistema, ou seja, tanto a quantidade como a identificação do(s) componente (es) que provocou (provocaram) a falha não são observáveis, mas apenas o valor do mínimo ou do máximo dos tempos de vida é observável. Este procedimento usado na literatura

é para criar novas distribuições de probabilidade fazendo uma mistura enumerável de uma família paramétrica de distribuições contínuas, isto é, uma combinação linear dos elementos dessa família, com pesos tirados da distribuição de uma variável discreta. Nos últimos anos várias distribuições de probabilidade têm sido propostas para análise de dados de sobrevivência deste tipo. Vários autores consideraram esse contexto, dentre eles podemos citar Marshall & Olkin (1997). Adamidis & Loukas (1998) propuseram a distribuição exponencial-geométrica (EG) com dois parâmetros, cuja motivação é a duração de um sistema em série. Especificamente, considera-se que tanto o número de componentes como a identificação do componente que provocou a falha não são observáveis e sim o valor do tempo de vida mínimo dentre eles é observado, em que o número de componentes segue distribuição geométrica e o tempo de duração de cada componente segue uma distribuição exponencial. Seguindo a mesma ideia da distribuição EG, Kus (2007) introduziu a distribuição exponencial-Poisson (EP) inserida também num sistema em série. Ela também tem dois parâmetros e apresenta a taxa de falha decrescente. Esta distribuição é obtida através da distribuição exponencial com uma distribuição de Poisson. Tahmasbi & Rezaei (2008) introduziram as distribuições exponenciais logarítmicas. Posteriormente, Chahkandi & Ganjali (2009) introduziram a exponencial série de potência (EPS), que contém as distribuições citadas EG, EP e as exponenciais logarítmicas como casos especiais. Uma vez que a distribuição Weibull generaliza a distribuição exponencial, foi natural que o mecanismo que foi feito com a distribuição exponencial seja utilizado também pela distribuição Weibull. Logo, Barreto-Souza *et al.* (2008) definiu a Weibull-Geométrica (WG) que generaliza as distribuições EG e Weibull. Louzada *et al.* (2011b) introduziram a distribuição geométrica exponencial complementar (CEG), que é uma complementação da ditribuição EG, cuja motivação é a duração de um sistema em paralelo $Y = \max(T_1, ..., T_N)$ em que a falha do sistema ocorre devido a falha de todos os componetes, em que o número de componentes segue distribuição geométrica e o tempo de duração de cada componente segue distribuição exponencial tal que nesse caso não há informações sobre qual fator foi responsável pela falha de um componente, mas apenas sabe-se o tempo de vida máximo para todos os riscos. Mais detalhes em: Goetghebeur & Ryan (1995), Reiser *et al.* (1995), Lu & Tsiatis (2001), Lu & Tsiatis (2005) e (Cooner *et al.*, 2006).

No capítulo 2, apresenta-se uma revisão de alguns conceitos em análise de dados de sobrevivência. Além de citar alguns critérios de seleção e análise de resíduos para os modelos estudados.

No Capítulo 3 é feita uma revisão da distribuição de probabilidade da Weibull-Poisson (WP). Proposta por Louzada *et al.* (2011a), generaliza as distribuições EP e Weibull e é deduzida na situação em que nem sempre é possível observar o valor exato da variável tempo de sobrevivência e sim o valor mínimo dos tempos. Esta distribuição é obtida por meio de uma suposição em que o número de fatores segue distribuição Poisson e o tempo de vida de cada componente segue distribuição de Weibull. O objetivo deste capítulo é apresentar um resultado relativo a consideração de um modelo de regressão log-linear para uma variável com distribuição WP. Antes disto é feita a revisão das propriedades da distribuição WP, recordando sua função de sobrevivência, a função quantil, a função densidade de probabilidade, a função de risco e a distribuição da *k-ésima* estatística de ordem. Os detalhes da estimação dos parâmetros da distribuição WP pelo método de máxima verossimilhança são apresentados de forma original, como também são apresentados resultados de um estudo de simulação com o objetivo de analisar as propriedades frequentistas dos estimadores de máxima verossimilhança dos parâmetros desta distribuição. Aplicações a conjunto de dados reais ilustram a aplicabilidade do modelo (WP). Ao final do Capítulo 3, propõe-se de forma original o modelo regressão log-linear para uma variável com distribuição WP. A justificativa para isso é o fato de ser comum na prática da pesquisa científica existirem características que podem influenciar a distribuição do tempo de sobrevivência. Essas características são chamadas de covariáveis e devem ser incluídas na análise estatística dos dados.

No Capítulo 4, aborda-se os modelos de sobrevivência de longa duração ou com fração de cura, que destinam-se a análise do tempo de sobrevivência em uma população quando grande parte das observações são censuradas, indicando que deve existir na população uma fração de indivíduos considerados imunes (curados) à doença estudada ou que não estão sujeitos ao evento de interesse. Nessa situação a função de sobrevivência associada é imprópria, ou seja, $\lim_{t\to\infty} S_T(t) = p$, em que p é proporção de imunes ou fração de cura e a proporção complementar $1-p$ é de não curados, o que diferencia dos modelos usuais de sobrevivência no qual considera que $\lim_{t\to\infty} S_T(t) = 0$. Muitos autores contribuíram para o desenvolvimento da teoria dos modelos de mistura de longa duração, dentre eles destaca-se

Berkson & Gage (1952). Posteriormente surgiram outros modelos na literatura, dentre eles os de Yakovlev *et al.* (1996), Maller & Zhou (1995), Chen & Ibrahim (2001). Rodrigues *et al.* (2009) propuseram uma teoria unificada dos modelos de sobrevivência de longa duração já existentes na literatura. Neste capítulo um resumo dos principais conceitos e trabalhos da área são apresentados e, a seguir, é proposta uma distribuição de longa duração Weibull-Poisson (LWP). Este modelo é proposto nessa dissertação e considera a distribuição Weibull-Poisson (WP) para o tempo de vida de indivíduos pertencentes a fração da população sujeita ao evento, em que supõe haver uma parcela p da população constituída de indivíduos imunes ao evento de interesse. Além disso, também estuda-se a LWP na situação em que as covariáveis são incluídas na análise na proporção de curados. Em seguida foi feito um estudo de simulação para verificar o comportamento dos estimadores dos seus parâmetros em relação ao vício e ao erro quadrático médio para diferentes tamanhos de amostras. Por fim, apresenta-se uma aplicação a conjunto de dados reais.

Capítulo 2

Metodologia

Como dito anteriormente, os dados de sobrevivência são compostos pelo tempo de sobrevivência e o tempo de censura, que pode ser classificado como censura: à direita, à esquerda e a intervalar. A censura à direita ocorre quando o tempo até a ocorrência do evento de interesse está à direita do tempo registrado. Como exemplo, em um estudo clínico em que o acontecimento de interesse é a morte de um indivíduo após lhe ter sido diagnosticado um determinado tumor maligno, nos indivíduos que estiverem vivos no final do estudo, a observação do tempo até a ocorrência é considerada como censurada à direita. Por outro lado, censura à esquerda ocorre quando o tempo registrado é maior que o tempo de falha. Por exemplo, em um estudo com um grupo de crianças visando determinar a idade em que cada criança aprendeu a ler, se alguma delas já sabiam ler no início do estudo e não lembrar a idade exata que tinham quando aprenderam a ler, a sua idade na época do estudo será uma observação censurada à esquerda da variável tempo até saber ler. A censura intervalar ocorre quando sabe-se apenas que o evento de interesse aconteceu em um determinado intervalo de tempo. A ocorrência da censura à direita é a mais frequente em estudos de sobrevivência e por este motivo a única abordada neste trabalho e pode acontecer de acordo com os seguintes mecanismos:

Censura tipo I: é aquela em que a coleta de dados do estudo termina após um período pré estabelecido de tempo de forma que o tempo completo de sobrevivência de um elemento é conhecido apenas se sua falha ocorre antes do final do estudo.

Censura tipo II: é aquela em que a coleta de dados termina após ter ocorrido o evento de interesse em um número pré - estabelecido de elementos da amostra.

Censura tipo III ou aleatória: é a que acontece quando os elementos deixam o estudo por qualquer motivo sem que o evento de interesse tenha ocorrido.

A censura aleatória é a que ocorre com mais frequência em estudos reais e nesta situação, geralmente se assume que o tempo de censura é uma variável aleatória independente do tempo completo de vida de um elemento.

Seja T for uma variável aleatória não negativa e contínua que representa o tempo de sobrevivência de um elemento e C ser uma variável aleatória, que representa o tempo de censura associado a este mesmo elemento, os dados obtidos são $t_i = \min(T_i, C_i)$ e δ_i a variável indicadora de ocorrência de falha, tal que:

$$\delta_i = \begin{cases} 1, & \text{se} \quad T \leq C, \qquad i=1,2,\ldots,n \\ 0, & \text{se} \quad T > C. \end{cases}$$

Para descrever o comportamento da variável tempo de sobrevivência T usa-se uma distribuição de probabilidade que pode ser especificada por sua função distribuição acumulada $F_T(t)$ ou, equivalentemente pela função de sobrevivência (ou função de confiabilidade), pela função densidade de probabilidade $f_T(t)$ ou ainda pela função de risco (ou função taxa de falha), $h_T(t)$. A função de sobrevivência do tempo t é definida como a probabilidade de um indivíduo só apresentar o evento após um tempo t, ou seja, $S_T(t) = 1 - F_T(t) = P(T > t)$, em que a função distribuição acumulada $F_T(t)$ é definida por: $F_T(t) = P(T \leq t) = \int_0^t f(u)du$. A função densidade de probabilidade é a derivada da função distribuição acumulada, ou seja,

$$f(t) = \frac{\partial F(t)}{\partial t}.$$

Como $F_T(t) = 1 - S_T(t)$, podemos escrever:

$$f_T(t) = \frac{\partial[1 - S_T(t)]}{\partial t} = -\partial[S_T(t)]/\partial(t).$$

A função taxa de falha ($h_T(t)$) é definida como o limite da probabilidade de que o indivíduo apresente o evento no intervalo de tempo $(t, t + \Delta t)$ dado que não tenha apresentado o evento até o tempo t dividido pelo comprimento do intervalo. Essa função é dada por:

$$h_T(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}.$$

A relação entre as funções de densidade, de sobrevivência e a de risco, é dada por:

$$h_T(t) = \frac{f_T(t)}{S_T(t)}.$$

Em outros tipos de estudos na área de sobrevivência, o tempo de vida observável T de um indivíduo é o seu tempo de sobrevivência quando ele está sujeito a diversas causas de falha ou de morte. Logo, o objetivo é determinar a distribuição do tempo até a falha de um sistema por qualquer uma das várias causas possíveis de serem identificadas. Este tipo de estudo ocorre tanto na área tecnológica de duração de equipamentos quanto na área médica. Na linguagem da área tecnológica, não é possível observar o tempo de vida de cada um dos componentes de um sistema com vários componentes seja em série ou em paralelo, ou em uma composição dos dois tipos de sistemas. O mesmo se aplica na área médica quando se considera que a morte de um indivíduo decorre da falência de um ou de muitos órgãos vitais.

Existem dois modelos bastante difundidos para o efeito de diversas causas. O primeiro deles chama-se modelo de riscos competitivos $(RC), T_{RC} = \min(T_1, ..., T_k),\ j = 1, ..., k$; sendo cada T_j o tempo de ocorrência até a falha pela causa j. Ou seja, o indivíduo apresenta a falha do evento quando ocorrer pelo menos uma dentre as k causas. O modelo de riscos complementares (RC^*) admite que $T_{RC^*} = \max(T_1, ..., T_k),\ j = 1, ..., k$; sendo cada T_j o tempo de ocorrência até a falha pela causa j. Neste modelo o indivíduo só apresenta o evento final se ocorrerem as falhas por todas as k causas. Podemos considerar os modelos RC e RC^* como associados ao tempo de duração de um circuito com k componentes em série e em paralelo, respectivamente.

Nos dois tipos de situações abordadas, quando se observa o mínimo ou o máximo de tempos de vida, com ou sem possibilidade de identificação da causa que foi a responsável pela falha, há interesse em criar novas distribuições de probabilidade que ampliem o quadro clássico de distribuições usáveis na modelagem de variáveis positivas e contínuas.

2.1 Inferência Estatística em análise de dados de sobrevivência

2.1.1 Estimador de Kaplan-Meier

Um dos objetivos em análise de sobrevivência é estimar a função de sobrevivência, $\widehat{S(t)}$ a partir de um conjunto de dados, com alguns deles possivelmente censurados. O estimador de Kaplan-Meier é o mais conhecido dentre os estimadores não paramétricos da função de sobrevivência. Esse estimador é chamado também de estimador limite produto e é definido como:

$$\widehat{S(t)} = \prod_{j:\ t_j<t} \left(\frac{n_j - d_j}{nj}\right) = \prod_{j:\ t_j<t} \left(1 - \frac{dj}{nj}\right),$$

em que:

$t_1 < t_2 \cdots < t_n$.

d_j: o número de falhas em t_j, j:1,...,k.

n_j: o número de indivíduos sob risco em t_j, ou seja, os elementos que não falharam e não foram censurados até o instante anterior a t_j.

De acordo com Colosimo & Giolo (2006) as principais propriedades do estimador de Kaplan-Meier são:

1) Ser assintoticamente não viciado para amostras grandes;

2) Ser fracamente consistente;

3) Convergir assintoticamente para um processo gaussiano;

4) Ser estimador de máxima verossimilhança de S(t).

2.1.2 Curva TTT (tempo total em teste)

A cada função de distribuição de probabilidade $F_T(t)$ de uma variável aleatória T está associada uma única função de risco $h_T(t)$. Informações sobre o formato do gráfico de uma estimativa da função $h_T(t)$ podem auxiliar na escolha de um modelo para $F_T(t)$. Existem várias formas que o gráfico da função de risco da variável T pode assumir e é importante utilizar uma metodologia para identificar o formato desta curva e a partir daí qual a distribuição é mais apropriada para esta variável. A curva TTT proposta por Barlow *et al.* (1972) é uma ferramenta útil para dar informações sobre a função de risco

que podem ser obtidas a partir de uma análise gráfica.

Para obter o gráfico TTT para o conjunto de dados não censurados utilizamos a proposta de Aarset (1987), em que a curva é obtida construindo o gráfico $G(r/n) = [(\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}]/(\sum_{i=1}^{r} T_{i:n})$ por r/n, em que $r = 1, ..., n$ e $T_{i:n}$ as estatísticas de ordem da amostra para $i = 1, 2..., n$ (Mudholkar *et al.*, 1996). Para o conjunto de dados com observações censuradas utilizamos a versão apresentada em Rinne (2009) e Roman (2013), em que $G(r/n) = (\sum_{i=1}^{r}(n-i+1)(T_{i:n} - T_{(i-1):n})$ e defini-se r^* como o indicador de ordem das observações não censuradas, $r^* = 1, 2, ..., n^*$. Selecionando os n^* valores de $G(r/n)$ que correspondem as observações não censuradas, que são denotadas por $G(r^*/n^*)$, o gráfico TTT corresponde a relação gráfica de r^*/n^* versus $TTT_{r^*} = \frac{G(r^*/n^*)}{G(n^*/n^*)}$.

Aarset (1987) mostrou que quando a curva TTT se apresentar graficamente como uma reta diagonal (curva A), há indícios de que a função de risco seja constante. Quando a curva é convexa (curva B) ou côncava (curva C), a função de risco é monotonicamente decrescente ou crescente, respectivamente, e se a curva é convexa e depois côncava (curva D), a função de risco é em forma de U e no caso reverso (curva E) ela é unimodal (ver Figura 2.1).

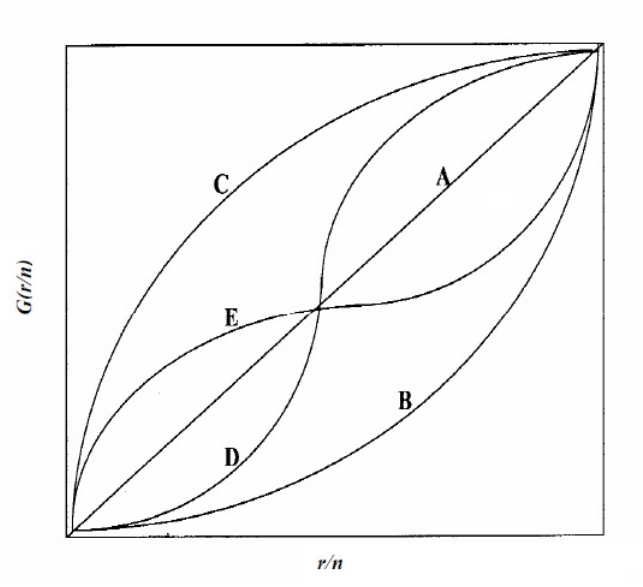

Figura 2.1: Gráficos ilustrativos de algumas curvas TTT.

2.1.3 Método de máxima verossimilhança

Existem vários métodos que podem ser utilizados para estimar os parâmetros dos modelos probabilísticos, sendo o mais comum o de máxima verossimilhança, pois é um método de estimação com propriedades importantes para os estimadores, uma delas é a distribuição assintótica, que permite encontrar intervalos de confiança para os parâmetros. Outra característica importante deste método é a inclusão de censuras no seu processo de estimação, o que não ocorre com outros métodos de estimação como o de mínimos quadrados. A ideia do método de máxima verossimilhança é encontrar uma estimativa para cada parâmetro que tem a maior verossimilhança de ter gerado a amostra.

Uma amostra aleatória de dados de sobrevivência, é composta por dois vetores $\mathbf{t} = (t_1, ..., t_n)$ e $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$, em que T é o tempo de sobrevivência, C o tempo de censura que é independente de T e $t_i = \min(T_i, C_i)$. δ é um vetor de variáveis aleatórias indicadoras de censura, tal que $\delta_i = 0$ indica que o indivíduo foi censurado e $\delta_i = 1$ o indivíduo não foi censurado. Considerando que os tempos de sobrevivência e de censura são independentes e que a censura é não informativa, a função de verossimilhança para todos os mecanismos de censura segundo Lawless (2003) é dada por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[f(t_i; \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i; \boldsymbol{\theta})\right]^{1-\delta_i},$$

em que $\boldsymbol{\theta}$ um vetor de parâmetros. A contribuição de cada indivíduo para a verossimilhança é dada pela função densidade no ponto t_i se o indivíduo apresentou o evento de interesse e pela função de sobrevivência no ponto t_i se o indivíduo foi censurado. O logaritmo da função de verossimilhança é dado por:

$$\ell(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \log \left[f(t_i; \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i; \boldsymbol{\theta})\right]^{1-\delta_i}.$$

O estimador de máxima verossimilhança para o vetor de parâmetros é calculado maximizando a função de verossimilhança $L(\boldsymbol{\theta})$ ou, equivalentemente, o logaritmo da função de verossimilhança $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$. Para encontrar os estimadores de máxima verossimilhança deve-se resolver o sistema de equações dado por:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

O sistema de equações dado por $U(\boldsymbol{\theta}) = 0$ é não linear, tornando-se necessário usar um algoritmo de otimização para resolvê-lo. Dessa forma, as estimativas desses parâmetros foram obtidas por meio de maximização numérica do logaritmo da função de verossimilhança, usando um processo iterativo, que foi o algoritmo quase-Newton baseado no método BFGS. No caso em que o tamanho da amostra é grande e sob certas condições de regularidade para a função de verossimilhança, intervalos de confiança e testes de hipóteses para os parâmetros podem ser obtidos usando o fato de que $\widehat{\boldsymbol{\theta}}$ tem distribuição assintótica Normal multivariada com média θ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$ estimado por $I(\boldsymbol{\theta}) = -E[L(\boldsymbol{\theta})]$ e $L(\boldsymbol{\theta}) = \left\{\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right\}$. Ou seja, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N_p(0, I(\boldsymbol{\theta})^{-1})$, em que $I(\boldsymbol{\theta})$ é a informação de fisher e q o número de parâmetros do modelo. Visto que o cálculo da $I(\boldsymbol{\theta})$ não é possível devido a presença de censuras, pode-se utilizar alternativamente a matriz de informação observada - $L(\boldsymbol{\theta})$ avaliada em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, em que esta matriz é um estimador consistente para $\boldsymbol{\Sigma}$(Mudholkar *et al.*, 1995). Para a construção de um teste de hipótese para os parâmetros é utilizada a diagonal principal da inversa da matriz -$L(\boldsymbol{\theta})$ como estimativa da matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$. Em relação ao intervalo de confiança para β_i em que $i = 1, ..., p$ com um nível de confiança ξ é dado por:

$$\widehat{\beta}_i \pm z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\beta}_i)}.$$

Sendo assim, a estatística para testar as hipóteses $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ é dada por $Z = \frac{\widehat{\beta}_i - \beta_i}{\sqrt{\widehat{V}(\beta_i)}} \sim N(0, 1)$.

2.1.4 Seleção de modelo

2.1.3.1 Critério de informação de Akaike (AIC) e Critério de informação de Bayes (BIC)

Como existem vários modelos probabilísticos é preciso escolher qual o mais adequado para um determinado conjunto de dados e é neste sentido que os critérios de seleção são utilizados. Neste trabalho foram utilizados: Critério de informação de Akaike (AIC) e o Critério de informação de Bayes (BIC) que são definidos por:

$$AIC = -2\log(L) + 2k; \qquad BIC = -2\log(L) + k\log(n),$$

em que L representa função de verossimilhança, k representa o número de parâmetros do modelo e n representa o número de observações. Tanto para o AIC quanto para BIC, o "melhor"modelo apontado pelo AIC (ou BIC) dentre vários outros, é o que tiver o menor AIC (ou BIC).

2.1.3.2 Teste assintótico da razão de verossimilhanças realizado no limite do espaço paramétrico

Dois modelos de probabilidade para um par de variáveis (Y, X) são ditos encaixados se um deles, o modelo nulo, for um caso especial do outro, o modelo alternativo. A escolha de um dentre esses dois modelos é a escolha entre duas hipóteses; H_0, que diz que o modelo adequado é o modelo nulo, e H_1 que afirma que o modelo alternativo é o adequado. Uma estatística de teste w_n denominada teste da razão de verossimilhanças $L(Y, X; H)$ que na verdade é igual ao logaritmo do quadrado da razão entre a função de verossimilhança sob H_1 e a função de verossimilhança sob H_0, é dada pela fórmula:

$$w_n = \log\left[\frac{L(Y, X; H_1)}{L(Y, X; H_0)}\right]^2 = -2\left[\{\log(L(Y, X; H_0))\} - \{\log(L(Y, X; H_1))\}\right]$$

A distribuição desta estatística sob H_0, quando n tende a infinito se aproxima (sob algmas condições de regularidade da distribuição de Y dado X) da distribuição qui-quadrado com número de graus de liberdade igual a diferença em valor absoluto entre o número de parâmetros dos dois modelos. O teste de hipótese H_0 versus H_1 com nível de significância ξ é feito comparando o valor observado da estatística de teste, w_n, com o quantil $(1 - \xi)$ da distribuição qui-quadrado com n graus de liberdade de referência e rejeitando H_0 se w_n superar esse quantil. O modelo Weibull $\sim$ W(β, γ) é encaixado no modelo Weibull-Poisson $\sim$ WP(α, β, γ) no limite do espaço paramétrico, ou seja, quando α tende a 0 a distribuição Weibull-Poisson WP (α, β, γ) se aproxima de uma distribuição Weibull (β, γ). Mais detalhes em Cancho *et al.* (2011).

Seja $\boldsymbol{\theta}$=(α, β, γ) o vetor de parâmetros da distribuição Weibull-Poisson dos tempos de vida para uma amostra de tamanho n. Para decidir se o modelo mais adequado é

WP(α, β, γ) ou W(β, γ), neste caso, testa-se a hipótese H_0: a distribuição da variável é W(β, γ) versus a hipótese alternativa H_1: a distribuição da variável é WP(α, β, γ). A estatística de teste nesse caso w_n é descrita por:

$$\omega_n = 2[\log(L)(\widehat{\theta_1}) - \log(L)(\widehat{\theta_0})],$$

em que $\widehat{\theta_0}$ é o estimador de máxima verossimilhança para vetor de parâmetros θ_0 do modelo definido por H_0 e $\widehat{\theta_1}$ é o estimador de máxima verossimilhança para o vetor de parâmetros θ_1 do modelo definido por H_1.

Na presença de dados de sobrevivência com longa duração, o modelo de função de sobrevivência geral é $S_{pop}(t) = p + (1-p)S_T(t)$. Essa função foi construída admitindo uma variável aleatória não negativa T que representa o tempo de vida de um indivíduo numa população na qual se admite indivíduos doentes (suscetíveis) com probabilidade $1-p$ e indivíduos curados (não suscetíveis) com probabilidade p. Desta forma, dada a função de sobrevivência populacional, $S_{pop}(t)$, temos que $\lim_{t\to\infty} S_{pop}(t) = p$, em que p é a proporção de indivíduos não suscetíveis ao evento de interesse. O $S_{pop}(t)$ é um caso particular de $S_T(t)$ quando ocorre $p = 0$, logo, é razoável proceder com um teste de hipótese para verificar se a proporção de indivíduos (ou imunes) na população é significativa, ou seja, estamos interessados em testar a hipótese nula $H_0 : p = 0$ versus $H_1 : p > 0$. Para testar se a proporção de indivíduos curados na população é nula, a estatística de teste da razão de verossimilhança tem a seguinte expressão:

$$\omega_n = 2[\log(L)(\widehat{\theta_1}, \widehat{p}) - \log(L)(\widehat{\theta_0}, 0)].$$

Sob certas condições de regularidades, Maller & Zhou (1995) mostraram que a distribuição da estatística da razão de verossimilhança sob H_0 é uma mistura com pesos (0.5 e 0.5) de uma distribuição qui-quadrado com um grau de liberdade com uma distribuição discreta com massa concentrada no valor 0, isto é,

$$P(\omega_n \leq w) = \frac{1}{2} + \frac{1}{2}P(\chi_1^2 \leq w). \tag{2.1}$$

O percentil de 95^0 da distribuição dada em 2.1, representado por $w_{0.95}$, é tal que,

$$\frac{1}{2} + \frac{1}{2}P(\chi_1^2 \leq w_{0.95}) = 0.95,$$

de forma que $w_{0.95} = 2,705543$. Portanto, rejeita-se H_0 a um nível de significância de 5% se $\omega_n > 2.705543$.

2.1.5 Análise de resíduos

Uma etapa importante após a formulação do modelo de regressão é a análise dos resíduos, destinada a avaliar se o modelo proposto está adequado, e também a identificar observações discrepantes. Técnicas gráficas são utilizadas para obter indícios de que o modelo está bem ajustado. De maneira geral, define-se resíduo referente a i-ésima observação por meio de uma função que depende da variável resposta e das estimativas dos parâmetros. Em relação à análise de sobrevivência são propostos na literatura vários resíduos. Ver por exemplo: Cox & Snell (1968), Colosimo & Giolo (2006). Para avaliar a qualidade do ajuste foi usado neste trabalho o resíduo de Cox-Snell ($\widehat{\epsilon_i}$).

2.2.5.1 Resíduo Cox-Snell

Esse resíduo é utilizado para verificar o ajuste global do modelo. Ele é descrito como :

$$\widehat{\epsilon_i} = \widehat{\Lambda}(t_i|x_i); i = 1, \ldots, n,$$

em que $\Lambda(.)$ é a função de risco acumulada estimada. De acordo com Lawless (2003) os resíduos devem seguir uma distribuição exponencial padrão se o modelo ajustado for adequado. Este resultado decorre do fato de que se T for uma variável do tipo contínua com função distribuição acumulada $F_T(t)$ e função de sobrevivência $S_T(t)$, as variáveis aleatórias $F_T(t)$ e $S_T(t)$ têm distribuição uniforme em $[0, 1]$ e em consequência disto $-\log(F_T(t))$ e $-\log(S_T(t)) = \Lambda(t)$ têm distribuição exponencial padrão. De acordo com Colosimo & Giolo (2006) uma forma de avaliar se o modelo proposto é adequado utilizando o resíduo Cox-Snell ($\widehat{\epsilon_i}$) é traçar o gráfico da curva de sobrevivência Kaplan-Meier calculada para os resíduos em relação ao modelo proposto e o gráfico da função de sobrevivência da distribuição exponencial com parâmetro 1. Quanto mais próximas as duas curvas se apresentarem, melhor é considerado o ajuste do modelo.

Capítulo 3

Modelo de regressão usando a distribuição Log Weibull-Poisson

Neste Capítulo a distribuição Weibull-Poisson (WP) proposta por Louzada *et al.* (2011a), é apresentada para fundamentar duas propostas novas que são nela baseadas. Esta distribuição pode modelar a distribuição de probabilidade de uma variável T que corresponde ao tempo de duração de um sistema em série, impondo condições sobre o número N de seus componentes. Admitindo uma distribuição de probabilidade para N e uma distribuição conjunta do tempo de duração de cada componente, para cada valor observado de N, resultando na determinação de uma nova distribuição que é a Weibull-Poisson. É deduzida sua função densidade de probabilidade $f_T(t)$, e a partir dela, são determinadas sua função distribuição acumulada $F_T(t)$, o quantil $Q_T(t)$, a função de sobrevivência $S_T(t)$ e a função de risco ou taxa de falha $h_T(t)$. São apresentados gráficos para alguns valores dos parâmetros dessas funções, mostrando que a função de risco pode apresentar diversas formas. O momento de ordem k da distribuição WP é expresso como uma série e é obtida a expressão em forma fechada da função densidade de probabilidade da k-ésima estatística de ordem de uma amostra de tamanho n da distribuição WP. Além disso, discutimos a similaridade entre a função densidade de probabilidade do mínimo e a do máximo de N variáveis aleatórias independentes com distribuição Weibull, quando a variável aleatória N segue uma distribuição Poisson truncado no zero. A partir desta revisão fizemos de forma orginal um estudo de simulação destinado a avaliar o efeito do percentual de censura nos dados bem como sobre o comportamento dos estimadores por máxima verossimilhança dos seus parâmetros, quanto ao vício e o erro quadrático médio, para diferentes tamanhos de amostras. Além disso, foi feita uma aplicação em dois conjuntos de dados extraídos da literatura. A ideia é mostrar a aplicabilidade dessa distribuição. Outra contribuição

original do autor neste capítulo é um modelo de regressão na forma de locação e escala de distribuições geradas pela distribuição Log Weibull-Poisson quando o parâmetro de locação é função linear de covariáveis. É apresentada uma aplicação a um conjunto de dados reais.

3.1 Distribuição Weibull-Poisson

Conforme já foi discutido na introdução, um procedimento usado na literatura para criar novas distribuições de probabilidade é fazer uma mistura enumerável de uma família paramétrica de distribuições contínuas, isto é, uma combinação linear dos elementos dessa família, com pesos tirados da distribuição de uma variável discreta. Um modelo para o tempo de vida de um sistema em série com N componentes, considera que o j-ésimo componente possui um tempo de duração aleatório Y_j. Se os tempos de duração dos diferentes componentes forem independentes entre si e identicamente distribuídos, a distribuição do tempo de duração do sistema em série dado $N = n$ é a distribuição do tempo mínimo entre os tempos de duração dos componentes, ou seja $(T|_{N=n}) = \min(Y_1, ..., Y_n)$. A função de sobrevivência desta nova variável $T|_{N=n}$ é $S|_{N=n}(y) = (S_Y(t))^n$, em que $S_Y(y)$, é a sobrevivência correspondente a variável Y. A função densidade de $(T|_{N=n})$, é $f|_{N=n}(y) = n[S_Y(t)]^{n-1} f_Y(t)$. A partir dessas considerações conseguimos encontrar a distribuição de probabilidade de $T = \min(Y_1, ..., Y_N)$, cuja função densidade é dada por $f_T(t) = E_N[N(S_Y(y))^{N-1} f_Y(y)]$. A criação da Weibull-Poisson (Louzada *et al.*, 2011a) foi motivada pela duração de um sistema em série, que é o tempo até que ocorra a primeira falha por alguma das N causas, em que cada uma delas com um tempo de vida Y_i, $i = 1, ..., N$. A falha do sistema ocorre no tempo $T = \min(Y_i, i = 1, ..., N)$. Especificamente, considera-se que tanto o número de componentes como a identificação do componente que provocou a falha não são observáveis e sim o valor do tempo de vida mínimo dentre eles é observado, considerando que este tipo de construção pode ser válido na modelagem da distribuição do tempo até a falha de um sistema em série quando vários fatores competem entre si para ocasionar a falha do sistema e muitas vezes não há possibilidade de identificar o fator ou causa da falha, sendo apenas observável o mínimo dentre os tempos de falha por cada componente.

3.1.1 Formulação do modelo

Sejam $Y_1, Y_2, ..., Y_N$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d) com distribuição Weibull $Y \sim \mathrm{W}(\beta, \gamma)$ com parâmetro de forma $\gamma > 0$ e de escala $\beta > 0$. A função densidade de probabilidade e sobrevivência, são respectivamente:

$$f_Y(y) = \gamma\beta^\gamma y^{\gamma-1} \exp\left\{-(\beta y)^\gamma\right\} I_{[0,\infty)}(y) \tag{3.1}$$

$$S_Y(y) = \begin{cases} 1, & y \leq 0 \\ \exp\left\{-(\beta y)^\gamma\right\}, & y > 0. \end{cases} \tag{3.2}$$

Seja N uma variável aleatória discreta com distribuição de Poisson truncada no zero de parâmetro $\alpha > 0$. A função de probabilidade $f_N(n)$ é dada por:

$$f_N(n) = \frac{\alpha^n}{n![\exp(\alpha) - 1]};\ n = 1, 2, 3.... \tag{3.3}$$

Teorema

Seja T uma variável aleatória não negativa definida como o mínimo de N variáveis aleatórias independentes $Y_i;\ i = 1, 2, ..., N,\ T = \min(Y_1, ..., Y_N)$, em que cada Y_i segue distribuição Weibull, isto é, $\mathrm{Y} \sim \mathrm{W}(\beta, \gamma)$ e N segue uma distribuição de Poisson(α) truncada no zero. Com isso a função densidade de $T = \min(Y_1, ..., Y_N)$ é dada por:

$$f_T(t) = \begin{cases} \dfrac{\alpha \exp\left\{\alpha \exp\left[-(\beta t)^\gamma\right] - (\beta t)^\gamma\right\} \beta^\gamma t^{\gamma-1}\gamma}{\exp(\alpha) - 1}; & t > 0,\ \alpha > 0,\ \beta > 0,\ \gamma > 0. \\ 0, & c.c. \end{cases}$$

Inicialmente para prova deste teorema será determinado que quando $N = n$, a variável T está condicionada a $N = n$, ou seja $T|_{N=n}$, que é dada por $(T|_{N=n}) = \min\{Y_1, ..., Y_n\}$. Logo, para encontrar a função densidade de T, neste primeiro momento encontra-se a função densidade condicional de $T|_{N=n}$, que é escrita da forma $f_{T|_{N=n}}(t) = \min\{Y_1, ..., Y_n\}$. Note que:

$$
\begin{aligned}
F_{T|_{N=n}}(t) &= [P\left((T|_{N=n}) \leq t\right)] \\
&= [P(\min\{Y_1, \ldots, Y_n\} \leq t)] \\
&= 1 - [P(\min\{Y_1, \ldots, Y_n\} > t)] \\
&= 1 - [P(Y_1 > t, Y_2 > t, Y_3 > t, \ldots, Y_n > t)] \\
&= 1 - \prod_{i=1}^{n} [P(Y_i > t)] \\
&= 1 - \prod_{i=1}^{n} [1 - P(Y_i =< t)] \\
&= 1 - \prod_{i=1}^{n} [1 - F_{Yi}(t)] \\
F_{T|_{N=n}}(t) &= 1 - [1 - F_{Yi}(t)]^n \\
1 - F_{T|_{N=n}}(t) &= [1 - F_{Yi}(t)]^n \\
S_{T|_{N=n}}(t) &= [S_{Yi}(t)]^n \\
&= [\exp\{-(\beta t)^{\gamma}\}]^n \\
&= [\exp\{-n(\beta t)^{\gamma}\}] \\
&= \left[\exp\left\{-\left(\beta t^{1/n}\right)^{\gamma}\right\}\right] \\
S_{T|_{N=n}}(t) &= \left[\exp\left\{-\left(\beta t^{1/n}\right)^{\gamma}\right\}\right].
\end{aligned}
$$

De acordo com a equação 3.2 observa-se que $S_{T|_{N=n}}(t)$ corresponde a função de sobrevivência de uma distribuição Weibull $(T|_{N=n}) \sim \mathrm{W}(\beta n^{1/\gamma}, \gamma)$. Logo a função densidade condicional de $f_{T|_{N=n}}(t) = \min\{Y_1, ..., Y_n\}$ é dada por:

$$
f_{T|N=n}(t) = \gamma(\beta n^{1/\gamma})^{\gamma} t^{\gamma-1} \exp\left\{-\left(n^{1/\gamma}\beta t\right)^{\gamma}\right\};\ t > 0, \quad \beta > 0,\ \gamma > 0. \tag{3.4}
$$

Para encontrar a função de densidade de $T = \min(Y_1, ..., Y_N)$, primeiramente utiliza-se que a função densidade conjunta de T e N é dada por:

$$
f_{T;N=n}(t) = f_{T|N=n}(t) f_N(n), \tag{3.5}
$$

substituindo $f_{T|_{N=n}}(t)$ e $f_N(n)$ dadas nas equações: 3.4 e 3.3 na equação 3.5, temos:

$$f_{T;N=n}(t) = \gamma(\beta n^{1/\gamma})t^{\gamma-1}\exp\left\{-\left(n^{1/\gamma}\beta t\right)^{\gamma}\right\}\frac{\alpha^n}{n![\exp(\alpha)-1]}.$$

Sabe-se que,

$$f_T(t) = \sum_{n=1}^{\infty} f_{T;N=n}(t) f_N(n),$$

resultado da seção 3.1, então:

$$f_T(t) = \sum_{n=1}^{\infty} \gamma(\beta n^{1/\gamma})t^{\gamma-1}\exp\left\{-\left(n^{1/\gamma}\beta t\right)^{\gamma}\right\}\frac{\alpha^n}{n![\exp(\alpha)-1]}$$

$$f_T(t) = \frac{1}{\exp(\alpha)-1}\sum_{n=1}^{\infty}\frac{\alpha^n}{n!}\gamma(n^{1/\gamma}\beta)^{\gamma}t^{\gamma-1}\exp\left\{-\left(n^{1/\gamma}\beta t\right)^{\gamma}\right\} \tag{3.6}$$

Substituindo $n = j+1$ na expressão 3.6, consegue-se obter:

$$f_T(t) = \frac{1}{\exp(\alpha)-1}\sum_{j=0}^{\infty}\frac{\alpha^{j+1}\gamma\left[\beta(j+1)^{1/\gamma}\right]^{\gamma}t^{\gamma-1}\exp\left\{-\left[\beta t\,(j+1)^{1/\gamma}\right]^{\gamma}\right\}}{(j+1)!}. \tag{3.7}$$

Simplificando a equação 3.7, obtem-se que:

$$\begin{aligned} f_T(t) =& \frac{1}{\exp(\alpha)-1}\sum_{j=0}^{\infty}\frac{\alpha^{j+1}\gamma\beta^{\gamma}t^{\gamma-1}\exp\left\{-(j+1)(\beta t)^{\gamma}\right\}}{j!} \\ & \frac{1}{\exp(\alpha)-1}\gamma\beta^{\gamma}t^{\gamma-1}\alpha\exp\left\{-(\beta t)^{\gamma}\right\}\sum_{j=0}^{\infty}\frac{\alpha^j\exp\left\{-[j(\beta t)^{\gamma}]\right\}}{j!}. \end{aligned} \tag{3.8}$$

Como $\exp(-\alpha) = \sum_{j=0}^{\infty}\frac{\alpha^j}{j!}$, então $\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]\right\} = \sum_{j=0}^{\infty}\frac{[\alpha\exp\{-(\beta t)^{\gamma}\}]^j}{j!}$.
Utilizando esse resultado, a equação 3.8 pode ser escrita como:

$$f_T(t) = \frac{\alpha\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\beta^{\gamma}t^{\gamma-1}\gamma}{\exp(\alpha)-1};\ t>0,\ \alpha>0,\ \beta>0,\ \gamma>0. \tag{3.9}$$

A função densidade de probabilidade correspondente a equação 3.9 chama-se Weibull-Poisson$\sim$WP(α,β,γ). A Figura 3.1 mostra a função densidade da WP para alguns valores dos parâmetros:

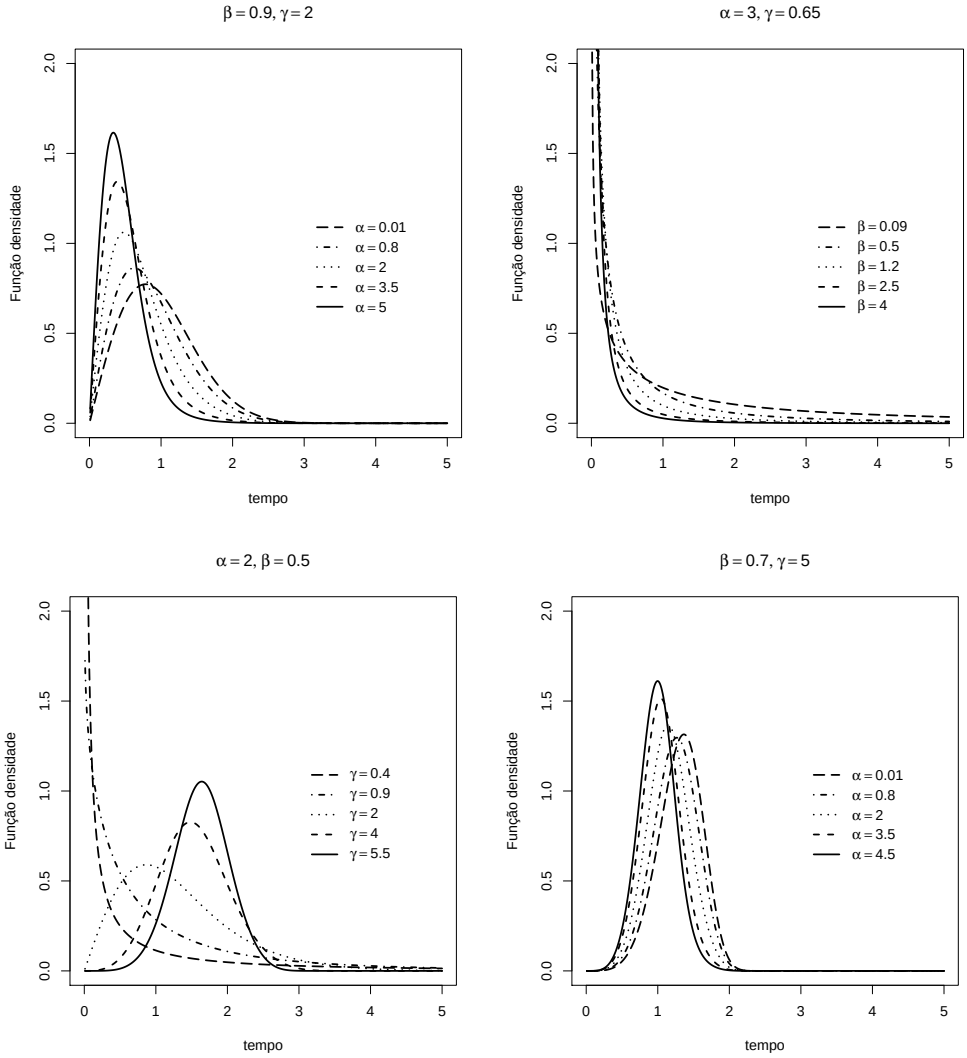

Figura 3.1: Gráficos ilustrativos da função densidade para a distribuição de probabilidade WP.

3.1.2 Propriedades da distribuição Weibull-Poisson

Seja T uma variável aleatória com distribuição Weibull-Poisson $\sim$WP(α, β, γ). A função distribuição acumulada $F_T(t)$, o Quantil desta função $Q_T(t)$, a função de sobrevivência $S_T(t)$ e a função de risco (taxa de falha) $h_T(t)$, são respectivamente:

$$F_T(t) = \frac{\exp(\alpha) - \exp\{\alpha \exp[-(\beta t)^\gamma]\}}{\exp(\alpha) - 1};\ t > 0,\ \beta > 0,\ \gamma > 0,\ \alpha > 0, \tag{3.10}$$

$$Q_T(t) = \frac{[\log(\alpha) - \log\{\log\{\exp(\alpha)(1-u)\} + u\}]^{1/\gamma}}{\beta};\ t > 0,\ \beta > 0,\ \gamma > 0,\ \alpha > 0,$$

$$S_T(t) = \frac{\exp\{\alpha \exp[-(\beta t)^\gamma]\} - 1}{\exp(\alpha) - 1};\ t > 0,\ \beta > 0,\ \gamma > 0,\ \alpha > 0 \text{ e} \tag{3.11}$$

$$h_T(t) = \frac{\alpha \exp\{\alpha \exp[-(\beta t)^\gamma] - (\beta t)^\gamma\} \beta^\gamma t^{\gamma-1} \gamma}{\exp\{\alpha \exp[-(\beta t)^\gamma]\} - 1};\ t > 0,\ \beta > 0,\ \gamma > 0,\ \alpha > 0.$$

A Figura 3.2 mostra a função de sobrevivência para alguns valores dos parâmetros:

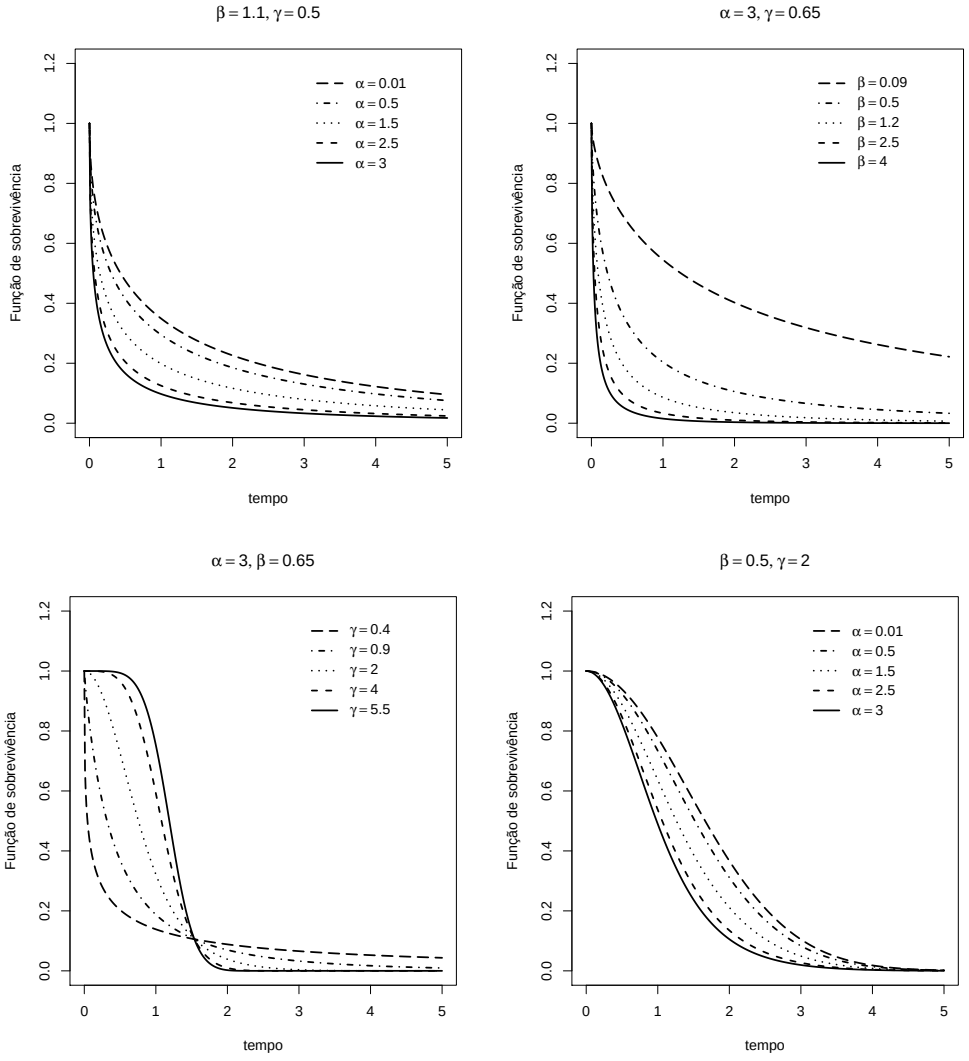

Figura 3.2: Gráficos ilustrativos da função de sobrevivência para a distribuição de probabilidade WP.

A Figura 3.3 mostra a função de risco da distribuição Weibull-Poisson (WP) para alguns valores dos parâmetros. Pode ser visto que a sua função de risco é bastante flexível e pode acomodar várias formas, como crescente, decrescente e unimodal.

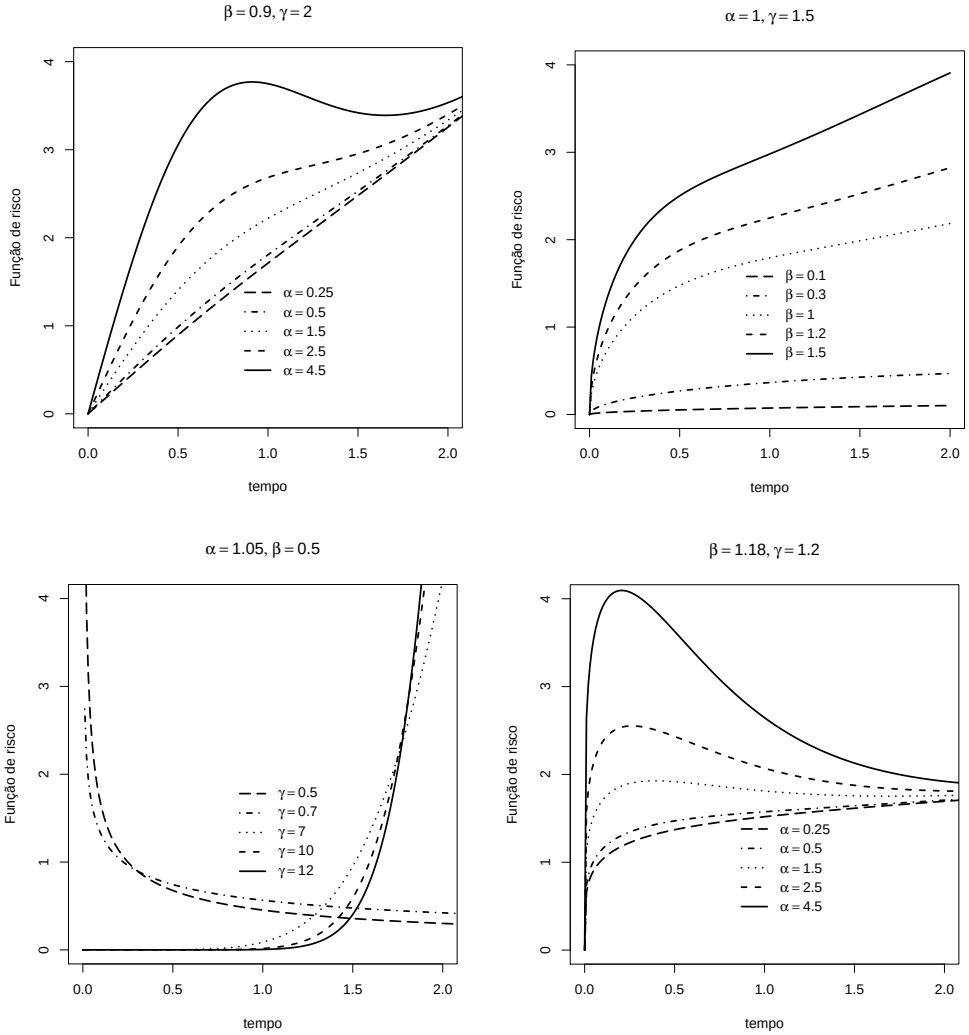

Figura 3.3: Gráficos ilustrativos da função de risco para a distribuição de probabilidade WP.

Teorema

O momento central de ordem r da distribuição Weibull-Poisson T é dado por:

$$E(T^r) = \frac{\gamma}{\exp(\alpha) - 1}\Gamma\left(1 + \frac{r}{\gamma}\right)\beta^{-r}\sum_{j=1}^{\infty} j^{-r/\gamma}\frac{\alpha^j}{j!}$$

Prova:

$$E(T^r) = \int_0^{\infty} t^r f_T(t)dt$$

$$E(T^r) = \int_0^{\infty} \frac{t^r \alpha \exp\left\{\alpha \exp\left[-(\beta t)^{\gamma}\right] - (\beta t)^{\gamma}\right\}\beta^{\gamma} t^{\gamma-1}\gamma}{\exp(\alpha) - 1}dt.$$

Sabe-se que:

$$\exp\{\alpha\exp[-(\beta t)^{\gamma}]\} = \sum_{j=0}^{\infty}\frac{[\alpha\exp\{-(\beta t)^{\gamma}\}]^{j}}{j!}.$$

Então:

$$E(T^{r}) = \int_{0}^{\infty}\frac{t^{r}}{\exp(\alpha)-1}\sum_{j=1}^{\infty}\frac{\alpha^{j}\gamma\left(\beta j^{1/\gamma}\right)^{\gamma}t^{\gamma-1}\exp\left\{-\left(\beta j^{1/\gamma}t\right)^{\gamma}\right\}}{j!}dt$$

$$E(T^{r}) = \frac{1}{\exp(\alpha)-1}\sum_{j=1}^{\infty}\frac{\alpha^{j}}{j!}\int_{0}^{\infty}t^{r}\gamma\left(\beta j^{1/\gamma}\right)^{\gamma}t^{\gamma-1}\exp\left\{-\left(\beta j^{1/\gamma}t\right)^{\gamma}\right\}dt. \tag{3.12}$$

Lembrando que o momento de ordem r da distribuição Weibull de parâmetros $(\beta j^{1/\gamma}, \gamma)$ é dado por:

$$\mu(r) = \int_{0}^{\infty}t^{r}\gamma\left(\beta j^{1/\gamma}\right)^{\gamma}t^{\gamma-1}\exp\left\{-\left(\beta j^{1/\gamma}t\right)^{\gamma}\right\}dt = (\beta j^{1/\gamma})^{-r}\Gamma(1+\frac{r}{\gamma}). \tag{3.13}$$

Substituindo 3.13 em 3.12, tem-se que:

$$E(T^{r}) = \frac{1}{\exp(\alpha)-1}\sum_{j=1}^{\infty}\frac{\alpha^{j}}{j!}(\beta j^{1/\gamma})^{-r}\Gamma(1+\frac{r}{\gamma}) = \frac{\gamma}{\exp(\alpha)-1}\Gamma\left(1+\frac{r}{\gamma}\right)\beta^{-r}\sum_{j=1}^{\infty}j^{-r/\gamma}\frac{\alpha^{j}}{j!}$$

■

A função densidade da k-ésima estatística de ordem de uma distribuição é dada por:

$$g_{k:n}(t) = \frac{1}{B(k,n-k+1)}f(t)(F(t))^{k-1}(S(t))^{n-k}, \tag{3.14}$$

em que $B(k,n-k+1) = \dfrac{(n-k)!(k-1)!}{n!}$.

Logo, substituindo as funções $f_T(t)$, $F_T(t)$ e $S_T(t)$ dadas em 3.9, 3.10 e 3.11 na equação 3.14 tem-se a função densidade da k-ésima estatística de ordem da WP, que é dada por:

$$\begin{aligned} g_{k:n}(t) =& \alpha\exp\{\alpha\exp[-(\beta t)^{\gamma}] - (\beta t)^{\gamma}\}\beta^{\gamma}t^{\gamma-1}\gamma\left[\exp(\alpha) - \exp\{\alpha\exp[-(\beta t)^{\gamma}]\}\right]^{k-1}\times \\ &\times\frac{[\exp\{\alpha\exp[-(\beta t)^{\gamma}]\}-1]^{n-k}}{B(k,n-k+1)[\exp(\alpha)-1]^{n}}. \end{aligned}$$

3.1.3 A similaridade entre as funções densidades do mínimo e do máximo

Nesta seção, discute-se a semelhança entre a função densidade de probabilidade do mínimo e a do máximo das variáveis aleatórias independentes com distribuição Weibull, quando a variável aleatória N segue uma distribuição Poisson truncado no zero. A equação é descrita por:

$$f_{T^*}(t) = \frac{-\alpha \exp\left\{-\alpha \exp\left[-(\beta t)^{\gamma}\right] - (\beta t)^{\gamma}\right\} \beta^{\gamma} t^{\gamma-1} \gamma}{\exp(-\alpha) - 1}; \; t > 0, \, \alpha > 0, \; \beta > 0, \; \gamma > 0. \quad (3.15)$$

É interessante verificar que a expressão 3.9 é válida para a variável $T^* = \max(Y_1, ..., Y_N)$ e $T = \min(Y_1, ..., Y_N)$ quando substituimos α por $-\alpha$ na expressão de função de densidade de T^*.

A prova deste resultado primeiramente baseia-se no fato de que quando $N = n$, a variável T^* condicionada a $N = n$, ou seja $T^*|_{N=n}$ é denotada por $(T^*|_{N=n}) = \max\{Y_1, ...Y_n\}$. Note que:

$$\begin{aligned}
F_{T^*|_{N=n}}(t) &= \left[P\left(T^* < t\right)|N = n\right] \\
&= \left[P(\max\{Y_1, \ldots, Y_n\} < t)\right] \\
&= \left[P(Y_1 < t, Y_2 < t, Y_3 < t, \ldots, Y_n < t)\right] \\
&= \prod_{i=1}^{n} \left[P(Y_i < t)\right] \\
&= \prod_{i=1}^{n} \left[F_{Yi}(t)\right] \\
&= \left[F_{Yi}(t)\right]^n \\
&= \left[1 - S_{Yi}(t)\right]^n \\
&= \left[1 - \exp\left\{-(\beta t)^{\gamma}\right\}\right]^n . \qquad (3.16)
\end{aligned}$$

Sabe-se que a função densidade $f_{T^*|N=n}(t)$ é:

$$f_{T^*|_{N=n}}(t) = \frac{\partial[F_{T^*|_{N=n}}(t)]}{\partial t}.$$

Então,

$$\begin{aligned} f_{T^*|_{N=n}}(t) &= n\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{n-1}\left[-\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[-\gamma(\beta t)^{\gamma-1}\beta\right] \\ &= n\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{n-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\gamma\beta^{\gamma-1}t^{\gamma-1}\beta \\ &= n\gamma\beta^{\gamma}t^{\gamma-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{n-1}. \end{aligned} \tag{3.17}$$

Para encontrar a função de densidade de $T^* = \max(Y_1, ..., Y_N)$ que é definida por $f_{T^*}(t)$, é obtida através da relação:

$$f_{T^*;N=n}(t) = f_{T^*|_{N=n}}(t)(f_N(n)). \tag{3.18}$$

Substituindo $f_{T^*|_{N=n}}(t)$ e $f_N(n)$ dadas nas equações 3.17 e 3.3 na equação 3.18, tem-se:

$$f_{T^*;N=n}(t) = n\gamma\beta^{\gamma}t^{\gamma-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{n-1}\frac{\alpha^n}{n![\exp(\alpha)-1]}.$$

Sabe-se que $f_{T^*}(t) = \sum_{n=1}^{\infty} f_{T;N=n}(t).f_N(n)$. Então,

$$\begin{aligned} f_{T^*}(t) &= \sum_{n=1}^{\infty} n\gamma\beta^{\gamma}t^{\gamma-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{n-1}\frac{\alpha^n}{n![\exp(\alpha)-1]} \\ &= \frac{1}{\exp(\alpha)-1}\sum_{n=1}^{\infty} n\gamma\beta^{\gamma}t^{\gamma-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{n-1}\frac{\alpha^n}{n!}. \end{aligned} \tag{3.19}$$

Fazendo $n = j + 1$ na expressão 3.19, obtem-se que:

$$f_{T^*}(t) = \frac{1}{\exp(\alpha)-1}\sum_{j=0}^{\infty}(j+1)\gamma\beta^{\gamma}t^{\gamma-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{j+1-1}\frac{\alpha^{j+1}}{(j+1)!} \tag{3.20}$$

Simplificando a equação 3.20, obtem-se que:

$$\begin{aligned} f_{T^*}(t) &= \frac{1}{\exp(\alpha)-1}\sum_{j=0}^{\infty}\gamma\beta^{\gamma}t^{\gamma-1}\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]^{j}\frac{\alpha^{j+1}}{j!} \\ &= \frac{1}{\exp(\alpha)-1}\gamma\beta^{\gamma}t^{\gamma-1}\alpha\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\sum_{j=0}^{\infty}\frac{[\alpha(1-\exp\left\{-(\beta t)^{\gamma}\right\})]^{j}}{j!}. \end{aligned} \tag{3.21}$$

Sabe-se que $\sum_{j=0}^{\infty}\frac{\alpha^j}{j!}$. Então:

$$\exp\left\{\alpha\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]\right\} = \sum_{j=0}^{\infty}\frac{[\alpha(1-\exp\left\{-(\beta t)^{\gamma}\right\})]^{j}}{j!},$$

Assim, a equação 3.21 pode ser escrita como:

$$\begin{aligned}
f_{T^*}(t) &= \frac{1}{\exp(\alpha)-1}\gamma\beta^{\gamma}t^{\gamma-1}\alpha\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\exp\left\{\alpha\left[1-\exp\left\{-(\beta t)^{\gamma}\right\}\right]\right\} \\
&= \frac{1}{\exp(\alpha)-1}\gamma\beta^{\gamma}t^{\gamma-1}\alpha\left[\exp\left\{-(\beta t)^{\gamma}\right\}\right]\exp(\alpha)\exp\left\{-\alpha\exp\left[-(\beta t)^{\gamma}\right]\right\} \\
&= \frac{1}{\exp(\alpha)-1}\gamma\beta^{\gamma}t^{\gamma-1}\alpha\exp\left\{-\alpha\exp\left[-(\beta\gamma)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\exp(\alpha) \\
&= \frac{1}{\exp(\alpha)-1}\gamma\beta^{\gamma}t^{\gamma-1}\alpha\exp\left\{-\alpha\exp\left[-(\beta\gamma)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\exp(\alpha)\times\frac{\left[-\exp(-\alpha)\right]}{\left[-\exp(-\alpha)\right]} \\
&= \frac{-\alpha\exp\left\{-\alpha\exp\left[-(\beta t)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\beta^{\gamma}t^{\gamma-1}\gamma}{\exp(-\alpha)-1};\ t>0,\ \alpha>0,\ \beta>0,\ \gamma>0.
\end{aligned}$$

3.1.4 Casos particulares da distribuição Weibull-Poisson

A distribuição Weibull-Poisson apresenta algumas distribuições como casos particulares que serão apresentadas a seguir:

- Para $\alpha \to 0$ na equação 3.9 a distribuição Weibull-Poisson se reduz a uma distribuição Weibull, com função densidade da forma:

$$f(t) = \gamma\beta^{\gamma}t^{\gamma-1}\exp\left\{-(\beta t)^{\gamma}\right\}$$

- Para $\gamma = 1$ na equação 3.9 a distribuição Weibull-Poisson se reduz a uma distribuição exponencial-Poisson (Kus, 2007), com função densidade da forma:

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

3.1.5 Estimação por máxima verossimilhança da Weibull-Poisson

Dada uma amostra aleatória de tamanho n composta por dois vetores $\mathbf{t} = (t_1, ..., t_n)$ e $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$. C tempo de censura que é indepedente de T, em que $t_i = \min(T_i, C_i)$. Sendo que para cada $i = 1, ..., n$, a variável T_i tem distribuição WP($\boldsymbol{\theta}$) com vetor de parâmetros $\boldsymbol{\theta}$=(α, β, γ) e δ é um vetor de variáveis aleatórias indicadoras de censura.

Sendo $f(.)$ a função densidade própria dado por 3.9 e $S(.)$ a função de sobrevivência própria dada a equação 3.11. O logaritmo da função de verossimilhança $l(\boldsymbol{\theta})$ para o vetor de parâmetros $\boldsymbol{\theta}$ considerando que os tempos de sobrevivência e de censura são independentes e que a censura é não informativa, pode ser escrito como:

$$\ell(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \delta_i \log(\alpha\gamma\beta^{\gamma}t_i^{\gamma-1}) + \sum_{i=1}^{n} \delta_i \left[\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\} - (\beta t_i)^{\gamma}\right] - \sum_{i=1}^{n} \delta_i \log(\exp(\alpha) - 1) + \\ + \sum_{i=1}^{n} (1-\delta_i) \log\left\{\exp\left\{\alpha \exp\left[-(\beta t_i)^{\gamma}\right]\right\} - 1\right\} - \sum_{i=1}^{n} (1-\delta_i) \log(\exp(\alpha) - 1).$$

Da seção 2.1.3, sabe-se que os estimadores de máxima verossimilhança são obtidos a partir da resolução do sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

Os componentes do vetor escore $U(\boldsymbol{\theta}) = U(\alpha, \beta, \gamma) = \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \alpha}; \frac{\partial \ell(\theta)}{\partial \beta}; \frac{\partial \ell(\boldsymbol{\theta})}{\partial \gamma}\right)$ no modelo WP, são dados por:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \alpha} = \frac{\sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} \delta_i \exp\left\{-(\beta t_i)^{\gamma}\right\}}{\alpha} - \frac{\sum_{i=1}^{n} \delta_i \exp(\alpha)}{\exp(\alpha) - 1} + \\ + \sum_{i=1}^{n} (1-\delta_i) \frac{\left(\exp\left\{\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\}\right\}\right) \exp\left\{-(\beta t_i)^{\gamma}\right\}}{\left(\exp\left\{\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\}\right\} - 1\right)} + \frac{\sum_{i=1}^{n} (1-\delta_i)}{\exp(\alpha) - 1}.$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta} = \sum_{i=1}^{n} \delta_i \frac{\gamma}{\beta} + \sum_{i=1}^{n} \delta_i \left(\frac{-\alpha(\beta t_i)^{\gamma}\gamma \exp\left\{-(\beta t_i)^{\gamma}\right\}}{\beta} - \frac{(\beta t_i)^{\gamma}\gamma}{\beta}\right) \\ - \sum_{i=1}^{n} (1-\delta_i) \frac{\alpha(\beta t_i)^{\gamma}\gamma \left(\exp\left\{\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\}\right\}\right)}{\beta \left(\exp\left\{\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\}\right\} - 1\right)}.$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \gamma} = \sum_{i=1}^{n} \delta_i \frac{\alpha\beta^{\gamma}t_i^{\gamma-1} + \alpha\gamma\beta^{\gamma}\log(\beta)t_i^{\gamma-1} + \alpha\beta^{\gamma}t_i^{\gamma-1}\log(t_i)}{\alpha\gamma\beta^{\gamma}t_i^{\gamma-1}} + \\ + \sum_{i=1}^{n} \delta_i \left(-\alpha(\beta t_i)^{\gamma}\log(\beta t_i) \exp\left\{-(\beta t_i)^{\gamma}\right\} - (\beta t_i)^{\gamma}\log(\beta t_i)\right) \\ - \frac{\sum_{i=1}^{n} (1-\delta_i)\alpha(\beta t_i)^{\gamma}\log(\beta t_i) \exp\left\{\alpha \exp\left[-(\beta t)^{\gamma}\right] - (\beta t)^{\gamma}\right\}}{\left(\exp\left\{\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\}\right\} - 1\right)}.$$

Como não existe uma forma analítica fechada para encontrar esses estimadores pode-se recorrer a métodos numéricos para resolver o sistema de equações. Dessa forma, as estimativas desses parâmetros foram obtidas por meio de maximização numérica do logaritmo da função de verossimilhança, usando um processo iterativo, que foi o algoritmo quase-Newton baseado no método BFGS.

Intervalos de confiança e testes de hipótetes podem ser obtidos usando a distribuição para grandes amostras dos estimadores de máxima verossimilhança como descrito na seção 2.1.3. Assim, a aproximação normal assintótica para $\widehat{\boldsymbol{\theta}}$ pode ser expressa por $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \sim N_3(0, L(\boldsymbol{\theta})^{-1})$. Para a distribuição Weibull-Poisson($\boldsymbol{\theta}$), a matriz de informação observada é dada por:

$$L(\theta) = -\begin{bmatrix} L_{\alpha\alpha} & L_{\alpha\beta} & L_{\alpha\gamma} \\ . & L_{\beta\beta} & L_{\beta\gamma} \\ . & . & L_{\gamma\gamma} \end{bmatrix}$$

3.1.6 Estudo de Simulação

Com objetivo de verificar se o estimador de máxima verossimilhança é adequado para diferentes tamanhos de amostra, foi feito nesta seção um estudo de simulação tal que o tempo da ocorrência da falha Y segue uma distribuição Weibull-Poisson com os parâmetros: $\alpha = \gamma = 2$, $\beta = 1$ e o tempo C da censura uma distribuição Weibull com parâmetros: $\gamma = 2$ e $\beta = 1$. Os tempos de sobrevivência observados foram obtidos fazendo: $t_i = \min(Y_i, C_i)$. Foram realizadas $B = 1000$ simulações e a partir destas amostras simuladas obteve-se as estimativas de máxima verossimilhança usando o recurso numérico *optim* que encontra-se no *Software R*. Através destas determinou-se as médias das estimativas de máxima verossimilhança, o viés e o erro quadrático médio (EQM) com diferentes tamanhos de amostra e 20% de censura. Os valores iniciais para o processo de otimização foram valores próximos dos verdadeiros valores dos parâmetros. A seguir é descrito o processo desta simulação:

1. Gerar $u_j \sim U(0, 1)$;

2. Determinar $y_j = F^{-1}(u_j) = \frac{(\log(\alpha)-\log(\log(e^{\alpha}(1-u_j)+u_j)))^{1/\gamma}}{\beta}$, em que F(.) é a função distribuição acumulada dada em 3.10;

3. Gerar a variável de censura $c_j \sim$Weibull(β, γ) descrita em 3.1 e fazer $t_j = \min(y_j, c_j)$;

4. Se $y_j < c_j$, então $\delta_j = 1$, caso contrário, $\delta_j = 0$, para $j = 1, ..., n$.

Na Tabela 3.1 encontra-se as médias das estimativas, o viés e o erro quadrático médio das $B = 1000$ simulações com diferentes tamanhos de amostra com 20% censura. De acordo com a Tabela 3.1 nota-se que as médias das estimativas se aproximam do verdadeiro valor do parâmetro quando aumentamos o tamanho da amostra. Outro fator importante é que o erro quadrático médio diminui à medida que o tamanho amostral aumenta, ou seja, as estimativas dos parâmetros melhoram tornando-se os estimadores cada vez menos viesados.

Tabela 3.1: Resultados das simulações com o modelo Weibull-Poisson com os parâmetros: $\alpha = 2$, $\beta = 1$ e $\gamma = 2$ com 20% de censura

Tamanho da amostra (n)	Parâmetros	Média	Vício	EQM
20	α	1.24900	-0.75099	2.21061
	β	1.20013	0.20013	0.11797
	γ	2.08093	0.08093	0.21695
60	α	1.59110	-0.40889	1.99688
	β	1.11343	0.11343	0.06855
	γ	1.98392	-0.01607	0.06287
100	α	1.67502	-0.32497	1.88176
	β	1.09450	0.09450	0.06278
	γ	1.97212	-0.02787	0.04259
600	α	2.03419	0.03419	1.06764
	β	1.01207	0.01207	0.02466
	γ	1.98018	-0.01981	0.00831
1000	α	2.02170	0.02170	0.72835
	β	1.00843	0.00843	0.01641
	γ	1.98679	-0.01321	0.00511

3.1.7 Aplicação

Nesta seção, vamos comparar a distribuição Weibull-Poisson com os casos particulares exponencial-Poisson e a Weibull em dois conjuntos de dados extraídos da literatura. A ideia é mostrar a aplicabilidade da distribuição, bem como verificar sua utilidade em outro conjunto de dados.

O primeiro conjunto de dados foi extraído de Lee & Wang (2003) e se refere ao tempo de sobrevivência de 137 pacientes com câncer em que o interesse é estudar o tempo até a sua remissão sendo que 7 % dos valores são dados censurados. O segundo conjunto de dados foi extraído de Pradhan & Kundu (2013) e consiste de 101 observações sobre o tempo de vida de um equipamento que corta alumínio que oscila em 18 ciclos por segundo, este conjunto não apresenta observações censuradas. Em primeiro lugar, a fim de verificar a forma da função de risco para esses dados foi feita uma análise gráfica usando a curva TTT, descrito na Seção 2.1.2. A Figura 3.4 ilustra esse resultado.

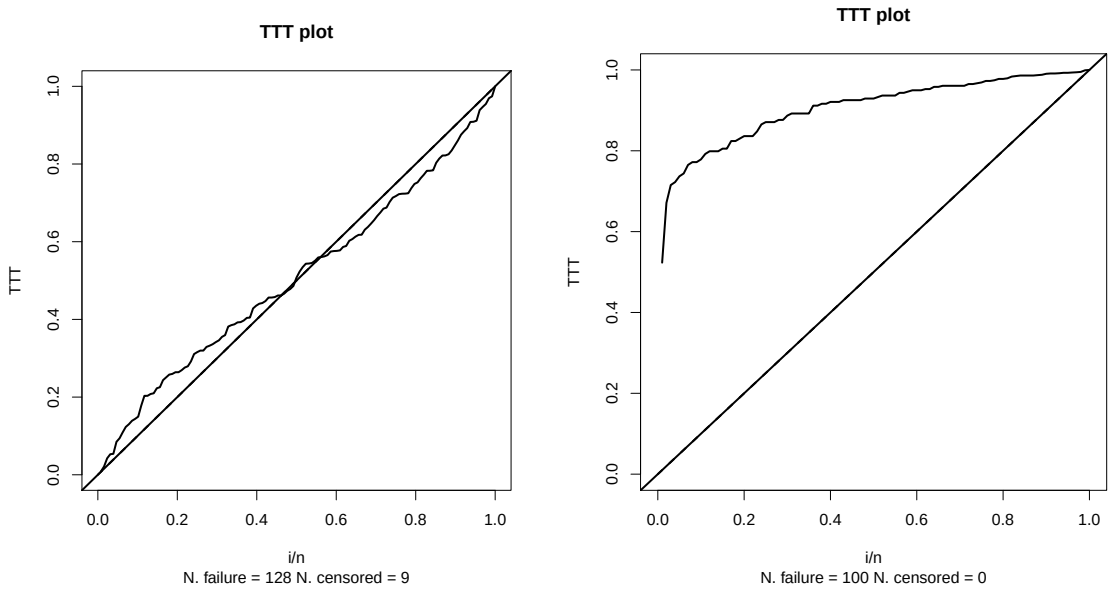

Figura 3.4: Painel esquerdo: Curva TTT dos dados reicidência de câncer; Painel direito: Curvas TTT dos dados de duração de equipamentos de alumínio.

De acordo com a Figura 3.4 pode-se notar que o formato da função de risco tem comportamento unimodal para os dados de câncer e para o segundo conjuntos de dados a função de risco é crescente. Logo, pode-se tentar utilizar a distribuição Weibull-Poisson para a modelagem dos dados pois o comportamento da sua taxa de falha também possui comportamento dentre outros unimodal e crescente.

Na Tabela 3.2 são apresentados as estimativas e o erro padrão das estimativas (entre parênteses) dos parâmetros do modelo WP e alguns casos particulares como são descritos na seção 3.5. Também foram usados os critérios de seleção AIC e BIC para identificar qual dos modelos é o mais apropriado em cada um dos casos. A Tabela 3.3 mostra os resultados desses critérios.

Tabela 3.2: Estimativas de máxima verossimilhança dos parâmetros das distribuições: exponencial-Poisson, Weibull e Weibull-Poisson dos dados

Dados	**Distribuições**	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\alpha}$
Câncer	exponencial-Poisson	0.10000	-	2.8926e-05
		(0.00884)	-	(0.05739)
	Weibull	0.09809	1.05354	-
		(0.00857)	(0.06813)	-
	Weibull-Poisson	0.03874	1.26282	3.93963
		(0.01433)	(0.08599)	(1.74958)
Equipamento	exponencial-Poisson	0.00748	-	0.00018
		(0.00074)	-	(0.00923)
	Weibull	0.00695	5.96088	-
		(0.00013)	(0.44859)	-
	Weibull-Poisson	0.00564	7.36313	5.89298
		(0.00042)	(0.51814)	(2.88154)

Tabela 3.3: Critérios de Seleção AIC e BIC para as distribuições: exponencial-Poisson, Weibull e Weibull-Poisson dos dados

Modelos	**Câncer**			**Equipamentos**		
	$\ell(\cdot)$	AIC	BIC	$\ell(\cdot)$	AIC	BIC
exponencial-Poisson	-422.7299	849.4598	855.2998	-595.4375	1194.8750	1200.1050
Weibull	-422.4147	848.8294	854.6694	-462.5549	929.1098	934.3400
Weibull-Poisson	-418.8944	843.7888	852.5487	-456.6857	919.3714	927.2168

Observar-se de acordo com a Tabela 3.3 que a distribuição Weibull-Poisson adquiriu o menor valor para os dois critérios de seleção, indicando que este modelo é mais adequado aos dados que as outras distribuições. Este resultado é confirmado pelos gráficos da Figura 3.5, pois verifica-se que a curvas da função de sobrevivência correspondente a Weibull-Poisson teve um comportamento mais similar à curva da função de sobrevivência estimado pelo método de Kaplan-Meier do que as outras distribuições.

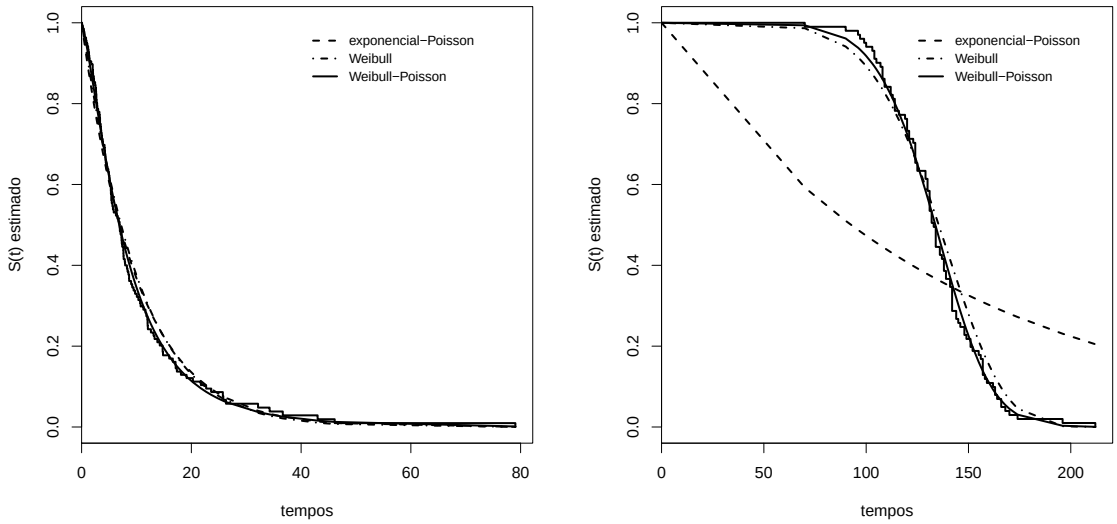

Figura 3.5: Curva da função de sobrevivência estimada pelo método de Kaplan-Meier e as curvas das funções de sobrevivência estimadas das distribuições: exponencial-Poisson, Weibull e Weibull-Poisson dos dados reicidência de câncer e dos equipamentos de alumínio.

Por fim, foi utilizado o teste da razão de verossimilhanças para selecionar os modelos (Weibull ou Weibull-Poisson) o que melhor se ajusta aos dados. O teste da razão de verossimilhanças no limite do espaço paramétrico foi utilizado pelo motivo de que esses modelos são encaixados, e a distribuição Weibull é o caso particular da distribuição Weibull-Poisson quando $\alpha \to 0$. O procedimento do teste encontra-se na seção 2.1.3.2. A hipótese nula estabelece que a distribuição Weibull é adequada, ou seja, $H_0 : \alpha \to 0$ e a hipótese alternativa é que a distribuição Weibull-Poisson é adequada, ou seja, $H_1 : \alpha > 0$. Para os dados de reincidência de câncer, o valor da estatística ω_n encontrado foi de 7.0406, que é maior do que $1/2 + 1/2\ P(\chi_1^2 \leq c) = 2.705543$, levando a uma forte evidência em favor da distribuição Weibull-Poisson com 5% de significância. O valor da estatística ω_n referente ao segundo conjunto de dados foi de 11.7384, que é maior do que $1/2 + 1/2$ $P(\chi_1^2 \leq c) = 2.705543$. Com isso há uma evidência em favor da distribuição Weibull-Poisson ao nível de significância de 5%.

3.2 Modelo Weibull-Poisson na presença de covariáveis

3.2.1 Introdução

É comum em situações práticas existirem características que podem influenciar no tempo de sobrevivência. Essas características são chamadas de covariáveis e devem ser incluídas na análise estatística dos dados. Um exemplo de covariável importante no estudo do tempo de sobrevivência de um determinado equipamento industrial, é o nível de voltagem a que o equipamento é submetido. Existem duas classes de modelos de regressão propostas na literatura: os modelos paramétricos e os semiparamétricos.

Os modelos paramétricos no qual se incluem os chamados modelos de tempo de vida acelerado são mais eficientes, porém menos flexíveis do que os modelos semiparamétricos. O segundo tipo de modelo também chamado de regressão de Cox, permite incorporar facilmente covariáveis dependentes do tempo, o que vem ocorrendo com bastante frequência em várias áreas de aplicação (Colosimo & Giolo, 2006). Muitos autores contribuíram para o desenvolvimento desses dois modelos, Kalbfleisch & Prentice (1980a), Cox & Oakes (1984), Lawless (2003), dentre outros. Neste trabalho o interesse está em um modelo de regressão paramétrico.

Para utilizar um modelo de regressão paramétrico deve-se associar uma distribuição de probabilidade ao tempo de sobrevivência. Pelo fato desta variável assumir valores em R_+ e apresentar uma distribuição assimétrica, não é recomendável usar a distribuição Normal para caracterizá-la. Neste trabalho foi utilizado o modelo de regressão Log Weibull-Poisson. O motivo de utilizá-lo é que este é um modelo de locação e escala que é descrito a seguir.

3.2.2 Modelo de locação e escala

No contexto de sobrevivência, uma forma de escrever um modelo de regressão paramétrico é usar distribuições que pertencem a família de locação e escala. Um modelo de locação e escala é descrito da forma:

$$Y = \mu + \sigma Z.$$

Considerando que Y pertença a família de distribuições que se caracteriza pelo fato de ter um parâmetro de locação μ $(-\infty < \mu < \infty)$ e um parâmetro de escala σ $(0 < \sigma < \infty)$. As distribuições que pertencem a essa família têm funções densidade de probabilidade e sobrevivência, dadas por:

$$f(y;\mu;\sigma) = \frac{1}{\sigma}\, g\left(\frac{y-\mu}{\sigma}\right)$$
$$S(y;\mu;\sigma) = G\left(\frac{y-\mu}{\sigma}\right)$$

Agora, considere que $Y = \log(T)$ e que o parâmetro de locação μ depende do vetor de covariáveis $\mathbf{X}$ e σ é o parâmetro de escala constante. Geralmente, o parâmetro de locação é escrito como $\mu(\underset{\sim}{x})=(\underset{\sim}{x}^T\boldsymbol{\beta})$, em que $\boldsymbol{\beta}=(\beta_0, ..., \beta_p)^T$ é o vetor de parâmetros desconhecidos. Nesse caso, um modelo de regressão que relaciona Y e o vetor de covariáveis $\mathbf{X}$ é o modelo de locação e escala que é descrito da forma:

$$Y = \mu(\mathbf{x}) + \sigma Z, \tag{3.22}$$

em que $Y = \log(T)$, $\mu(\underset{\sim}{x})=(\underset{\sim}{x}^T\boldsymbol{\beta})$ e Z é erro aleatório. Vale ressaltar que esse modelo é log linear para T, pois é um modelo de regressão linear para Y. Outra observação importante desse modelo é que o vetor de covariáveis $\mathbf{X}$ tem efeito multiplicativo em T, ou seja, $T = \exp(\mu(x))\exp(\sigma Z)$. Logo tem efeito linear em Y. Além disso, a função de sobrevivência para Y dado x tem a forma $G\left(\frac{y-\mu(x)}{\sigma}\right)$, em que $G(z)$ é a função de sobrevivência de Z. Mais detalhes em Lawless (2003).

3.2.3 Modelo de regressão Log Weibull-Poisson

Seja T uma variável com distribuição Weibull-Poisson, com função de densidade dada na equação 3.9 com a seguinte reparametrização: $\gamma = 1/\sigma$ e $\beta = \exp(-\mu)$. Nesse caso, por meio do método Jacobiano, a função de densidade $Y = \log(T)$ é dada por:

$$\begin{aligned} f(y,\alpha,\mu,\sigma) =& \frac{\alpha}{(\exp(\alpha)-1)\sigma} \exp\left\{\alpha \exp\left[-\left(\exp\left(\frac{y-\mu}{\sigma}\right)\right)\right] + \left(\frac{y-\mu}{\sigma}\right)\right\} \times \\ & \times \exp\left\{-\left[\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}, \end{aligned} \tag{3.23}$$

em que $-\infty < y < \infty$, $\alpha > 0$, $\sigma > 0$ e $-\infty < \mu < \infty$.

A equação 3.23 representa a função densidade da variável aleatória Y, que é chamada de distribuição Log Weibull-Poisson. A função de sobrevivência é dada por:

$$S(y,\alpha,\mu,\sigma)=\frac{\exp\left\{\alpha\exp\left[-\left(\exp\left(\frac{y-\mu}{\sigma}\right)\right)\right]\right\}-1}{\exp(\alpha)-1}. \tag{3.24}$$

O modelo de locação e escala dado na equação 3.22, considerando que Y dado $\mathbf{X}$ tem uma distribuição log Weibull Poisson, dado em 3.23 pode ser representado por:

$$y_i=(\underset{\sim}{x}^T\boldsymbol{\beta})+\sigma.z_i \quad i=1,2,....,n, \tag{3.25}$$

em que $\boldsymbol{\beta}$=$(\beta_0,...,\beta_p)^T$, $\sigma>0$ e $\alpha>0$ são os parâmetros desconhecidos, $x_i^T=(1,x_{i1},...x_{ip})$ é o vetor de covariáveis, e z_i que tem função densidade dada por:

$$f(z,\alpha)=\frac{\alpha\exp\left\{z+\alpha\exp\left[-\exp(z)\right]+\exp\left[-\exp\left(z\right)\right]\right\}}{\exp(\alpha)-1},\ -\infty<z<\infty;\alpha>0.$$

A função de sobrevivência de $Y|x$ é dado por:

$$S(y|x,\alpha,\sigma)=\frac{\exp\left\{\alpha\exp\left[-\left(\exp\left(\frac{y-x_i^T\beta}{\sigma}\right)\right)\right]\right\}-1}{\exp(\alpha)-1}. \tag{3.26}$$

3.2.4 Casos particulares do modelo de regressão Log Weibull-Poisson

O modelo de regressão Log Weibull-Poisson apresenta como casos particulares os modelos de regressão, que serão apresentados a seguir:

- Para $\alpha\rightarrow 0$ na equação 3.24 o modelo de regressão Log Weibull-Poisson se reduz ao modelo de regressão Log Weibull (ou de valor extremo) com função de sobrevivência da forma:

$$S(y)=\exp\left\{-\left[\exp\left(\frac{y-x^T\beta}{\sigma}\right)\right]\right\}$$

- Para $\sigma=1$ na equação 3.24 o modelo de regressão Log Weibull-Poisson se reduz ao modelo de regressão Log exponencial-Poisson com função de sobrevivência da forma:

$$S(y) = \frac{\exp\left\{\alpha \exp\left[-\exp\left(\frac{y-x^T\beta}{\sigma}\right)\right]\right\} - 1}{\exp(\alpha) - 1} \tag{3.27}$$

3.2.5 Estimação por máxima verossimilhança do modelo Log Weibull-Poisson

Dada uma amostra aleatória de tamanho n composta por dois vetores $\mathbf{t} = (t_1, ..., t_n)$ e $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$. C tempo de censura que é indepedente de T, em que $y_i = \min(\log(T_i), \log(C_i))$. δ é um vetor de variáveis aleatórias indicadoras de censura e $\underset{\sim}{x_i}$ o vetor de covariáveis associado ao i-ésimo elemento da amostra. Se a função densidade de $Y = \log(T)$, $f(.)$, é dada na equação 3.23 e a função de sobrevivência de Y, $S(.)$, é dada por 3.24, e supondo ainda que a censura é não informativa. O logaritmo da função de verossimilhança para o modelo de locação e escala na equação 3.22, é dado por:

$$\begin{aligned}\ell(\boldsymbol{\theta}) \propto & \sum_{i=1}^{n}(\delta_i)\ln(\alpha) - \sum_{i=1}^{n}(\delta_i)\exp\left(\frac{y_i - x_i^T\beta}{\sigma}\right) + \sum_{i=1}^{n}(\delta_i)\left[\alpha\exp\left\{-\exp\left(\frac{y_i - x_i^T\beta}{\sigma}\right)\right\} + \frac{y_i - x_i^T\beta}{\sigma}\right] \\ & - \sum_{i=1}^{n}(\delta_i)\ln\left[(\exp(\alpha) - 1)\,\sigma\right] - \sum_{i=1}^{n}(1-\delta_i)\ln\left[\exp\left\{\alpha\exp\left(-\exp(\frac{y_i - x_i^T\beta}{\sigma})\right)\right\} - 1\right] \\ & - \sum_{i=1}^{n}(1-\delta_i)\ln(\exp(\alpha) - 1).\end{aligned}$$

Para encontrar os estimadores de máxima verossimilhança deve-se resolver o sistema de equações $U(\boldsymbol{\theta}) = 0$ que é não linear, tornando-se necessário usar um algoritmo de otimização para o vetor de parâmetros $(\boldsymbol{\theta}) = (\alpha, \sigma, (\boldsymbol{\beta})^T)^T$. Dessa forma, as estimativas desses parâmetros foram obtidas por meio de maximização numérica do logaritmo da função de verossimilhança usando um processo iterativo. Neste trabalho, o algoritmo quase-Newton baseado no método BFGS foi usado. Intervalos de confiança e testes de hipótetes podem ser obtidos usando a distribuição para grandes amostras dos estimadores de máxima verossimilhança como descrito na seção 2.1.3. Assim, a aproximação normal assintótica para $\widehat{\boldsymbol{\theta}}$ pode ser expressa por $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N_{p+2}(0, L(\boldsymbol{\theta})^{-1})$. Para a distribuição Log Weibull-Poisson$(\boldsymbol{\theta})$, a matriz de informação observada é dada por:

$$L(\theta) = -\begin{bmatrix} L_{\alpha\alpha} & L_{\alpha\sigma} & L_{\alpha\beta_j} \\ . & L_{\sigma\sigma} & L_{\sigma\beta_j} \\ . & . & L_{\beta_j\beta_s} \end{bmatrix}$$

3.2.6 Estudo de Simulação

Para examinar o desempenho dos estimadores dos parâmetros do modelo de regressão Log Weibull-Poisson, foi feito um estudo de simulação do modelo de regressão $Y = \log(T) = \beta_0 + \beta_1 x_i + Z\sigma$ com a variável $Y = \log(T)$ logaritmo do tempo de sobrevivência T seguindo uma distribuição Log Weibull-Poisson, com função densidade dada na equação 3.22. Os valores dos coeficientes escolhidos foram $\beta_0 = 6$, $\beta_1 = 5$, em que os valores de x foram obtidos de uma amostra de tamanho n da distribuição $U(0, 1)$. Os valores dos parâmetros da Log Weibull-Poisson foram $\alpha = 3$ e $\sigma = 0.8$. Os tempos de censura C foram amostrados de uma distribuição Log exponencial-Poisson com os parâmetros $\beta_0 = 6$, $\beta_1 = 5$ e $\alpha = 3$. Os tempos de sobrevivência observados foram obtidos fazendo: $y_i = \min(\log(T_i, C_i))$.

Foram realizadas $B = 1000$ simulações e a partir destas amostras simuladas obteve-se as estimativas de máxima verossimilhança usando o recurso numérico *optim* que encontra-se no *Software R*. Através destas determinou-se as médias das estimativas de máxima verossimilhança, o viés e o erro quadrático médio (EQM) com diferentes tamanhos de amostra e 20% de censura. Os valores iniciais para o processo de otimização foram valores próximos dos verdadeiros valores dos parâmetros. A seguir é descrito o processo desta simulação:

1. Gerar $u_j \sim U(0, 1)$;
2. Determinar $y_j = F^{-1}(u_j) = \mu_j + \sigma \log\left\{-\log\left[\frac{\log(\exp(\alpha)(1-u_j)+u_j)}{\alpha}\right]\right\}$;
3. Gerar a variável de censura c_j; $c_j = F^{-1}(u_j) = \mu + \log\left\{-\log\left[\frac{\log(\exp(\alpha)(1-u)+u)}{\alpha}\right]\right\}$;
4. Encontrar o mínimo $y_j = \min(y_j, c_j)$;.
5. Se $y_j < c_j$, então $\delta_j = 1$, caso contrário, $\delta_j = 0$, para $j = 1, ..., n$.

Na Tabela 3.4 encontra-se as médias das estimativas, o viés e o erro quadrático médio das $B = 1000$ simulações com diferentes tamanhos de amostra com 20% de censura. Nota-se que as médias das estimativas se aproximam do verdadeiro valor do parâmetro quando aumenta-se o tamanho da amostra. Outro fator importante é que o erro quadrático médio diminui à medida que o tamanho amostral aumenta, ou seja, as estimativas dos parâmetros melhoram tornando-se os estimadores cada vez menos viesados.

Tabela 3.4: Resultados das simulações do modelo Log Weibull-Poisson com os parâmetros: $\beta_0 = 6$, $\beta_1 = 5$, $\alpha = 3$ e $\sigma = 0.8$ com 20% de censura

Tamanho da amostra (n)	Parâmetros	Média	Vício	EQM
40	β_0	5.58248	-0.41752	0.42791
	β_1	4.98904	-0.01096	0.38172
	α	1.50937	-1.49063	4.38738
	σ	0.80980	0.00980	0.01497
100	β_0	5.77158	-0.22842	0.28095
	β_1	5.00703	0.00703	0.15284
	α	2.29682	-0.70318	4.34622
	σ	0.81716	0.01716	0.00712
600	β_0	5.99460	-0.00540	0.12380
	β_1	4.99835	-0.00165	0.02493
	α	3.09801	0.09801	2.70200
	σ	0.81031	0.01031	0.00124
1000	β_0	6.00132	0.00132	0.09431
	β_1	5.00277	0.00277	0.01519
	α	3.09357	0.09357	1.97114
	σ	0.80825	0.00825	0.00070

3.2.7 Aplicação

O conjunto de dados foi extraído do Apêndice D da dissertação de Couto (2010), e refere-se ao estudo de isolamento elétrico considerando três níveis de voltagem: 52.5 Kv, 55.0 Kv e 57.5Kv. Foram testadas 20 amostras para cada um dos três níveis sendo que 10 % dos valores são dados censurados.

Inicialmente para identificar um modelo apropriado para os dados, estimou-se os parâmetros via máxima verossimilhança sem considerar a presença de covariáveis. Em seguida, ainda sem a presença das covariáveis, foi feita uma análise gráfica usando a curva TTT e o método de Kaplan-Meier como mostra a Figura 3.6. O painel esquerdo da Figura 3.6 mostra a curva TTT, o qual indica que a função de risco é unimodal. Logo, pode-se tentar utilizar a distribuição Weibull-Poisson para a modelagem dos dados pois o comportamento da sua função de risco também possui comportamento, dentre outros, unimodal. Este resultado é confirmado pelo painel direito da Figura 3.6, pois verifica-se que a curva da função de sobrevivência correspondente as distribuições Log Weibull-Poisson e Log exponencial-Poisson conseguiram obter um comportamento mais similar à curva da função de sobrevivência estudado por Kaplan-Meier do que a Log Weibull.

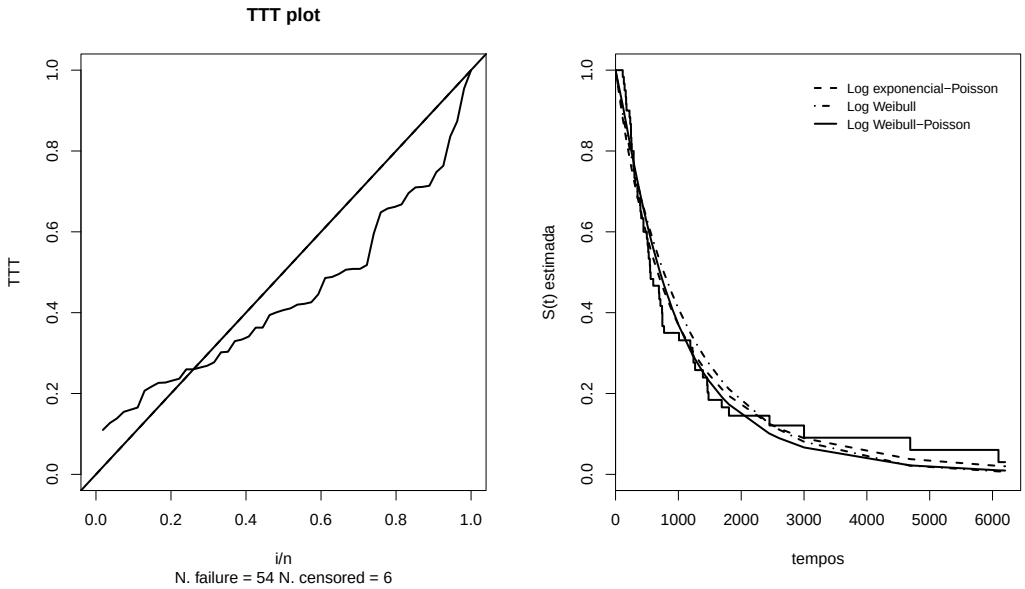

Figura 3.6: Painel esquerdo: Curva TTT; Painel direito: Curva da função de sobrevivência estimada pelo método de Kaplan-Meier e das funções de sobrevivência estimadas pelas distribuições Log exponencial-Poisson, Log Weibull e Log Weibull-Poisson.

A próxima etapa foi considerar as distribuições mais adequados a fim de realizar o ajuste considerando a presença de covariáveis. Dessa forma, foram escolhidas os modelos Log Weibull-Poisson e Log exponencial-Poisson para a modelagem com covariáveis. Neste caso o modelo em estudo é expresso da forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \sigma Z_i, i = 1, ...60,$$

em que Z_i é o erro aleatório independente e identicamente distribuído e a resposta y_i denota o logaritmo do tempo de sobrevivência. A Tabela 3.5 mostra as estimativas de máxima verossimilhança e os respectivos erros padrão para os parâmetros dos dois modelos com os valores dos critérios de seleção.

Tabela 3.5: Estimativas dos parâmetros dos modelos de regressão exponencial-Poisson e Weibull-Poisson.

	Log exponencial-Poisson			**Log Weibull-Poisson**		
Parâmetros	Estimativas	E.P.	p-valor	Estimativas	E.P.	p-valor
α	0.01570	0.15392	-	3.36176	1.76967	-
σ	-	-	-	0.72043	0.07757	-
β_0	21.13660	3.57326	<0.0001	20.87550	3.10312	<0.0001
β_1	-0.25887	0.06498	<0.0001	-0.24033	0.05628	<0.0001
$-\ell(\cdot)$	84.84902			81.58688		
AIC	175.6980			171.1738		
BIC	181.9811			179.5511		

E.P.=erro padrão

Pode-se observar de acordo com a Tabela 3.5 que o modelo Log Weibull-Poisson obteve o menor valor para os dois critérios de seleção em relação ao modelo Log exponencial-Poisson, indicando que este modelo se ajusta melhor aos dados. Os resultados das estimativas dos parâmetros e os respectivos erros padrão dos dois modelos são semelhantes, logo, a variável nível de voltagem (x_{i1}) é significativa em ambos modelos ao nível de 5%. O próximo passo após as conclusões feitas anteriormente sobre os modelos foi a análise de resíduos, que é útil para verificar a adequabilidade do modelo. A Figura 3.7 apresenta o gráfico de resíduo Cox-Snell para os modelos Log Weibull-Poisson e Log exponencial-Poisson, respectivamente:

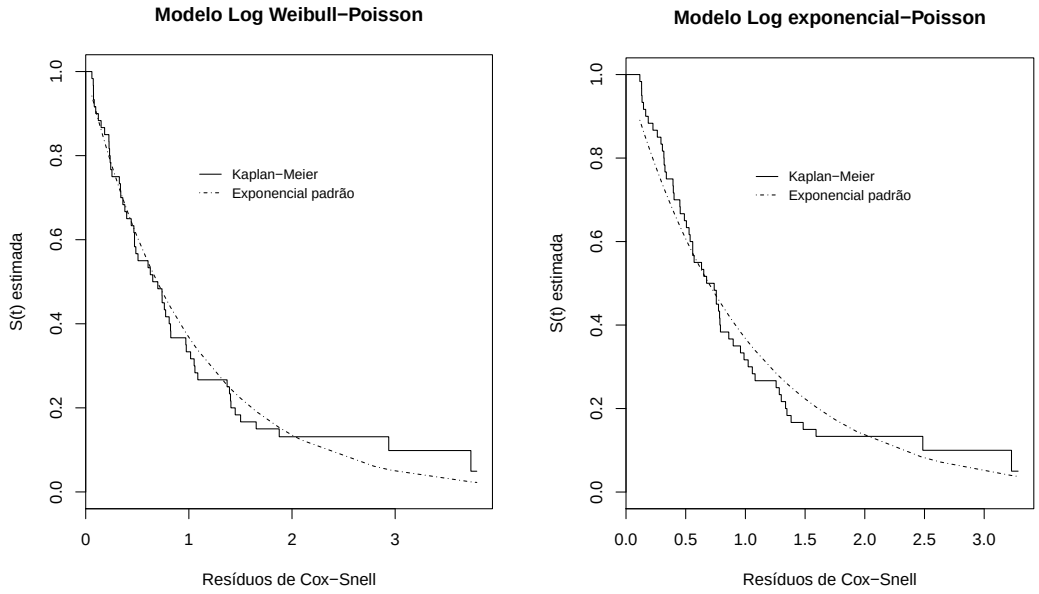

Figura 3.7: Painel esquerdo: Curva da função de sobrevivência do resíduo estimada pelo método de Kaplan-Meier e a curva de sobrevivência estimada da exponencial padrão das distribuições Log exponencial-Poisson e Log Weibull-Poisson.

Da Figura 3.7 em relação ao resíduo de Cox-Snell, nota-se que para o modelo de regressão Log Weibull-Poisson, a curva exponencial está mais próxima da curva de sobrevivência estimada em relação ao modelo de regressão Log exponencial-Poisson, dando indícios que o modelo se ajusta bem aos dados. Logo, a partir das considerações citadas, o modelo de regressão Log Weibull-Poisson final é descrito da forma:

$$y = \log(t) = 20.87550 - 0.24033X_1$$

Como a variável foi transformada para a logaritmica, a interpretação não é de uma forma direta como é feita na regressão linear, logo, uma possível proposta é baseada na razão de tempos medianos (Hosmer & Lemeshow, 1999). De acordo com o modelo final, a interpretação é que o tempo mediano de sobrevivência estimado deve diminuir aproximadamente em 27.16% ([exp(0.24033) x 100%]) quando a variável X_1 (nível de voltagem) aumenta em uma unidade.

3.3 Considerações finais

Neste capítulo foi apresentada a distribuição Weibull-Poisson (WP) proposta por Louzada *et al.* (2011a) a qual generaliza a distribuição exponencial-Poisson proposta por Kus (2007) e a Weibull. Também foram discutidos suas propriedades e mostradas as diversas formas que a função de risco pode ter. Um estudo de simulação foi realizado para verificar o comportamento das estimativas por máxima verossimilhança desta distribuição. Também foi feita uma análise desta distribuição com a inclusão de covariáveis. Por fim, foram feitas aplicações a conjuntos de dados reais para verificar a adequabilidade deste modelo e através das análises da curva TTT, dos valores observados dos critérios AIC e BIC, do Kaplan-Meier e do teste TRV, pode-se notar que a distribuição WP ajustou-se bem aos dados. Em relação a inclusão de covariáveis no modelo foram feitas as análises de resíduos e a distribuição também ajustou-se bem aos dados. A distribuição WP foi formulada na situação em que nem sempre é possível observar o valor da variável tempo de sobrevivência e sim o valor mínimo dos tempos, mas pode ser usada em qualquer outra ocasião desde que se adeque aos dados. Logo se espera que esta distribuição seja útil em outro conjunto de dados.

Capítulo 4

Modelo de longa duração Weibull-Poisson (LWP)

Nesta seção apresenta-se a distribuição de longa duração Weibull-Poisson (LWP). Discute-se no início da seção o modelo de mistura padrão e como ele é construído, a origem da LWP, suas propriedades e as diversas formas de sua função risco. Posteriormente são deduzidas a função geradora de momentos e a função densidade da k-ésima estatística de ordem desta distribuição. Um estudo de simulação foi realizado para verificar o comportamento dos estimadores de seus parâmetros via máxima verossimilhança, em relação ao vício e ao erro quadrático médio, para diferentes tamanhos de amostras. Além disso, essa distribuição foi analisada na situação em que as covariáveis são inclusas no estudo. Aplicações aos conjuntos de dados reais ilustram a aplicabilidade desse modelo.

4.1 Modelo de longa duração

4.1.1 Introdução

O interesse em estudos de sobrevivência pode estar centrado na distribuição do tempo até a morte de um paciente, a ocorrência de uma doença, ou até o tempo de vida de um componente eletrônico, dentre outros. Para alguns elementos da amostra, o evento de interesse nunca ocorre, dando evidências de que uma parcela da população estudada não está sujeita ao evento de interesse especificado, que constitui a parcela de indivíduos considerados como imunes (ou curados) ao evento. Essa característica aparece em estudos em que os dados tem uma grande quantidade de observações censuradas. Nesta situação, recomenda-se os modelos de sobrevivência de longa duração.

Existem situações em que a classificação individual de curados ou não curados numa amostra não é possível, entretanto, se o gráfico do estimador da função de sobrevivência Kaplan-Meier apresentar uma cauda longa num nível visivelmente acima de zero, temos uma forte indicação de que as observações censuradas determinam uma fração de cura maior do que zero. O que por sua vez, caracteriza a presença de indivíduos curados (imunes) na população. Ou seja, a função de sobrevivência Kaplan-Meier converge para um $p > 0$ quando $t \to \infty$, como ilustrado na Figura 4.1:

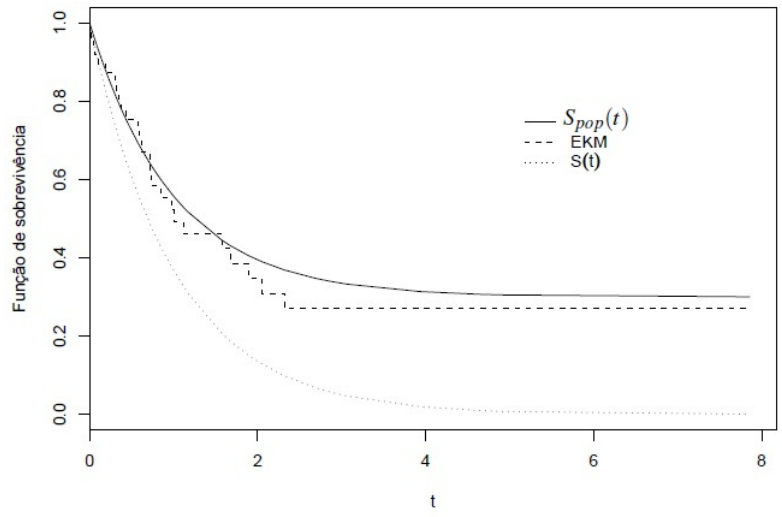

Figura 4.1: Gráfico de uma função de sobrevivência própria $S(t)$ e uma função de sobrevivência imprópria $S_{pop(t)}$

Muitos autores contribuíram para a teoria dos modelos de mistura de longa duração, dentre eles destaca-se Berkson & Gage (1952) que propuseram um modelo de mistura com o objetivo de estimar a proporção de curados numa população submetida a um tratamento de câncer de estômago. Alternativamente, existem modelos baseados em estrutura de riscos competitivos, nesse caso, é definida uma variável aleatória como o número de causas competindo para ocorrência do evento de interesse, que por sua vez, é associada a variável resposta do modelo. Por exemplo, em estudos de câncer, as células cancerígenas competem entre si para dar origem ao tumor visível e isto é relacionado a uma variável que marca o tempo até a origem do tumor. Assim, a variável resposta é definida como o menor dentre estes tempos. Rodrigues *et al.* (2009) propuseram um modelo de teoria unificada que engloba o modelo de mistura padrão assim como outros modelos propostos na literatura. Este capítulo está organizado da seguinte forma: na seção 4.2 é apresentado o modelo de mistura padrão de Berkson & Gage (1952). Na seção

4.3 é apresentada a distribuição Weibull-Poisson no contexto de longa duração, desde a formulação até a parte inferencial do modelo. Os resultados de estudo da simulação são apresentados, mostrando o comportamento dos estimadores dos parâmetros quanto ao vício e ao erro quadrático médio, para diferentes tamanhos de amostras da distribuição. Na seção 4.6 estuda-se a presença de covariáveis na proporção de curados. Por fim, é mostrada uma aplicação a conjunto de dados reais.

4.2 Modelo de mistura padrão

O modelo de mistura padrão proposto por Berkson & Gage (1952) é um dos mais conhecidos na análise de sobrevivência para ajustar dados de longa duração. Este consiste em uma mistura de duas distribuições paramétricas, sendo uma função de sobrevivência imprópria considerada para a população total (curados e não curados) e uma função de sobrevivência própria para a parte da população formada pelos não curados. O modelo de mistura padrão é derivado considerando uma variável de Bernoulli indicadora não observável M_i, que representa se o indivíduo se encontra ou não em risco. Ou seja:

$$M_i = \begin{cases} 1 & \text{se o i-ésimo indivíduo não está em risco;} \\ 0 & \text{se o i-ésimo está em risco.} \end{cases}$$

com $P(M_i = 0) = p$ e $P(M_i = 1) = (1 - p)$; $i = 1, 2, ..., n$.

A função de sobrevivência do tempo de vida T na sub-população de indivíduo não curado é indicada por $S_T(t)$, uma função de sobrevivência própria. No entanto, na sub-população de indivíduo curado a função de sobrevivência de T é imprópria, já que o seu tempo de vida é infinito. Desta forma, a função de sobrevivência da variável aleatória não negativa e contínua T, representando o tempo de vida de um indivíduo condicional ao valor de M é:

$$P(T > t|M_i = 1) = S_T(t) \quad \text{e} \quad P(T > t|M_i = 0) = 1$$

A probabilidade de o tempo de vida ser maior que um determinado tempo t, independente do grupo a que ele pertença é dada por:

$$\begin{aligned} S_{pop}(t) &= P(T > t) \\ &= P(T > t|M_i = 0).P(M_i = 0) + P(T > t|M_i = 1).P(M_i = 1) \end{aligned} \tag{4.1}$$

Como $P(T > t|M_i = 1) = S_T(t)$; $P(T > t|M_i = 0) = 1$; $P(M_i = 0) = p$ e $P(M_i = 1) = (1 - p)$, a função $S_{pop}(t)$ pode ser escrita por

$$S_{pop}(t) = p + (1 - p)S_T(t) \tag{4.2}$$

A equação 4.2 satisfaz as seguintes propriedades:

1. Se $p = 0$, $S_{pop}(t) = S_T(t)$;

2. $S_{pop}(t)$ é decrescente;

3. $\lim_{t\to\infty} S_{pop}(t) = p$.

A propriedade 3 retrata o fato de que a função de sobrevivência populacional é imprópria, pois a curva de sobrevivência se estabiliza em p, justamente a proporção de curados da população.

Existem relações entre as funções de sobrevivência imprópria ($S_{pop}(t)$) com as funções de densidade e de taxa de falha. A função densidade de probabilidade imprópria também é definida como a derivada da função densidade acumulada imprópria, ou seja:

$$f_{pop}(t) = \frac{\partial F_{pop}(t)}{\partial t},$$

como $F_{pop}(t) = 1 - S_{pop}(t)$, então:

$$f_{pop}(t) = \frac{\partial[1 - S_{pop}((t)]}{\partial t} = -\partial[S_{pop}((t)]/\partial(t) = (1 - p)f(t), \tag{4.3}$$

em que $f(.)$ representa a função de densidade própria relativa ao grupo dos indivíduos em risco. A função de risco da população total é dada por:

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} \tag{4.4}$$

4.3 Modelo de longa duração Weibull-Poisson (LWP)

4.3.1 Formulação do modelo

Considere uma variável aleatória não negativa T que representa o tempo de vida com distribuição WP(α, β, γ). Também considere uma população na qual se admite indivíduos doentes (suscetíveis) com probabilidade $1-p$ e indivíduos curados (não suscetíveis) com probabilidade p. O modelo é caracterizado pela função de sobrevivência imprópria:

$$S_{pop}(t) = p + (1-p)S_T(t). \tag{4.5}$$

A função $S_T(t)$ pode ser especificada por funções de sobrevivências usais como Weibull, exponencial, entre outras.

Neste trabalho utilizamos a função de sobrevivência da distribuição Weibull-Poisson (3.11) dada por:

$$S_T(t) = \frac{\exp\{\alpha \exp[-(\beta t)^\gamma]\} - 1}{\exp(\alpha) - 1};\ t > 0,\ \beta > 0,\ \gamma > 0,\ \alpha > 0.$$

Então, a equação 4.5 pode ser escrita como:

$$\begin{aligned} S_{pop}(t) &= p + \frac{(1-p)(\exp\{\alpha \exp[-(\beta t)^\gamma]\} - 1)}{\exp(\alpha) - 1} \\ &= \frac{p\exp(\alpha) - p + (1-p)(\exp\{\alpha \exp[-(\beta t)^\gamma]\} - 1)}{\exp(\alpha) - 1} \\ &= \frac{p\exp(\alpha) - p + (\exp\{\alpha \exp[-(\beta t)^\gamma]\}) - 1 - p(\exp\{\alpha \exp[-(\beta t)^\gamma]\}) + p}{\exp(\alpha) - 1}. \end{aligned}$$

Logo:

$$S_{pop}(t) = \begin{cases} 1, & t < 0 \\ \dfrac{p\exp(\alpha) - p + (\exp\{\alpha \exp[-(\beta t)^\gamma]\}) - 1 - p(\exp\{\alpha \exp[-(\beta t)^\gamma]\}) + p}{\exp(\alpha) - 1}, & t \geq 0 \end{cases} \tag{4.6}$$

A função de sobrevivência imprópria (4.6) corresponde a função de sobrevivência da distribuição de longa duração padrão Weibull-Poisson LWP$(\alpha, \beta, \gamma, p)$ para $\alpha > 0$, $\beta > 0$, $\gamma > 0$ e $0 < p < 1$. A Figura 4.2 mostra a função de sobrevivência imprópria para alguns valores dos parâmetros da LWP:

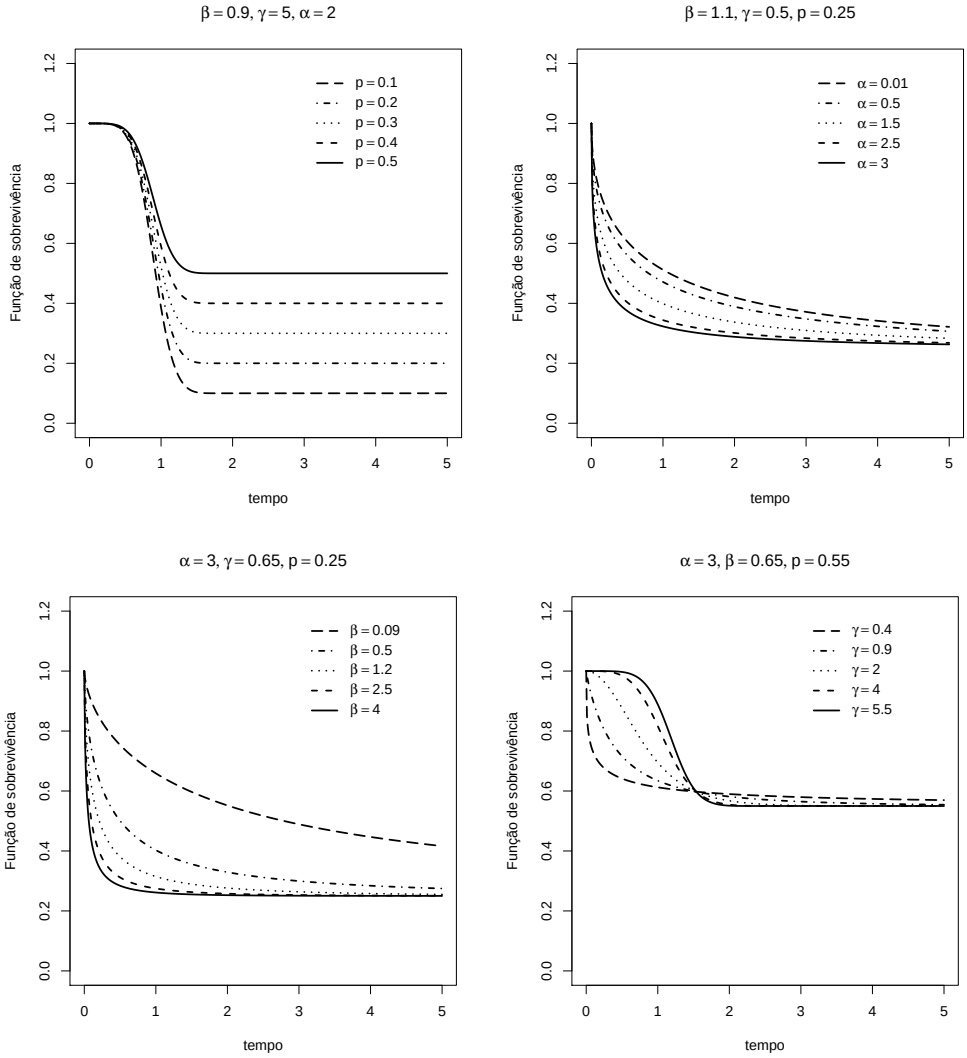

Figura 4.2: Gráficos ilustrativos da função de sobrevivência imprópia da LWP.

A função densidade de probabilidade populacional é obtida diretamente considerando que $f_{pop}(t) = -dS_{pop}(t)/dt$ é dada por:

$$f_{pop}(t) = \frac{(1-p)\alpha \exp\left\{\alpha \exp\left[-(\beta t)^{\gamma}\right] - (\beta t)^{\gamma}\right\} \beta^{\gamma} t^{\gamma-1} \gamma}{\exp(\alpha) - 1}. \tag{4.7}$$

A Figura 4.3 mostra a função de densidade para alguns valores dos parâmetros da LWP:

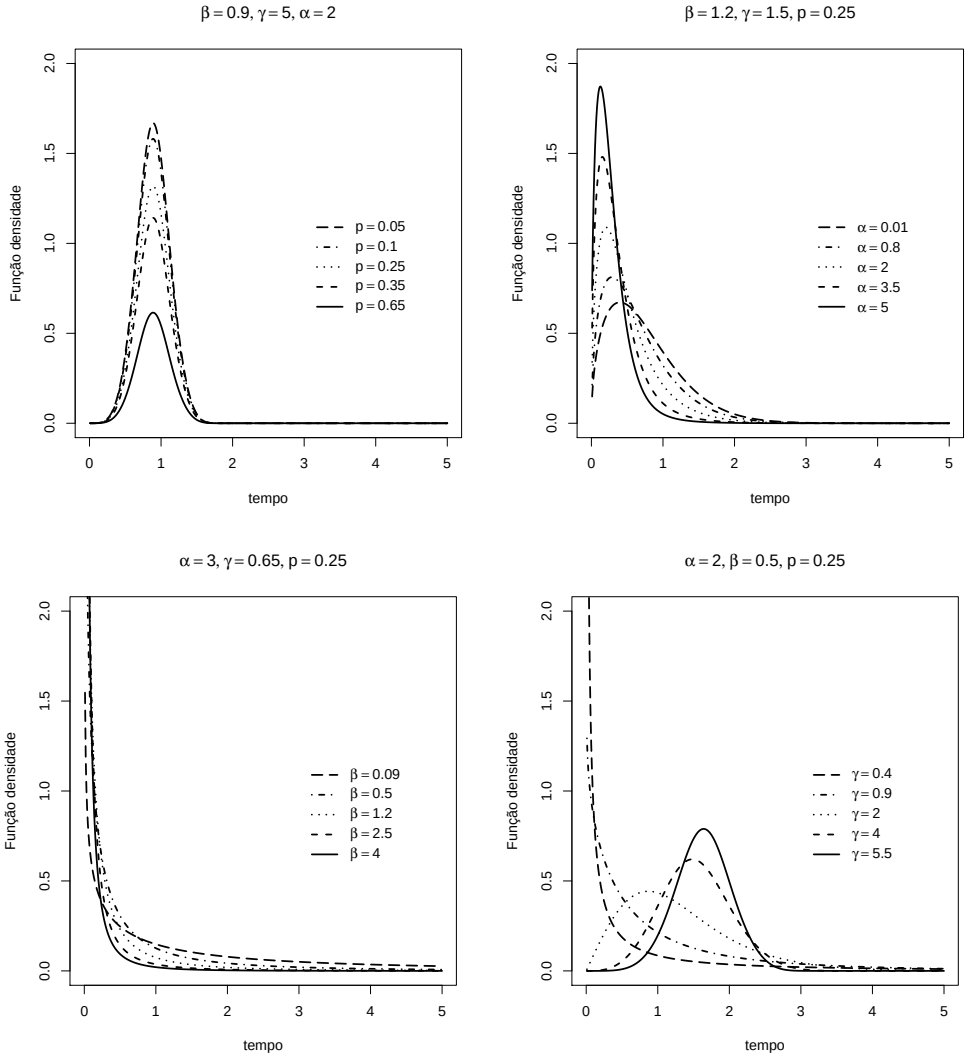

Figura 4.3: Gráficos ilustrativos da Função de densidade imprópia da LWP.

4.3.2 Propriedades da LWP

Nesta seção, apresentamos a Função de distribuição acumulada, a função de risco, a função quantil, os momentos e a distribuição da k-ésima estatística de ordem da LWP. Seja T uma variável aleatória com distribuição de longa duração Weibull-Poisson $(\alpha, \beta, \gamma, p)$. A função distribuição acumulada $F_{pop}(t)$ e o quantil desta função $Q_{pop}(t)$ para $t > 0$, $\alpha > 0$, $\beta > 0$, $\gamma > 0$ e $0 < p < 1$, são respectivamente:

$$F_{pop}(t) = \frac{(\exp\{\alpha \exp[-(\beta t)^{\gamma}]\} - \exp(\alpha))(-1+p)}{\exp(\alpha) - 1},$$

$$Q_{pop}(t) = \frac{1}{\beta}\left[\log(\alpha) - \log\left(\log(\exp(\alpha)(1-p-u)+u) - \log(1-p)\right)\right]^{1/\gamma}.$$

A função de risco populacional da LWP, obtida através da relação $h_{pop}(t) = f_{pop}(t)/S_{pop}(t)$, é dada por:

$$h_{pop}(t) = \frac{(1-p)\alpha\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right] - (\beta t)^{\gamma}\right\}\beta^{\gamma}t^{\gamma-1}\gamma}{p\exp(\alpha) - p + \left(\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]\right\}\right) - 1 - p\left(\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]\right\}\right) + p}.$$

O comportamento de $h_{pop}(t)$ para alguns valores de α, β, γ e p é apresentado na Figura 4.4:

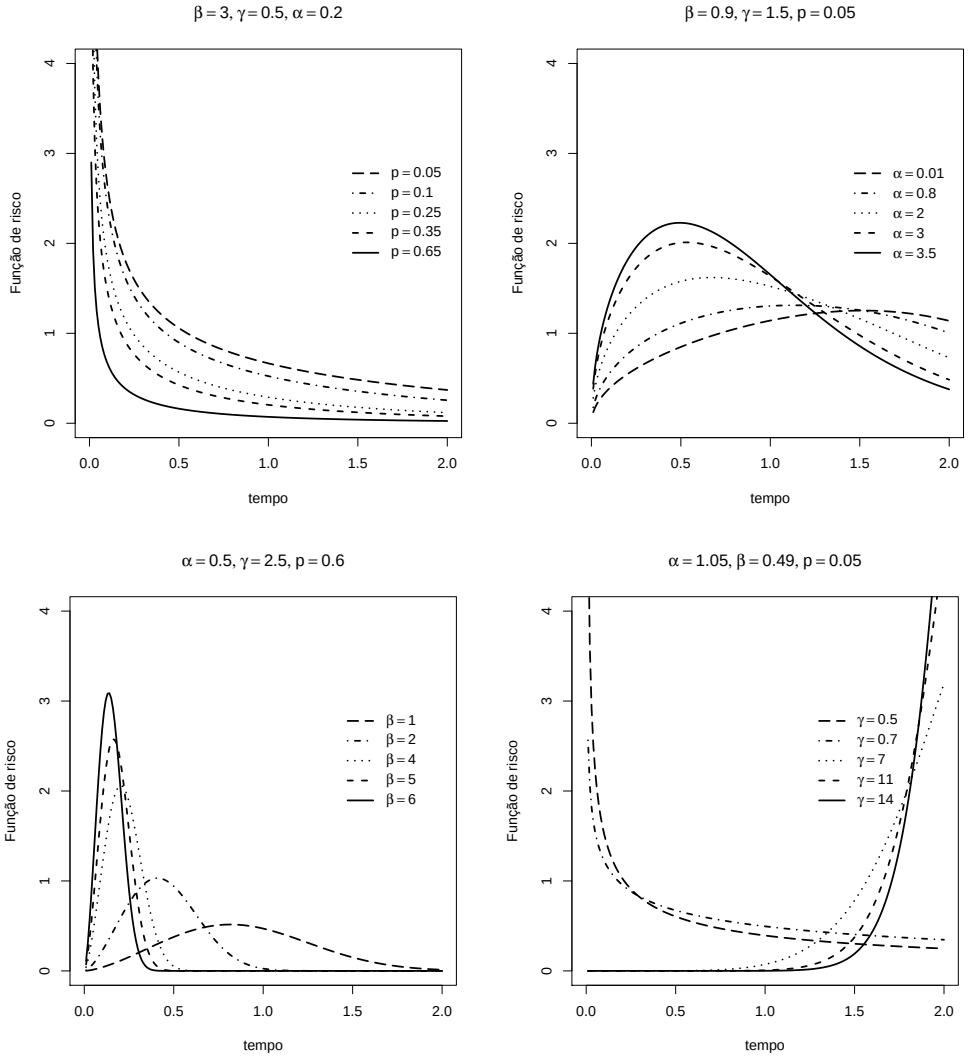

Figura 4.4: Gráficos ilustrativos da Função de risco imprópia da LWP.

Pode ser visto através da Figura 4.4 que a função de risco da LWP é bastante flexível e pode ter várias formas, como crescente, decrescente e unimodal. O momento central de ordem r da distribuição LWP é dado por:

$$E(T^r) = (1-p)\frac{1}{\exp(\alpha)-1}\sum_{j=1}^{\infty}\frac{\alpha^j}{j!}(\beta j^{1/\gamma})^{-r}\Gamma(1+\frac{r}{\gamma})$$

Prova:

$$E(T^r) = \int_0^{\infty} t^r f_{pop}(t)dt$$

$$E(T^r) = \int_0^{\infty} \frac{t^r(1-p)\alpha\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\beta^{\gamma}t^{\gamma-1}\gamma}{\exp(\alpha)-1}dt$$

$$E(T^r) = (1-p)\int_0^{\infty} \frac{t^r\alpha\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\beta^{\gamma}t^{\gamma-1}\gamma}{\exp(\alpha)-1}dt. \tag{4.8}$$

Lembrar que:

$$\mu(r) = \int_0^{\infty} \frac{t^r\alpha\exp\left\{\alpha\exp\left[-(\beta t)^{\gamma}\right]-(\beta t)^{\gamma}\right\}\beta^{\gamma}t^{\gamma-1}\gamma}{\exp(\alpha)-1}dt. \tag{4.9}$$

$\mu(r)$ é o momento de ordem r da distribuição Weibull-Poisson (α, β, γ), tal que:

$$\mu(r) = \frac{\gamma}{\exp(\alpha)-1}\Gamma\left(1+\frac{r}{\gamma}\right)\beta^{-r}\sum_{j=1}^{\infty}j^{-r/\gamma}\frac{\alpha^j}{j!}. \tag{4.10}$$

Utilizando a equação 4.9 na equação 4.7 temos:

$$E(T^r) = (1-p)\frac{1}{\exp(\alpha)-1}\sum_{j=1}^{\infty}\frac{\alpha^j}{j!}(\beta j^{1/\gamma})^{-r}\Gamma(1+\frac{r}{\gamma}).$$

■

A função densidade de probabilidade da k-ésima estatística de ordem da LWP é dada por:

$$f_{k:n}(t) = g_{k:n}(t)(1-p)^k \left(\frac{p\exp(\alpha) + (\exp\{\alpha\exp\{-(\beta t)^\gamma\}\})(1-p) - 1}{(\exp\{\alpha\exp\{-(\beta t)^\gamma\}\}) - 1} \right)^{n-k},$$

em que $g_{k:n}(t)$ é a função densidade de probabilidade da k-ésima estatística da distribuição WP dada na equação 3.13.

A prova desse resultado parte primeiramente da seguinte definição:

$$f_{k:n}(t) = \frac{1}{B(k, n-k+1)} f(t)(F(t))^{k-1}(S(t))^{n-k},$$

em que

$$B(k, n-k+1) = \frac{(n-k)!(k-1)!}{n!}.$$

Usando a definição, temos:

$$\begin{aligned}
f_{k:n}(t) &= \frac{1}{B(k,n-k+1)} f_{pop}(t)(F_{pop}(t))^{k-1}(S_{pop}(t))^{n-k} \\
&= \frac{1}{B(k,n-k+1)} \frac{(1-p)\alpha\exp\{\alpha\exp[-(\beta t)^\gamma] - (\beta t)^\gamma\}\beta^\gamma t^{\gamma-1}\gamma}{\exp(\alpha)-1} \times \\
&\left[\frac{(\exp\{\alpha\exp[-(\beta t)^\gamma]\} - \exp(\alpha))(-1+p)}{\exp(\alpha)-1}\right]^{k-1} \times \\
&\left[\frac{p\exp(\alpha) - p + (\exp\{\alpha\exp[-(\beta t)^\gamma]\}) - 1 - p(\exp\{\alpha\exp[-(\beta t)^\gamma]\}) + p}{\exp(\alpha)-1}\right]^{n-k} \\
&= \frac{\alpha\exp\{\alpha\exp[-(\beta t)^\gamma] - (\beta t)^\gamma\}\beta^\gamma t^{\gamma-1}\gamma(1-p)^k}{B(k,n-k+1)(\exp(\alpha)-1)^n} (\exp\{\alpha\exp\{-(\beta t)^\gamma\}\} - \exp(\alpha))^{k-1} \times \\
&\times [p\exp(\alpha) + (\exp\{\alpha\exp[-(\beta t)^\gamma]\})(1-p) - 1]^{n-k} \\
&= \alpha\exp\{\alpha\exp[-(\beta t)^\gamma] - (\beta t)^\gamma\}\beta^\gamma t^{\gamma-1}\gamma[\exp(\alpha) - \exp\{\alpha\exp[-(\beta t)^\gamma]\}]^{k-1} \times \\
&\frac{[\exp\{\alpha\exp[-(\beta t)^\gamma]\} - 1]^{n-k}}{B(k,n-k+1)[\exp(\alpha)-1]^n}(1-p)^k \left(\frac{p\exp(\alpha) + (\exp\{\alpha\exp\{-(\beta t)^\gamma\}\})(1-p) - 1}{(\exp\{\alpha\exp\{-(\beta t)^\gamma\}\}) - 1} \right)^{n-k}.
\end{aligned}$$

Mas a estatística de ordem da distribuição WP é dada por:

$$\begin{aligned}
g_{k:n}(t) =& \alpha\exp\{\alpha\exp[-(\beta t)^\gamma] - (\beta t)^\gamma\}\beta^\gamma t^{\gamma-1}\gamma[\exp(\alpha) - \exp\{\alpha\exp[-(\beta t)^\gamma]\}]^{k-1} \times \\
&\times \frac{[\exp\{\alpha\exp[-(\beta t)^\gamma]\} - 1]^{n-k}}{B(k,n-k+1)[\exp(\alpha)-1]^n}.
\end{aligned}$$

Então,

$$f_{k:n}(t) = g_{k:n}(t)(1-p)^k \left(\frac{p\exp(\alpha) + (\exp\{\alpha\exp\{-(\beta t)^\gamma\}\})(1-p) - 1}{(\exp\{\alpha\exp\{-(\beta t)^\gamma\}\}) - 1} \right)^{n-k}$$

■

4.3.3 Casos particulares da distribuição LWP

A distribuição LWP apresenta algumas distribuições como casos particulares que serão apresentadas a seguir:

- Para $\alpha \to 0$ na equação 4.6, a distribuição LWP se reduz a uma distribuição de longa duração Weibull (LW) com função de sobrevivência imprópria da forma:

$$S_{pop}(t) = p + (1-p)\exp\{-(\beta t)^\gamma\}$$

- Para $\gamma = 1$ na equação 4.6, a distribuição LWP se reduz a uma distribuição de longa duração exponencial-Poisson (LEP) com função de sobrevivência imprópria da forma:

$$S_{pop}(t) = p + (1-p)\frac{\exp\{\alpha\exp[-(\beta t)]\} - 1}{\exp(\alpha) - 1}$$

- Para $p = 0$ na equação 4.6, a distribuição LWP se reduz a uma distribuição WP com função de sobrevivência da forma:

$$S_{pop}(t) = \frac{\exp\{\alpha\exp[-(\beta t)^\gamma]\} - 1}{\exp(\alpha) - 1}$$

4.3.4 Estimação por máxima verossimilhança da LWP

Dada uma amostra aleatória de tamanho n composta por dois vetores $\mathbf{t} = (t_1, ..., t_n)$ e $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$. C tempo de censura que é indepedente de T, em que $t_i = \min(T_i, C_i)$. Sendo que para cada $i = 1, ..., n$, a variável T_i tem distribuição LWP($\mathbf{v}$) com vetor de parâmetros $\mathbf{v} = (\alpha, \beta, \gamma, p)$ e δ é um vetor de variáveis aleatórias indicadoras de censura. Considerando $f_{pop}(.)$ a função densidade dado na equação 4.7 e $S_{pop}(.)$ dado na equação 4.6. O logaritmo da função de verossimilhança $l(\mathbf{v})$ para o vetor de parâmetros $\mathbf{v} = (\alpha, \beta, \gamma, p)$ considerando que os tempos de sobrevivência e de censura são independentes e que a censura é não informativa, pode ser escrito como:

$$\begin{aligned}\ell(v) \propto & \sum_{i=1}^{n} \delta_i \log(\alpha\gamma\beta^{\gamma} t_i^{\gamma-1}(1-p)) + \sum_{i=1}^{n} \delta_i \left[\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\} - (\beta t_i)^{\gamma}\right] \\ & - \sum_{i=1}^{n} \delta_i \log(\exp(\alpha) - 1) + \sum_{i=1}^{n} (1-\delta_i) \log\left(\exp(\alpha)\left(\exp\left\{\alpha \exp\left[-(\beta t)^{\gamma}\right]\right\}\right)(1-p) - 1\right) \\ & - \sum_{i=1}^{n} (1-\delta_i) \log(\exp(\alpha) - 1),\end{aligned}$$

As estimativas de máxima verossimilhança de $\widehat{v} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{p})$ podem ser obtidas através da resolução do sistema de equações:

$$U(\mathbf{v}) = \frac{\partial l(\mathbf{v})}{\partial\ \mathbf{v}} = 0. \tag{4.11}$$

Os componentes do vetor escore $U(\mathbf{v}) = U(\alpha, \beta, \gamma, p) = \left(\frac{\partial \ell(\mathbf{v})}{\partial \alpha}; \frac{\partial \ell(\mathbf{v})}{\partial \beta}; \frac{\partial \ell(\mathbf{v})}{\partial \gamma}; \frac{\partial \ell(\mathbf{v})}{\partial p}\right)$ são dados por:

$$\begin{aligned}\frac{\partial \ell(\mathbf{v})}{\partial \alpha} = & \frac{\sum_{i=1}^{n} \delta_i}{\alpha} + \sum_{i=1}^{n} \delta_i \left(\exp\left\{\alpha \exp\left[-(\beta t_i)^{\gamma}\right]\right\}\right) - \frac{\sum_{i=1}^{n} \delta_i \exp(\alpha)}{\sum_{i=1}^{n} (1-\delta_i) \exp(\alpha) - 1} + \\ & + \frac{\exp\left\{\alpha \exp\left[-(\beta t_i)^{\gamma}\right] - (\beta t_i)^{\gamma}\right\} \beta^{\gamma} t_i^{\gamma-1} (1-p) p \exp(\alpha) + \exp\left\{-(\beta t_i)^{\gamma}\right\} (1-p) p \exp(\alpha)}{\exp\left\{\alpha \exp\left[-(\beta t_i)^{\gamma}\right]\right\} (1-p) p (\exp(\alpha) - 1)} + \\ & + \frac{\sum_{i=1}^{n} (1-\delta_i)}{\exp(\alpha) - 1}.\end{aligned}$$

$$\frac{\partial \ell(\mathbf{v})}{\partial \beta} = \frac{\sum_{i=1}^{n} \delta_i \gamma}{\beta} + \sum_{i=1}^{n} \delta_i \left(\frac{-\alpha(\beta t_i)^{\gamma}\gamma \exp\left\{-(\beta t_i)^{\gamma}\right\}}{\beta} - \frac{(\beta t_i)^{\gamma}\gamma}{\beta} \right)$$
$$\frac{-\sum_{i=1}^{n}(1-\delta_i)\alpha(\beta t_i)^{\gamma}\gamma \exp\left\{-(\beta t_i)^{\gamma}\right\} \exp\left\{\alpha \exp\left[-(\beta t_i)^{\gamma}\right]\right\}(1-p)p\exp(\alpha)}{\beta[\exp\left\{\alpha \exp\left[-(\beta t_i)^{\gamma}\right]\right\}(1-p)p(e^{\alpha}-1)]}.$$

$$\frac{\partial \ell(\mathbf{v})}{\partial \gamma} = \frac{\sum_{i=1}^{n} \delta_i [\alpha\beta^{\gamma}(t_i)^{\gamma-1}(1-p) + \alpha\gamma\beta^{\gamma}\log(\beta)t_i^{\gamma-1}(1-p) + \alpha\gamma\beta^{\gamma}t_i^{\gamma-1}\log(t_i)(1-p)]}{\alpha\gamma\beta^{\gamma}t_i^{\gamma-1}(1-p)} +$$
$$+ \sum_{i=1}^{n} \delta_i \left(-\alpha(\beta t_i)^{\gamma}\log(\beta t_i)\exp\left\{-(\beta t_i)^{\gamma}\right\} - (\beta t_i)^{\gamma}\log(\beta t_i)\right) +$$
$$\frac{-\sum_{i=1}^{n}(1-\delta_i)\alpha(\beta t_i)^{\gamma}\log(\beta t_i)\exp\left\{-(\beta t_i)^{\gamma}\right\}\exp\left\{\alpha\exp\left[-(\beta t_i)^{\gamma}\right]\right\}(1-p)p\exp(\alpha)}{\exp\left\{\alpha\exp\left[-(\beta t_i)^{\gamma}\right]\right\}(1-p)p(\exp(\alpha)-1)}.$$

$$\frac{\partial \ell(\mathbf{v})}{\partial p} = \frac{\sum_{i=1}^{n} \delta_i}{1-p} +$$
$$+ \frac{\sum_{i=1}^{n}(1-\delta_i)\left(-\exp\left\{\alpha\exp\left[-(\beta t_i)^{\gamma}\right]\right\}\exp(\alpha)p + \exp\left\{\alpha\exp\left[-(\beta t_i)^{\gamma}\right](1-p)\exp(\alpha)\right\}\right)}{\exp\left\{\alpha\exp\left[-(\beta t_i)^{\gamma}\right]\right\}(1-p)p(\exp(\alpha)-1)}.$$

O sistema de equações dado em (4.11) é não linear, tornando-se necessário usar um algoritmo de otimização para resolvê-lo. Dessa forma, as estimativas desses parâmetros foram obtidas por meio de maximização numérica do logaritmo da função de verossimilhança usando um processo iterativo. O processo iterativo foi o algoritmo quase-Newton baseado no método BFGS. Intervalos de confiança e testes de hipótetes podem ser obtidos usando a distribuição para grandes amostras dos estimadores de máxima verossimilhança como descrito na seção 2.1.3. Assim, a aproximação normal assintótica para $\widehat{\mathbf{v}}$ pode ser expressa por $\sqrt{n}(\widehat{\mathbf{v}} - \mathbf{v}) \sim N_4(0, L(\mathbf{v})^{-1})$. Para a distribuição LWP($\mathbf{v}$), a matriz de informação observada é dada por:

$$L(\mathbf{v}) = -\begin{bmatrix} L_{\alpha\alpha} & L_{\alpha\beta} & L_{\alpha\gamma} & L_{\alpha p} \\ . & L_{\beta\beta} & L_{\beta\gamma} & L_{\beta p} \\ . & . & L_{\gamma\gamma} & L_{\gamma p} \\ . & . & . & L_{pp} \end{bmatrix}$$

4.3.5 Estudo de Simulação

Com objetivo de verificar se o estimador de máxima verossimilhança é adequado para diferentes tamanhos de amostra da LWP, fizemos um estudo de simulação tal que o tempo da ocorrência da falha Y segue uma distribuição de longa duração Weibull-Poisson (LWP) com os parâmetros $\alpha = 3$, $\gamma = 2$, $\beta = 1$ e $p = 0.10$. Consideramos o tempo da censura C uma variável aleatória com distribuição Weibull com parâmetros $\gamma = 2$ e $\beta = 1$. Os tempos de sobrevivência observados foram obtidos fazendo $t_i = \min(Y_i, C_i)$. Foram realizadas $B = 1500$ simulações e a partir destas amostras simuladas obteve-se as estimativas de máxima verossimilhança usando o recurso numérico *optim* que encontra-se no *Software R*. Através destas determinou-se as médias das estimativas de máxima verossimilhança, o viés e o erro quadrático médio (EQM) com diferentes tamanhos de amostra e 20% de censura. Os valores iniciais para o processo de otimização foram valores próximos dos verdadeiros valores dos parâmetros. A seguir é descrito o processo desta simulação:

1. Gerar $u_j \sim U(0,1)$

2. Gerar y_j, tal que:

$$y_j = \begin{cases} \infty, & \text{se} \quad u_j \leq 1-p; \quad j=1,2,\ldots,n \\ F^{-1}(u_j) = \dfrac{1}{\beta}\left[\ln(\alpha) - \ln\left(\ln(e^{\alpha}(1-p-u)+u) - \ln(1-p)\right)\right]^{1/\gamma}, & \text{se} \quad u_j > 1-p; \end{cases}$$

em que F(.) é a função distribuição acumulada da distribuição de longa duração Weibull-Poisson LWP;

3. Gerar a variável de censura $c_j \sim$Weibull(β, γ) e fazer $t_j = \min(y_j, c_j)$;

4. Se $y_j > c_j$ ou $y_j = \infty$, então $\delta_j = 0$, caso contrário, $\delta_j = 1$, para $j = 1, ..., n$.

Os resultados dessa simulação são apresentados na Tabela 4.1:

Tabela 4.1: Resultados das simulações da LWP com os parâmetros: $\alpha = 3$, $\beta = 1$, $\gamma = 2$ e $p = 0.10$ com 20% de censura

Tamanho da amostra (n)	Parâmetros	Média	Vício	EQM
20	α	1.17659	-1.82341	5.06912
	β	1.58461	0.58461	0.52577
	γ	2.16158	0.16158	0.29105
	p	0.07547	-0.02453	0.00372
60	α	1.23464	-1.76536	5.79931
	β	1.54350	0.54350	0.43918
	γ	1.95488	-0.04512	0.08350
	p	0.07781	-0.02219	0.00147
100	α	1.70900	-1.29099	4.74682
	β	1.23600	0.23600	0.65703
	γ	1.68329	-0.31670	0.07795
	p	0.05905	-0.04094	0.00213
600	α	3.08024	0.08024	4.01806
	β	1.18182	0.18182	0.16040
	γ	1.97306	-0.02694	0.02092
	p	0.08077	-0.01923	0.00046
1000	α	2.81550	-0.18449	4.00939
	β	1.23606	0.23606	0.14352
	γ	1.94644	-0.05356	0.01971
	p	0.08039	-0.01961	0.00010

Nota-se de acordo com a Tabela 4.1 que as médias das estimativas se aproximam do verdadeiro valor do parâmetro quando aumentamos o tamanho da amostra. Outro fator importante é que o erro quadrático médio diminui à medida que o tamanho amostral aumenta, ou seja, as estimativas dos parâmetros melhoram tornando-se os estimadores cada vez menos viesados. Esses resultados são observados especialmente para tamanho da amostra maiores que 100. Já para o tamanho de amostra inferior a 100, notamos que a média das estimativas não se encontram tão próximas, principalmente em relação ao parâmetro α. Dessa forma é recomendável a utilização de outros métodos de estimação para os parâmetros com o intuito de verificar se eles apresentam o melhor desempenho para tamanhos de amostra inferiores a 100.

4.3.6 Aplicação

A ideia é mostrar a aplicabilidade desta nova distribuição, bem como verificar a sua utilidade em dois conjuntos de dados. O primeiro conjunto de dados consiste no tempo (em dias) até a reincidência ao crime de 477 indivíduos em regime semi-aberto no qual aproximadamente 60% dos indivíduos apresentam censura. O segundo conjunto de dados foi extraído de Kalbfleisch & Prentice (1980b) e consiste em dados de sobrevivência de 195 pacientes com carcinoma epidermóide em que aproximadamente 30% dos tempos de sobrevivência são censurados. Segundo o estudo, os 30% de pacientes sobreviveram ao tempo de análise e alguns deles foram perdidos, pois alguns se mudaram. O objetivo é estudar o tempo de vida desses conjuntos de dados. Em primeiro lugar, a fim de verificar a forma da função de risco para esses dados foi feita uma análise gráfica usando a curva TTT como mostra a Figura 4.5.

De acordo com a Figura 4.5 nota-se que há evidências do formato unimodal da função de risco nos dois conjuntos de dados. Logo, a distribuição LWP pode ser utilizada para a modelagem dos dados, pois o comportamento da sua função de risco também possui comportamento, dentre outros, unimodal. Na Tabela 4.2 são apresentadas as estimativas e o erro padrão (entre parênteses) a partir do método de máxima verossimilhança com o recurso do *Software R* por meio do comando *optim* para os dois conjuntos de dados em estudo. Também foram usados os critérios de seleção AIC e BIC para identificar qual dessas distribuições é a mais apropriada para modelar o tempo de sobrevivência desse conjunto de dados. A Tabela 4.3 mostra os resultados.

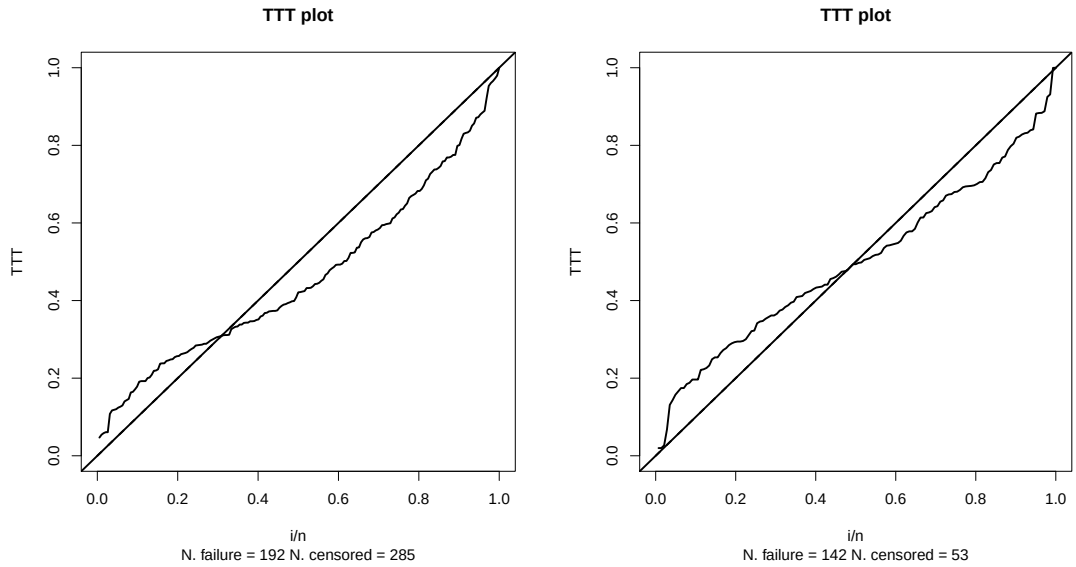

Figura 4.5: Painel esquerdo: curva TTT dos dados de Crime; Painel direito: curva TTT dos dados de Carcinoma.

Tabela 4.2: Estimativas de máxima verossimilhança dos parâmetros das distribuições: LEP, LW e LWP dos dados

Dados	**Distribuições**	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\alpha}$	$\widehat{p}$
Crime	LEP	0.00076	-	2.1423e-06	0.42087
		(0.00015)	-	(3.6140e-05)	(0.06187)
	LW	0.00118	1.56319	-	0.55103
		(7.5894e-05)	(0.10409)	-	(0.02688)
	LWP	0.00076	1.73567	2.22860	0.54457
		(0.00023)	(0.12531)	(1.45938)	(0.02886)
Carcinoma	LEP	0.00179	-	4.7397e-05	0.13131
		(0.00025)	-	(0.00271)	(0.04808)
	LW	0.00209	1.45394	-	0.20875
		(0.00015)	(0.10384)	-	(0.03495)
	LWP	0.00109	1.65632	3.45871	0.20106
		(0.00039)	(0.13287)	(2.01304)	(0.03656)

Tabela 4.3: Critérios de Seleção AIC e BIC para as distribuições: LEP, LW e LWP dos dados

Modelos	**Crime**			**Carcinoma**		
	$\ell(\cdot)$	AIC	BIC	$\ell(\cdot)$	AIC	BIC
LEP	-1727.518	3461.036	3473.539	-1082.608	2171.216	2181.035
LW	-1712.804	3431.608	3444.111	-1072.699	2151.398	2161.217
LWP	-1710.762	3429.524	3446.194	-1070.570	2149.140	2162.232

Observa-se de acordo com a Tabela 4.3 que as distribuições de longa duração Weibull-Poisson (LWP) e Weibull (LW) tiveram os menores valores para os dois critérios de seleção. Este resultado é confirmado pela Figura 4.6, pois se verifica que a curva da função de sobrevivência correspondente as distribuições LWP e LW tiveram um comportamento mais similar à curva da função de sobrevivência dado por Kaplan-Meier.

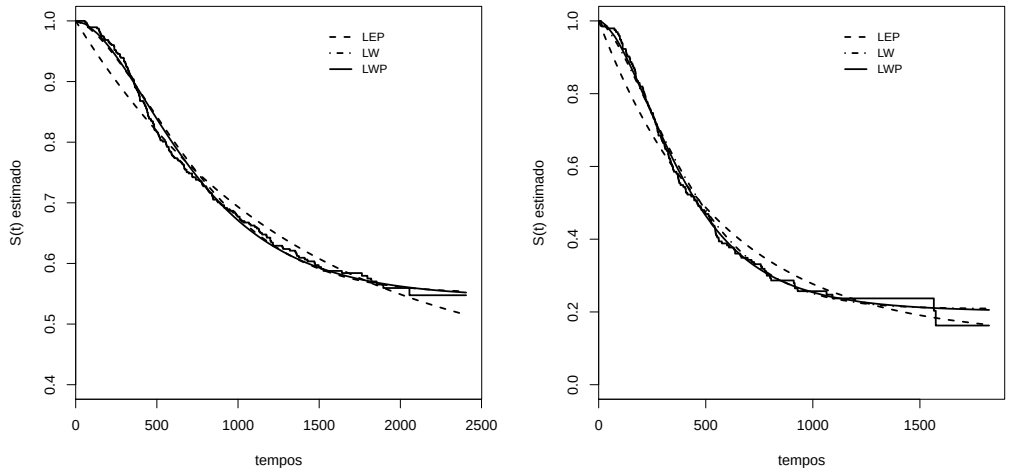

Figura 4.6: Curva da função de sobrevivência estimada pelo método de Kaplan-Meier e as curvas das funções de sobrevivência estimadas dos dados seguido das distribuições LEP, LW e LWP para os dados de Crime e Carcinoma.

Por fim, foi utilizado o teste da razão de verossimilhanças no limite do espaço paramétrico para os dois conjuntos de dados para verificar qual dos dois modelos de longa duração (LW ou LWP) se ajustam melhor nos dois conjuntos de dados. O teste da razão de verossimilhanças no limite do espaço paramétrico foi utilizado pelo motivo de que esses modelos são encaixados, e as distribuições LW e WP são casos particulares da distribuição LWP quando $\alpha \to 0$ e $p = 0$, respectivamente. O procedimento do teste encontra-se na seção 2.1.3.2. Para o conjunto de dados referente ao tempo de reincidência ao crime, partimos do fato de que a hipótese nula retrata que a distribuição LW é adequada, ou seja, $H_0 : \alpha \to 0$ e a hipótese alternativa é que a distribuição LWP é adequada, ou seja, $H_1 : \alpha > 0$. O ω_n encontrado de 4.084 é maior do que $1/2 + 1/2\ P(\chi^2_1 \leq c) = 2.705543$, apresentando forte evidência em favor da distribuição LWP com 5% de significância. Além disso, comparamos a distribuição LWP com a distribuição WP para verificar se a proporção p de indivíduos curados (ou imunes) é significativa utilizando o mesmo procedimento do teste da razão de verossimilhanças no limite do espaço paramétrico. A hipótese nula

nesse caso estabelece que a distribuição WP é adequada, ou seja, $H_0 : p = 0$ e a hipótese alternativa que a distribuição LWP é adequada, ou seja, $H_1 : p > 0$. O valor encontrado de w_n de 43.31 foi maior do que $1/2 + 1/2\ P(\chi_1^2 \leq c) = 2.705543$. Este fato, em nível de significância de 5%, evidencia um melhor ajuste pela distribuição LWP. O ω_n referente ao segundo conjunto de dados foi de 4.258, logo, é maior do que $1/2 + 1/2\ P(\chi_1^2 \leq c) = 2.705543$. Com isso há uma evidência em favor da distribuição LWP ao nível de significância de 5%. Também comparamos a distribuição LWP com a distribuição WP para verificar se a proporção de curados é significativa. O valor w_n de 19.65 foi maior do que $1/2 + 1/2\ P(\chi_1^2 \leq c) = 2.705543$. Este fato, ao nível de significância de 5%, também evidencia um melhor ajuste considerando a distribuição LWP.

4.4 Modelo de longa duração Weibull-Poisson (LWP) com covariáveis

4.4.1 Introdução

É comum na prática que existam características associadas ao tempo de sobrevivência que acabam influenciando na proporção de indivíduos curados em estudo. Uma maneira de estudar essas influências é através da inclusão de covariáveis na proporção de curados através da função de ligação logística.

4.4.2 Estimação por máxima verossimilhança da LWP com covariáveis

Dada uma amostra aleatória em que os t_i são provenientes de uma distribuição LWP($\mathbf{v}$) com covariáveis e δ_i uma variável aleatória indicadora de censura. Considerando que $f_{pop}(.)$ representa a função de densidade imprópria relativa ao grupo dos indivíduos em risco dado na equação 4.7 e $S_{pop}(.)$ representa a função de sobrevivência imprópria relativa ao grupo dos indivíduos em risco dado na equação 4.6, a função de log-verossimilhança pode ser escrita como:

$$\ell(\mathbf{v}) \propto \sum_{i=1}^{n} \delta_i \log(\alpha\gamma\beta^{\gamma} t_i^{\gamma-1}(1-p_i)) + \sum_{i=1}^{n} \delta_i \left[\alpha \exp\left\{-(\beta t_i)^{\gamma}\right\} - (\beta t_i)^{\gamma}\right]$$
$$- \sum_{i=1}^{n} \delta_i \log(\exp(\alpha)-1) + \sum_{i=1}^{n} (1-\delta_i) \log\left(p_i \exp(\alpha) \left(\exp\left\{\alpha \exp\left[-(\beta t)^{\gamma}\right]\right\}\right)(1-p_i) - 1\right)$$
$$- \sum_{i=1}^{n} (1-\delta_i) \log(\exp(\alpha)-1),$$

em que $p_i = \frac{\exp({x_i}^T\beta)}{1+\exp({x_i}^T\beta)}$; $\boldsymbol{\beta}=(\beta_0, ..., \beta_p)^T$ é o vetor de parâmetros desconhecidos a serem estimados e $x_i^T = (1, x_{i1}, ..., x_{ip})$ é o vetor de covariáveis. Observe que a probabilidade de cura varia de indivíduo para indivíduo e a ligação logística mantém cada p_i estritamente entre 0 e 1.

Para encontrar os estimadores de máxima verossimilhança deve-se resolver o sistema de equações $U(\mathbf{v}) = 0$ que é não linear, tornando-se necessário usar um algoritmo de otimização para o vetor de parâmetros $\mathbf{v} = (\alpha, \beta, \gamma, (\boldsymbol{\beta})^T)^T$. Dessa forma, as estimativas desses parâmetros foram obtidas por meio de maximização numérica do logaritmo da função de verossimilhança usando um processo iterativo. Neste trabalho, o algoritmo quase-Newton baseado no método BFGS foi usado. Intervalos de confiança e testes de hipótetes podem ser obtidos usando a distribuição para grandes amostras dos estimadores de máxima verossimilhança como descrito na seção 2.1.3. Assim, a aproximação normal assintótica para $\widehat{\mathbf{v}}$ pode ser expressa por $\sqrt{n}(\widehat{\mathbf{v}}-\mathbf{v}) \sim N_{p+3}(0, L(\mathbf{v})^{-1})$. Para a distribuição LWP com covariáveis, a matriz de informação observada é dada por:

$$L(\mathbf{v}) = -\begin{bmatrix} L_{\alpha\alpha} & L_{\alpha\beta} & L_{\alpha\gamma} & L_{\alpha\beta_j} \\ . & L_{\beta\beta} & L_{\beta\gamma} & L_{\beta\beta_j} \\ . & . & L_{\gamma\gamma} & L_{\sigma\beta_j} \\ . & . & . & L_{\beta_j\beta_s} \end{bmatrix}$$

4.4.3 Estudo de Simulação

Para examinar o desempenho dos estimadores da distribuição LWP na presença de covariáveis, foi feito nesta seção um estudo de simulação tal que o tempo de sobrevivência T segue uma distribuição LWP com inclusão de covariáveis na proporção de curados através da função de ligação logística, ou seja, $p_i = \frac{\beta_0+\beta_1 x_i}{1+\beta_0+\beta_1 x_i}$ em que x_i foi obtido a partir de

uma distribuição $U(0,1)$. Os valores dos parâmetros foram: $\beta_0 = -5.5$, $\beta_1 = 7.0$, $\alpha = 3$, $\beta = 1$ e $\gamma = 2$. Os tempos de censura C segue uma distribuição Weibull(β, γ) com valores $\beta = 1$ e $\gamma = 2$. Os tempos de sobrevivência considerados foram obtidos fazendo: $t_i = \min(T_i, C_i)$. Foram realizadas $B = 1500$ simulações e a partir destas amostras simuladas obteve-se as estimativas de máxima verossimilhança usando o recurso numérico *optim* que encontra-se no *Software R*. Através destas determinou-se as médias das estimativas de máxima verossimilhança, o viés e o erro quadrático médio (EQM) com diferentes tamanhos de amostra e 30% de censura. Os valores iniciais para o processo de otimização foram valores próximos dos verdadeiros valores dos parâmetros. A seguir é descrito o processo desta simulação:

1. Gerar $u_j \sim U(0,1)$
2. Gerar y_j, tal que:

$$y_j = \begin{cases} \infty, & \text{se} \quad u_j \leq 1-p \\ F^{-1}(u_j) = \dfrac{1}{\beta}\left[\ln(\alpha) - \ln\left(\ln(e^{\alpha}(1-p-u)+u) - \ln(1-p)\right)\right]^{1/\gamma}, & \text{se} \quad u_j > 1-p. \end{cases}$$

Para j=1,2,...,n em que F(.) é a Função distribuição acumulada da distribuição de longa duração Weibull-Poisson LWP.

3. Gerar a variável de censura $c_j \sim$Weibull(β, γ) e fazer $t_j = \min(y_j, c_j)$
4. Se $y_j > c_j$ ou $y_j = \infty$, então $\delta_j = 0$, caso contrário, $\delta_j = 1$, para $i = 1, ..., n$

São apresentados na Tabela 4.4 as médias das estimativas e o erro quadrático médio das $B = 1500$ simulações para diferentes tamanhos de amostra com 30% de censura. Nota-se de acordo com a Tabela 4.4 que as médias das estimativas se aproximam do verdadeiro valor do parâmetro quando aumentamos o tamanho da amostra. Outro fator importante é que o erro quadrático médio diminui à medida que o tamanho amostral aumenta, ou seja, as estimativas dos parâmetros melhoram tornando-se os estimadores cada vez menos viesados. Esses resultados são observados especialmente para tamanho da amostra maiores que 100. Já para o tamanho de amostra inferior a 100, notamos que a média das estimativas não se encontram tão próximas, principalmente em relação ao parâmetro α. Dessa forma é recomendável a utilização de outros métodos de estimação para os parâmetros com o intuito de verificar se eles apresentam o melhor desempenho para tamanhos de amostra inferiores a 100.

Tabela 4.4: Resultados das simulações da LWP na presença de covariáveis com os parâmetros: $\beta_0 = -5.5$, $\beta_1 = 7.0$, $\alpha = 3$, $\beta = 1$ e $\gamma = 2$ com 30% de censura

Tamanho da amostra (n)	Parâmetros	Média	Vício	EQM
40	α	2.00191	-0.99809	5.86277
	β	1.80263	0.80263	0.93759
	γ	2.27731	0.27731	0.23161
	β_0	-7.16321	-1.66321	8.86753
	β_1	9.00639	2.00639	17.8342
100	α	2.42670	-0.57330	5.85291
	β	1.34269	0.34269	0.27817
	γ	2.10561	0.10561	0.06915
	β_0	-4.81221	0.68779	2.29201
	β_1	5.34149	-1.65851	5.94269
600	α	4.30770	1.30770	4.51799
	β	1.00746	0.00746	0.05313
	γ	2.13136	0.13136	0.02585
	β_0	-4.48685	1.01315	1.23138
	β_1	4.93887	-2.06113	4.62448
1000	α	2.92586	-0.07414	1.26713
	β	0.98589	-0.01411	0.02080
	γ	1.94212	-0.05788	0.00718
	β_0	-4.57976	0.92024	0.97730
	β_1	5.09710	-1.90290	3.86841

4.4.4 Aplicação

Para ilustrar esta aplicação, foi usado o conjunto de dados com 862 pacientes com câncer internados na UTI do INCA (Instituto Nacional do Câncer) em que aproximadamente 42% dos tempos dos indivíduos possuem censura. O conjunto de dados encontra-se em Carvalho *et al.* (2011). Além da variável tempo de sobrevivência e da variável indicadora de censura, outras variáveis foram observadas:

- X_1= Idade (anos completos).
- X_2= Sexo (0 = feminino; 1 = masculino).
- X_3= Tipo de tumor (0 = sólido localizado; 1 = metastático; 2 = hematológico).
- X_4= Perda de peso recente acima de >10% ou IMC <18 (0 = não; 1 = sim).
- X_5= Comorbidades severas presentes (0 = não ; 1 = sim).
- X_6= Leucopenia presente (0 = não; 1= sim).

De início, para identificar um modelo apropriado para o tempo de sobrevivência, estimou-se os parâmetros via verossimilhança sem considerar a presença de covariáveis. Em seguida, ainda sem a presença das covariáveis, foi feita uma análise gráfica usando a curva TTT e o método de Kaplan-Meier como mostra a Figura 4.7. O painel esquerdo da Figura 4.9 mostra o gráfico TTT, o qual indica que a função de risco é decrescente. Logo, pode-se tentar utilizar a distribuição LWP pra a modelagem dos dados pois o comportamento da sua função de risco também possui comportamento, dentre outros, decrescente. Este resultado é confirmado pelo painel direito da Figura 4.7, pois verifica-se que a curva da função de sobrevivência correspondente as distribuições de longa duração Weibull-Poisson (LWP) e exponencial-Poisson (LEP) tiveram um comportamento mais similar à curva da função de sobrevivência estudado por Kaplan-Meier do que a distribuição de longa duração Weibull (LW).

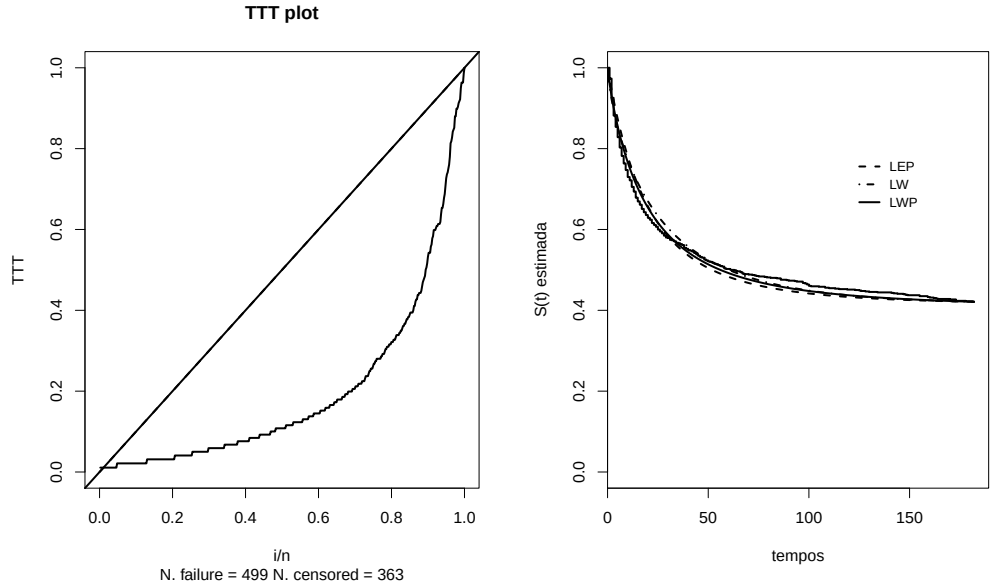

Figura 4.7: Painel esquerdo: TTT plot; Painel direito: Curva da função de sobrevivência estimada pelo método de Kaplan-Meier e das funções de sobrevivência estimadas das distribuições: LEP, LW e LWP.

A segunda etapa deste trabalho foi considerar as distribuições mais adequadas a fim de realizar o ajuste considerando a presença de covariáveis. Dessa forma, foram escolhidas as distribuições LWP e LEP para a modelagem com covariáveis. A Tabela 4.5 mostra as estimativas de máxima verossimilhança e os respectivos erros padrão (E.P.) para os parâmetros dos dois modelos com os valores dos critérios de seleção.

Tabela 4.5: Valores estimados dos parâmetros dos modelos LEP e LWP com covariáveis.

	LEP			**LWP**		
Parâmetros	Estimativas	E.P.	p-valor	Estimativas	E.P.	p-valor
α	2.77111	0.43536	-	2.63174	0.57968	-
β	0.01695	0.00233	-	0.01549	0.00333	-
γ	-	-	-	0.88842	0.03520	-
β_0	0.91642	0.57807	0.11289	0.63550	0.47696	0.18273
β_1	-0.02138	0.00571	0.00018	-0.01865	0.00554	0.00075
β_2	-0.03908	0.01410	0.00560	-0.04775	0.02047	0.01967
β_3	-0.59662	0.14148	<0.0001	-0.56492	0.11740	<0.0001
β_4	-0.54145	0.27946	0.05268	-0.51363	0.26435	0.05201
β_5	-0.64586	0.30463	0.03399	-0.61759	0.28856	0.03233
β_6	-1.36854	0.47478	0.00394	-1.34459	0.48249	0.00532
$-log(L(v))$	2653.419			2648.255		
AIC	5324.838			5316.510		
BIC	5367.671			5364.103		

E.P.=erro padrão

Pode-se observar de acordo com a Tabela 4.5 que a distribuição de longa duração Weibull-Poisson (LWP) com as covariáveis obteve o menor valor para os dois critérios de seleção em relação a distribuição de longa duração Exponencial-Poisson (LEP), indicando que este modelo é mais adequado aos dados. Os resultados das estimativas dos parâmetros e os respectivos erros padrão das duas distribuições são semelhantes, logo, só a variável X_4 (Perda de peso recente acima de >10% ou IMC <18) não foi significativa em ambos modelos ao nível de 5%. Partindo do fato que o estimador tem uma distribuição assintótica Normal, foi encontrado a média do valor estimado e o erro padrão para proporção de curados para as duas distribuições ao nível de 5%. A Tabela 4.6 mostra o resultado:

Tabela 4.6: Valores estimados da proporção de curados dos modelos LEP e LWP com covariáveis.

	LEP			**LWP**		
Parâmetro	Média da estimativa	E.P.	p-valor	Média da estimativa	E.P.	p-valor
p	0.41739	0.18056	0.02079	0.41113	0.18063	0.02284

E.P.=erro padrão

Através da Tabela 4.7 pode-se notar que os resultados foram bem próximos em relação a média das estimativas e o erro padrão dos dois modelos, logo, levando a mesma conclusão para a significância da proporção de curados no estudo. Os intervalos de confiança da média da proporção de curados das distribuições LWP e LEP que foram, respectivamente (0.0634907; 0.7713083) e (0.0570926; 0.7651692).

4.5 Considerações finais

Neste capítulo foi discutido o modelo de longa duração Poisson-Weibull (LWP). Este modelo provém da distribuição Weibull-Poisson (WP) proposta por Louzada *et al.* (2011a) que generaliza a distribuição Weibull-Poisson proposta por Kus (2007) e Weibull após considerar uma estrutura de longa duração. Também foram discutidas suas propriedades assim como as diversas formas em relação a função de risco. Um estudo de simulação com um percentual de censura foi realizado para verificar o comportamento dos estimadores desta distribuição via máxima verossimilhança. Além disso, essa distribuição foi analisada na situação em que as covariáveis são inclusas no estudo e através das análises como a curva TTT plot, o critério AIC, o Kaplan-Meier e o TRV no limite do espaço paramétrico pode-se notar que em relação ao conjunto de dados em estudo, a distribuição LWP ajustou-se bem aos dados. Logo espera-se que esta distribuição seja útil em outros conjuntos de dados de longa duração.

Capítulo 5

Conclusões e Trabalhos futuros

Neste trabalho discute-se sobre a distribuição Weibull-Poisson (WP), sua origem, construção do modelo, função de sobrevivência, função densidade de probabilidade, função de risco, função quantil e a função densidade de probabilidade da k-ésima estatística de ordem. Além disso, aborda-se a similaridade entre as funções densidades do mínimo e do máximo com a inclusão da parte inferencial do modelo. Também realiza-se um estudo de simulação com um percentual de censura a fim de observar o comportamento dos estimadores desta distribuição para diferentes tamanhos de amostra. Ainda, analisa-se a situação em que covariáveis são inclusas no estudo.

Também estuda-se o modelo de longa duração Weibull-Poisson (LWP), além de abordar um estudo de simulação com um percentual de censura como utilizado na WP. Além disso, analisa-se também a distribuição LWP na presença de covariáveis.

Com o objetivo de ilustrar a aplicabilidade dos modelos WP e LWP na presença ou ausência de covariáveis, considerou-se conjuntos de dados reais, em que as estimativas dos parâmetros foram determinadas através da abordagem de máxima verossimilhança. Com o intuito de selecionar o modelo mais adequado foram utilizados os critérios AIC, BIC, Kaplan-Meier, TRV no limite do espaço paramétrico. Concluiu-se que as distribuições WP e LWP adequaram-se melhor aos conjuntos de dados segundo esses critérios quando comparadas com as distribuições exponencial-Poisson e Weibull.

Extensão deste trabalho seria utilizar a mesma metodologia considerando o máximo dos tempos, $T^* = \max(Y_1, ..., Y_N)$, além de estudar outros tipos de resíduos e medidas de influência tanto no contexto sem longa duração quanto no contexto de longa duração.

Referências Bibliográficas

Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, **2**, 106–108.

Adamidis, K. & Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statistics Probability Latters*, **39**, 35–42.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. John Wiley & Sons.

Barreto-Souza, W., Morais, A. L. d. & Cordeiro, G. M. (2008). The Weibull-Geometric Distribution. *Journal of Statistical Computation and Simulation*, **00**, 1–14.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Cancho, V. G., Louzada-Neto, F. & Barriga, G. D. C. (2011). The poisson-exponential lifetime distribution. *Computational Statistics & Data Analysis*, **55**(1), 677–686.

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. Z. & Shimakura, S. E. (2011). *Análise de sobrevivência: teoria e aplicações em saúde*. Editora Fiocruz.

Chahkandi, M. & Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Computational Statistics & Data Analysis*, **53**(1), 4433–4440.

Chen, H. M. & Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, **57**, 43–52.

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Editora Edgard Blucher.

Cooner, F., Banerjee, S. & McBean, A. M. (2006). Modelling geographically referenced survival data with a cure fraction. *Statistical Methods in Medical Research*, **15**(4), 307–324.

Couto, E. (2010). *Modelo de regressão log-gamma generalizado exponenciado com dados censurados*. Dissertação, ESALQ- Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

Cox, D. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society B*, **30**(2), 248–275.

Goetghebeur, E. & Ryan, L. (1995). A modified log rank test for competing risks with missing failure type. *Biometrika*, **77**, 207–211.

Hosmer, D. & Lemeshow, S. (1999). *Applied Survival Analysis*. John Wiley and Sons.

Kalbfleisch, J. & Prentice, R. (1980a). *The Statistical Analysis of Failure Time Data*. Wiley & Sons.

Kalbfleisch, J. & Prentice, R. (1980b). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Canada.

Kus, C. (2007). A new lifetime distribution. *Computation Statist. Data Analysis*, **51**, 4497–4509.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, second edition.

Lee, E. T. & Wang, J. W. (2003). *Statistical Methods for Survival a Data Analysis - Third Edition* . Wiley, New Jersey.

Louzada, F., Bereta, E. & Franco, M. (2011a). The poisson-weibull distribution. *Advances and Applications in Statistics*, **22**, 117–118.

Louzada, F., Roman, M. & Cancho, V. (2011b). The complementary exponential geometric distribution: Model, properties, and a comparison with its counterpart. *Computational Statistics & Data Analysis*, **55**, 2516–2524.

Lu, K. & Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, **57**(4), 1191–1197.

Lu, K. & Tsiatis, A. A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis*, **11**(1), 29–40.

Maller, R. & Zhou, X. (1995). Testing for the Presence of Immune or Cured Individuals in Censored Survival Data. *Biometrics*, **51**, 1197–1205.

Marshall, A. W. & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the Exponential and Weibull families. *Biometrika,*, **84**(3), 641–652.

Mudholkar, G., Srivastava, D. & Freimer, M. (1995). The exponentiated weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, **37**(4), 436–445.

Mudholkar, G. S., Srivastava, S. K. & Kollia, G. D. (1996). A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data. **91**(436), 1575–1583.

Pradhan, B. & Kundu, D. (2013). Inference and optimal censoring schemes for progressively censored birnbaum-saunders distribution. *Journal of Statistical Planning and Inference*, **143**, 1098–1108.

Reiser, B., Guttman, I., Lin, D., Guess, M. & Usher, J. (1995). Bayesian inference for masked system lifetime data. *Applied Statistics*, **44**(1), 79–90.

Rinne, H. (2009). *The Weibull Distribuition: A handbook*. Chapman & Hall/CRC, Boca Raton, FL.

Rodrigues, J., V., C., de Castro, M. & Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics and Probability Letters*, **79**(2), 753–759.

Roman, M. (2013). *Modelos para dados de sobrevivência na presença de diferentes esquemas de ativação baseados na distribuição geométrica*. Tese, UFSCAR-Universidade Federal de São Carlos, São Carlos.

Tahmasbi, R. & Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate. *Computational Statistics Data Analysis*, **52**(8), 3889–3901.

Yakovlev, A., Tsodikov, A. & Asselain, B. (1996). Stochastic models of tumor latency and their biostatistical applications. *World Scientific Pub Co Inc*.