Regressão de dados binários: Distribuição Weibull

Renault Caron

Regressão de dados binários: Distribuição Weibull

Renault Caron

Orientador: Prof. Dr. Adriano Polpo de Campos

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

UFSCar - São Carlos

Abril/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C293rd

Caron, Renault.
Regressão de dados binários : distribuição Weibull / Renault Caron. -- São Carlos : UFSCar, 2010.
51 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Análise de regressão. 2. Dados categóricos. 3. Modelos não-lineares (Estatística). 4. Regressão logística. 5. SAS (Programa de computador). 6. Biometria. I. Título.

CDD: 519.536 (20ª)

Renault Caron

Regressão de dados binários: Distribuição Weibull

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 12 de março de 2010.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Adriano Polpo de Campos (DEs-UFSCar/Orientador)

1° Examinador _____
Prof. Dr. Francisco Louzada Neto (DEs-UFSCar)

2° Examinador _____
Prof. Dr. Ronaldo Dias (IMECC-UNICAMP)

Agradecimentos

Agradeço primeiramente a Deus por toda sorte e fortúnios que concede em minha vida;

À minha mãe, Marlene Ribeiro da Silva, pelos sábios conselhos, incentivos e por sempre acreditar em mim. Ao meu pai, Renaldo Luiz Caron, pelo exemplo de inteligência auto-didata, humildade e pelo investimento em meus estudos. Ao meu tio, Carlos Alberto Ribeiro Diniz, por me acolher em sua casa na minha vinda à São Carlos e por todo apoio e companheirismo até hoje;

Aos meus irmãos Renan Caron, Renê Caron e Anita Reane Caron, pelos exemplos, apoio e por tudo que me ensinaram;

Ao meu orientador Adriano Polpo de Campos, pelos conselhos, incentivo e por defender artigos frutos desta dissertação. Ao professor Carlos Alberto de Bragança Pereira pelos comentários utilizados no artigo publicado no 29º MaxEnt;

À banca de defesa: Francisco Louzada Neto e Ronaldo Dias pelas valiosas sugestões. À banca de qualificação: Marcio Alves Diniz e Juan Carlos Ruilova Teran pela indicação de referências bibliográficas importantes nesta fase;

Aos amigos que fiz em São Carlos, Bruno, Vitor, Matheus, Thaysa, Natália, Marita e Lilian pela companhia nos momentos de diversão. Aos companheiros da turma de 2008 do Mestrado, Lorene, Thiago, Jhon, Bruno, Saulo, Vitor, Ana Cláudia e demais alunos das turmas de graduação em estatística;

À agência financiadora Capes, pela bolsa que recebi durante o mestrado.

À todas as pessoas que de alguma forma me ajudaram na vinda a São Carlos e/ou na conclusão deste trabalho, eu agradeço do fundo do meu coração.

Resumo

Neste trabalho propõe-se um novo modelo, para conjunto de dados com variável resposta binária, baseado na função densidade acumulada Weibull. Apresenta-se um resumo das funções de ligação mais conhecidas da literatura. Esta classe de modelos possui como caso especial o modelo complementar log-log e boas aproximações aos modelos logístico e probito. Três conjunto de dados reais são utilizados para comparar o modelo proposto com vários outros modelos. Em um dos conjuntos de dados o modelo é expandido para suportar variável resposta multinomial, isto é, variável discreta com mais de dois eventos de interesse. Os resultados obtidos são muito bons, pois a estimação dos parâmetros é razoavelmente simples e o modelo mostrou-se extremamente eficientes.

Palavras-chave: Resposta binária, função de ligação Weibull, regressão logística.

Abstract

In this work a new class of models for binary data based on Weibull distribution is introduced. A review is made of the most known linkage functions. This class of models has as special case the complementary log-log model and approximates well the logit and probit models. Three real data sets are given to compare the proposed model with many others. In one of these data sets the model is extended to allow multinomial data, that is, a discrete variable with more than two outcomes. The results are very good, because the estimation of parameters is quite simple and the model has shown to be very efficient.

Keywords: Binary response, Weibull link function, logit regression.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A estimação da proporção de um determinado evento em um conjunto de dados com variável resposta binária tem recebido um grande número de soluções estatísticas. Os modelos lineares generalizados foram formulados por Nelder e Wedderburn [11] como uma maneira de ampliar as opções para a distribuição da variável resposta. Assim, para dados binários pode-se supor que a variável resposta tenha distribuição Binomial e que a relação funcional entre a proporcão esperada de sucessos é dada por uma função de ligação. O modelo mais utilizado para este propósito é o modelo de regressão logística. Como no caso binomial o parâmetro de interesse sempre é uma proporção, é muito razoável que outras funções densidade acumulada sejam utilizadas para gerarem novas ligações.

Este problema vem sendo abordado há bastante tempo. Existem várias funções de ligação assimétricas na literatura e a mais conhecida é a complementar log-log, vide Paula [12]. Apresenta-se no capítulo 2, por ordem cronológica, algumas funções de ligação: Prentice [15] introduziu uma função de ligação bi-paramétrica que contêm os modelos logito, probito e complementar log-log e outras funções de ligação assimétricas como casos particulares; Aranda-Ordaz [2] propôs uma função de ligação que contêm como caso particular os modelos logito e complementar log-log; Stukel [19] definiu uma classe de funções de ligação bi-paramétrica que generaliza o modelo logito; Chen, Dey & Shao [6] apresentaram um modelo probito assimétrico utilizando abordagem bayesiana. Porém, em todos estes casos, os estimadores pontuais de máxima verossimilhança não são obtidos

analiticamente, sendo necessário o uso de métodos numéricos para a obtenção dos estimadores.

Neste contexto, no capitulo 3 é proposto uma nova função de ligação baseada na distribuição Weibull, Caron & Polpo [5], para modelar variáveis binárias. O interesse é obter uma função de ligação que se adeque bem aos mais diversos casos e que seu proceso de estimação não seja complicado. Apresenta-se também a interpretação dos parâmetros, um teste de significância baseado em razão de verossimilhanças e três medidas de qualidade do ajuste.

Três estudo de casos a conjunto de dados reais são apresentados no capitulo 4. Inicialmente, aborda-se o conjunto de dados sobre a *Mortalidade de Besouros* expostos a disulfeto de carbono gasoso, publicado por Bliss [3], para mostrar que a função de ligação complementar log log é um caso particular da função de ligação Weibull; na seção a seguir, estuda-se o conjunto de dados sobre a ocorrência do início da menstruação em *Garotas de Varsóvia*, apresentado por Milicer e Szczotka [10], neste caso define-se uma transformação baseada na distribuição Weibull Refletida, Cohen [7]; e, por fim, um conjunto sobre a *Mutação no DNA de caramujos* expostos à radiação gama, por Grazeffe et al. [9], para exemplificar uma aplicação a dados com resposta multinomial.

Um estudo extensivo é feito no capítulo 5 com o objetivo de mostrar que os modelos logito e probito podem ser bem aproximados pelo modelo binário Weibull e que o oposto não ocorre.

Por fim, no capítulo 6, apresenta-se as principais conlusões e algumas opções de pesquisas futuras.

Capítulo 2

Modelos para Dados Binários

Dados binários são aqueles que admitem dois resultados possíveis para a variável resposta. Eles são utilizados em diversas áreas do conhecimento. Para ilustar, seguem alguns exemplos práticos em que esse tipo de resposta aparece: (i) concessão de crédito de um banco, aprovado ou não aprovado; (ii) resultado do diagnóstico de um exame laboratório, positivo ou negativo; (iii) intenção de voto de um eleitor em relação ao candidato A, vota ou não vota; (iv) inspenção de uma peça recém-fabricada, defeituosa ou não-defeituosa; (v) teste da publicidade de um novo produto, vendeu ou não-vendeu, etc. Neste caso considera-se um problema de sucesso e fracasso, sendo sucesso o resultado mais importante da resposta.

A pesquisa em dados binários intensificou-se a partir da década de 50. Um dos primeiros estudos, revisado por Richardson [17], abordava um problema de uma Tabela de contigência 2 x 2 aplicado em epidemiologia. Muitos trabalhos desenvolvidos nas décadas de 50 e 60 são utilizados até hoje, principalmente na análise descritiva dos dados.

Por muitos anos a regressão linear normal era usada para explicar a maioria dos fenômenos aleatórios. Mesmo quando não era razoável assumir normalidade utilizava-se algum tipo de transformação para alcançar a normalidade desejada. Um dos métodos mais utilizados para este fim é a transformação de Box-Cox [4].

Com o desenvolvimento computacional a partir da década de 70, alguns modelos que exigiam a utilização de processos iterativos para a estimação dos parâmetros começaram a ser mais utilizados. A proposta mais interessante e, pode-se dizer, inovadora no assunto, foi apresentada por Nelder e Wedderburn [11] que propuseram os modelos lineares generalizados (MLGs). A idéia básica consiste em abrir o leque de opções para a distribuição da variável resposta, incluindo todas distribuições que pertençam à família exponencial, bem como dar maior flexibilidade para a relação funcional entre a média da variável resposta e a parte linear do modelo. Assim, para dados binários pode-se supor que a variável resposta tenha distribuição Binomial e que a relação funcional entre a proporção esperada de sucessos é dada por uma função de ligação. Esta deve garantir para quaisquer valores dos parâmetros do modelo um valor entre 0 e 1 para a proporção estimada.

Na próxima seção apresenta-se um resumo das funções de ligação mais conhecidas da literatura.

2.1 Resumo das Funções de Ligação

Ao invés de usar um modelo linear na proporção de sucessos em relação a variáveis explicativas (variável selecionada como previsora e potencial variável de explicação da variável resposta) usa-se uma transformação do intervalo da proporção de sucessos (0;1) para $(-\infty; \infty)$. Um modelo linear é então adotado para o valor transformado da proporção de sucessos. Esta transformação é denominada função de ligação.

Logito

A distribuição logito tem densidade dada por

$$f(y) = \frac{exp(y)}{(1 + exp(y))^2} \tag{2.1}$$

em que -$\infty < y < \infty$. Sua função densidade acumulada é dada por

$$F(y) = \frac{exp(y)}{1 + exp(y)}. \tag{2.2}$$

O modelo logístico binomial é obtido substituindo a notação $F(y)$ pela representação da proporção π e y pela representação do componente linear ($X\beta = \eta$) na expressão 2.2. Note que para qualquer valor de η no intervalo ($-\infty$; ∞) existe um valor de π em (0;1). Assim se $\eta \to -\infty$ tem-se que $\pi \to 0$, se $\eta \to \infty$ tem-se que $\pi \to 1$, e para $\eta = 0$, $\pi = 0{,}5$. Como no caso binomial o parâmetro de interesse sempre é uma proporção, é muito razoável que outras funções de distribuição acumuladas sejam utilizadas para gerarem novas ligações e consequentemente novos modelos.

O modelo binomial com ligação logito é definido por

$$\pi = \frac{exp(\eta)}{1 + exp(\eta)} \tag{2.3}$$

ou, equivalentemente,

$$log\left(\frac{\pi}{1-\pi}\right) = \eta. \tag{2.4}$$

Probito

A ligação probito é definida por

$$\Phi^{-1}(\pi) = \eta \tag{2.5}$$

em que $\Phi(.)$ é a função densidade acumulada da normal padrão. Assim, vale ressaltar que para qualquer valor de η no intervalo ($-\infty$, ∞) há um valor da função probito de π no intervalo (0,1). Observe que para $\eta = 0$ tem-se $\pi = 0{,}5$.

Complementar Log-Log

A função de ligação complementar log-log é derivada da distribuição do valor extremo e dada por

$$\pi = 1 - exp\left(-exp(\eta)\right). \tag{2.6}$$

Na Figura 2.1 apresenta-se a $F(\eta)$ no intervalo da distribuição logística e da distribuição do valor extremo para valores de η variando no intervalo [-3;3]. Note que a curva logística é simétrica em torno de $F(\eta) = 1/2$, enquanto que a curva do valor extremo apresenta comportamentos distintos para $F(\eta) \leq 1/2$ e

$F(\eta) > 1/2$.

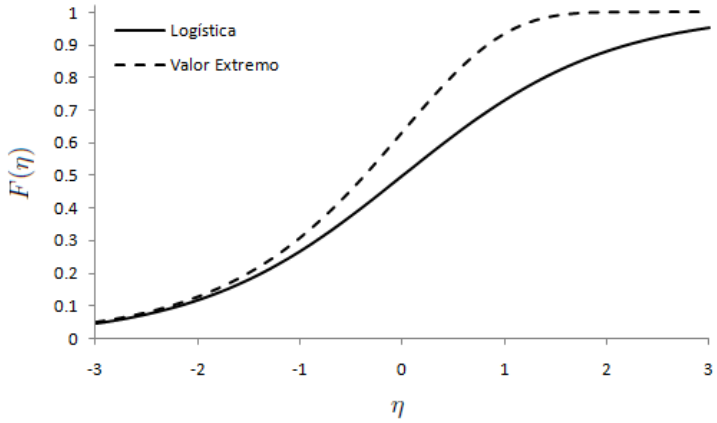

FIGURA 2.1: Função distribuição das curvas logística e valor extremo

Prentice

A função de ligação proposta por Prentice [15] abrange os modelos logito, probito e algumas ligações assimétricas como casos limites (por exemplo, complementar log-log). Ele utilizou a função densidade acumulada do $\log(F_{2m_1,2m_2})$, em que $F_{2m_1,2m_2}$ é uma variável aleatória com distribuição F-Snedecor com parâmetros $2m_1$ e $2m_2$, dada por

$$f(y) = \frac{exp(ym_1)(1+exp(y))^{-(m_1+m_2)}}{\mathcal{B}(m_1,m_2)}, \tag{2.7}$$

em que $\mathcal{B}$ representa a função beta. Obtêm-se a ligação logito tomando $m_1 = m_2 = 1$, probito com $m_1 \to \infty$ e $m_2 \to \infty$, valor mínimo extremo ($m_1 = 1$, $m_2 \to \infty$) e do valor máximo extremo ($m_1 \to \infty$, $m_2 = 1$). Convenientemente obtêm-se dois modelos uniparamétricos

$$\pi = \left[\frac{exp(\eta)}{1+exp(\eta)}\right]^{m_1} \tag{2.8}$$

e

$$\pi = 1 - \left[\frac{exp(-\eta)}{1+exp(-\eta)}\right]^{m_2}, \tag{2.9}$$

fixando os parâmetros $m_2 = 1$ e $m_1 = 1$, respectivamente.

Os parâmetros dos modelos (2.8) e (2.9) são calculados pelo método de máxima-verossimilhança utilizando o algoritmo Newton-Rapson. Porém, devido a dificuldade em calcular a função densidade acumulada da densidade dada em (2.7), a estimativa dos parâmetros para o modelo bi-paramétrico é muito difícil de ser obtida.

Aranda-Ordaz

Uma outra transformação importante, proposta por Aranda-Ordaz [2], é uma função de ligação uni-paramétrica assimétrica que tem como casos particulares os modelos logito e complementar log-log, dada por

$$\eta = log\left[\frac{(1-\pi)^{-\alpha}-1}{\alpha}\right], \tag{2.10}$$

em que $0 < \pi < 1$ e α é uma constante desconhecida. Quando $\alpha = 1$ tem-se a ligação logito e, $\alpha \to 0$ obtém-se a ligação complementar log-log. Na Figura 2.2 tem-se o comportamento de π para alguns valores de α. Em muitas situações práticas o interesse pode ser testar se o modelo logístico é apropriado ou se há necessidade de uma transformação na ligação.

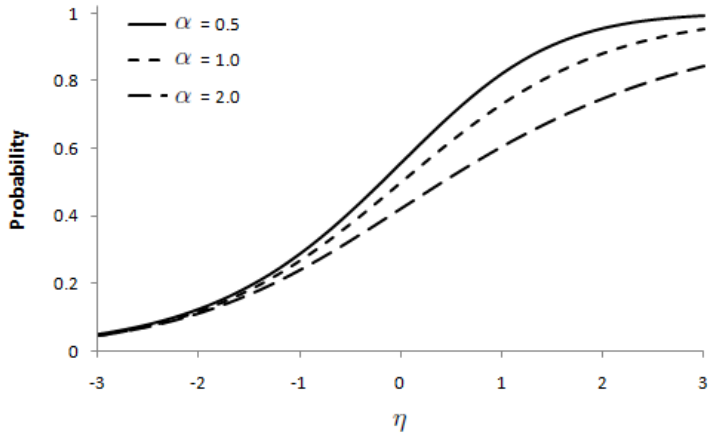

FIGURA 2.2: Transformação de Aranda-Ordaz para alguns valores de α

Stukel

Stukel [19] definiu uma classe de ligações bi-paramétricas que generaliza o modelo logístico. O modelo proposto por Stukel aproxima várias distribuições importantes, como a probito, a complementar log-log e outras funções de ligação

assimétrica. A generalização proposta é

$$log\left(\frac{\pi}{1-\pi}\right) = h(\eta), \tag{2.11}$$

em que $h(\eta)$ é uma função não linear estritamente crescente indexada por dois parâmetros de forma a_1 e a_2. Esta função é definida a seguir.

Para $\eta > 0$

$$h(\eta) = \begin{cases} \left[exp\left(a_1|\eta|\right) - 1\right]/a_1, & \text{para } a_1 > 0 \\ \eta, & \text{para } a_1 = 0 \\ -\left[log\left(1 - a_1|\eta|\right)\right]/a_1, & \text{para } a_1 < 0 \end{cases} . \tag{2.12}$$

Para $\eta < 0$

$$h(\eta) = \begin{cases} -\left[exp\left(a_2|\eta|\right) - 1\right]/a_2, & \text{para } a_2 > 0 \\ \eta, & \text{para } a_2 = 0 \\ \left[log\left(1 - a_2|\eta|\right)\right]/a_2, & \text{para } a_2 < 0 \end{cases} . \tag{2.13}$$

Comumente utiliza-se uma abordagem bayesiana para ajustar o modelo da Stukel.

Probito Assimétrica

Chen et al [6] definem a probito assimétrica (skew-probit) que tem como caso particular o modelo probito. O modelo segue de

$$\pi = \Phi_{SN}(X\beta; \mu, \sigma^2, \lambda), \tag{2.14}$$

em que $\Phi_{SN}(.; \mu, \sigma^2, \lambda)$ é a função densidade acumulada da normal assimétrica. Chen et al. [6] sugere $\mu = 0$, $\sigma^2 = 1 + \lambda^2$ e $-\lambda$ para o parâmetro de assimetria. Para $\lambda = 0$ tem-se modelo probito.

Neste caso uma abordagem bayesiana é sugerida para ajustar o modelo probito-assimétrico. Deve-se ressaltar que o tempo necessário para calcular a estimativa dos parâmetros para 3 covariáveis no exemplo apresentado por Chen et al. [6] foi de 45 minutos calculando 20000 interações utilizando o software Gibbs.

No próximo capítulo apresenta-se uma nova função de ligação flexível que acomoda vários modelos.

Capítulo 3

Regressão Binária Weibull

Métodos de regressão são fundamentais em análises de dados onde há interesse em descrever o relacionamento entre uma variável resposta e uma ou mais variáveis explicativas. Neste capítulo propõe-se uma função de ligação baseada na função densidade acumulada Weibull. A Figura 3.1 mostra a flexibilidade da distribuição Weibull para diferentes valores do parâmetro de forma (γ). Dá-se, na seção 3.1, a construção da função de ligação Weibull. Na seção 3.2, aborda-se o procedimento de estimação dos parâmetros do modelo. Discute-se o *Odds* e *Odds Ratio* (Razão de Chances) na seção 3.3. O procedimento do teste da razão de verossimilhanças é dado na seção 3.4 e, na seção 3.5, algumas medidas de qualidade de ajuste do modelo.

3.1 Função de Ligação Weibull

Considere uma variável aleatória W de uma distribuição Weibull com três parâmetros. Sua função densidade acumulada é dada por

$$F_W(w) = 1 - \exp\left\{-\left(\frac{w-\mu}{\lambda}\right)^{\gamma}\right\}, \tag{3.1}$$

em que $w > \mu$, $\mu \in \mathbb{R}$, $\lambda > 0$ e $\gamma > 0$.

Agora, suponha que $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ é um vetor de n variáveis aleatórias binárias independentes e $\mathbf{X} = (X_1, \ldots, X_k)$ é uma matriz composta por k vetores de tamanho n. O vetor $\mathbf{Y}_{n\times 1}$ é a variável resposta e a matriz $\mathbf{X}_{n\times k}$ contém as

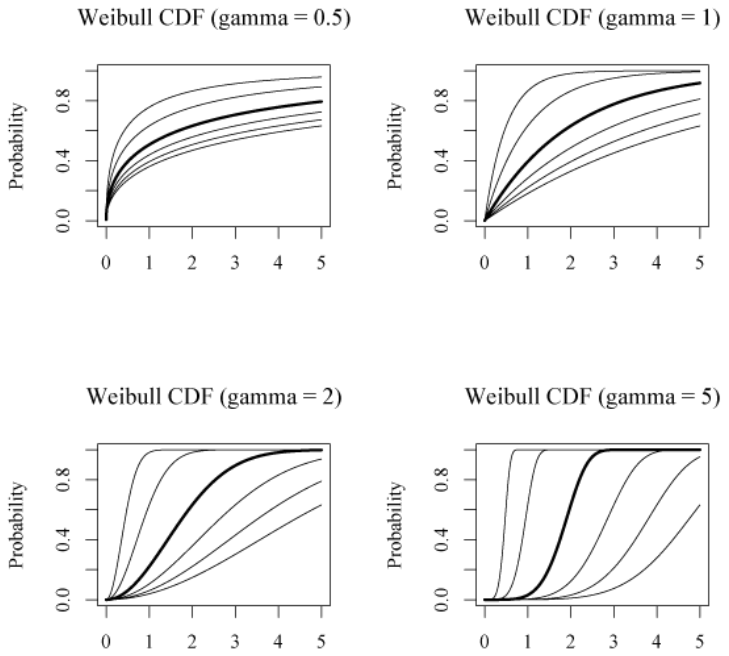

FIGURA 3.1: Formas da distribuição Weibull.

variáveis explicativas. Define-se então X_i, $i = 1, \ldots, n$, como a i-ésima linha da matriz $\mathbf{X}$, ou seja, X_i é um vetor de variáveis explicativas de tamanho k relacionadas a variável resposta Y_i. Suponha também, que $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_k^*)$ é um vetor de parâmetros. Tomando $W_i = \mu^* + \boldsymbol{\beta}^* X_i'$, tem-se

$$\begin{aligned}
\Pr(Y_i = 1 \mid W_i = \mu^* + \boldsymbol{\beta}^* x_i') &= F_W(\mu^* + \boldsymbol{\beta}^* x_i') \\
&= 1 - \exp\left\{ - \left(\frac{\mu^* + \boldsymbol{\beta}^* x_i'}{\lambda} \right)^{\gamma} \right\} \\
&= 1 - \exp\left\{ - \left(\frac{\mu^* + \beta_1^* x_{i1} + \cdots + \beta_k^* x_{ik}}{\lambda} \right)^{\gamma} \right\} \\
&= 1 - \exp\left\{ - \left(\frac{\mu^*}{\lambda} + \frac{\beta_1^*}{\lambda} x_{i1} + \cdots + \frac{\beta_k^*}{\lambda} x_{ik} \right)^{\gamma} \right\} \\
&= 1 - \exp\left\{ - \left(\mu + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \right)^{\gamma} \right\} \\
\Pr(Y_i = 1 \mid X_i = x_i) &= 1 - \exp\left\{ - \left(\mu + \boldsymbol{\beta} x_i' \right)^{\gamma} \right\}, \qquad (3.2)
\end{aligned}$$

em que $\mu + \boldsymbol{\beta} x_i' > 0$, $\gamma > 0$, $\mu = \frac{\mu^*}{\lambda}$ e $\boldsymbol{\beta} = \left(\frac{\beta_1^*}{\lambda}, \ldots, \frac{\beta_k^*}{\lambda} \right)$.

Para definir a função de ligação considera-se que π_i é a proporção de $Y_i = 1$. Então iguala-se π_i com $F_W(\mu + \boldsymbol{\beta} x_i')$, e obtém-se a função de ligação

$$\pi_i = 1 - exp\left\{ - \left(\mu + \boldsymbol{\beta} x_i' \right)^{\gamma} \right\}, \qquad (3.3)$$

em que $\mu + \boldsymbol{\beta} x_i' > 0$ e $\gamma > 0$. A Equação (3.3) pode ser escrita em termos de π_i como

$$g(x) = ln\left[\frac{1}{1-\pi_i}\right] = (\mu + \boldsymbol{\beta} x_i')^{\gamma} \tag{3.4}$$

A motivação de propor esta função de ligação foi a de obter um modelo mais geral do que os modelos mais usados atualmente (logito, probito, complementar log-log) e que não fosse muito complicado calcular os estimadores dos parâmetros.

3.2 Interpretação dos parâmetros

Note, na Equação (3.1), que usa-se a distribuição Weibull com três parâmetros e após as simplificações, como apresentado na Equação (3.2), precisa-se apenas estimar os parâmetros μ, $\boldsymbol{\beta}$ e γ. No entanto, é importante mencionar que a interpretação dos parâmetros $\boldsymbol{\beta}$ não é igual a interpretação no modelo logito ou em modelos de regressão. Isto ocorre porque o parâmetro de locação μ e todos os $\boldsymbol{\beta}$'s são divididos pelo parâmetro de escala λ. Isto foi feito para simplificar o modelo proposto, uma vez que o objetivo principal é estimar a frequência π de cada observação usando implicitamente a distribuição Weibull com três parâmetros. Vale ressaltar que a transformação realizada nos parâmetros não altera as propriedades da distribuição Weibull.

Se o parâmetro β_j, $j = 1, \ldots, k$, for positivo e a variável X_j associada a ele aumentar seu valor, aumentará o valor da proporção π. No caso do β_j ser negativo ocorre o efeito contrário, isto é, se X_j aumenta π diminui.

Na próxima seção, apresenta-se o processo de estimação baseado em máxima verossimilhança.

3.3 Estimação dos parâmetros

Primeiro, descreve-se a função de verossimilhança relacionada ao problema. Quando Y_i é uma variável binária e a proporção de $Y_i = 1$ é π_i, então

assume-se que Y_i tem distribuição Bernoulli com probabilidade de sucesso igual a π_i, isto é, $Y_i \sim \text{Bernoulli}(\pi_i)$. Considerando que Y_i, $i = 1, \ldots, n$ são variáveis aleatórias independentes e que observamos $\mathcal{D} = \{n, \mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}\}$, podemos escrever a função de verossimilhança como

$$\begin{aligned} L(\beta, \gamma \mid \mathcal{D}) &\propto \prod_{i=1}^{n} {\pi_i}^{y_i} (1-\pi_i)^{1-y_i} \\ &\propto \prod_{i=1}^{n} \{1 - exp[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}]\}^{y_i} \{exp[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}]\}^{1-y_i} \end{aligned} \quad (3.5)$$

e a log-verossimilhança como

$$\begin{aligned} l(\beta, \gamma \mid \mathcal{D}) &\propto \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\} \\ &\propto \sum_{i=1}^{n} \Bigg\{ y_i \log\{1 - exp[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}]\} + \\ &\qquad (1-y_i)\log\{exp[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}]\} \Bigg\} \\ &\propto \sum_{i=1}^{n} \{y_i \log\{1 - exp[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}]\} + (y_i - 1)(\mu + \boldsymbol{\beta} x_i)^{\gamma}\}. \end{aligned} \quad (3.6)$$

Para encontrar os estimadores de máxima verossimilhança deve-se derivar a Equação (3.6) em relação a cada um dos parâmetros de interesse. Assim obtêm-se as Equações (3.7), (3.8) e (3.9). Igualando as derivadas a zero, encontram-se as condições de 1ª ordem para a obtenção dos estimadores de máxima verossimilhança para os parâmetros do modelo. Porém, não obtém-se uma solução algébrica para este sistema de equações, necessitando assim a utilização de métodos numéricos para a obtenção dos estimadores.

As derivadas da Equação (3.6) são dadas por

$$\frac{\delta l}{\delta \mu} = \sum_{i=1}^{n} \left\{ \frac{y_i \gamma (\mu + \boldsymbol{\beta} x_i')^{\gamma-1} [y - \pi(x_i')]}{\pi(x_i')} \right\}, \quad (3.7)$$

$$\frac{\delta l}{\delta \beta_i} = \sum_{i=1}^{n} \left\{ \frac{y_i \gamma x_i' (\mu + \boldsymbol{\beta} x_i')^{\gamma-1} [y - \pi(x_i')]}{\pi(x_i')} \right\}, \quad (3.8)$$

$$\frac{\delta l}{\delta \gamma} = \sum_{i=1}^{n} \left\{ \frac{y_i ln[\mu + \boldsymbol{\beta} x_i'](\mu + \boldsymbol{\beta} x_i')^{\gamma} [y - \pi(x_i')]}{\pi(x_i')} \right\}, \quad (3.9)$$

em que

$$\pi(x_i') = 1 - exp\left[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}\right]. \tag{3.10}$$

No software SAS 9.2 [18] utilizamos a *procedure* NLP para obter as estimativas de máxima verossimilhança. O método numérico de otimização é o Newton-Rapson Ridge, indicado pelo comando "*tech=nrridg*".

A técnica Newton-Rapson Ridge (NRRIDG) usa o gradiente e a matriz Hessiana e, portanto, requer que a função objetivo tenha derivadas contínuas de primeira e segunda ordem dentro da região plausível.

Este algorítmo usa um passo puramente de Newton quando a matriz Hessiana é positiva definida e quando reduz o valor da função objetivo com sucesso. Se pelo menos uma dessas duas condições não forem satisfeitas, um múltiplo da matriz identidade é adicionado a matriz Hessian.

A seguir um código básico para o ajuste de um modelo binário weibull.

```
proc nlp data=DATA cov=2 vardef=n tech=nrridg;
max fx;
parms mu, b1, bn, k;
bounds 0 < k ;
y = RESP;
xb = mu + b1*X1 + bn*Xn ;
p = 1 - exp(-(xb)**k) ;
fx = y*log(p)+(1-y)*log(1-p);
run;
```

Adapta-se o código a qualquer conjunto de dados substituindo "DATA" pelo nome do conjunto de dados; "RESP" pelo nome da variável resposta binária; "X1" e "Xn" pelas variáveis explicativas. Observe que não importa o número de parâmetros e covariáveis no modelo, desde que o nome de todos os parâmetros sejam indicados ao lado do comando *parms* separados por vírgula e que o componente "xb" esteja escrito como uma combinação linear das variáveis explicativas.

3.4 *Odds* e *Odds Ratio*

Em alguns casos é importante descrever a probabilidade de sucesso de uma variável em termos da *odds* (chance) daquele evento. A *odds* de um sucesso é definido como a razão da probabilidade de sucesso pela probabilidade de falha. Portanto se p é a verdadeira probabilidade de sucesso, a *odds* de um sucesso é $p/(1-p)$. No contexto deste trabalho, o sucesso é uma característica de interesse da variável resposta (por exemplo, Y=1) e o interesse está em estimar a chance deste sucesso ocorrer em relação ao fracasso (se sucesso é Y=1, então o fracasso é representado por Y=0). Considerando a estimação da proporção de sucesso $P(Y_i = 1) = \pi(x_i')$, dada na Equação (3.10), pode-se estimar a chance de sucesso por

$$\pi(x_i')/[1 - \pi(x_i')] = \frac{1 - exp\left[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}\right]}{exp\left[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}\right]} = exp\left[-(\mu + \boldsymbol{\beta} x_i')^{\gamma}\right] - 1. \quad (3.11)$$

Quando dois conjuntos são comparados, uma medida que relaciona a chance de sucesso de um conjunto em relação a outro é chamada de *odds ratio* (razão de chances). Suponha que p_1 e p_2 são as probabilidades de sucesso em dois conjuntos, então a *odds* de um sucesso no r-ésimo conjunto é $p_r/(1-p_r)$, r = 1, 2 e a *odds ratio* de um sucesso de um conjunto em relação a outro é denotada por

$$\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \quad (3.12)$$

Dado o método de estimação por máxima verossimilhança e o processo numérico para a obtenção dos estimadores descreve-se, a seguir, como calcular a *odds* e a *odds ratio* utilizando o modelo Weibull.

Considerando o ajuste de um modelo Weibull a um problema com a variável resposta (Y) ocorrência de tumor em ratos (sim ou não) e a variável explicativa (X) exposição de fumaça de cigarro (sim ou não), pode-se escrever a *odds ratio* da ocorrência de tumor entre ratos expostos à fumaça de cigarro e não expostos a fumaça por

$$\psi = \frac{exp\left[(\mu + \beta_1 x)^{\gamma}\right] - 1}{exp\left[(\mu)^{\gamma}\right] - 1}. \quad (3.13)$$

Quando a chance de um sucesso em cada um dos conjuntos for idêntica, ψ é igual a 1. Isto acontece quando as duas probabilidades de sucesso são iguais.

Valores de ψ menores que 1 sugerem que a chance de um sucesso é menor no primeiro conjunto do que no segundo, enquanto que uma *odds ratio* maior que 1 indica que a *odds* de um sucesso é maior no primeiro conjunto de dados. A *odds ratio* é uma medida da diferença entre duas probabilidades de sucesso que pode assumir qualquer valor positivo, ao contrário da diferença entre duas probabilidade de sucesso, $p_1 - p_2$, que é restringida no intervalo (-1,1).

Na seção seguinte, é abordado o problema da significância dos parâmetros do modelo através do procedimento do teste da razão de verossimilhanças.

3.5 Teste de significância dos parâmetros: Razão de Verossimilhanças

Após o cálculo dos estimadores deve-se avaliar se as variáveis incluídas no modelo são significativas em relação à variável resposta. Para isto, é necessário formular e testar uma hipótese estatística. A execução deste teste é feita a partir de um método geral que varia de um modelo para outro apenas em seus detalhes específicos.

Uma aproximação para testar a significância do coeficiente de uma variável em qualquer modelo é dada através da comparação dos valores observados na variável resposta para os dois modelos (com e sem a variável em questão). Utiliza-se o teste da razão de verossimilhanças para testar as hipóteses H_0: $\beta_j = 0$ para j = 1, 2, ...,k e H_0: $\gamma = 1$ sob cada hipótese definida. Se os valores preditos são melhores quando a variável está presente isso é um indício de que obtém-se uma variável significante.

É necessário comparar os valores observados da variável resposta com valores preditos a partir de modelos com e sem a variável em questão. Para isso utiliza-se na comparação o log da função de verossimilhança. A comparação dos valores observados com os valores preditos é realizada através da razão de verossimilhanças na expressão (3.14).

$$R = -2ln\left[\frac{L\left(\hat{\beta};\mathcal{D}\right)}{L\left(\beta^0;y\right)}\right], \tag{3.14}$$

em que $L\left(\hat{\beta};\mathcal{D}\right)$ é a função de verossimilhança do modelo ajustado sobre o conjunto de dados $\mathcal{D} = \{y, X\}$ e $L\left(\beta^0;y\right)$ é a função de verossimilhança do modelo saturado, ou seja, do modelo em que a estimativa da proporção $\hat{\pi}$ é a variável resposta y. A estatística R, possui distribuição assintótica qui-quadrado com graus de liberdade dependendo da hipótese formulada. Da Equação (3.6), e da (3.14) tem-se que

$$R = -2ln\left[\frac{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}{\prod_{i=1}^{n}y_i^{y_i}(1-y_i)^{(1-y_i)}}\right]. \tag{3.15}$$

Neste caso, como a variável resposta é binária, a verossimilhança do modelo saturado é igual a 1. Segue da definição de modelo saturado que $\hat{\pi}_i = y_i$ e, assim, a função de verossimilhança é

$$L\left(\beta^0;y\right) = \prod_{i=1}^{n}y_i^{y_i}(1-y_i)^{(1-y_i)} = 1. \tag{3.16}$$

Sendo assim,

$$R = -2ln\left[L\left(\hat{\beta};\mathcal{D}\right)\right]. \tag{3.17}$$

Para testar a significância da variável independente deve-se comparar o valor de R em (3.17) para o modelo com e sem a variável independente na equação. A mudança em R devido à inclusão da variável independente no modelo é obtida por

$$S = R(\textit{modelo sem a variável}) - R(\textit{modelo com a variável}). \tag{3.18}$$

Como a função de verossimilhança do modelo saturado possui a mesma forma para ambos os valores de R, pode-se reescrever S como

$$S = -2ln\left[\frac{L\left(\hat{\beta}_{-i};\mathcal{D}\right)}{L\left(\hat{\beta};\mathcal{D}\right)}\right], \tag{3.19}$$

em que $L\left(\hat{\beta};\mathcal{D}\right)$ é a função de verossimilhança do modelo com todas as variáveis em estudo e $L\left(\hat{\beta}_{-i};\mathcal{D}\right)$ é a função de verossimilhança do modelo com todas as variáveis exceto a variável a ser testada (X_i). Sob a hipótese H_0: $\beta_i = 0$, a estatística S dada em (3.19) segue assintoticamente uma distribuição Qui-quadrado com 1 grau de liberdade.

A seguir são dadas algumas medidas para avaliar a qualidade do ajuste do modelo.

3.6 Medidas de qualidade do ajuste para o modelo binário Weibull

Após o ajuste de um modelo a um conjunto de dados é natural comparar as estimativas do modelo com os valores observados. Se os valores estimados são próximos dos valores observados podemos dizer que o modelo é aceitável. Caso contrário, certamente o modelo não será aceito e precisará ser revisado. Os aspectos de adequabilidade de um modelo são conhecidos como **qualidade do ajuste** (*goodness of fit*).

Existem várias métodos estatísticos para medir a discrepância entre a proporção binomial observada, y_i/n_i, e a proporção ajustada, $\hat{p}_i$. Dentre estes métodos o mais usado é baseado na função de verossimilhança.

3.6.1 AIC

O *Akaike's information criterion* (AIC), desenvolvido por Hirotsugu Akaike [1], é uma medida de qualidade de um modelo estatístico. O AIC é uma função da log-verossimilhança maximizada. Não há teste de hipótese para medida AIC, mas há uma medida para comparar diferentes modelos. Dado um conjunto de dados, vários modelos competitivos podem ser ordenados pelo seus AIC, assim o modelo que tiver o menor valor de AIC é o melhor.

No caso geral, o AIC é

$$AIC = 2 * k - 2 * ln\left(L\left(\hat{\beta}; \mathcal{D}\right)\right), \tag{3.20}$$

em que k é o número de parâmetros.

Aumentando o número de parâmetro no modelo melhora-se a qualidade do ajuste. Assim o AIC não apenas recompensa a qualidade do ajuste como inclui uma penalidade que é uma função crescente do número de parâmetros estimados. Esta penalidade desencoraja um modelo com mais parâmetros que o necessário. A metodologia AIC tenta achar um modelo que melhor explica os dados com um menor número de parâmetros.

3.6.2 *Deviance*

O *Deviance* mede o quão distante o modelo está dos dados, comparando o modelo ajustado com o modelo saturado. Se o número de sucessos estimado pelo modelo é dado por $\hat{y}_i = n_i\hat{p}_i$, a estatística D, deviance, pode ser escrita como

$$D = 2\sum_i \left\{ y_i log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right\} \tag{3.21}$$

Assim, observa-se que esta estatística compara as observações y_i com seus correspondentes valores $\hat{y}_i$ obtidos pelo modelo ajustado. De acordo com Cohen [7] o *deviance* é distribuido assintóticamente como qui-quadrado com $(n - k)$ graus de liberdade, onde n é o número de observações binomiais e k é o número de parâmetros desconhecidos incluídos no modelo.

3.6.3 Estatística X^2 de Pearson

Uma estatística muito utilizada para avaliar a qualidade do ajuste é a estatística X^2 de Pearson [13] definida por

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1 - \hat{p}_i)}. \tag{3.22}$$

Tanto o *deviance* quanto esta estatística X^2 tem a mesma distribuição assintótica qui-quadrado com (n-k) graus de liberdade. Os valores numéricos das duas estatísticas diferem mas raramente isto tem importância prática. Grandes diferenças entre as duas estatísticas podem ser tomadas como um indicador de que a aproximação qui-quadrado não é adequada para o deviance ou a estatística X^2.

Uma vez dado o modelo Weibull, um método de estimação, teste de significância para o parâmetro, estatísticas para avaliar a qualidade do ajuste, na próxima seção apresenta-se alguns exemplos do uso deste novo modelo.

Capítulo 4

Estudo de Casos

Foram usados três conjuntos de dados reais para mostrar algumas aplicações da nova função de ligação. Inicialmente, na seção 4.1, aborda-se o conjunto de dados sobre a *Mortalidade de Besouros* expostos a disulfeto de carbono gasoso (Bliss [3]) para mostrar que a função de ligação complementar log-log é um caso particular da função de ligação Weibull; na seção 4.2, o conjunto de dados sobre a ocorrência do início da menstruação nas *Garotas de Varsóvia* (Milicer e Szczotka [10]) é usado para mostrar que uma pequena adaptação a função de ligação Weibull permite que sua adequabilidade a diferentes conjuntos de dados seja mais razoável; e, por fim, na seção 4.3, um conjunto sobre a *Mutação no DNA de caramujos* expostos à radiação gama (Grazeffe et al. [9]) para exemplificar uma aplicação a dados com resposta multinomial.

4.1 Mortalidade de Besouros

Considerando a função de ligação complementar log-log, dada na Equação (2.6), mostra-se que esta pode ser escrita como um caso limite da função de ligação Weibull. Da Equação (3.1) tem-se que:

$$\pi = 1 - \exp\left\{-\left(\frac{w-\mu}{\lambda}\right)^{\gamma}\right\}, \tag{4.1}$$

e tomando $-\mu = \lambda = \gamma = k$, obtém-se

$$\pi = 1 - \exp\left\{-\left(\frac{w+k}{k}\right)^k\right\} = 1 - \exp\left[-\left(1+\frac{w}{k}\right)^k\right]. \tag{4.2}$$

Aplicando o limite de $k \to \infty$ em (4.2) tem-se

$$\lim_{k\to\infty}\left\{1 - \exp\left[-\left(1+\frac{w}{k}\right)^k\right]\right\} = 1 - \exp\left[-\exp\left(w\right)\right], \tag{4.3}$$

que é a função complementar log-log dada em (2.6).

Considerando o problema apresentado por Bliss [3], *Mortalidade de Besouros*, é conhecido que o modelo complementar log-log é bem adequado. Ajusta-se a este conjunto de dados agora o modelo Weibull, o qual produz resultados muito semelhantes ao do modelo complementar log-log.

O objetivo de Bliss [3] é obter um inseticida eficaz contra besouros. Foi realizado um experimento com 481 besouros para verificar a taxa de mortalidade destes insetos adultos após a exposição de 5 horas de gás carbono disulfídio (CS_2). Um total de 291 besouros morreram no experimento. Este conjunto de dados é mostrado nas três primeiras colunas da Tabela 4.1, sendo a proporção de besouros mortos a variável resposta e o log da dose de CS_2 a variável explicativa. Este conjunto é frequentemente citado por representar um problema em que os modelos logito probito não são adequados. Isso ocorre pois estes modelos não comportam uma relação assimétrica, como é o caso da complementar log-log e da Weibull.

TABELA 4.1: Mortalidade de besouros expostos a disulfeto de carbono gasoso

log(Dose)	Nº de besouros		Estimativa		
CS_2	Expostos	Mortos	Weibull	Logito	Probito
1.6907	59	6	5.59	3.45	3.27
1.7242	60	13	11.28	9.84	10.89
1.7552	62	18	20.96	22.45	23.65
1.7842	56	28	30.37	33.89	33.88
1.8113	63	52	47.78	50.10	49.60
1.8369	59	53	54.14	53.29	53.28
1.861	62	61	61.11	59.22	59.63
1.8839	60	60	59.95	58.74	59.21

Assim, aplica-se a função de ligação definida em (3.3) e utiliza-se o algoritmo de Newton-Raphson Ridge, disponível no software SAS [18], para obtenção dos estimadores dos parâmetros. Apresenta-se na Tabela 4.2 suas estimativas e a estimativa da matriz de variâncias e covariâncias dos parâmetros na Equação (4.4). Nas três últimas colunas da Tabela 4.1 as estimativas do número esperado de besouros mortos tanto do modelo binário Weibull quanto dos modelos logito e probito são apresentadas. A estatística razão de verossimilhança de *goodness-of-fit* em relação ao modelo logito é $\chi^2_1 = 7.78$, indicando que o modelo Weibull tem melhor ajuste em relação ao modelo logito. Os modelos assimétricos propostos por Prentice [15], Aranda-Ordaz [2] e Stukel [19] indicam resultados similares. Um resumo com a log-verossimilhança dos outros modelos propostos é mostrado na Tabela 4.3.

O valor da função de log-verossimilhança maximizada é de -182.34, que coincide com o valor obtido pelo modelo complementar log-log e é muito próximo ao reportado por Prentice [15]. Aranda Ordaz [2] mostrou que a transformação complementar log-log é a mais apropriada para estes dados. De fato, o ajuste alcançado pelo modelo Weibull mostra uma boa predição dos dados, conforme mostra a Figura 4.1.

A linha tracejada representa o modelo logito, a pontilhada o probito e a linha contínua o modelo Weibull. Os dados são indicados pelos losângulos.

TABELA 4.2: Estimativas dos parâmetros - Mortalidade de Besouros

Parâmetro	Estimativa	g.l.	Qui-quadrado	Pr > Qui
μ	0.83993			
β_1	0.08915	1	280.744	< 0.0001
γ	247.049	1	56.916	< 0.0001

$$V_{\hat{\mu},\hat{\beta}_1,\hat{\gamma}} = \begin{bmatrix} 1.46 * 10^{-5} & -8.08 * 10^{-6} & -1.05 * 10^{-9} \\ -8.08 * 10^{-6} & 4.49 * 10^{-6} & 5.85 * 10^{-10} \\ -1.05 * 10^{-9} & 5.85 * 10^{-10} & 7.63 * 10^{-14} \end{bmatrix}. \quad (4.4)$$

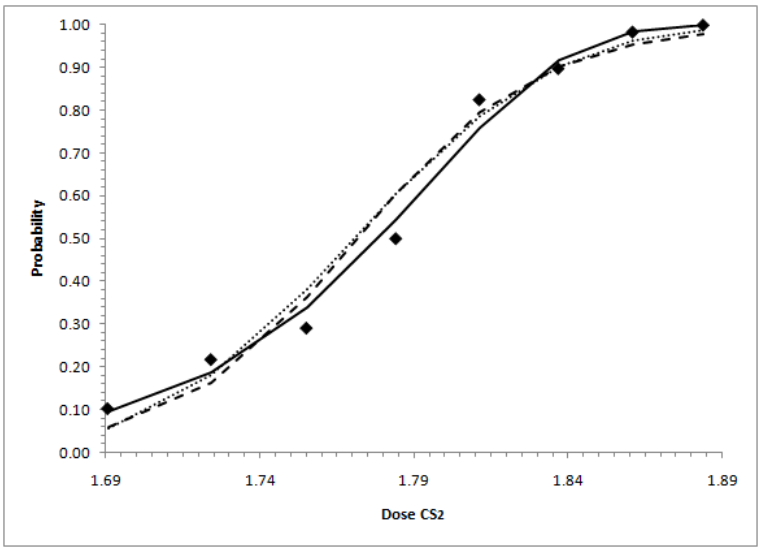

FIGURA 4.1: Dados observados (Mortalidade de Besouros) e curva das probabilidades ajustadas.

TABELA 4.3: Comparação das Log-verossimilhanças - Mortalidade de Besouros

Modelo	Log-verossimilhança	AIC	X^2 de Pearson	*Deviance*
Prentice	-182.25	370.5	3.06	1.42
C.Log-log	-182.34	368.68	3.25	1.48
Aranda-Ordaz	-182.34	370.68	3.25	1.48
Weibull	-182.34	370.68	3.29	1.49
Stukel	-182.68	371.36	5.48	3.06
Probito	-185.85	375.7	9.86	4.54
Logito	-186.23	376.48	10.04	4.88

Observa-se a partir da Tabela 4.3 que os modelos de Prentice, Weibull, Aranda-Ordaz e Complementar log-log possuem verossimilhanças praticamente idênticas. Isto indica que estes modelos estão se aproximando muito bem do modelo complementar log-log que possui menor AIC, por ter menos parâmetros. Por outro lado, os modelos Logito e Probito não possuem um bom ajuste.

4.2 Garotas de Varsóvia

Para tratar o problema das Garotas de Varsóvia utiliza-se uma pequena alteração na função de ligação Weibull. Então, primeiramente, introduz-se esta transformação e na sequência apresenta-se o problema.

Considere a distribuição Weibull Refletida, dada por Cohen [7], originada de uma transformação linear da variável W, com distribuição Weibull.

$$\text{V - } \mu = \text{- (W - } \mu) = \mu \text{ - W.}$$

Isto nos leva a uma reflecção da distribuição Weibull no eixo vertical em w = μ resultando na função densidade acumulada de V, dada por

$$F_V(v) = \exp\left\{-\left(\frac{\mu - v}{\lambda}\right)^{\gamma}\right\}, \tag{4.5}$$

em que $v < \mu$, $\mu \in \mathbb{R}$, $\lambda > 0$ e $\gamma > 0$.

Fazendo $V_i = \mu^* + \boldsymbol{\beta}^* X_i'$ e igualando π_i com $F_V(\mu + \boldsymbol{\beta} x_i')$ obtém-se, analogamente ao demonstrado na seção 3.1, a função de ligação

$$\pi_i = exp\left\{-(\mu + \boldsymbol{\beta} x_i')^{\gamma}\right\}, \tag{4.6}$$

em que $\mu + \boldsymbol{\beta} x_i' > 0$, $\gamma > 0$, $\mu = \frac{\mu^*}{\lambda}$ e $\boldsymbol{\beta} = \left(\frac{-\beta_1^*}{\lambda}, \ldots, \frac{-\beta_k^*}{\lambda}\right)$. A Equação (4.6) pode ser escrita em termos de π_i como

$$g(x) = ln\left[\frac{1}{\pi_i}\right] = (\mu + \boldsymbol{\beta} x_i')^{\gamma} \tag{4.7}$$

Note que esta transformação equivale, na prática, a trocar os valores de zeros por uns e de uns por zeros da variável resposta. Pode-se visualizar isto na Figura 4.2 que para diferentes valores de γ mostra a função densidade acumulada Weibull, linha contínua, e Weibull Refletida, linha tracejada.

Dada a transformação na função de ligação Weibull, estuda-se o problema das Garotas de Varsóvia. Milecer e Szczotka [10] investigaram a idade da primeira menstruação em 3918 garotas de Varsóvia. Para 25 médias de idade observou-se a ocorrência (Y = 1) ou não (Y = 0) do início de períodos de menstruação nas adolescentes. Verificou-se que um total de 2308 garotas haviam entrado

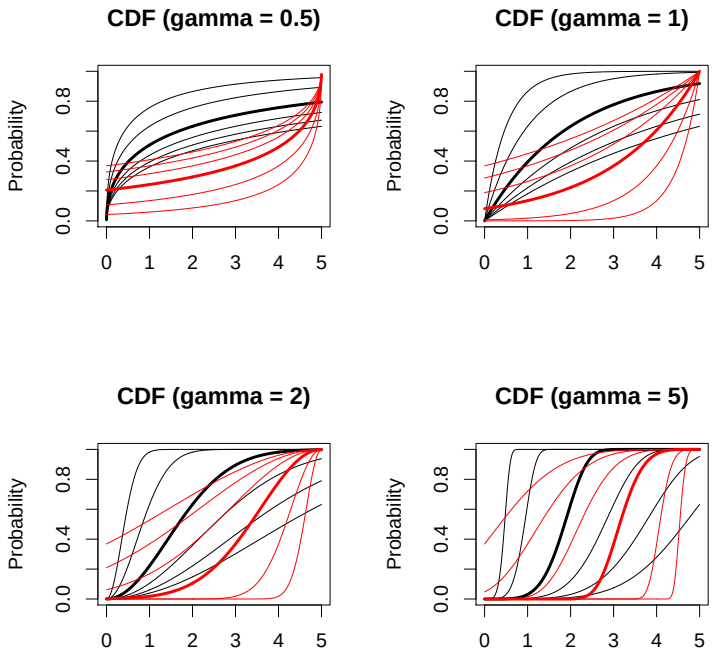

FIGURA 4.2: Formas da distribuição Weibull e Weibull Refletida.

no período de mestruação até o fim do experimento. O interesse é estimar a proporção de adolescentes que iniciaram o período de mestruação utilizando como variável explicativa a idade média em 12 categorias. Os dados são apresentados na Tabela 4.4

TABELA 4.4: Ocorrência do início da menstruação em garotas de Varsóvia.

	Número de garotas			Número de garotas	
Idade	Menstruadas	Entrevistadas	Idade	Menstruadas	Entrevistadas
9,21	0	376	13,08	47	99
10,21	0	200	13,33	67	106
10,58	0	93	13,58	81	105
10,83	2	120	13,83	88	117
11,08	2	90	14,08	79	98
11,33	5	88	14,33	90	97
11,58	10	105	14,58	113	120
11,83	17	111	14,83	95	102
12,08	16	100	15,08	117	122
12,33	29	93	15,33	107	111
12,58	39	100	15,58	92	94
12,83	51	108	15,83	112	114
			17,53	1049	1049

Foi utilizado neste conjunto de dados a função de ligação baseada na distribuição Weibull refletida, definida na Equação (4.6). As estimativas dos parâmetros tal como seus testes de significância são mostrados na Tabela 4.5. A estimativa da matriz de variâncias e covariâncias dos parâmetros é dada na Equação (4.8).

TABELA 4.5: Estimativas dos parâmetros - Garotas de Varsóvia

Parâmetro	Estimativa	g.l.	Qui-Quadrado	Pr > Qui
μ	-2.969			
β_1	0.157	1	3679.8	< 0.0001
γ	6.269	1	2337.6	< 0.0001

$$V_{\hat{\mu},\hat{\beta}_1,\hat{\gamma}} = \begin{bmatrix} 0.157 & -0.012 & 0.540 \\ -0.012 & 0.001 & -0.043 \\ 0.540 & -0.043 & 1.906 \end{bmatrix}. \tag{4.8}$$

De acordo com Stukel [19] a estatística de razão de verossimilhança para verificar a qualidade do ajuste do modelo logito é $\chi^2_{23} = 26.7$; porém um exame dos resíduos mostra que um melhor ajuste pode ser feito nas caudas.

A estatística para testar o ajuste do modelo com a função de ligação Weibull em relação ao modelo logístico é $\chi^2_1 = 12.66$ (p = 0.0004). A melhora do ajuste pode ser percebido, particularmente na cauda inferior, como mostra Figura 4.3.

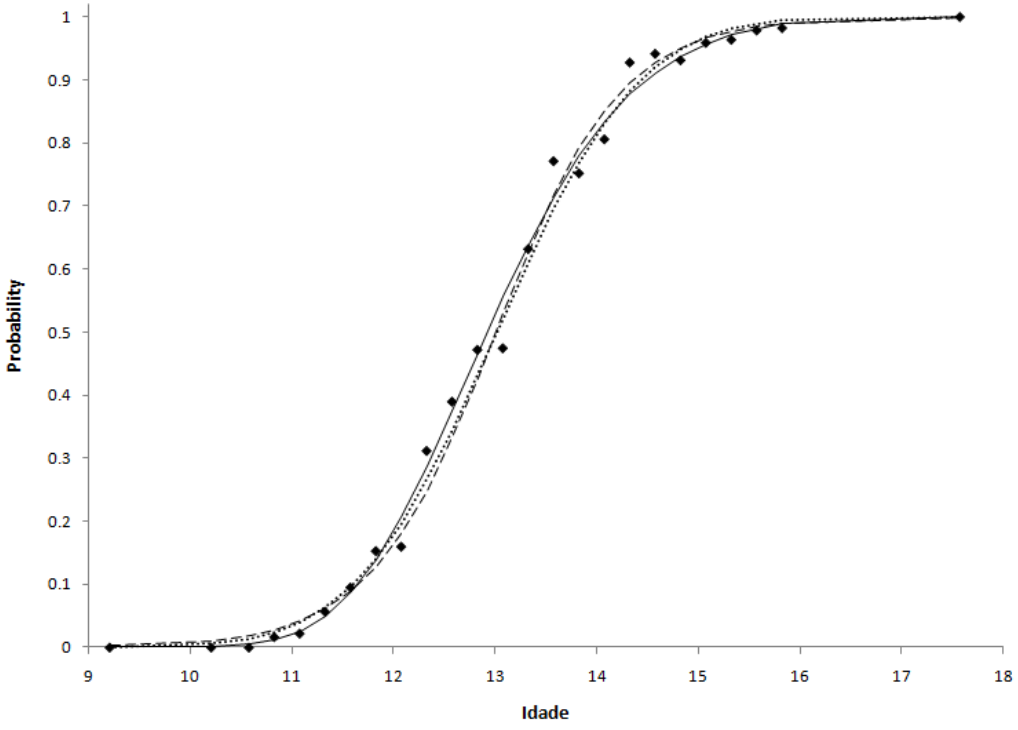

FIGURA 4.3: Dados observados (Garotas de Varsóvia) e curva das probabilidades ajustadas.

Observe que os ajustes das curvas logito (linha tracejada) e probito (em pontilhado) não ajustam tão bem as caudas inferior e superior em relação ao ajuste dado pela curva Weibull (linha contínua).

Um resumo com a log-verossimilhança dos outros modelos propostos é mostrado na Tabela 4.6.

TABELA 4.6: Comparação das Log-verossimilhanças - Garotas de Varsóvia

Modelo	Log-verossimilhança	AIC	X^2 de Pearson	*Deviance*
Weibull	-813.32	1632.64	12.99	6.10
Stukel	-814.5	1635	15.01	8.13
Prentice	-815.77	1637.54	15.92	8.23
Aranda-Ordaz	-817.58	1641.16	21.78	9.86
Probito	-817.74	1639.48	21.90	9.94
Logito	-819.65	1643.3	21.86	11.59
C.Log-log	-866.65	1737.3	95.73	52.42

De acordo com a Tabela 4.6 conclui-se que o modelo Weibull é o modelo que faz o melhor ajuste, pois este apresenta maior valor da função de log-verossimilhança maximizada e menor valor de AIC, X^2 de Pearson e *Deviance*.

4.3 Mutação no DNA de caramujos

Este exemplo foi apresentado por Grazeffe et al. [9] que em seu trabalho usaram um ajuste por regressão logística. Aqui trabalha-se com o mesmo problema utilizando a função de ligação Weibull e comparando com os ajustes obtidos por Grazeffe et al. [9]. Como o objetivo aqui é exemplificar um problema com variável resposta multinomial utilizando a função de ligação Weibull e devido as dificuldades de estimação dos modelos já citados, compara-se os resultados apenas com a regressão logística.

As águas são o destinatário da maior parte das substâncias químicas produzidas nas indústrias. Essas substâncias genotóxicas podem causar mutações no DNA de seres aquáticos, uma perda da diversidade ecológica e afetar na capacidade reprodutiva das espécies, o que pode acarretar em um grande problema na cadeia alimentar.

O Ensaio Cometa (EC) é um teste realizado em células individuais que detecta danos causados por substâncias genotóxicas. Essa técnica vem sendo amplamente utilizada, pois detecta muitos danos no DNA e tem algumas vantagens

sobre os outros métodos existentes, dentre estas está o fato de não ser necessário utilizar muitas células para a sua aplicação.

O caramujo de água doce foi escolhido para a realização deste estudo. Foram utilizados caramujos da mesma espécie, criados em laboratório com pelo menos 2 meses de idade e 10mm de diâmetro. Esses caramujos foram expostos a radiações ionizantes, em diferentes dosagens, sendo elas: 0, 2.5, 5, 10 e 20. As mutações foram classificadas como mostra a Tabela 4.7. A frequência absoluta e relativa (%) do número de células em cada classe de mutação no DNA estão dispostos na Tabela. 4.8

TABELA 4.7: Níveis de Mutação no DNA

Nível	**Legenda**
C0	Não houve migração no DNA
C1	O DNA migrou pouco
C2	O DNA migrou de forma intermediária
C3	O DNA migrou muito

TABELA 4.8: Mutação no DNA de caramujos expostos à radiação gama

Dose de	Classes de Mutação no DNA (%)				Número de
radiação	C0	C1	C2	C3	células (%)
0	654 (59.5)	125 (11.4)	72 (6.5)	249 (22.6)	1100 (100)
2.5	442 (49.1)	178 (19.8)	105 (11.7)	175 (19.4)	900 (100)
5	197 (21.4)	253 (28.1)	173 (19.2)	277 (30.8)	900 (100)
10	159 (15.9)	296 (29.6)	264 (26.4)	281 (28.1)	1000 (100)
20	58 (6.4)	49 (5.4)	133 (14.8)	660 (73.3)	900 (100)

As quantidades de interesse são as proporções de migração de cada nível. Defini-se π_0 a proporção de células em que não houve migração, π_1 a proporção de células em que o DNA migrou pouco, π_2 a proporção de células em que o DNA migrou de forma intermediária e π_3 é a proporção de células em que o DNA migrou muito.

Como neste caso a variável resposta não é binária, é necessário dividir o conjunto de dados em três partes, construindo três modelos auxiliares. Mais detalhes sobre o método de divisão pode ser encontrado em Stern e Pereira [14]. O processo de divisão consiste em transformar, de forma apropriada, o problema multinomial em um problema de dados binários.

Em uma primeira etapa considera-se uma variável Y_0 binária, em que Y_0 é igual a um se o nível de mutação foi C0 e zero, caso contrário. Na segunda etapa, descarta-se as observações cujo o nível de mutação foi C0 e constroi-se uma variável binária Y_1, em que Y_1 é igual a um se o nível de mutação for igual a C1 e zero, caso contrário. Por fim, na terceira etapa, descarta-se as observações com níveis de mutação C0 e C1, e constroi-se uma variável binária Y_2, em que Y_2 é igual a um se o nível de mutação for igual a C2 e zero se for C3.

Desta forma é construido três modelos, um para cada etapa, em que o interesse é estimar as quantidades θ_0, θ_1 e θ_2, definidas como: θ_0 é a proporção de não migração no DNA ($P(Y_0 = 1) = \theta_0$); θ_1 é a proporção de DNA migrou pouco ($P(Y_1 = 1) = \theta_1$), neste modelo o nível "Não houve migração no DNA" não foi usado; e θ_2 é a proporção de DNA migrou de forma intermediária ($P(Y_2 = 1) = \theta_2$), neste modelo além do nível "Não houve migração no DNA", o nível "O DNA migrou pouco" também não foi usado. Apresenta-se um esquema na Tabela 4.9 para mostrar como os três modelos foram construidos.

TABELA 4.9: Definição das variáveis Y_i

Níveis de Mutação no DNA	Y_0	Y_1	Y_2
C0	1	-	-
C1	0	1	-
C2	0	0	1
C3	0	0	0
Número de observações	4800	3290	2389

Uma vez construído os três modelos, e as proporções θ_0, θ_1 e θ_2 estimadas, é necessária uma nova transformação para retornarmos as quantidades

de interesse originais (π_0, π_1, π_2 e π_3). Esta transformação é obtida através das Equações (4.9), (4.10), (4.11) e (4.12).

$$\pi_0 = \theta_0, \tag{4.9}$$

$$\pi_1 = (1-\theta_0) * \theta_1, \tag{4.10}$$

$$\pi_2 = (1-\theta_0) * (1-\theta_1) * \theta_2, \tag{4.11}$$

$$\pi_3 = (1-\theta_0) * (1-\theta_1) * (1-\theta_2). \tag{4.12}$$

A Tabela 4.10 mostra os valores dos coeficientes da regressão binária Weibull para as proporções θ, os valores dos graus de liberdade (g.l.), a estatística qui-quadrado de razão de verossimilhança, como apresentada na seção 3.5 e os respectivos p-valores. Na Tabela 4.11 tem-se a frequência absoluta e relativa (%) do número esperado de mutações no DNA pelo modelo binário Weibull.

TABELA 4.10: Coeficientes do modelo condicional da probabilidade de Mutação no DNA

Resposta	Coeficientes		g.l.	qui-quadrado	Pr > qui
θ_0	Constante	0.092			
	Dose	2.409	1	60.38	<0.001
	$\sqrt{Dose}$	-3.683	1	42.96	<0.001
	γ	0.274	1	49.39	<0.001
θ_1	Constante	0.034			
	Dose	-0.016	1	29.05	<0.001
	$\sqrt{Dose}$	0.064	1	15.18	<0.001
	γ	0.327	1	5.23	0.022
θ_2	Constante	0.232			
	Dose	0.079	1	58.75	<0.001
	$\sqrt{Dose}$	-0.004	1	62.31	<0.001
	γ	0.937	1	120.85	<0.001

TABELA 4.11: Número esperado de mutações no DNA pelo modelo binário Weibull

Dose de radiação	Classes de Mutação no DNA (%) C0	C1	C2	C3
0	654 (59.5)	125 (11.4)	72 (6.5)	249 (22.6)
2.5	441 (49)	170 (18.9)	101 (11.2)	188 (20.9)
5	211 (23.4)	257 (28.5)	182 (20.2)	250 (27.8)
10	136 (13.6)	303 (30.3)	263 (26.3)	299 (29.9)
20	68 (7.6)	49 (5.4)	131 (14.6)	652 (72.4)

Os valores da Tabela 4.11 podem ser comparados com os valores reais e os valores estimados pelo modelo logito através de forma gráfica como apresentado nas Figuras 4.4, 4.5, 4.6 e 4.7. Nestas quatro figuras, a linha tracejada representa o modelo logito e a linha contínua o modelo Weibull. Os dados são indicados pelos losângulos.

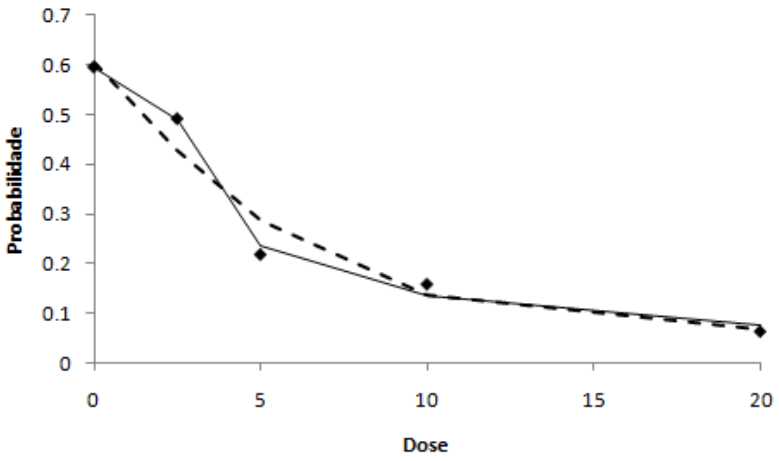

FIGURA 4.4: Dados observados C0 e curva das probabilidades ajustadas.

Observa-se pelas Figuras 4.4 e 4.7 que o modelo Weibull apresentou estimativas mais próximas dos valores reais que o modelo Logito e que nas Figuras 4.5 e 4.6 os dois modelos apresentaram estimativas muito próximas.

Apresenta-se na Tabela 4.12 o valor dos resíduos *Deviance* e X^2 de Pearson, descritos na seção 3.5, para os diferentes níveis de migração no DNA,

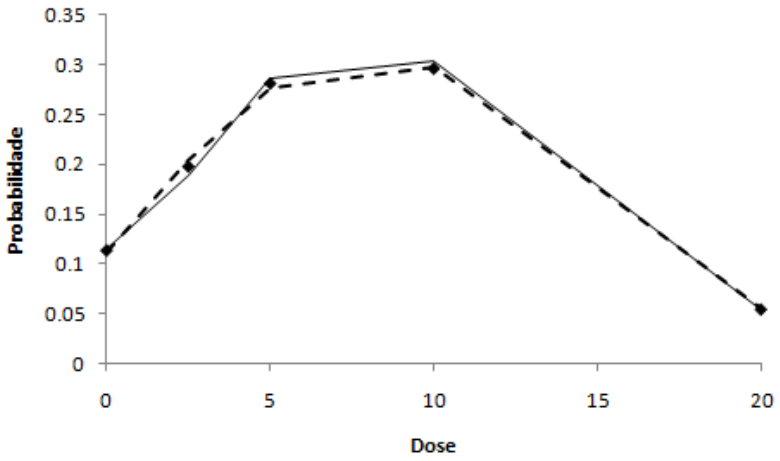

FIGURA 4.5: Dados observados C1 e curva das probabilidades ajustadas.

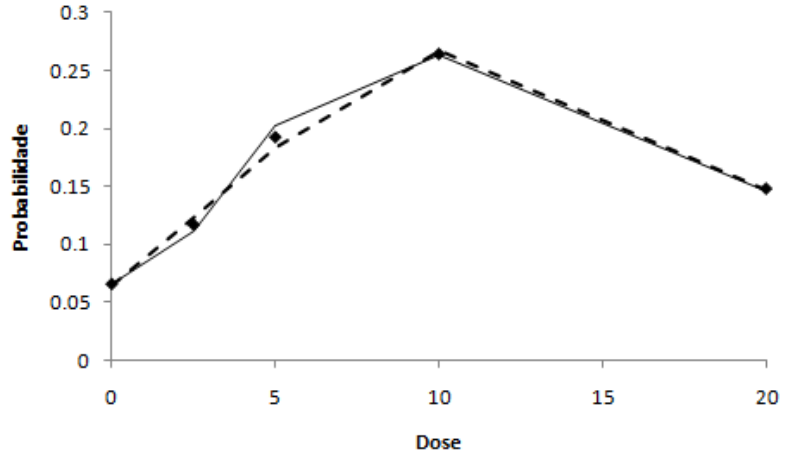

FIGURA 4.6: Dados observados C2 e curva das probabilidades ajustadas.

assim como a soma de todos os níveis. Observamos que nos níveis C0 e C3 o modelo binário Weibull produziu um erro muito menor que o modelo Logito. Percebe-se, também, que a soma dos resíduos para ambas estatísticas é em torno de quatro vezes maior para o modelo Logito. Para obter-se as estimativas de todas as categorias (C0, C1, C2, C3) da variável resposta foram utilizados três modelos condicionais. Portanto, o valor da medida AIC e o valor maximizado da log-verossimilhança, definidas para um único modelo, não se aplicam e por isso não foram utilizados.

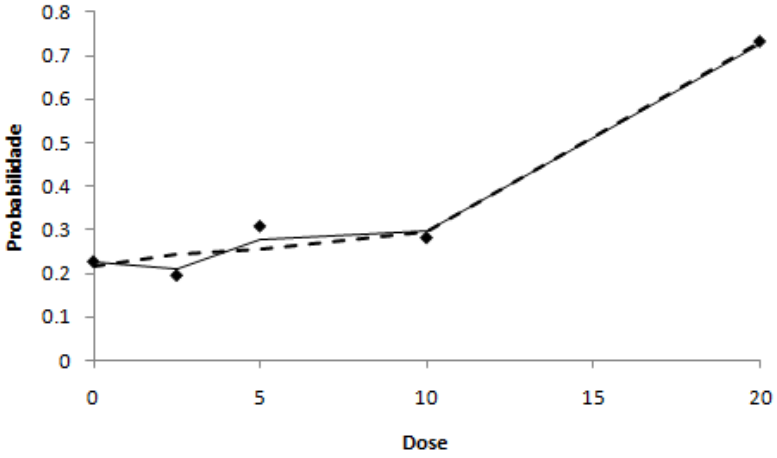

FIGURA 4.7: Dados observados C3 e curva das probabilidades ajustadas.

TABELA 4.12: Estatísticas Deviance e X^2 de Pearson para os modelos Logito e Weibull

Estatística	Modelo	**C0**	C1	C2	C3	Soma
Deviance	Weibull	7.44	0.76	0.81	7.02	16.03
	Logístico	38.31	0.31	0.90	27.22	66.74
X^2 **de Pearson**	Weibull	7.54	0.76	0.81	7.06	16.18
	Logístico	37.53	0.31	0.90	27.13	65.87

Tomando os resultados dos três exemplos apresentados neste capítulo, pode-se considerar que as estimativas proporcionadas pela função de ligação Weibull são muito promissoras. No próximo capítulo estuda-se a relação do modelo Weibull com o Probito e Logito.

Capítulo 5

Comparando o modelo Weibull com o Probito e Logito

No capítulo 4 o modelo Weibull mostrou-se eficaz sendo que, na seção 4.1, estudou-se a relação entre o modelo Weibull e complementar log-log. Como visto, o modelo complementar log-log pode ser tratado como um caso particular do modelo Weibull. Estes resultados motivaram um estudo em relação aos dois principais modelos utilizados atualmente para ajustar dados com resposta binária, Probito e Logito, comparando-os com o modelo Weibull.

5.1 Weibull e Probito

Ao comparar a distribuição Weibull com a distribuição Normal Padrão obtém-se apenas relações aproximadas. De acordo com Dubey [8], a qualidade da aproximação depende do critério que será escolhido para se igualar estas distribuições. Assim ele sugeriu quatro aproximações:

- $\gamma \approx 3{,}60235$ para assimetria $= 0$,
- $\gamma \approx 3{,}43954$ para média $=$ mediana,
- $\gamma \approx 3{,}31247$ para média $=$ moda,
- $\gamma \approx 3{,}25889$ para moda $=$ mediana.

A respeito da kurtose, tem-se dois valores de γ ($\gamma \approx 2{,}25200$ e $\gamma \approx 5{,}77278$) que fornecem kurtose = 3, ou seja, o mesmo valor de kurtose da distribuição Normal padrão. Considere agora a função densidade acumulada Normal padrão

$$\Phi(\eta) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} exp[-t^2/2]dt. \tag{5.1}$$

A seguir, considere a função densidade acumulada de uma variável aleatória, T, Weibull padronizada,

$$T = \frac{W - E[W]}{\sqrt{Var[W]}} = \frac{W - (\mu + \lambda\Gamma_1)}{\lambda\sqrt{\Gamma_2 - \Gamma_1^2}}, \Gamma_i = \Gamma\left(1 + \frac{i}{\gamma}\right), \tag{5.2}$$

dada por

$$\begin{aligned}
F_T(\eta) &= Pr\left(\frac{W - E[W]}{\sqrt{Var[W]}} \leq \eta\right) && (5.3)\\
&= Pr\left(W \leq E[W] + \eta\sqrt{Var[W]}\right) && (5.4)\\
&= 1 - exp\left[-\left(\frac{E[W] + \eta\sqrt{Var[W]} - \mu}{\lambda}\right)^{\gamma}\right] && (5.5)\\
&= 1 - exp\left[-\left(\frac{\mu + \lambda\Gamma_1 + \eta\lambda\sqrt{\Gamma_2 - \Gamma_1^2} - \mu}{\lambda}\right)^{\gamma}\right] && (5.6)\\
&= 1 - exp\left[-\left(\Gamma_1 + \eta\sqrt{\Gamma_2 - \Gamma_1^2}\right)^{\gamma}\right], && (5.7)
\end{aligned}$$

o qual só depende do parâmetro de forma γ.

Deseja-se explorar (5.7) em comparação com (5.1) para os valores de γ dados na Tabela 5.1.

TABELA 5.1: Valores de gama

γ	Γ_1	$\sqrt{\Gamma_2 - \Gamma_1^2}$	Nota
2,25200	0,88574	0,41619	kurtose = 3
3,25889	0,89645	0,30249	moda = mediana
3,31247	0,89719	0,29834	média = moda
3,43954	0,89892	0,28897	média = mediana
3,60235	0,90114	0,27787	assimetria = 0
5,77278	0,92573	0,18587	kurtose = 3

Usando a Tabela 5.1 as seis funções de distribuição Weibull, que foram comparadas com $\Phi(\eta)$, são dadas por

$$F_W^{(1)}(\eta) = 1 - exp\left[-(0,41619\eta + 0,88574)^{2,25200}\right], \quad (5.8)$$

$$F_T^{(2)}(\eta) = 1 - exp\left[-(0,30249\eta + 0,89645)^{3,25889}\right], \quad (5.9)$$

$$F_T^{(3)}(\eta) = 1 - exp\left[-(0,29834\eta + 0,89719)^{3,31247}\right], \quad (5.10)$$

$$F_T^{(4)}(\eta) = 1 - exp\left[-(0,28897\eta + 0,89892)^{3,43954}\right], \quad (5.11)$$

$$F_T^{(5)}(\eta) = 1 - exp\left[-(0,27787\eta + 0,90114)^{3,60235}\right], \quad (5.12)$$

$$F_T^{(6)}(\eta) = 1 - exp\left[-(0,18587\eta + 0,92573)^{5,77278}\right]. \quad (5.13)$$

Rinne [16] comparou as diferenças absolutas $\Delta^{(i)}(\eta) = F_T^{(i)}(\eta) - \Phi(\eta)$, para i=1,2, ...,6, entre as funções de distribuição Normal padrão e Weibull padronizada. Para estes seis valores de γ e η, variando de -3 a 3 com amplitude de 0,1, ele observou que, em geral, as probabilidades associadas a cauda inferior (superior) da distribuição normal pode ser aproximada satisfatoriamente usando uma distribuição Weibull com parâmetro de forma $\gamma \approx 3.60235$ ($\gamma \approx 3.25889$).

Para ilustrar esta aproximação construiu-se um estudo de simulação da seguinte forma: a variável explicativa (X) foi definida de -2 a 2 com amplitude de 0,02 para todos os conjuntos de dados. Foram considerados 33 valores de μ e 33 valores de β_1, ambos variando de -2.5 a 2.5 com amplitude de 0.15625. Assim, define-se a variável resposta como $Y = \Phi(\mu + \beta_1 X)$, para toda a combinação de valores de μ e β_1, totalizando 33*33=1089 variáveis resposta. Desta forma, foram gerados 1089 conjuntos de dados com 201 observações cada.

Em cada conjunto de dados é ajustado o modelo Probito e Weibull. Para cada modelo registra-se o valor de sua log-verossimilhança maximizada e com estes valores constroi-se um gráfico de dispersão, Figura 5.1.

Observe que os pontos alinhados indicam que a maioria dos modelos ajustados possuem valores iguais. Os pontos aglomerados entre -60 e -30 indicam que o modelo Probito obteve um ajuste ligeiramente melhor por ter uma log-verossimilhança maximizada maior do que o modelo Weibull para um mesmo conjunto de dados.

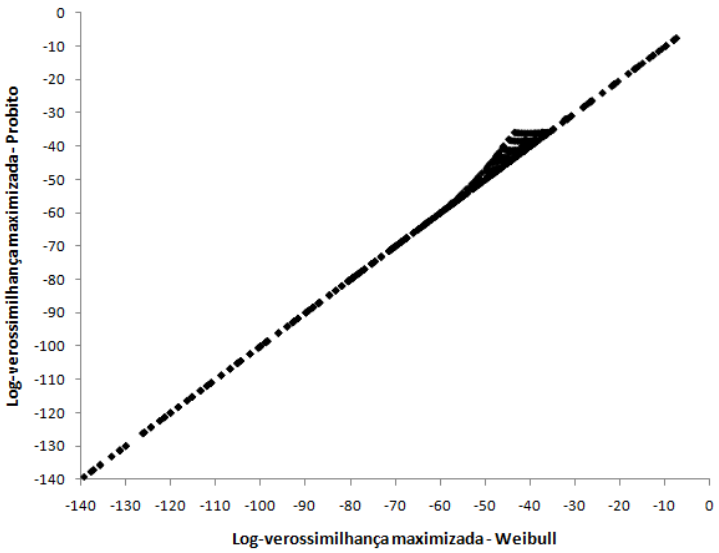

FIGURA 5.1: Gráfico de dispersão do ajuste dos Modelos Weibull e Probito em dados provenientes de um modelo logístico

Deseja-se agora comparar estes modelos ajustando um modelo Probito em dados provenientes de um modelo Weibull. Para tal, foram gerados 1089 conjuntos de dados com 801 observações cada. A variável explicativa (X) é definida de -4 a 4 com amplitude de 0,005 para todos os conjuntos de dados. Considere $\gamma = 1$, μ variando de 8 a 34 e β_1 variando de -4 a 4. Assim, define-se a variável resposta como $Y = 1 - exp(-(\mu + \beta_1 X))$ para toda a combinação de valores de μ e β_1.

Em cada conjunto de dados é ajustado um modelo Probito e um Weibull. Novamente, para cada modelo registra-se o valor de sua log-verossimilhança maximizada e constroi-se um gráfico de dispersão, Figura 5.2. Os pontos fora da reta indicam que o modelo Weibull obteve um ajuste melhor por ter uma log-verossimilhança maximizada maior do que o modelo Probito para um mesmo conjunto de dados. Em alguns casos esse ajuste foi muito superior, como por exemplo para um caso em que o modelo Weibull obteve valor de log-verossimilhança maximizada próximo a -60 o Probito obteve valor próximo a -120.

Os resultados obtidos indicam que o modelo Weibull aproxima bem o modelo Probito como esperado. Porém, não pode-se afirmar o mesmo no caso contrário.

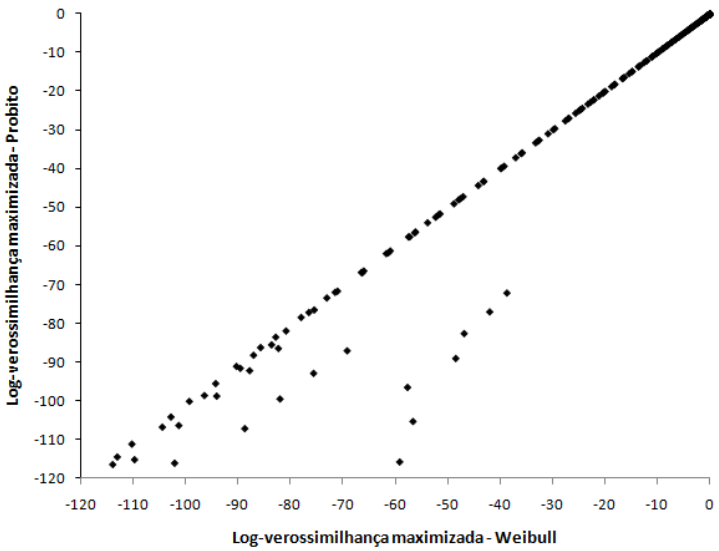

FIGURA 5.2: Gráfico de dispersão do ajuste dos Modelos Weibull e Probito em dados provenientes de um modelo Weibull

5.2 Weibull e Logito

As distribuições logística e normal são muito similares, sendo ambas simétricas, mas a distribuição logística tem mais kurtose (kurtose = 4.2). É sugestivo aproximar a distribuição logística a uma versão simétrica da distribuição Weibull.

Comparando a função densidade acumulada logística reduzida

$$F(\eta) = \frac{1}{1 + exp(-\pi\eta/\sqrt{3})}, \tag{5.14}$$

com a função densidade acumulada Weibull reduzida nas Equações (5.9), (5.10) e (5.11), tem-se resultados piores do que comparando com a distribuição normal. A Figura 5.3 mostra o melhor ajuste da função densidade acumulada Weibull (γ = 3,60235 e simetria = 0) em comparação com as funções de distribuição logística e normal.

Para ilustrar esta aproximação construiu-se um estudo de simulação da seguinte forma: a variável explicativa (X) foi definida de -4 a 4 com amplitude de 0,04 para todos os conjuntos de dados. Foram considerados 33 valores de μ e 33 valores de β_1, ambos variando de -4 a 4 com amplitude de 0.25. Assim, define-se a variável resposta como $Y = \Phi(\mu + \beta_1 X)$, para toda a combinação de valores de

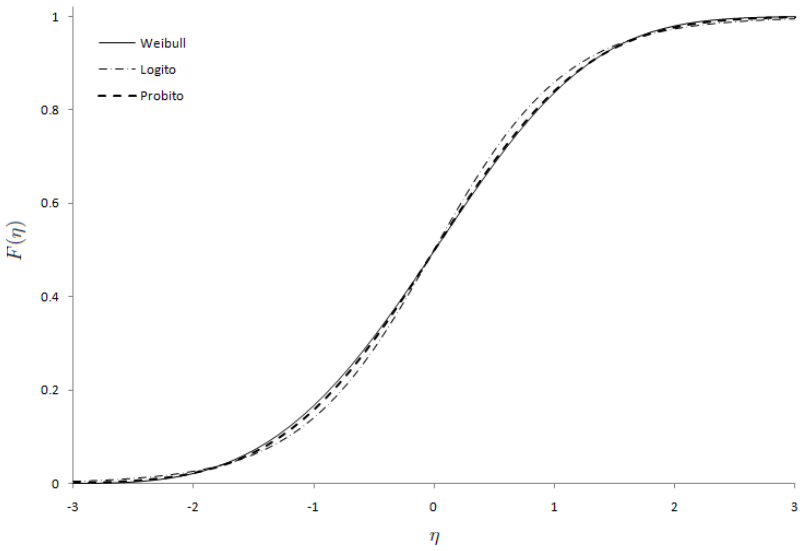

FIGURA 5.3: Função densidade acumulada Logística, Normal e Weibull padronizadas.

μ e β_1, totalizando 33*33=1089 variáveis resposta. Desta forma, foram gerados 1089 conjuntos de dados com 201 observações cada.

Em cada conjunto de dados é ajustado um modelo Logito e um Weibull. Para cada modelo registra-se o valor de sua log-verossimilhança maximizada e constroi-se um gráfico de dispersão, Figura 5.4.

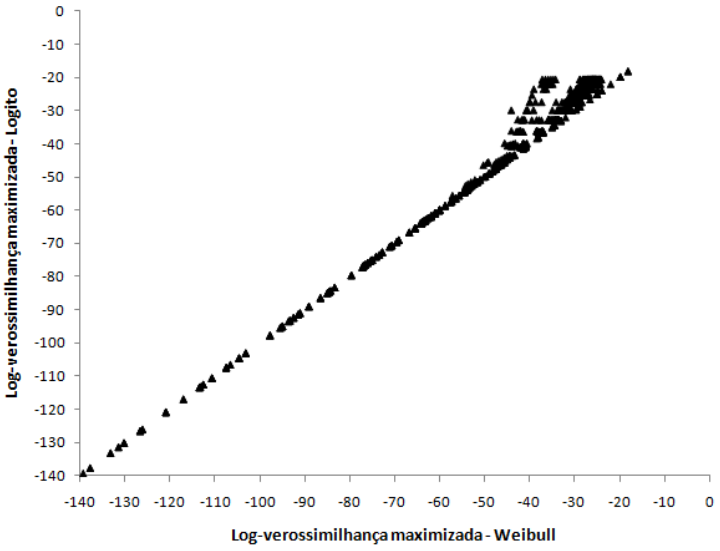

FIGURA 5.4: Gráfico de dispersão do ajuste dos Modelos Weibull e Logito em dados provenientes de um modelo logístico

Observe que os pontos alinhados indicam que a maioria dos modelos ajustados possuem valores iguais. Os pontos aglomerados entre -50 e -20 indicam que o modelo Logito obteve um ajuste um pouco melhor por ter uma log-verossimilhança maximizada maior do que o modelo Weibull para um mesmo conjunto de dados.

Para gerar a Figura 5.5 foram utilizados os mesmos dados da Figura 5.2, porém é mostrado o ajuste do modelo Logito ao invés do modelo Probito.

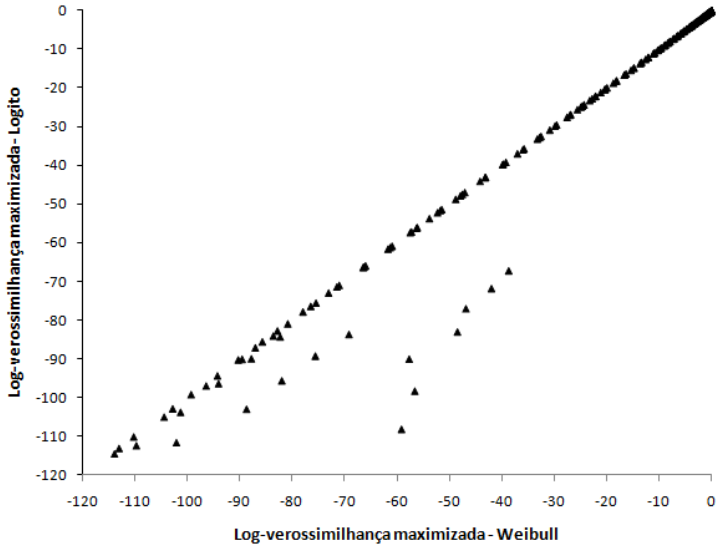

FIGURA 5.5: Gráfico de dispersão do ajuste dos Modelos Weibull e Logito em dados provenientes de um modelo Weibull

Este estudo de simulações reforça a ideia de que o modelo Weibull aproxima muito bem os modelos Logito e Probito e que o oposto não pode ser afirmado. O objetivo é vizualizar as relações entre o modelo Weibull para com os modelos Logito e Probito. Ainda é necessário um estudo de simulação mais amplo considerando: as outras funções de ligação citadas no trabalho; diferentes tamanhos de amostra, não apenas com o tamanho de 201 observações; e também diferentes proporções de 1 (uns) e 0 (zeros) na variável resposta (Y).

Capítulo 6

Conclusões

Neste trabalho foi proposto um novo modelo que apresentou-se muito interessante pelos resultados obtidos. A ideia de utilizar a distribuição Weibull permite uma grande flexibilidade na estimação com um modelo simples e muito eficiente.

A aplicação do modelo Weibull nos conjuntos de dados reais, apresentada no capítulo 4, mostrou que o modelo é capaz de estimar várias situações. Comparando os ajustes obtidos nas seções 4.1 e 4.2 com os modelos mais sofisticados (Aranda-Ordaz, Prentice e Stukel), o modelo Weibull sempre apresentou bons resultados.

Para o conjunto de dados sobre Mortalidade de Besouros, na seção 4.1, os ajustes dos modelos Prentice, Weibull e Aranda-Ordaz foram muito similares e o modelo complementar log-log foi ligeiramente superior pelo critério AIC. No conjunto de dados Garotas de Varsóvia, na seção 4.2, um modelo Weibull mostrou-se o melhor entre todos. Por fim, no caso do conjunto de dados sobre Mutação no DNA de caramujos, na seção 4.3, a comparação foi feita apenas com o modelo Logito e o resultado do modelo Weibull foi muito superior.

A eficiência do modelo Weibull deve-se ao fato de que esta distribuição é flexível. Tal fato pode ser comprovado quando das relações com alguns modelos. Estudou-se estas relações com os modelos complementar log-log (na seção 4.1), Probito (na seção 5.1) e Logito (na seção 5.2). Mostrou-se que o modelo Weibull

é capaz de se adaptar a estes casos. Já o contrário não pode ser observado. O motivo principal disto é que os modelos Probito e Logito têm funções de ligação simétrica e que o modelo Weibull, por ter um parâmetro a mais que estes, possui uma função de ligação mais genérica.

Ainda é necessário mais estudos pra comprovar a eficiência do modelo, principalmente em relação aos modelos mais sofisticados. Nos exemplos com dados reais pode-se visualizar a qualidade do modelo Weibull em relação aos modelos Aranda-Ordaz, Prentice e Stukel. Estes estudos não são conclusivos, logo a principal proposta futura é estudar estas relações. Outras propostas futuras incluem: otimizar o processo de estimação, estudar seleção de variáveis e influência de valores extremos.

O Modelo Weibull é muito promissor e, se nós pudermos escolher somente um único modelo, escolheríamos o modelo Weibull.

Apêndice A

Códigos em SAS

Código em linguagem SAS utilizado para o ajuste do modelo Weibull aos dados: *Mortalidade de Besouros*.

```
proc nlp data=beetle cov=2 vardef=n outest=est maxfunc=200 maxiter=200
tech=nrridg;
max fx;
parms mu=0.9, b1=0.155, k=3.6978;
bounds 0<k;
xb = mu + b1*x;
p= 1 - exp(-(xb)**k) ;
fx = y*log(p)+(1-y)*log(1-p);
run;
```

Código em linguagem SAS utilizado para o ajuste do modelo Weibull aos dados: *Garotas de Varsóvia*.

```
proc nlp data=menarche cov=2 vardef=n outest=est pcov phes tech=nrridg ;
max fx;
parms mu =0.9 , b1 =0.155, k = 3.6978;
bounds 0 < k;
xb = mu + b1*x;
p = exp(-(xb)**k);
```

```
fx = y*log(p)+(1-y)*log(1-p);
run;
```

Código em linguagem SAS utilizado para o ajuste do modelo Weibull aos dados: *DNA Damage*.

```
/*** Passo 1 ********************************************************/
data teta0;
set mutation;
teta0 = C0;run;
proc nlp data=teta0 cov=2 vardef=n outest=opar0 maxfunc=150 maxiter =
250 pcov phes tech=nrridg ;
max fx; parms mu0=0.9 , b10=0.01, b20=0.04, k0=3.6978;
bounds 0 < k0; y = teta0;
xbeta0 = mu0 + b10*x +b20*rootx;
pi= exp(-(xbeta0)**k0);
fx = y*log(pi)+(1-y)*log(1-pi);
run;
/*** Passo 2 ********************************************************/
data teta1;
set mutation;
If C0 eq 1 then delete;
teta1 = C1;run;
proc nlp data=teta1 cov=2 vardef=n outest=opar01 maxfunc=200 maxiter =
300 pcov phes tech=nrridg ;
max fx; parms mu1=0.9, b11=0.01,b21=0.04, k1=1;
bounds 0 < k1; y = teta1;
xbeta1 = mu1 + b11*x + b21*rootx;
pi= 1-exp(-(xbeta1)**k1);
fx = y*log(pi)+(1-y)*log(1-pi);
run;
/*** Passo 3 ********************************************************/
data teta2;
```

```
set mutation;
If C0 eq 1 then delete;
If C1 eq 1 then delete;
teta2 = C2;run;
proc nlp data=teta2 cov=2 vardef=n outest=opar02 maxfunc=150 maxiter =
250 pcov phes tech=nrridg ;
max fx; parms mu2=0.9, b12=0.04,b22=0.01, k2=1;
bounds 0 < k2; y = teta2;
xbeta2 = mu2 + b12*x +b22*x2;
pi=1- exp(-(xbeta2)**k2) ;
fx = y*log(pi)+(1-y)*log(1-pi);
run;
/*** Calculando a proporção estimada ********************************/
data stat0;
set opar0;
where _TYPE_ = 'PARMS';
do i=1 to 4800;
output; end; run;
data stat1;
set opar01;
where _TYPE_ = 'PARMS';
do i=1 to 3290;
output; end; run;
data stat2;
set opar02;
where _TYPE_ = 'PARMS';
do i=1 to 2389;
output; end; run;
data pred_teta0;
merge teta0 stat0;
xb = mu0 + b10*x +b20*rootx;
p_wei0 = exp(-(xb)**k0);
```

```
run;
data pred_teta1;
merge teta1 stat1;
xb = mu1 + b11*x + b21*rootx;
p_wei1 =1- exp(-(xb)**k1);
run;
data pred_teta2;
merge teta2 stat2;
xb = b02 + b12*x +b22*x2;
p_wei2 =1- exp(-(xb)**k2);
run;
proc means data=pred_teta0 n sum mean ;
class x; var p_wei0 ;
output out=tab0 mean=media0 n=n0; run;
proc means data=pred_teta1 n sum mean ;
class x; var p_wei1 ;
output out=tab1 mean=media1 n=n1; run;
proc means data=pred_teta2 n sum mean ;
class x; var p_wei2 ;
output out=tab2 mean=media2 n=n2; run;
data tabfinal;
merge tab0 tab1 tab2;
p0 = media0;
p1 = (1-media0)*media1;
p2 = (1-media0)*(1-media1)*media2;
p3 = (1-media0)*(1-media1)*(1-media2);
run;
proc means data=tabfinal n sum mean ;
class x; var p0 p1 p2 p3; run;
```

Código em linguagem SAS utilizado para gerar os 1089 conjuntos de dados proveniente de um modelo Probito e para ajustar os modelos Weibull e Probito em todos os conjuntos de dados.

```
%macro simulated;
%do j = 0 %to 32;
%do l = 0 %to 32;
data simulated&j.&l.;
do i = 0 to 200;
x = i/50 - 2;
mu = -2.5 + &j./6.4;
b1 = -2.5 + &l./6.4;
y =probnorm(mu + b1*x);
output; end; run; %end; %end;
%mend; %simulated;
%macro weib;
data parms;
_TYPE_ = 'INITIAL' ; muw = .; b1w= .;k = .; _RHS_ = .;
output;run;
data inprob;
_TYPE_ = 'INITIAL' ; mu = .; b1= .;k = .; _RHS_ = .;
output;run;
%do j = 0 %to 32;
%do l = 0 %to 32;
proc nlp data=simulated&j.&l. cov=2 vardef=n outest=inest&j.&l.;
max fx;
parms mu , b1;
xb = mu + b1*x;
pi= probnorm(xb);
fx = y*log(pi)+(1-y)*log(1-pi);
run;
proc nlp data=simulated&j.&l. cov=2 vardef=n outest=est&j.&l. tech=nrridg;
max fx;
parms muw=0.9, b1w=0.155,k=3.6978;
bounds 0< k ;
xb = muw + b1w*x;
```

```
pi= 1 - exp(-(xb**k)) ;
fx = y*log(pi)+(1-y)*log(1-pi); run;
data inprob;
set inprob inest&j.&l. ;
where _TYPE_='PARMS';
keep _TYPE_ mu b1 _RHS_;run;
data parms;
set parms est&j.&l.;
where _TYPE_ = 'PARMS';
keep _TYPE_ muw b1w k _RHS_;
run; %end; %end;
data parmprob;
set inprob ;
RHSprob = _RHS_;
drop _RHS_; run;
data final;
merge parms parmprob;
run;%mend;%weib;
```

Referências Bibliográficas

[1] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6): 716-723. doi:10.1109/TAC.1974.1100705. MR0423716

[2] Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data, Biometrika, 68, 357-363.

[3] Bliss, C. I. (1935). The calculation of the dosage-mortality curve, Ann. Appl. Biol. 22, pp. 134-167

[4] Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). Journal of the Royal Statistical Society B 26, 211-252.

[5] Caron, R. e Polpo, A. (2009). Binary data regression: Weibull distribution. AIP Conf. Proc. 1193, 187, DOI:10.1063/1.3275613.

[6] Chen, M-H, Dey, D. K. & Shao, Q-M. (1999). A new skewed link model for dichotomous quantal response data. Journal of the American Statistical Association, 94, 448, 1172-1186.

[7] Cohen, A.C. (1973). The reflected Weibull distribution. Technometrics 15, 867-873.

[8] Dubey, S.D. (1967a): Normal and Weibull distributions; Naval Research Logistics Quarterly 14,67-79.

[9] Grazeffe, V. S., Tallarico L. F., Pinheiro, A. S., Kawano, T., Suzuki, M. F., Okazaki, K., Pereira, C. A. B., Nakano, E. (2008). Establishment of the comet assay in the freshwater snail *Biomphalaria glabrata*. Mutation Research 654, 58-63.

[10] Milicer, H. & Szczotka, F. (1966). Age at menarche in Warsaw girls in 1965. Human Biology 38, 199-203.

[11] Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society A 135, 370-384.

[12] Paula, G. A. (2004). Modelos de Regressão com apoio computacional. Instituto de Matemática e Estatística. Universidade de São Paulo.

[13] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of a variables in such that in can be reasonably supposed to have arisen from random sampling. Philos. Mag., v.50, p.157-172.

[14] Pereira, C. A. B. & Stern, J. M. (2007). An essay on the role of Bernoulli and Poisson processes in bayesian statistics. Technical Report, RT-MAC-2007-06, p.39.

[15] Prentice, R. L. (1976). Generalization of the probit and logit models. Biometrics 32, 761-768.

[16] Rinne, H. (2009). The Weibull distribution : a handbook. Boca Raton, CRC Press.

[17] Richardson, J. (1994). The analysis of 2 x 1 and 2 x 2 contingency tables: an historical review. Statistical Methods in Medical Research, Vol. 3, No. 2, 107-133.

[18] Statistical Analysis Software (SAS), versão 9.2, 2008.

[19] Stukel, T. A. (1988). Generalized logistic models. Journal of the American Statistical Association, 83, 402, 426-431.