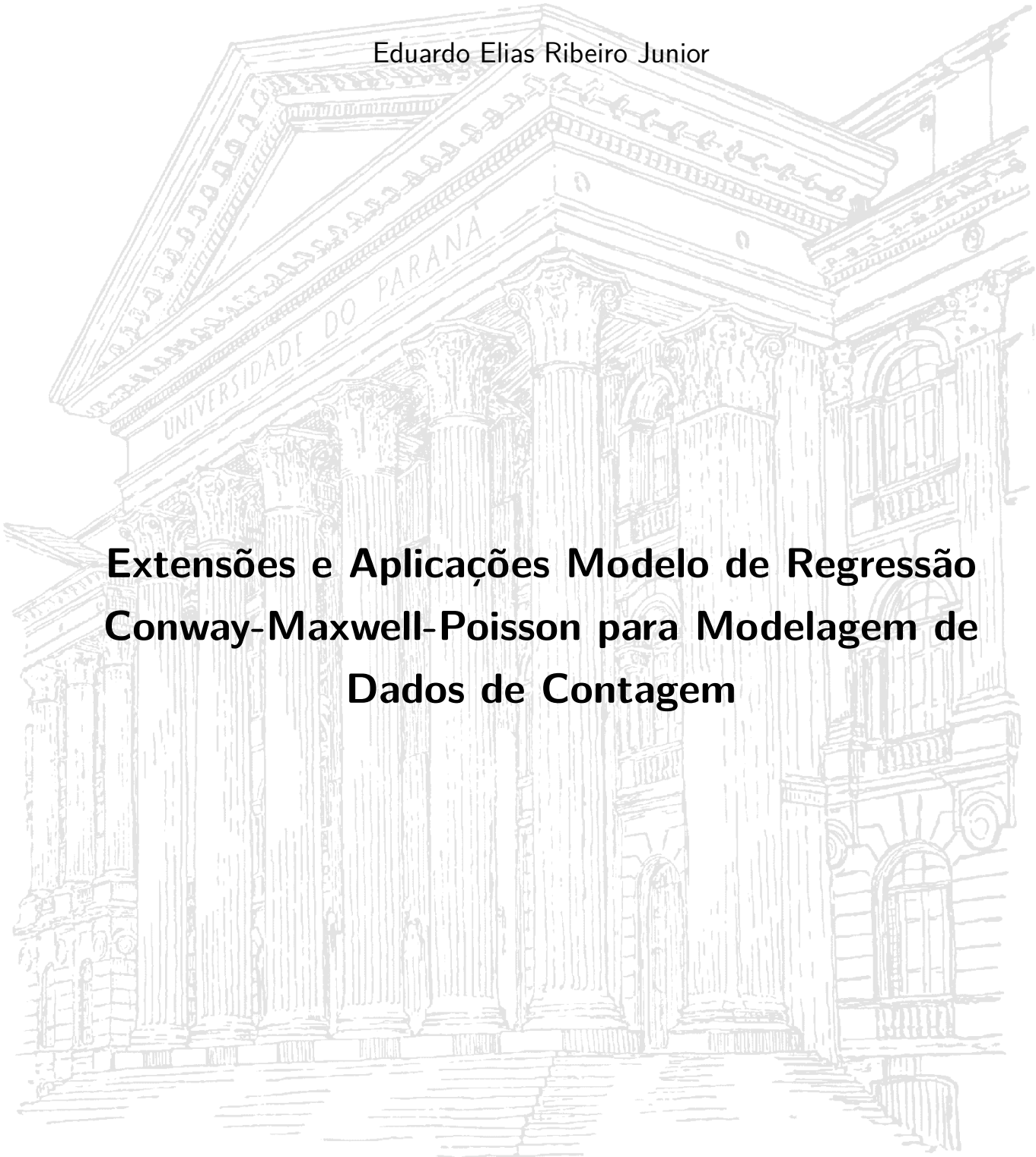
Universidade Federal do Paraná

Eduardo Elias Ribeiro Junior



Extensões e Aplicações Modelo de Regressão Conway-Maxwell-Poisson para Modelagem de Dados de Contagem

Curitiba

2015

Eduardo Elias Ribeiro Junior

Extensões e Aplicações Modelo de Regressão Conway-Maxwell-Poisson para Modelagem de Dados de Contagem

Projeto de Pesquisa apresentado à disciplina Laboratório A do Curso de Graduação em Estatística da Universidade Federal do Paraná, como requisito para elaboração do Trabalho de Conclusão de Curso

Universidade Federal do Paraná

Setor de Ciências Exatas

Departamento de Estatística

Orientador: Prof. Dr. Walmes Marques Zeviani

Curitiba

2015

Sumário

1 Introdução

Modelos estatísticos de regressão são fundamentalmente os principais métodos suporte para a prática de Estatística aplicada. Diversas áreas do conhecimento são familiares a estes modelos, pois eles objetivam descrever a relação entre uma variável dependente, de interesse, com variáveis independentes, a fim de compreender este relacionamento e realizar predições através do modelo estabelecido, principal interesse de pesquisas aplicadas. Um modelo estatístico, na sua forma univariada consiste no estabelecimento de um equação matemática que relaciona uma variável de interesse Y com variáveis preditoras X_i, considerando uma distribuição de probabilidades para Y condicionalmente à X e de média associada a um preditor linear dos efeitos das variáveis independentes.

Podemos destacar o modelo linear normal, com distribuição Normal associada à $[Y \mid X]$ e preditor linear da forma $X\beta$, como o modelo predominante dentre as análises estatísticas aplicadas. Até a década de 70, para situações em que a variável resposta Y não se apresentava de forma contínua com domínio nos reais ou ainda não se verificava os pressupostos do modelo, a alternativa mais utilizada era encontrar alguma forma de transformação da variável resposta para atingir a normalidade (PAULA, 2013). Porém, com a introdução dos da teoria do modelos lineares generalizados (MLG) por Nelder e Wedderburn (1972) e com o avanço computacional, a análise de dados não normais passou a ser corriqueiramente via MLG's. Esta nova classe de modelos flexibiliza a distribuição condicional associada, permitindo outras distribuições pertencentes à família exponencial de distribuições o que contempla as distribuições Normal, Poisson, Binomial, Gama entre outras bem conhecidas na literatura.

Com isso a modelagem de dados passou a ser mais fiel a natureza dos dados em observação. Neste contexto, a análise de variáveis aleatórias de contagem, que representam o número de ocorrências de um evento de interesse em um intervalo de tempo ou espaço específico, foi enriquecida expressivamente. Pois via o modelo linear normal as estimativas contêm erros padrão inconsistentes e podem produzir predições negativas para o número de eventos (KING, 1989).

Para análise estatística destas variáveis temos o modelo probabilístico de Poisson, já consolidado na literatura, sendo amplamente utilizado. Este modelo possui apenas um parâmetro, denotado por λ que atua na média e na variância a partir de uma relação identidade ($\lambda = E[Y] = V[Y]$). Essa propriedade, chamada de equidispersão, é uma particularidade do modelo Poisson que pode não ser adequada a diversas situações. Quando aplicado sem verificação desta suposição, o modelo Poisson acarreta em erros padrões inconsistentes.

Algumas abordagens para modelagem de dados de contagens com fuga de equidispersão foram propostas na literatura. O caso de superdispersão, quando a variância é maior que a média é o mais comum e tem uma gama de métodos para análise mais extensa. A superdispersão pode ocorrer pela ausência de covariáveis importantes, excesso de zeros, diferentes amplitudes de domínio (*offset*) não considerado, heterogeneidade de unidades amostrais, excesso de zeros, entre outros (Ribeiro Jr et al., 2012). Para estes casos a abordagem mais comum é a adoção de modelos com efeitos aleatórios que capturam a variabilidade extra. Um caso particular dos modelos Poisson de efeitos aleatórios, muito adotado no campo aplicado da Estatística, ocorre quando consideramos distribuição Gama para os efeitos aleatórios, nesta situação temos expressão fechada para a função de probabilidade marginal que assume a forma Binomial Negativa. Contudo outra manifestação de fuga da suposição de equidispersão é a subdispersão, situação menos comum na literatura. O processo que reduz a variabilidade das contagens Poisson não é tão claro quanto o que produz variabilidade extra e é muito particular do experimento em estudo. São poucas as abordagens descritas na literatura que compreendem o tratamento de subdispersão, uma vez que efeitos aleatórios só capturam avariabilidade extra. Podemos citar os modelos de quase-verossimilhança como a abordagem mais utilizada, todavia não é possível recuperar a verdadeira distribuição da variável resposta nesta abordagem pois aqui a modelagem é baseada apenas nos dois primeiros momentos (PAULA, 2013).

Anteriormente à formalização dos MLG's, em um contexto de filas Conway e Maxwell (1962) propuseram uma distribuição de probabilidades que generaliza a distribuição Poisson com a adição de mais uma parâmetro, denotado por ν, contemplando agora os casos de sub e superdispersão. Posteriormente esta distribuição foi nomeada como COM-Poisson (Conway-Maxwell-Poisson em homenagem à Richard W. Conway, William L. Maxwell seus autores). A função distribuição de probabilidade COM-Poisson assume a forma

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}, \qquad y \in \mathbb{Z}_+ \tag{1.1}$$

onde $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^s}{(s!)^\nu}$, $\lambda > 0$ e $\nu \geq 0$. A flexibilidade da COM-Poisson é verificada pela razão de probabilidades consecutivas, que se caracteriza não necessariamente linear em y ($P(Y = y-1)/P(Y = y) = y^\nu/\lambda$) sendo governada pelo parâmetro extra ν (SHMUELI et al., 2005). Uma característica bastante relevante é que esta distribuição pertence à família exponencial e possui como casos particulares as distribuições Poisson, Geométrica e Binomial. Portanto, empregando a distribuição COM-Poisson na estrutura de um MLG obtemos um modelo de regressão sem a imposição de equidispersão.

A aplicação do modelo de regressão COM-Poisson não apresenta forma analítica para estimação, portanto métodos numéricos são empregados nesta tarefa. Ainda a

constante normalizadora $Z(\lambda, \nu)$ em 1.1 se apresenta como uma série infinita que computacionalmente deve ser truncada. Algumas aplicações do modelo de regressão COM-Poisson são apresentadas em Shmueli et al. (2005) e Sellers e Shmueli (2010).

Pela similaridade da função de distribuição COM-Poisson em 1.1 com a Poisson vários aspectos podem ser estendidos. Por exemplo, citamos a inclusão de efeitos aleatórios para acomodar superdispersão, porém há situações em que o delineamento do experimento sugere a correlação entre observações de parcelas da amostra e independência entre parcela. São efeitos não observáveis que agem no nível de observação e isso pode ser incorporado no modelo de regressão COM-Poisson. Da mesma forma excesso de zeros podem ser introduzidos a esta distribuição da mesma forma com o que ocorre para o modelo Poisson, através de truncamento (modelos Hurdle) ou inflação (modelos de mistura). Estas extensões citadas para o modelo COM-Poisson não são bem consolidadas na literatura e são raras suas aplicações. Uma constatação do fato é que não há implementações destas extensões nos principais softwares estatísticos.

Na literatura brasileira a aplicação dos modelos COM-Poisson são escassos. Foram encontrados apenas aplicações na área de Análise de Sobrevivência mais especificamente em modelos com fração de cura. Portanto a presente pesquisa visa colaborar com a literatura estatística brasileira apresentando e explorando alternativas para modelagem de dados de contagem e suas extensões para situações comuns em estudos experimentais e observacionais.

2 Objetivos

2.1 Objetivos Gerais

Apresentar o modelo de regressão COM-Poisson, alternativa paramétrica não comumente utilizada pela comunidade de Estatística aplicada, trazendo discussões sobre aspectos inferenciais deste modelo. Estender as aplicações do modelo COM-Poisson para situações específicas (efeitos aleatórios e excesso de zeros).

2.2 Objetivos Específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelos de regressão para dados de contagem;
- Avaliar as propriedades de soluções numéricas para 1) cálculo da densidade de probabilidade do modelo e 2) para estimação de modelos de regressão de efeito fixo;
- Propor e implementar uma extensão do modelo de regressão COM-Poisson para acomodar efeitos aleatórios;
- Propor e implementar uma extensão do modelo de regressão COM-Poisson para acomodar contagens com excesso de zeros.
- Fazer a aplicação do modelo COM-Poisson e das extensões desenvolvidas à dados reais ou simulados. Fazer comparações com os modelos disponíveis para as situações estudadas: Poisson, Quase-Poisson, Binomial Negativo, Poisson de efeito aleatório.

3 Materiais e Métodos

3.1 Materiais

3.1.1 Dados para análise

Este projeto de pesquisa via o estudo e avaliação do modelo de regressão COM-Poisson para modelagem de dados de contagem. Portanto pretende-se analisar vários conjuntos de dados, preferencialmente os já analisados na literatura via outras técnicas, para efeitos de comparação.

Inicialmente temos o conjunto `defoliation`, disponível no *software* R através do pacote `legTools`[1]. Este conjunto de dados contém 125 observações provenientes de um experimento em casa de vegetação inteiramente casualizado com 5 repetições, cujo plantas de algodão (*Gossypium hirsutum*) foram submetidas à níveis de desfolha artificial (5 níveis) combinados com o estágio fenológico da planta na aplicação da desfolha (5 níveis), observou-se como variável resposta o número de capulhos produzidos ao final do ciclo cultura. Este conjunto de dados foi analisado considerando o modelo de contagem Gamma por Zeviani et al. (2014).

Ainda serão utilizados dados simulados para avaliação da acurácia dos métodos de estimação e comparação de abordagens distintas em diferentes cenários.

3.1.2 Recursos Computacionais

Para análise e elaboração do trabalho será utilizado o *software R*, na versão 3.2 (R Core Team, 2015). Atualmente há três bibliotecas desenvolvidas em R dedicadas à distribuição COM-Poisson. São eles `COMPoissonReg` (SELLERS; LOTZE, 2011), este conjunto de funções foi desenvolvido com base nas análises apresentadas no artigo *A flexible regression model for count data* (SELLERS; SHMUELI, 2010); `compoisson` as funções desta biblioteca se dedicam apenas ao modelo probabilístico; e `CompGLM` esta é mais recente biblioteca de funções com ênfase no modelo COM-Poisson, escrita em `C++` permite a estimação de modelos de regressão além de funções probabilísticas.

Outros recursos e bibliotecas, principalmente para otimização de funções e elaboração de gráficos, serão utilizadas.

[1] Em desenvolvimento pelo Laboratório de Estatística e Geoinformação da UFPR <http://git.leg.ufpr.br/leg/legTools>

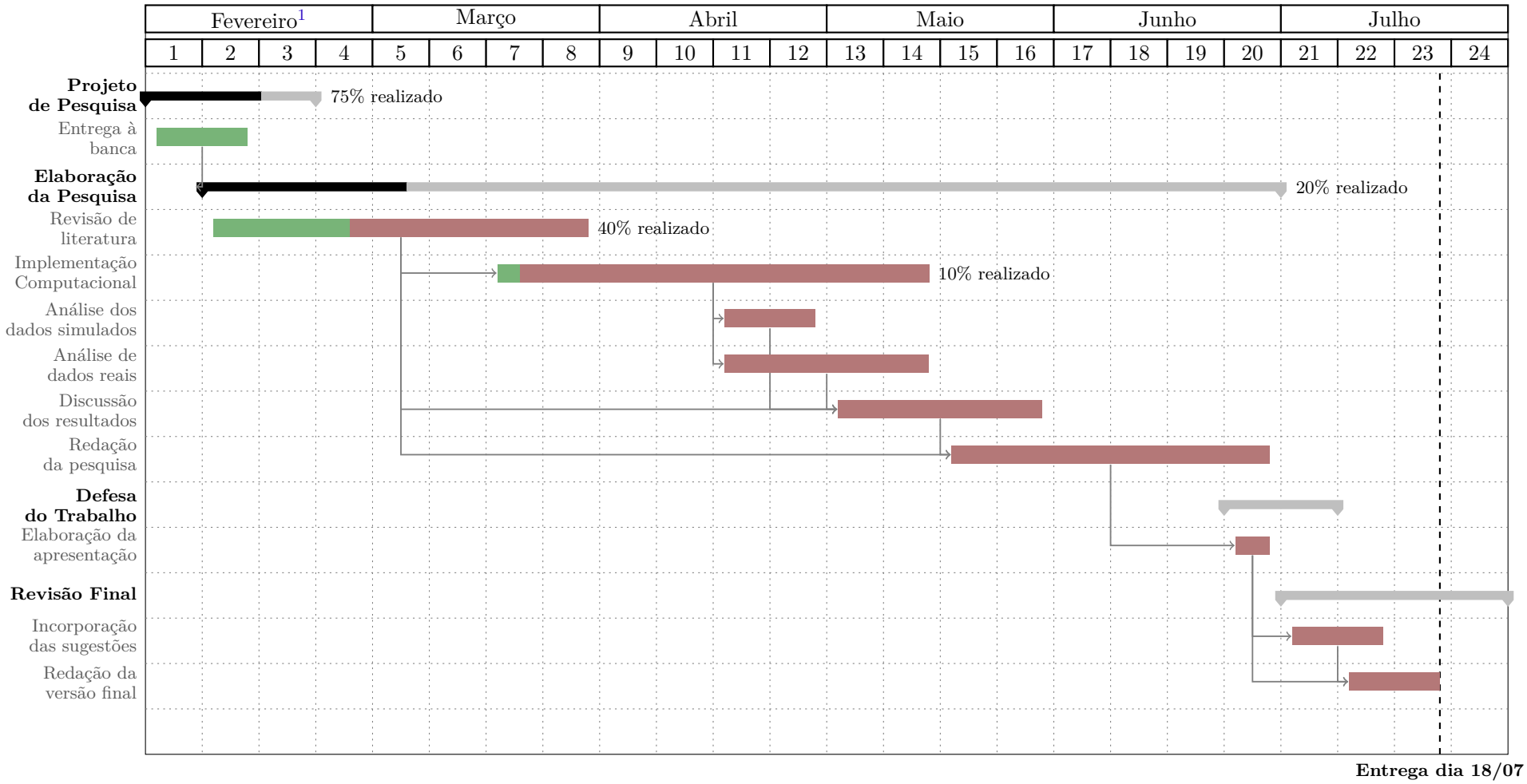
3.2 Métodos

Neste trabalho se fará uso da teoria de Modelos Lineares Generalizados descritas inicialmente por Nelder e Wedderburn (1972), cujo referência nacional é dada por Paula (2013) que também aborda modelos para dados de contagem com excesso de zeros. Para estimação dos modelos COM-Poisson de efeitos aleatórios pretende-se utilizar os métodos de integração apresentados no material de Ribeiro Jr et al. (2012) com discussão sobre aspectos computacionais e implementação em R.

Para comparação de modelos será considerado a avaliação da log-verossimilhança no ponto $\hat{\underline{\theta}}$ e os critérios de informação AIC, proposto inicialmente para modelos de séries temporais por Akaike em 1974 e estendido para qualquer modelo estimado por máxima verossimilhança e o critério BIC, modificação do AIC proposto por Schwarz em 1978 que tende a ser mais rigoroso na indicação de modelos mais complexos.

Ainda aspectos inferenciais computacionalmente intensivos serão avaliados como intervalos de confiança para estimativas dos parâmetros via perfis da verossimilhança e intervalos de confiança para os valores preditos baseados em *bootstrap*, necessários quando o número de observações é pequeno.

4 Cronograma de Atividades



REFERÊNCIAS

CONWAY, R. W.; MAXWELL, W. L. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, v. 12, p. 132—-136, 1962. Citado na página 4.

KING, G. Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator. *American Journal of Political Science*, v. 33, n. 3, p. 762—-784, aug 1989. ISSN 00925853. Disponível em: <http://www.jstor.org/stable/2111071?origin=crossref>. Citado na página 3.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, p. 370–384, 1972. Citado 2 vezes nas páginas 3 e 8.

PAULA, G. A. *Modelos de regressão com apoio computacional*. IME-USP São Paulo, 2013. Disponível em: <https://www.ime.usp.br/~giapaula/textoregressao.htm>. Citado 3 vezes nas páginas 3, 4 e 8.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <http://www.R-project.org/>. Citado na página 7.

Ribeiro Jr, P. J. et al. Métodos computacionais para inferência com aplicações em R. In: *20º Simpósio Nacional de Probabilidade e Estatística*. [s.n.], 2012. p. 282. Disponível em: <http://leg.ufpr.br/doku.php/cursos:mcie>. Citado 2 vezes nas páginas 4 e 8.

SELLERS, K.; LOTZE, T. *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*. [S.l.], 2011. R package version 0.3.4. Disponível em: <http://CRAN.R-project.org/package=COMPoissonReg>. Citado na página 7.

SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. *Annals of Applied Statistics*, v. 4, n. 2, p. 943–961, 2010. ISSN 19326157. Citado 2 vezes nas páginas 5 e 7.

SHMUELI, G. et al. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, v. 54, n. 1, p. 127–142, 2005. ISSN 00359254. Citado 2 vezes nas páginas 4 e 5.

ZEVIANI, W. M. et al. The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, n. October, p. 1–11, 2014. ISSN 0266-4763. Disponível em: <http://dx.doi.org/10.1080/02664763.2014.922168>. Citado na página 7.