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

Eduardo Elias Ribeiro Junior
Orientação: Prof. Dr. Walmes Marques Zeviani

Projeto de Pesquisa - Laboratório A
Departamento de Estatística (DEST)
Universidade Federal do Paraná (UFPR)

28 de novembro de 2015

Sumário

Sumário

- Disciplinas:
  - Análise de Regressão Linear (CE071 - 2014/1S)
  - Modelos Lineares Generalizados (CE225 - 2014/2S)
  - Estatística Computacional II (CE089 - 2014/2S)

- Disciplinas:
  - Análise de Regressão Linear (CE071 - 2014/1S)
  - Modelos Lineares Generalizados (CE225 - 2014/2S)
  - Estatística Computacional II (CE089 - 2014/2S)

- Trabalho proposto na disciplina CE089:
  - Distribuição Conway-Maxwell-Poisson
  - Simulação, métodos de estimação, função de verossimilhança, inferência estatística

- Disciplinas:
  - Análise de Regressão Linear (CE071 - 2014/1S)
  - Modelos Lineares Generalizados (CE225 - 2014/2S)
  - Estatística Computacional II (CE089 - 2014/2S)

- Trabalho proposto na disciplina CE089:
  - Distribuição Conway-Maxwell-Poisson
  - Simulação, métodos de estimação, função de verossimilhança, inferência estatística

- Sugestão de leitura do artigo *The Gamma-count distribution in the analysis of experimental underdispersed data* por Zeviani et al., (2014):
  - Apresentação da distribuição *Count-gama* (concorrente à *Conway-Maxwell-Poisson*)
  - Análise de dados utilizando um modelo de regressão
  - Discussão de aspectos inferenciais

Sumário

Dados de contagem

Representam o número de ocorrências de um evento de interesse em um domínio específico.

Se Y é uma v.a de contagem, $y \in \mathbb{Z}_+$, ou seja, $y = 0, 1, 2, ...$

Dados de contagem

Representam o número de ocorrências de um evento de interesse em um domínio específico.

Se Y é uma v.a de contagem, $y \in \mathbb{Z}_+$, ou seja, $y = 0, 1, 2, ...$

Exemplos:

- Número de filhos por casal;
- Número de indivíduos infectados por uma doença;
- Número de insetos mortos após k dias da aplicação de inseticida;
- ...

Modelos de regressão

Permitem a inclusão de variáveis independentes (covariáveis) para:

- Descrever a relação entre a variável resposta e as variáveis preditoras; e
- Realizar predições por meio do modelo estabelecido.

Modelos de regressão

Permitem a inclusão de variáveis independentes (covariáveis) para:

- Descrever a relação entre a variável resposta e as variáveis preditoras; e
- Realizar predições por meio do modelo estabelecido.

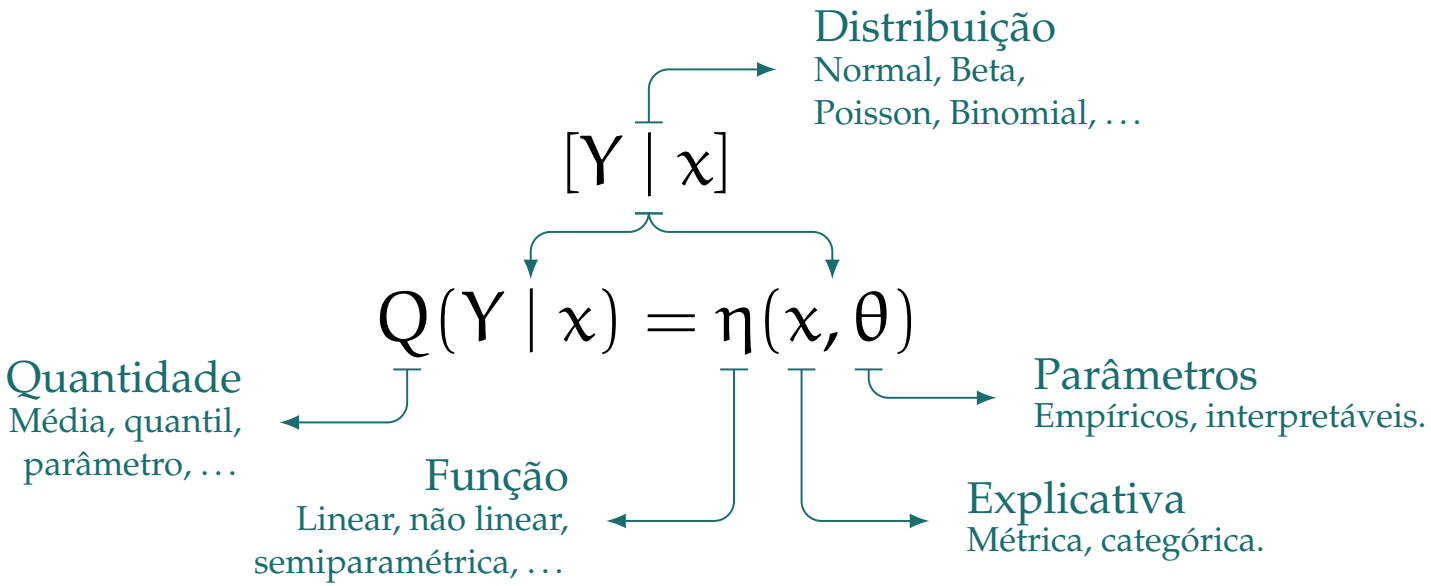

Figure 1: Representação esquemática de um modelo de regressão

Modelo Poisson

Densidade de probabilidade

$$\Pr(Y = y) = \frac{\lambda^y}{y! e^{\lambda}} \quad y \in \mathbb{Z}_+ \tag{1}$$

Modelo Poisson

Densidade de probabilidade

$$\Pr(Y = y) = \frac{\lambda^y}{y! e^{\lambda}} \quad y \in \mathbb{Z}_+ \tag{1}$$

Propriedades

- $\frac{P(Y=y-1)}{P(Y=y)} = \frac{y}{\lambda}$
- $E(Y) = \lambda$
- $V(Y) = \lambda$

Equidispersão

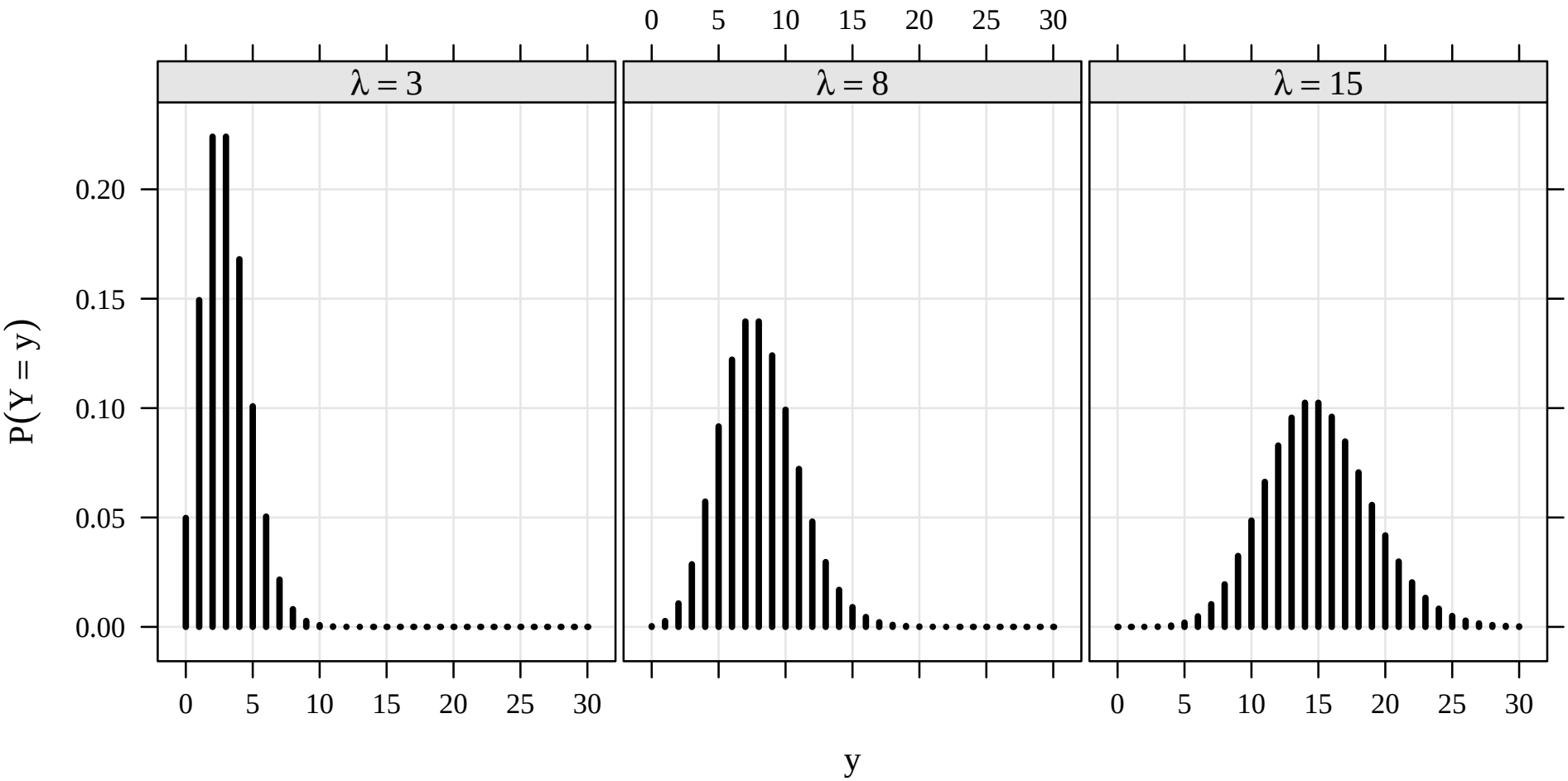

Figure 2: Densidade de probabilidade da distribuição Poisson

Abordagens para fuga da suposição

- **Modelo quase-Poisson**

  $V(Y) = \phi V(\mu)$
  Nesta abordagem estima-se ϕ separadamente:

  - Produz as mesmas estimativas pontuais do que o modelo Poisson;
  - Corrige os erros-padrão das estimativas;
  - Não é possível recuperar a verdadeira distribuição de Y;

Abordagens para fuga da suposição

- **Modelo quase-Poisson**

  $V(Y) = \phi V(\mu)$
  Nesta abordagem estima-se ϕ separadamente:

  - Produz as mesmas estimativas pontuais do que o modelo Poisson;
  - Corrige os erros-padrão das estimativas;
  - Não é possível recuperar a verdadeira distribuição de Y;

- **Modelo de efeitos aleatórios**

  $g(\mu) = X\underline{\beta} + Z\underline{b}$
  Onde b são efeitos aleatórios, variáveis não observadas (latentes) provenientes de uma distribuição de probabilidades.

  - Contemplam a estrutura de delineamento experimentada;
  - Capturam (somente) a variabilidade extra especificada pelo modelo;
  - São computacionalmente intensivos;

Modelo COM-Poisson

Densidade de probabilidade

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} \quad y \in \mathbb{Z}_+ \tag{2}$$

$$\text{onde } Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}; \text{ e} \qquad \lambda > 0 \text{ e } \nu \geqslant 0$$

Modelo COM-Poisson

Densidade de probabilidade

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} \quad y \in \mathbb{Z}_+ \tag{2}$$

$$\text{onde } Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}; \text{ e} \qquad \lambda > 0 \text{ e } \nu \geqslant 0$$

Propriedades

- $\frac{P(Y=y-1)}{P(Y=y)} = \frac{y^\nu}{\lambda}$
- $E(Y) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu}$
- $V(Y) \approx \frac{1}{\nu} E(Y)$
- $E(Y^\nu) = \lambda$

Modelo COM-Poisson

Densidade de probabilidade

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} \quad y \in \mathbb{Z}_+ \tag{2}$$

$$\text{onde } Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}; \text{ e} \qquad \lambda > 0 \text{ e } \nu \geqslant 0$$

Propriedades

- $\frac{P(Y=y-1)}{P(Y=y)} = \frac{y^\nu}{\lambda}$
- $E(Y) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu}$
- $V(Y) \approx \frac{1}{\nu} E(Y)$
- $E(Y^\nu) = \lambda$

Casos particulares

- Distribuição Poisson, quando $\nu = 1$
- Distribuição Bernoulli, quando $\nu \to \infty$
- Distribuição Geométrica, quando $\nu = 0,\ \lambda < 1$

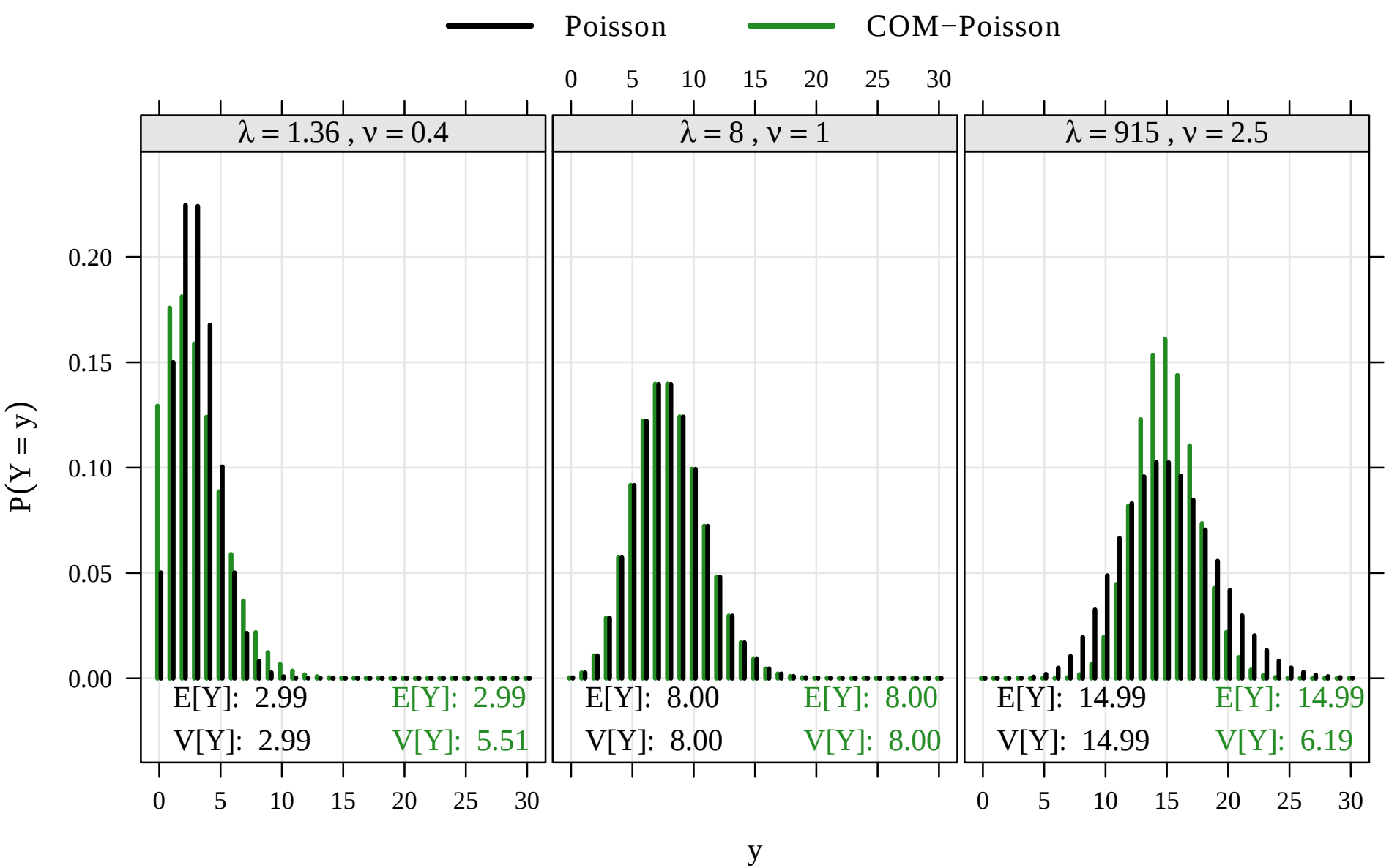

Figure 3: Densidade de probabilidade da distribuição COM-Poisson comparada com a Poisson

Extensões do modelo de regressão COM-Poisson

- **Excesso de zeros**
  O mecanismo gerador das variáveis aleatórias de contagem é proveniente de duas distribuições.

Extensões do modelo de regressão COM-Poisson

- **Excesso de zeros**
  O mecanismo gerador das variáveis aleatórias de contagem é proveniente de duas distribuições.

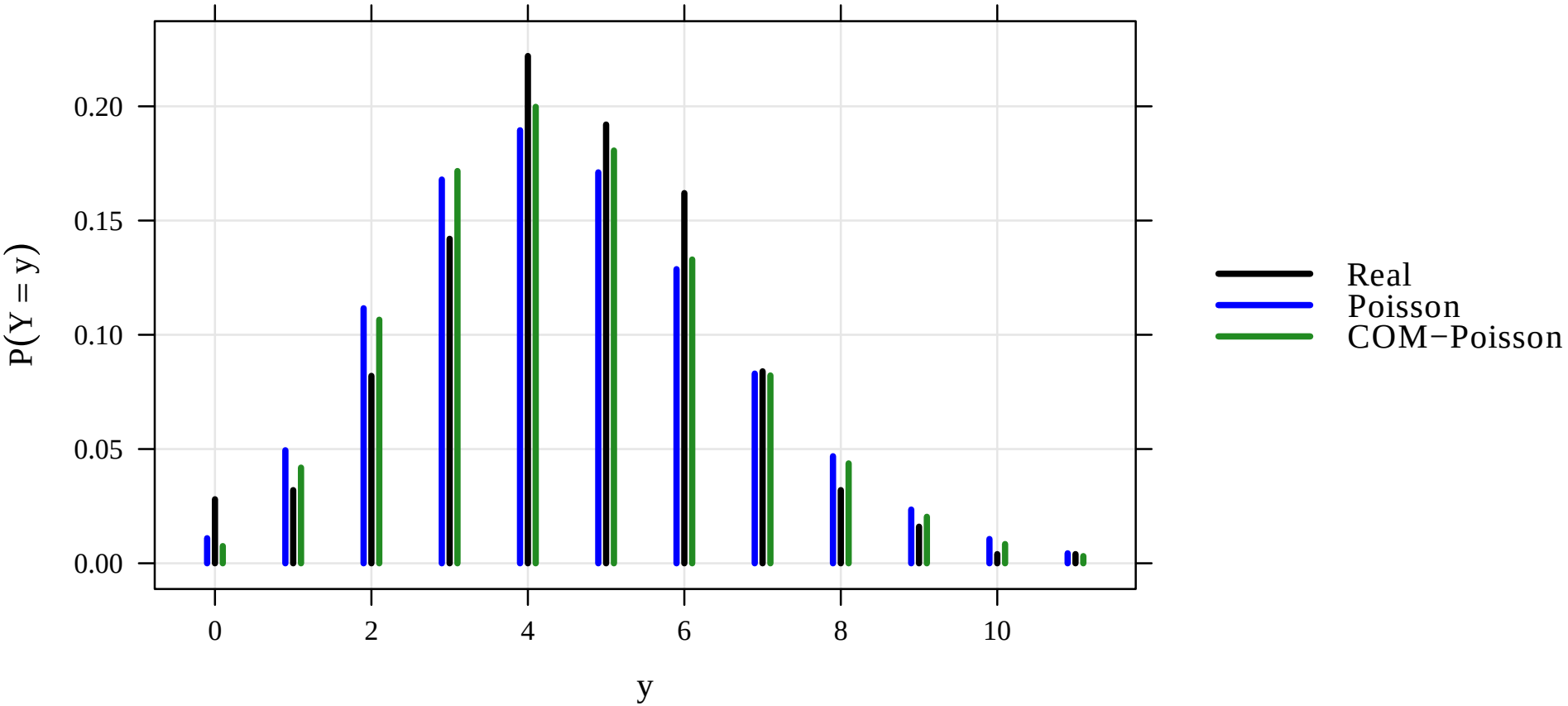

Figure 4: Contagens que apresentam excesso de zeros

Extensões do modelo de regressão COM-Poisson

- **Efeitos aleatórios**
  Correlação entre grupos de indivíduos induzida pelo delineamento experimental ou estrutura do problema.

Extensões do modelo de regressão COM-Poisson

- **Efeitos aleatórios**
  Correlação entre grupos de indivíduos induzida pelo delineamento experimental ou estrutura do problema.

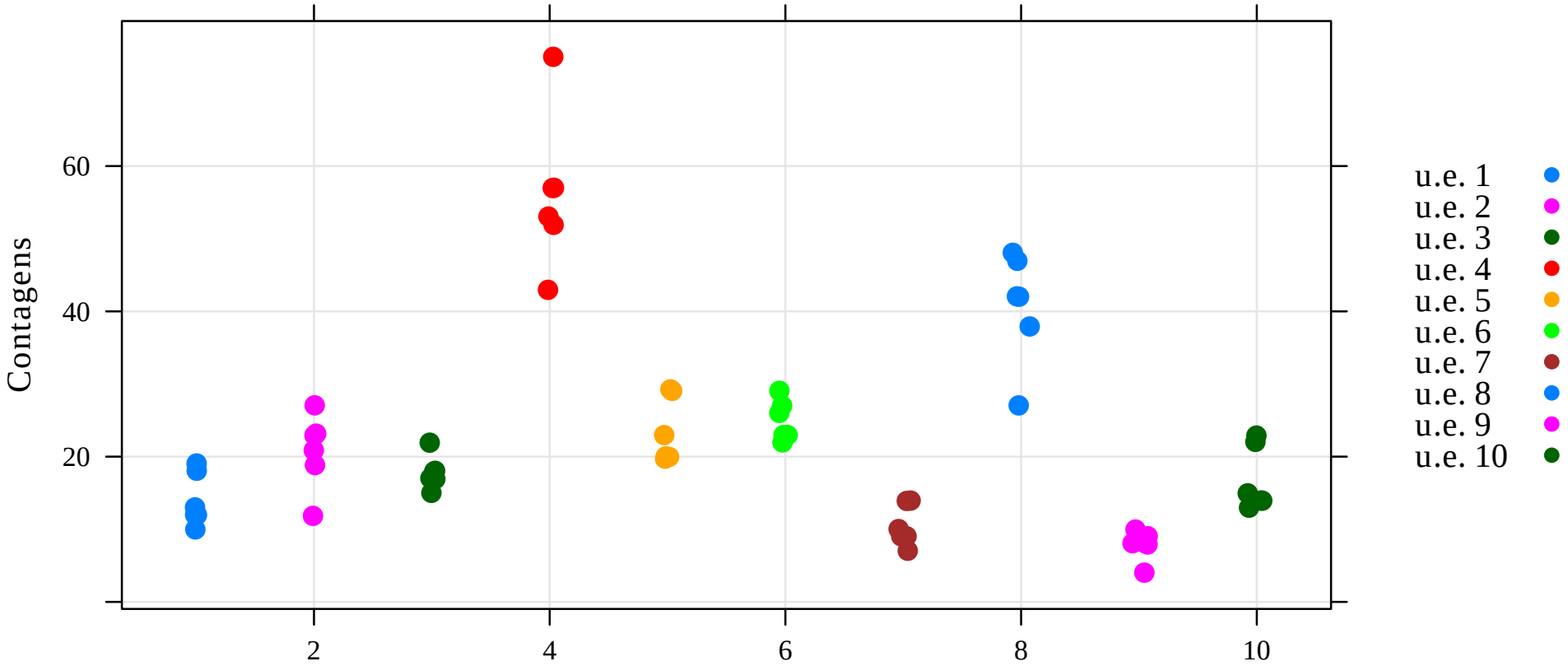

Figure 5: Contagens que apresentam um efeito aleatório da unidade experimental (u.e.)

Sumário

Objetivos gerais

- Apresentar o modelo de regressão COM-Poisson com discussão sobre aspectos inferenciais;

Objetivos gerais

- Apresentar o modelo de regressão COM-Poisson com discussão sobre aspectos inferenciais;
- Estender as aplicações para situações específicas como efeitos aleatórios e excesso de zeros; e

Objetivos gerais

- Apresentar o modelo de regressão COM-Poisson com discussão sobre aspectos inferenciais;
- Estender as aplicações para situações específicas como efeitos aleatórios e excesso de zeros; e
- Contribuir para a comunidade Estatística, principalmente aplicada, com aplicações e discussões de uma abordagem paramétrica flexível para dados de contagem.

Objetivos específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelagem de dados de contagem;

Objetivos específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelagem de dados de contagem;
- Avaliar as propriedades de soluções numéricas para i) cálculo da densidade de probabilidade e ii) estimação dos modelos de regressão de efeito fixo;

Objetivos específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelagem de dados de contagem;
- Avaliar as propriedades de soluções numéricas para i) cálculo da densidade de probabilidade e ii) estimação dos modelos de regressão de efeito fixo;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar efeitos aleatórios;

Objetivos específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelagem de dados de contagem;
- Avaliar as propriedades de soluções numéricas para i) cálculo da densidade de probabilidade e ii) estimação dos modelos de regressão de efeito fixo;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar efeitos aleatórios;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar contagens com excesso de zeros;

Objetivos específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelagem de dados de contagem;
- Avaliar as propriedades de soluções numéricas para i) cálculo da densidade de probabilidade e ii) estimação dos modelos de regressão de efeito fixo;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar efeitos aleatórios;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar contagens com excesso de zeros;
- Fazer aplicação do modelo COM-Poisson e suas extensões desenvolvidas à dados reais e simulados; e

Objetivos específicos

- Apresentar e discutir aspectos da distribuição COM-Poisson para modelagem de dados de contagem;
- Avaliar as propriedades de soluções numéricas para i) cálculo da densidade de probabilidade e ii) estimação dos modelos de regressão de efeito fixo;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar efeitos aleatórios;
- Propor e implementar a extensão do modelo de regressão COM-Poisson para acomodar contagens com excesso de zeros;
- Fazer aplicação do modelo COM-Poisson e suas extensões desenvolvidas à dados reais e simulados; e
- Fazer comparações com as abordagens já utilizadas para as situações estudadas: Poisson, Quase-Poisson, Binomial Negativo, Poisson de efeito aleatório.

Sumário

Conjunto de dados

A pesquisa tem como um dos objetivos a avaliação do método, portanto pretende-se utilizar vários conjuntos de dados.

Dados de desfolha

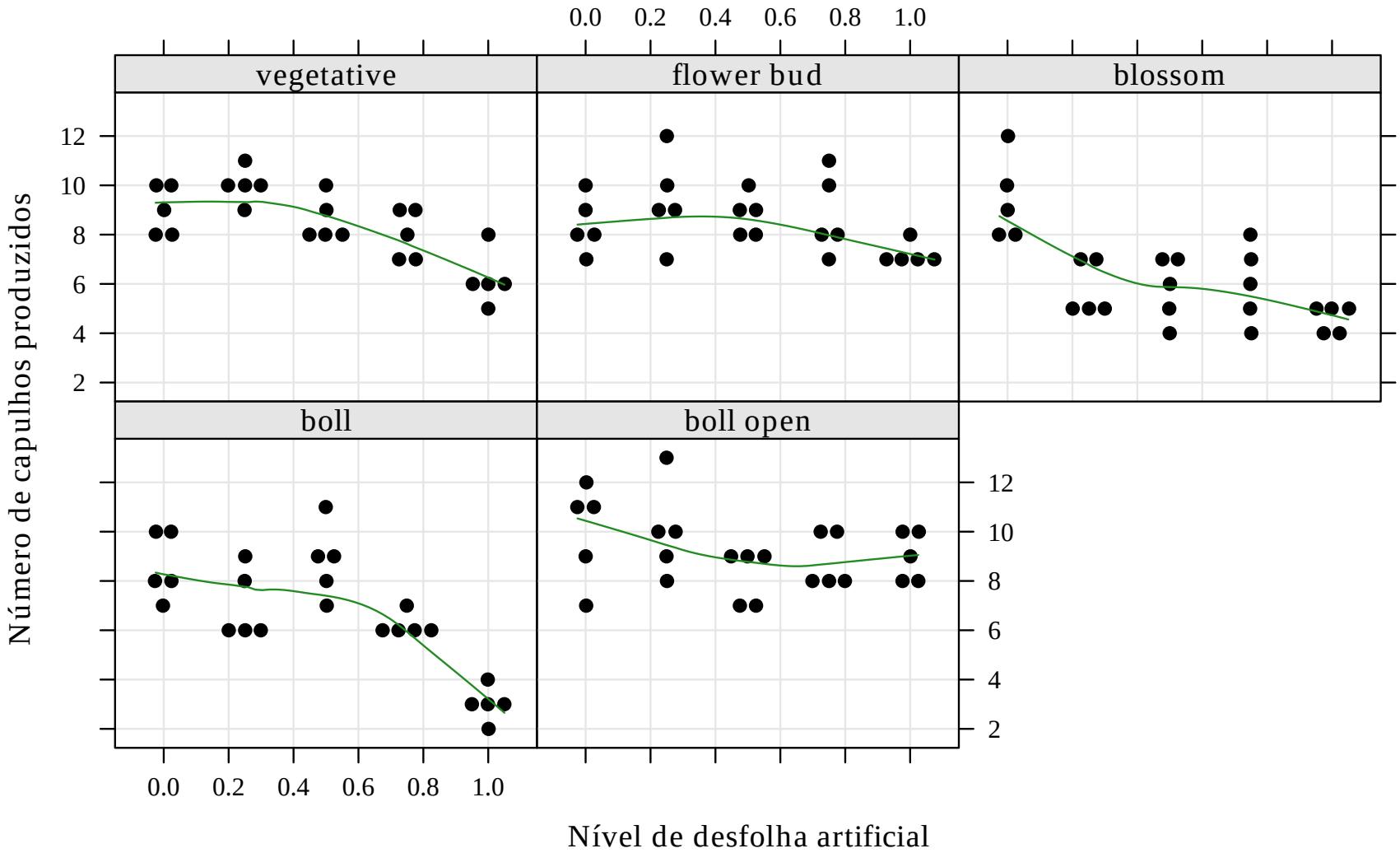

Figure 6: Número de capulhos produzidos pelo nível de desfolha estratificado por estágio da planta

Resursos Computacionais

Software R, versão 3.2

Pacotes:

- `COMPoissonReg` (SELLERS; LOTZE, 2011);
- `compoisson` (DUNN, 2012);
- `CompGLM` (POLLOCK, 2014);
- Bibliotecas para elaboração de gráficos e otimização de funções;

Modelos para excesso de zeros

- **Modelos de Barreira (condicionais ou truncados)**
  Hurdle models

$$\Pr(Y = y) = \begin{cases} \pi & \text{se } y = 0, \\ (1-\pi)\dfrac{f_*(y)}{1 - f_*(0)} & \text{se } y = 1, 2, \ldots \end{cases} \tag{3}$$

Modelos para excesso de zeros

- **Modelos de Barreira (condicionais ou truncados)**
  Hurdle models

$$\Pr(Y = y) = \begin{cases} \pi & \text{se } y = 0, \\ (1-\pi)\dfrac{f_*(y)}{1 - f_*(0)} & \text{se } y = 1, 2, \ldots \end{cases} \tag{3}$$

- **Modelos Inflacionados de Zeros (mistura)**
  e.g. Zero Inflated Poisson Regression (ZIP)

$$\Pr(Y = y) = \begin{cases} \pi + (1-\pi)f_*(0) & \text{se } y = 0, \\ (1-\pi)f_*(y) & \text{se } y = 1, 2, \ldots \end{cases} \tag{4}$$

Modelos para excesso de zeros

- **Modelos de Barreira (condicionais ou truncados)**
  Hurdle models

$$\Pr(Y = y) = \begin{cases} \pi & \text{se } y = 0, \\ (1-\pi)\dfrac{f_*(y)}{1 - f_*(0)} & \text{se } y = 1, 2, \ldots \end{cases} \tag{3}$$

- **Modelos Inflacionados de Zeros (mistura)**
  e.g. Zero Inflated Poisson Regression (ZIP)

$$\Pr(Y = y) = \begin{cases} \pi + (1-\pi)f_*(0) & \text{se } y = 0, \\ (1-\pi)f_*(y) & \text{se } y = 1, 2, \ldots \end{cases} \tag{4}$$

Modelos de efeitos aleatórios

$$
\begin{aligned}
Y \mid b &\sim f_*(\mu, \phi) \\
g(\mu) &= \beta_0 + b_i \\
b_i &\sim N(0, \sigma^2)
\end{aligned}
$$

$$
Pr(Y = y) = \int_{-\infty}^{\infty} [Y \mid X, b_i][b_i] db_i \tag{5}
$$

Modelos de efeitos aleatórios

$$
\begin{aligned}
Y \mid b &\sim f_*(\mu, \phi) \\
g(\mu) &= \beta_0 + b_i \\
b_i &\sim N(0, \sigma^2)
\end{aligned}
$$

$$
Pr(Y = y) = \int_{-\infty}^{\infty} [Y \mid X, b_i][b_i] db_i \tag{5}
$$

- Métodos de integração numérica, discutidos em Ribeiro Jr et al., (2012)
  - Aproximação de Laplace
  - Quadratura Gaussiana
  - Monte Carlo (e.g. MCMC)

Métodos de estimação

- **Máxima Verossimilhança**

$$L(\underline{\theta} \mid \underline{Y}) = \prod_{i=1}^{n} f_*(y_i \mid \underline{\theta})$$

$$\hat{\underline{\theta}} \implies \max(\log(L(\underline{\theta} \mid \underline{Y})))$$

Métodos de estimação

- **Máxima Verossimilhança**

$$L(\underline{\theta} \mid \underline{Y}) = \prod_{i=1}^{n} f_*(y_i \mid \underline{\theta})$$

$$\hat{\underline{\theta}} \implies \max(\log(L(\underline{\theta} \mid \underline{Y})))$$

- **Mínimos Quadrados Ponderados Iterativamente**
  Iterative Reweighted Least Squares (IRLS)

$$\partial ll/\partial \beta_j \approx (y_i - E[Y_i]x_{ij})$$
$$\partial ll/\partial \nu \approx \log(y_i!) + E[\log(Y_i!)]$$

Métodos de estimação

- **Máxima Verossimilhança**

$$L(\underline{\theta} \mid \underline{Y}) = \prod_{i=1}^{n} f_*(y_i \mid \underline{\theta})$$

$$\hat{\underline{\theta}} \implies \max(\log(L(\underline{\theta} \mid \underline{Y})))$$

- **Mínimos Quadrados Ponderados Iterativamente**
  Iterative Reweighted Least Squares (IRLS)

$$\partial ll/\partial \beta_j \approx (y_i - E[Y_i]x_{ij})$$
$$\partial ll/\partial \nu \approx \log(y_i!) + E[\log(Y_i!)]$$

Critérios para Comparação

- **Critério de Informação de Akaike**
  Akaike Information Criterion (AIC)

$$AIC = 2k - 2\log(L(\hat{\underline{\theta}} \mid \underline{Y}))$$

Critérios para Comparação

- **Critério de Informação de Akaike**
  Akaike Information Criterion (AIC)

$$AIC = 2k - 2\log(L(\hat{\underline{\theta}} \mid \underline{Y}))$$

- **Critério de Informação Bayesiano**
  Bayesian Information Criterion (BIC)

$$BIC = \log(n)k - 2\log(L(\hat{\underline{\theta}} \mid \underline{Y}))$$

Critérios para Comparação

- **Critério de Informação de Akaike**
  Akaike Information Criterion (AIC)

$$AIC = 2k - 2\log(L(\hat{\underline{\theta}} \mid \underline{Y}))$$

- **Critério de Informação Bayesiano**
  Bayesian Information Criterion (BIC)

$$BIC = \log(n)k - 2\log(L(\hat{\underline{\theta}} \mid \underline{Y}))$$

- **Teste de razão de verossimilhanças (TRV)**

$$TRV = 2\log(L(\hat{\underline{\theta}}_p, \underline{y})) - \log(L(\hat{\underline{\theta}}_q), \underline{y})$$
$$TRV \sim \chi^2_{p-q}$$

Sumário

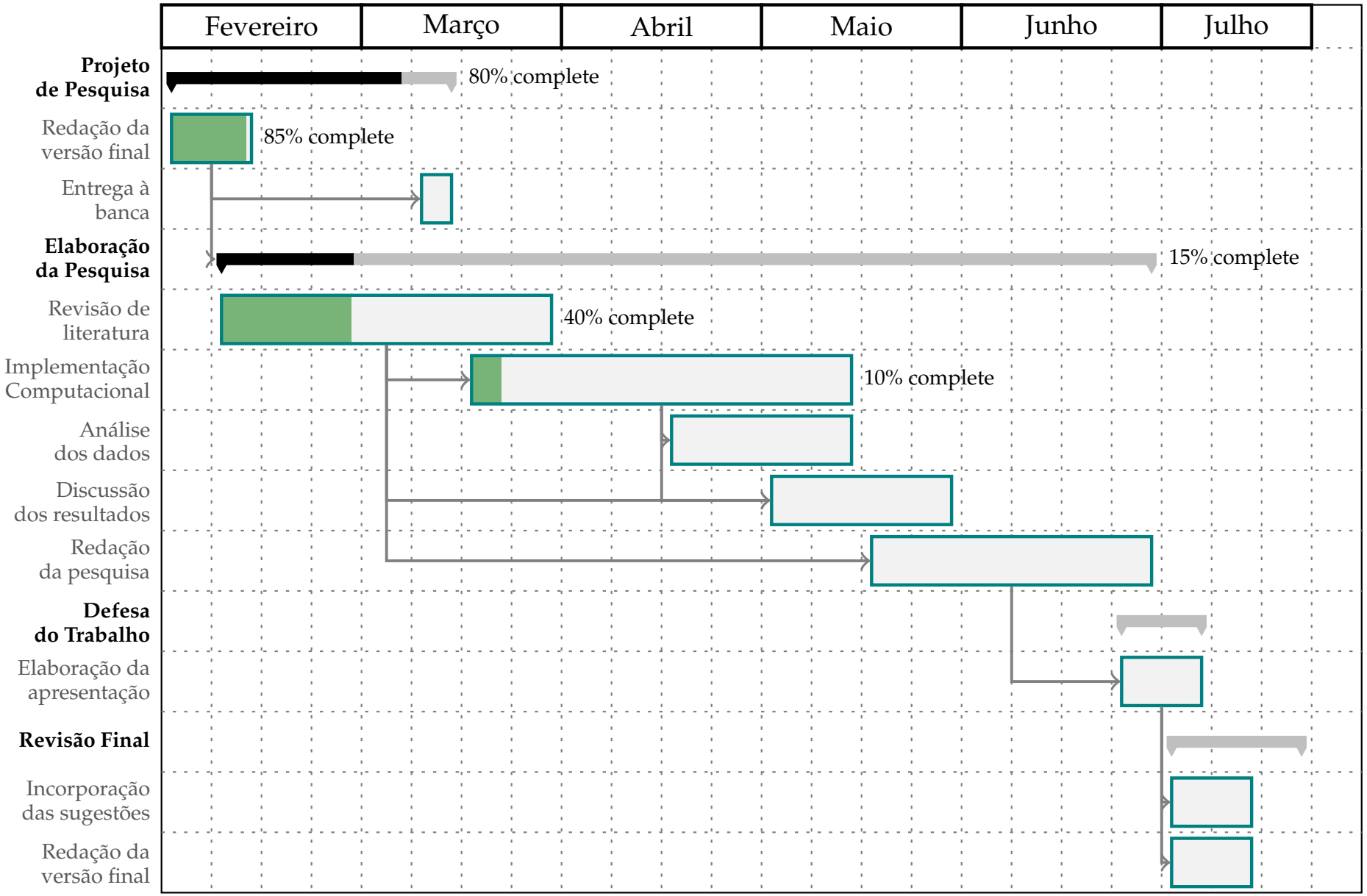

Figure 7: Cronograma de atividades para 2016

Sumário

Referências

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, ago. 1989.

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.

PAULA, G. A. **Modelos de regressão com apoio computacional**. Traducao. [s.l.] IME-USP São Paulo, 2013.

RIBEIRO JR, P. J. et al. **Métodos computacionais para inferência com aplicações em R**20º simpósio nacional de probabilidade e estatística. **Anais**...2012Disponível em: <http://leg.ufpr.br/doku.php/cursos:mcie>

SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. **Annals of Applied Statistics**, v. 4, n. 2, p. 943–961, 2010.

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.

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.