Data Science and Big Data

Cesar Augusto Taconeli

24 de agosto, 2018

Modelos lineares generalizados

Modelos lineares generalizados

Objetivo

Apresentar a teoria e aplicações dos Modelos Lineares Generalizados, propostos originalmente em Nelder e Wedderburn (1972), que configuram extensões dos modelos lineares clássicos (com erros normalmente distribuídos) e que permitem analisar a relação funcional entre um conjunto de variáveis independentes e uma variável aleatória dependente com distribuição pertencente à família exponencial de distribuições.

Sumário

Parte 1- Introdução

Uma breve reflexão...

George Box

All models are wrong but some are useful

Richard Feynman

No matter how beautiful your theory, no matter how clever you are or what your name is, if it disagrees with experiment, it's wrong.

John W. Tukey

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

Modelos Lineares

- Exemplos de modelos lineares:
  - Modelos de regressão linear;
  - Modelos de análise de variância;
  - Modelos de análise de covariância.
- Vamos usar o termo **regressão** de forma genérica, contemplando toda a classe de modelos lineares (generalizados).

Modelos Lineares

- Modelos lineares descrevem a relação entre uma variável aleatória (resposta) e um conjunto de variáveis (fatores) explicativas.

- Algumas restrições se aplicam aos modelos lineares:
  - A relação entre as variáveis (reposta e explicativas) é descrita por um conjunto de parâmetros, por meio de uma função linear;
  - Condicional aos valores das variáveis explicativas, as respostas são independentes, tem distribuição Normal e igual variância.

- Tais suposições nem sempre são verificadas na prática, tornando necessária a utilização de modelos mais flexíveis.

Modelos Lineares Generalizados

- **Origem:** Nelder e Wedderburn (1972): "Generalized Linear Models", publicado no *Journal of the Royal Statistical Society*;
- Extensão dos modelos lineares, incorporando, sob uma teoria unificada, diversos outros modelos propostos até então;
- Tais modelos permitem analisar, num contexto de análise de regressão, variáveis respostas pertencentes à **família exponencial** de distribuições;
- Como casos particulares da família exponencial temos as distribuições binomial, Poisson, Normal, Gama e Normal Inversa, dentre outras.

Ilustração - alguns problemas abordados em MLG

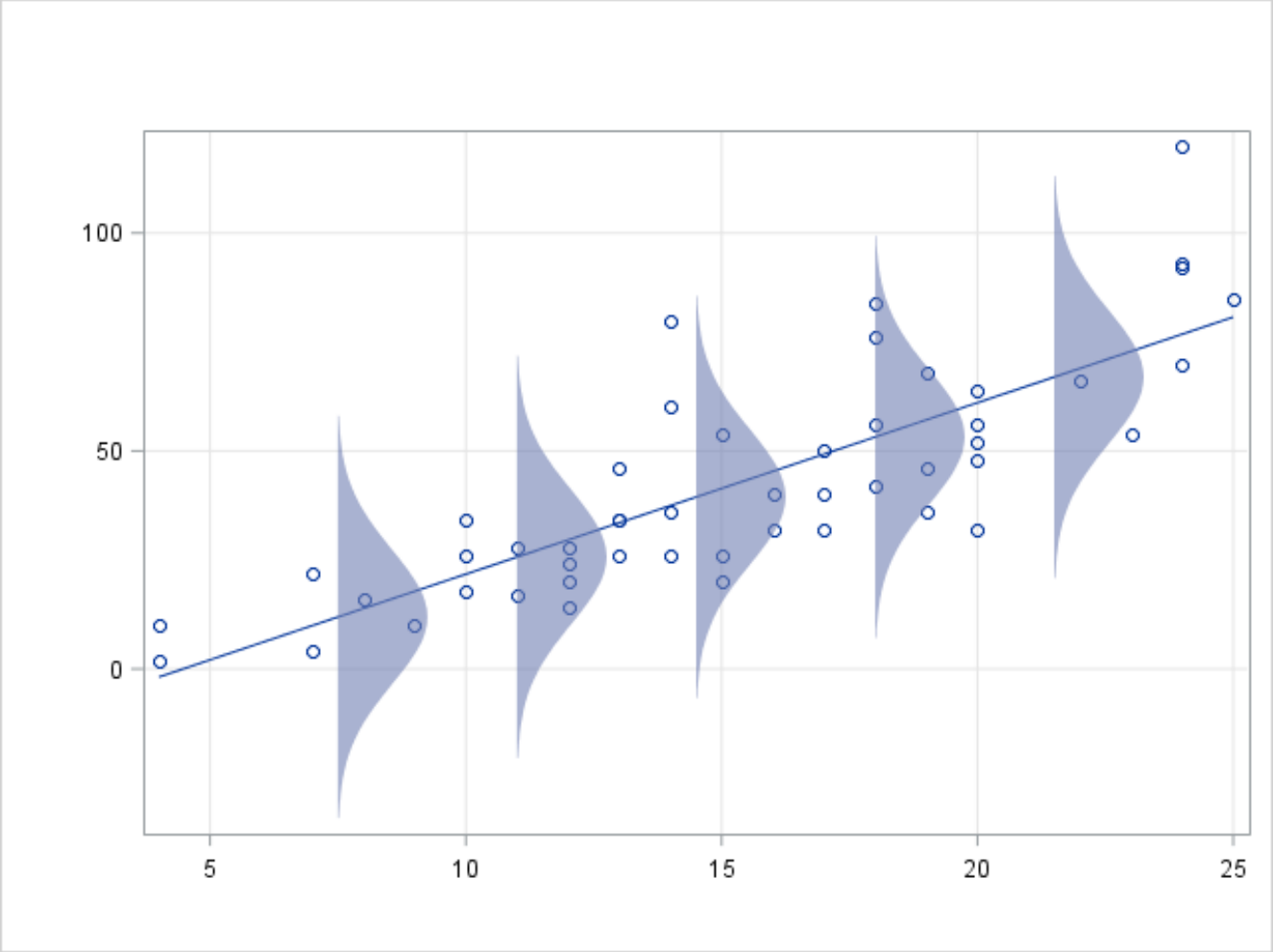


Figura 1: Regressão com erros normais - I

Ilustração - alguns problemas abordados em MLG

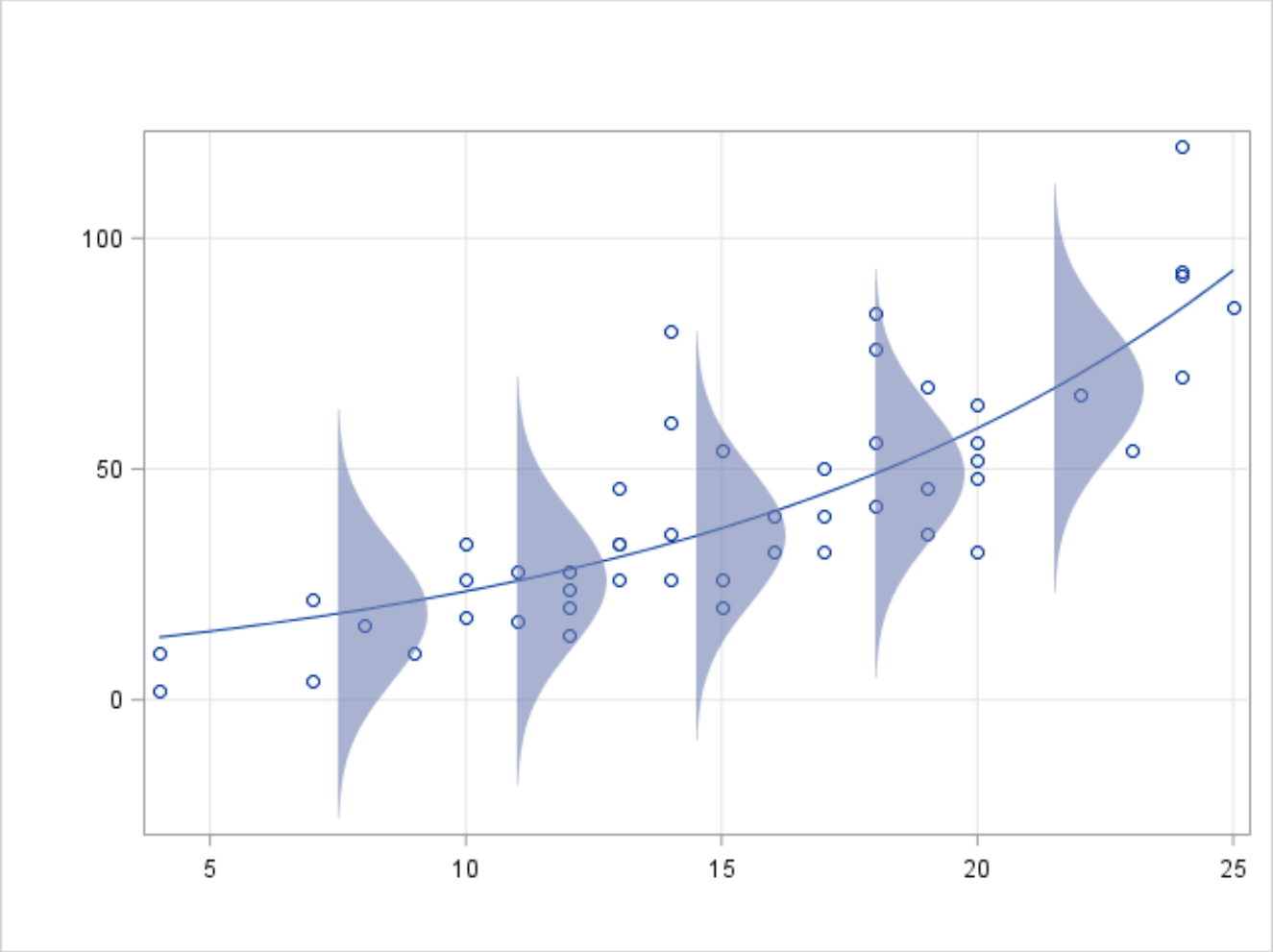


Figura 2: Regressão com erros normais - II

Ilustração - alguns problemas abordados em MLG

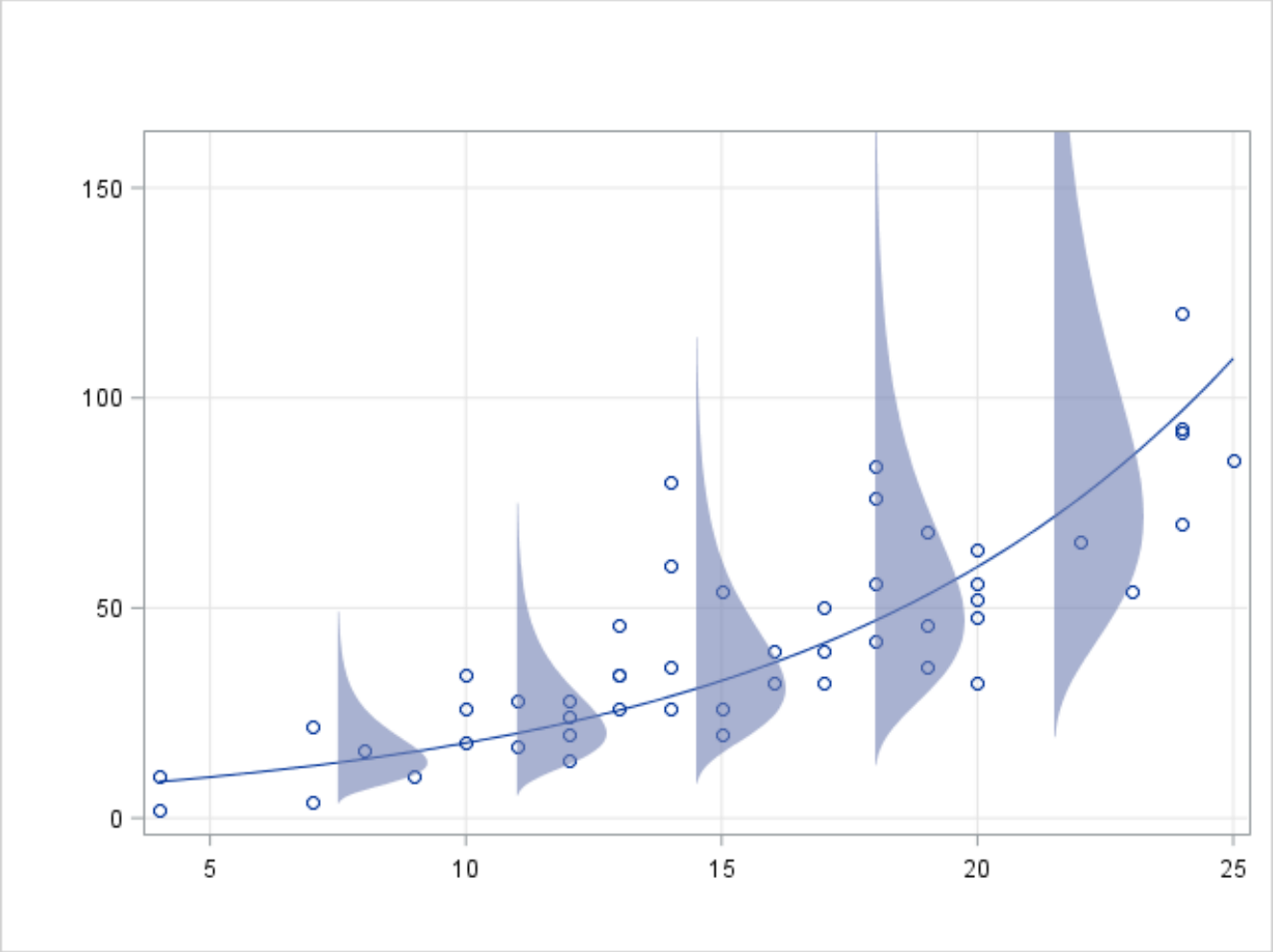


Figura 3: Regressão para dados contínuos assimétricos

Ilustração - alguns problemas abordados em MLG

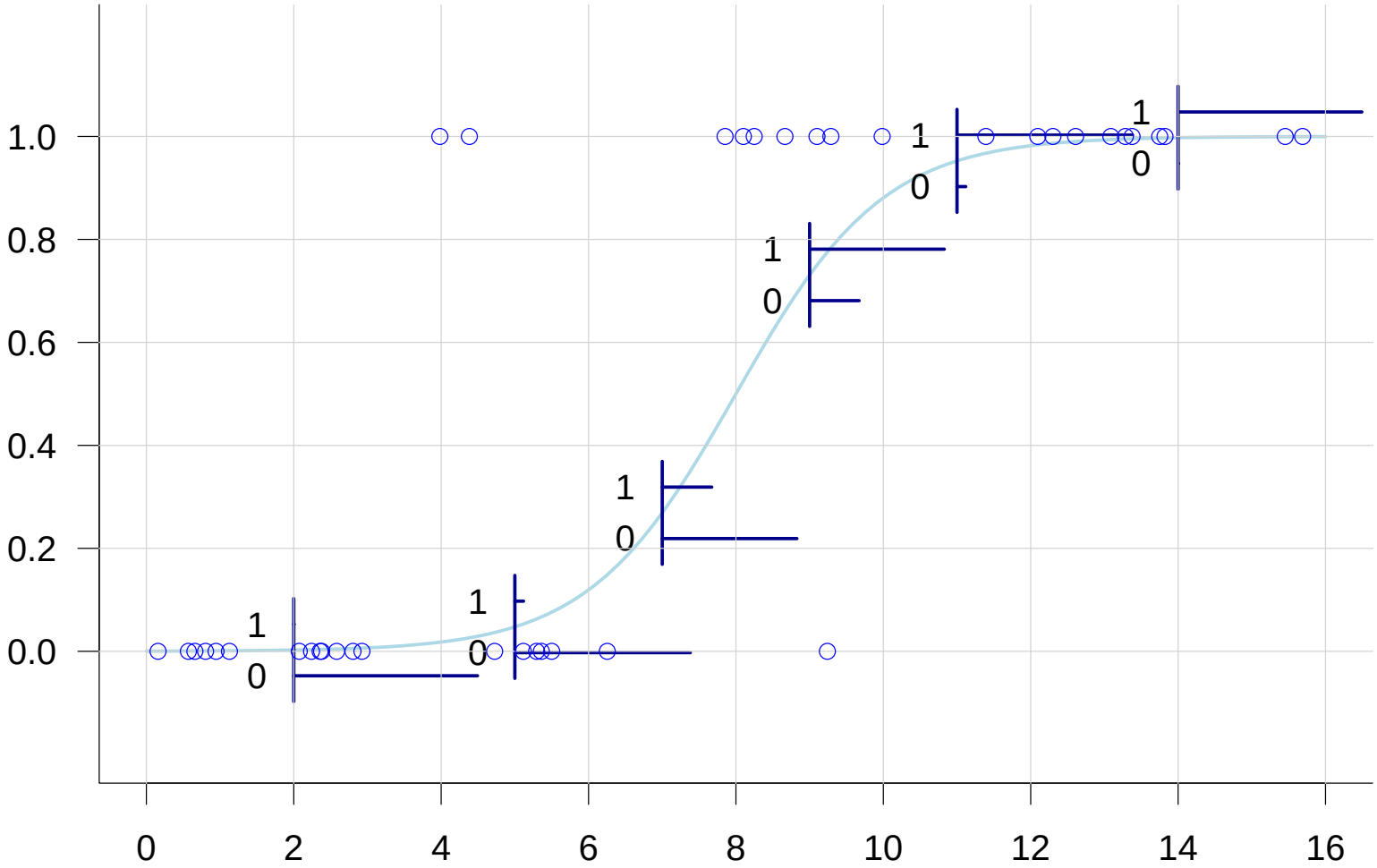
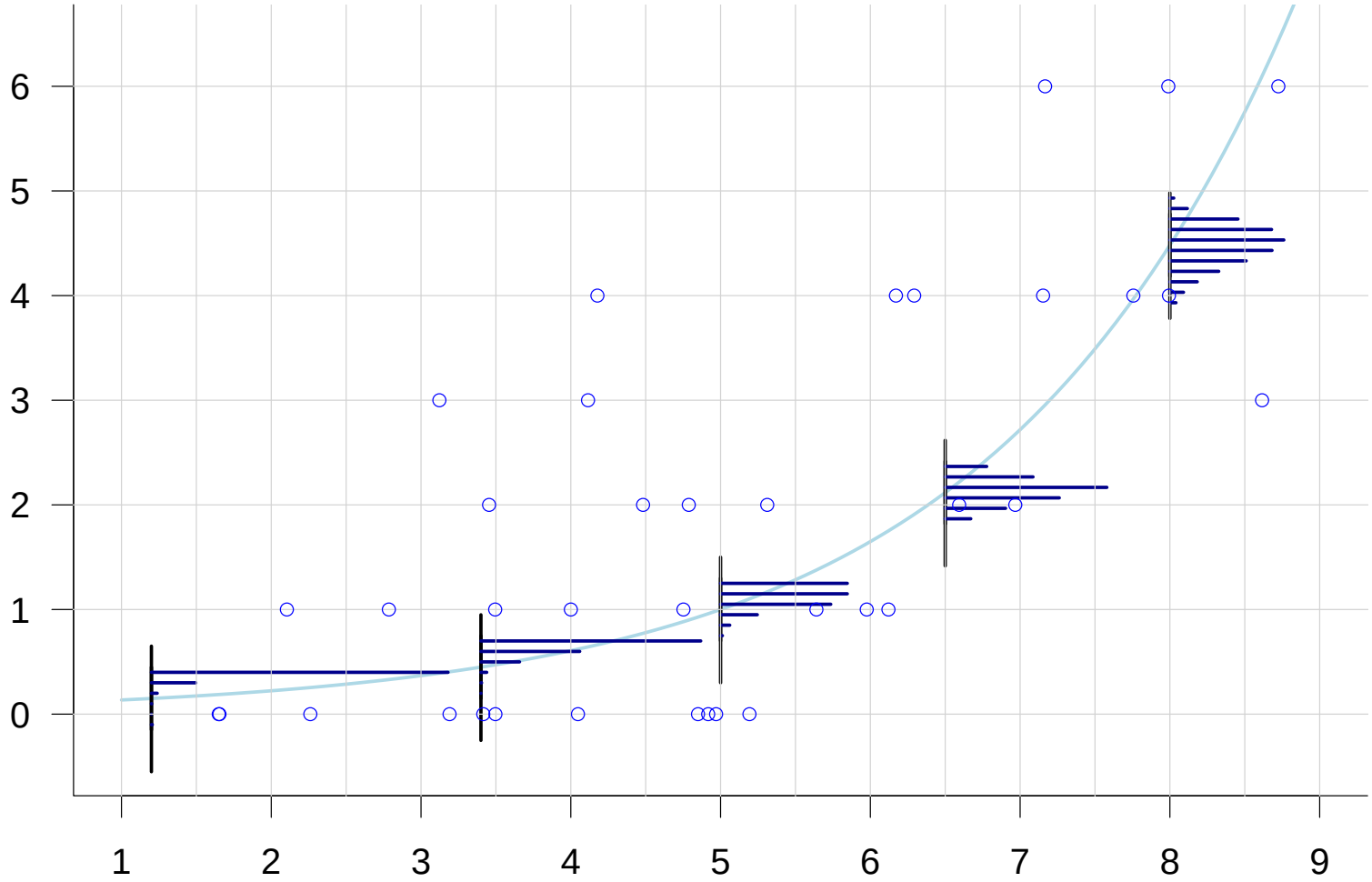


Ilustração - alguns problemas abordados em MLG



Parte 2- Família exponencial de distribuições

Componente aleatório de um modelo linear generalizado

- O componente aleatório de um modelo linear generalizado consiste em uma variável aleatória y, por meio de um conjunto de observações independentes $y_1, y_2, ..., y_n$, com distribuição pertencente à *família exponencial*.

- Mais especificamente, assumimos que a função (densidade) de probabilidades de y possa ser expressa na forma:

$$f(y_i; \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)\right\}, \quad (1)$$

sendo usualmente chamada de *forma canônica da família exponencial*.

Componente aleatório de um modelo linear generalizado

- O parâmetro θ_i é chamado *parâmetro natural* (ou *parâmetro canônico*) e ϕ o *parâmetro de dispersão* da distribuição.

- Em geral, temos $a(\phi) = \phi$ ou $a_i(\phi) = \frac{\phi}{\omega_i}$, sendo ω_i um peso particular a cada observação.

- A família exponencial de dispersão contempla diversas distribuições uni e bi-paramétricas pertencentes à família exponencial, por exemplo as distribuições binomial, poisson, normal, gama e normal inversa.

Algumas propriedades da família exponencial de dispersão

- Para distribuições pertencentes à família exponencial de dispersão, expressões para $E(y_i)$ e $Var(y_i)$ são dadas por:

$$E(y_i) = \mu_i = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} \tag{2}$$

e

$$Var(y_i) = a(\phi) \times b''(\theta_i) = a(\phi) \times \frac{\partial \mu_i}{\partial \theta_i}. \tag{3}$$

Algumas propriedades da família exponencial de dispersão

- Assim, a variância de y_i pode ser fatorada em dois componentes:
  - O primeiro ($a(\phi)$) é função de um parâmetro (ϕ) que está associado exclusivamente à dispersão de y_i (não à sua média);
  - O segundo, usualmente denotado por $V(\mu_i) = b''(\theta_i)$ e chamado *função de variância*, é função da média da distribuição, e exprime a relação variância-média de y.
- Cada distribuição pertencente à família exponencial tem sua particular função de variância e vice-versa (unicidade).

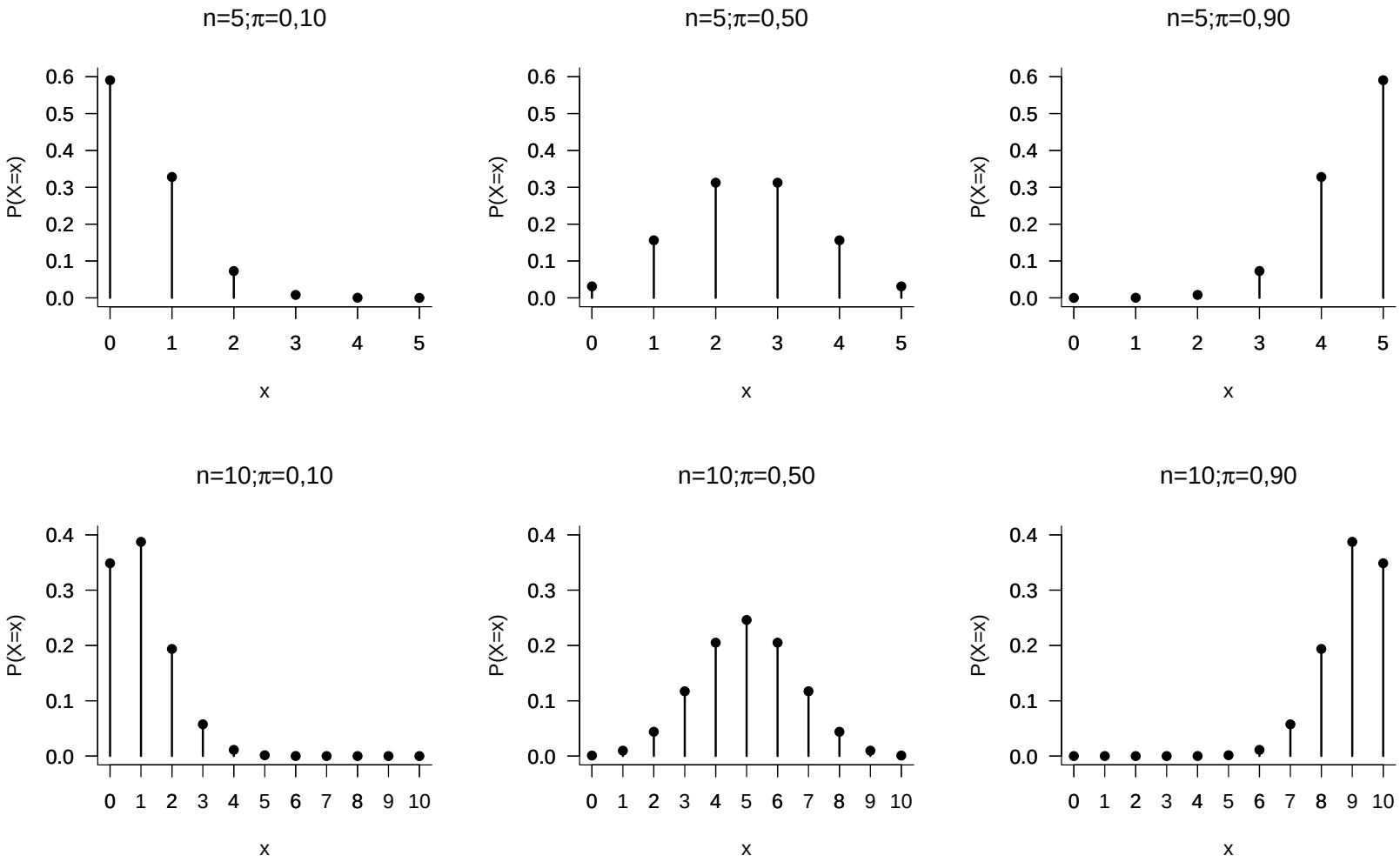
Distribuição binomial

- Uma variável aleatória x_i tem distribuição binomial se sua função de probabilidades é dada por:

$$f(x_i; n_i, \pi_i) = \binom{n_i}{x_i} \pi_i^{x_i} (1-\pi_i)^{n_i - x_i}; \quad x_i = 0, 1, 2, ..., n_i; \quad 0 < \pi_i < 1, \quad (4)$$

em que x_i corresponde à contagem de *sucessos* em n_i observações independentes de um experimento binário.

Distribuição binomial



Distribuição binomial

- Podemos expressar a distribuição binomial, de maneira alternativa, pela variável $y_i = \frac{x_i}{n_i}$, a fração amostral de sucessos, com função de probabilidades:

$$f(y_i; n_i, \pi_i) = \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i - (n_i y_i)}; y_i = 0, \frac{1}{n_i}, \frac{2}{n_i}, ..., 1; 0 < \pi_i < 1. \tag{5}$$

Distribuição binomial

- Neste caso, temos que:
  - $\theta_i = log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = log(1 + e^{\theta_i})$;
  - $a(\phi) = \frac{1}{n_i}$, $c(y_i, \phi) = \binom{n_i}{n_i y_i}$;
  - $E(y_i) = \mu_i = b'(\theta_i) = \pi_i$;
  - $V(\mu_i) = b''(\theta_i) = \mu_i(1 - \mu_i)$;
  - $Var(y_i) = \frac{\mu_i(1-\mu_i)}{n_i}$.

Distribuição binomial

- Algumas notas sobre a distribuição binomial:

* A distribuição binomial é usada na modelagem de dados binários ou de proporções discretas;

* É bem aproximada pela distribuição $Normal(\pi, \frac{\pi(1-\pi)}{m})$ quando $m\pi > 0,5$ e $0,1 \leq \pi \leq 0,9$ ou $m\pi > 25$, para qualquer valor de π.

Distribuição Poisson

- Uma variável aleatória discreta y_i tem distribuição de Poisson se sua função de probabilidades é dada por:

$$f(y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \tag{6}$$

com $y_i = 0, 1, 2, ...$ e $\mu_i > 0$.

- A distribuição Posson é frquentemente usada na modelagem da contagem de eventos de interesse em unidades de tempo ou espaço.

Distribuição Poisson

- Neste caso, temos que:
  - $\theta_i = log\left(\mu_i\right), b(\theta_i) = e^{\theta_i}$;
  - $a(\phi) = 1,\ c(y_i, \phi) = -log(y_i!)$;
  - $E(y_i) = b'(\theta_i) = \mu_i$;
  - $V(\mu_i) = b''(\theta_i) = \mu_i$;
  - $Var(y_i) = a(\phi)V(\mu_i) = \mu_i$.

Distribuição Poisson

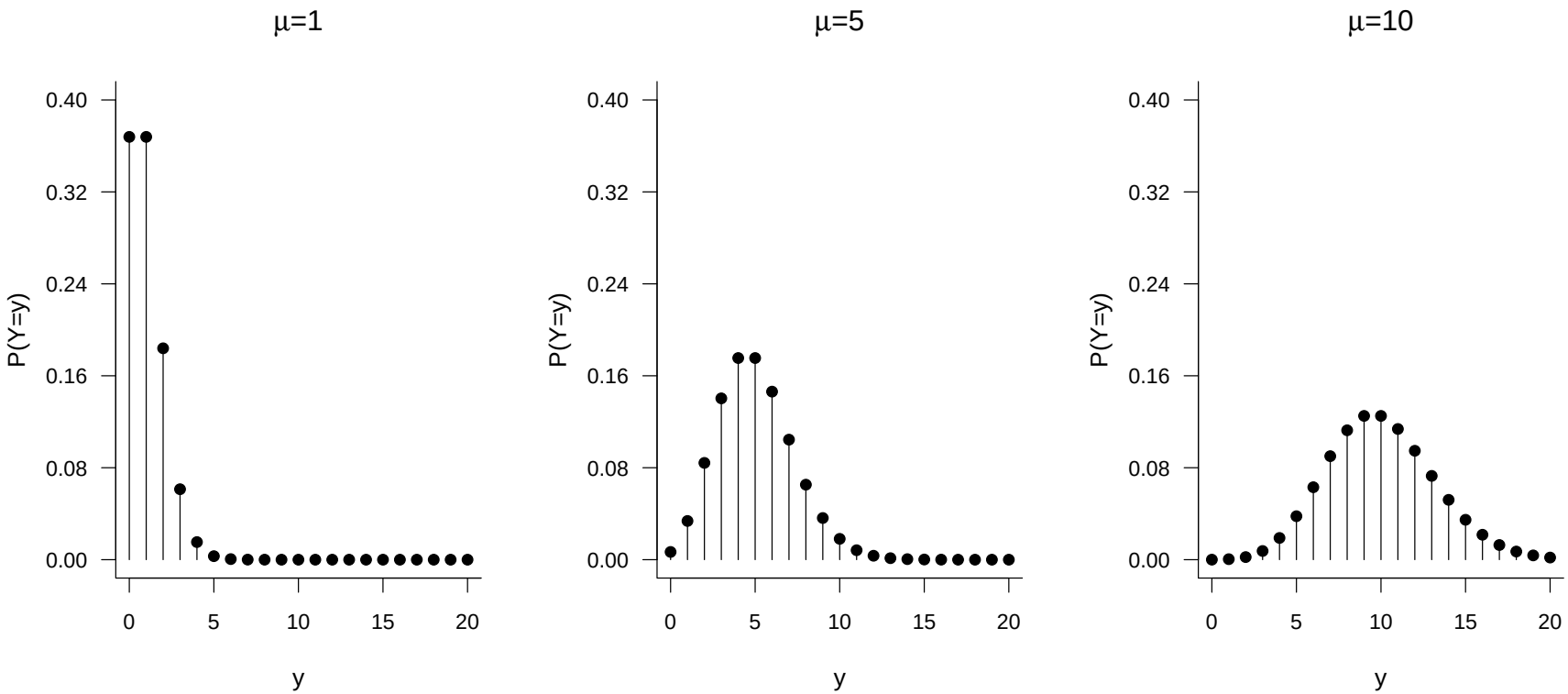


Figura 7: Ilustração - distribuição de Poisson

Distribuição Poisson

- Algumas notas sobre a distribuição de Poisson:
  - Se eventos ocorrem independente e aleatoriamente no tempo (ou espaço), com taxa média de ocorrência constante, o modelo atribui probabilidades ao número de eventos por intervalo de tempo (ou região do espaço);
  - Proporciona, em geral, uma descrição satisfatória de dados cuja variância é proporcional à média;
  - Surge como caso limite para a distribuição binomial quando $n \to \infty$ e $\pi \to 0$ (matendo fixo $\mu = n\pi$);
  - É bem aproximada pela distribuição $Normal(\mu, \mu)$ para μ suficientemente grande.

Distribuição normal

- Uma variável aleatória contínua y_i tem distribuição normal se sua função densidade de probabilidade é dada por:

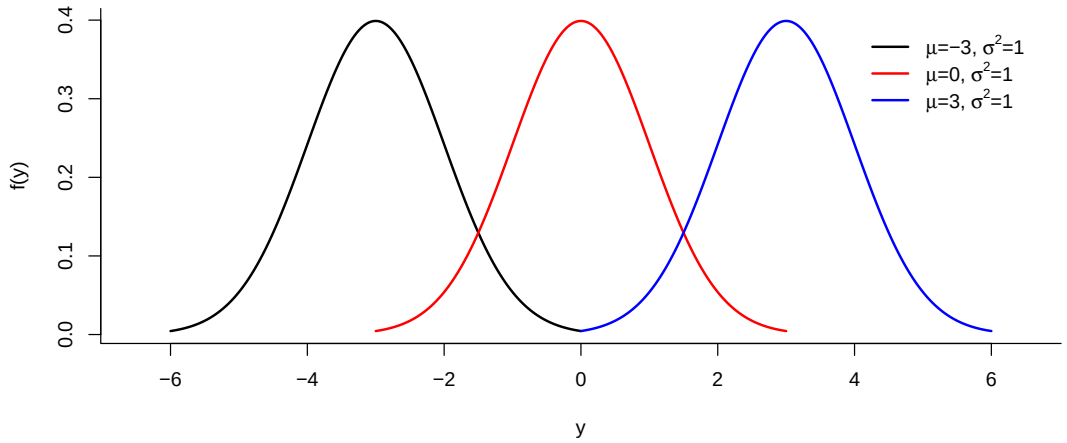
$$f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\}, \tag{7}$$

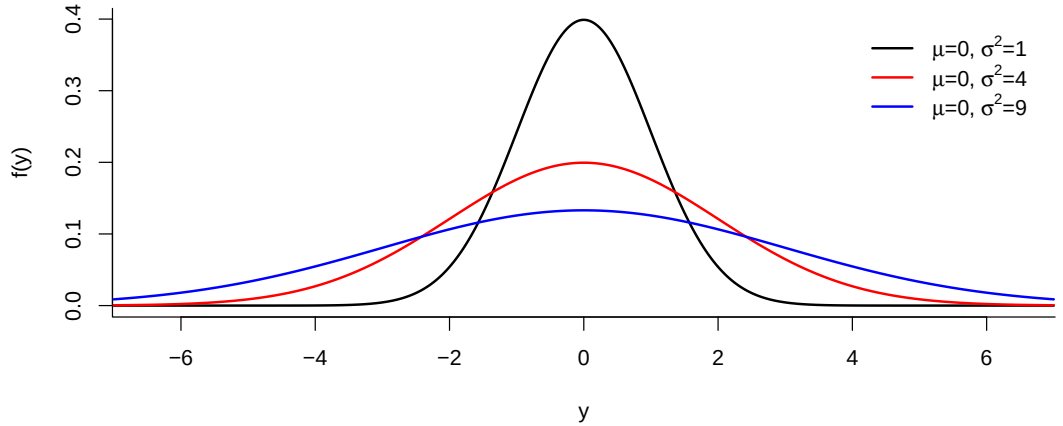
com $-\infty < y_i < \infty$; $-\infty < \mu_i < \infty$; $\sigma > 0$.

Distribuição Normal

- Neste caso, temos que:
  - $\theta_i = \mu_i$, $b(\theta_i) = \frac{\theta_i^2}{2}$;
  - $a(\phi) = \sigma^2$, $c(y_i, \phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right]$
  - $E(y_i) = b'(\theta_i) = \mu_i$;
  - $V(\mu_i) = b''(\theta_i) = 1$;
  - $Var(y_i) = a(\phi)V(\mu_i) = \sigma^2$.

Distribuição normal





Distribuição gama

- Uma variável aleatória contínua y_i tem distribuição gama se sua função densidade de probabilidade é dada por:

$$f(y_i;\mu_i,\nu)=\frac{\left(\frac{\nu}{\mu_i}\right)^{\nu}}{\Gamma(\nu)}y_i^{\nu-1}\exp\left\{-\frac{y_i\nu}{\mu_i}\right\}, \tag{8}$$

com $y_i>0$, $\mu_i>0$, $\nu>0$ e $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$.

- Uma das parametrizações alternativas da distribuição gama é a seguinte:

$$f(y;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}y^{\alpha-1}exp\left\{-\beta y\right\}, \tag{9}$$

tal que a equivalência das duas parametrizações decorre de $\mu=\frac{\alpha}{\beta}$ e $\nu=\alpha$.

Distribuição gama

- Neste caso, temos que:
  - $\theta_i = -\frac{1}{\mu_i}$, $b(\theta_i) = -log(-\theta_i)$;
  - $a(\phi) = \nu^{-1}$, $c(y_i, \phi) = \nu log(\nu y_i) - log(y_i) - log(\Gamma(\nu))$
  - $E(y_i) = b'(\theta_i) = \mu_i$;
  - $V(\mu_i) = b''(\theta_i) = \mu_i^2$;
  - $Var(y_i) = a(\phi)V(\mu_i) = \nu^{-1}\mu_i^2$.

Distribuição gama

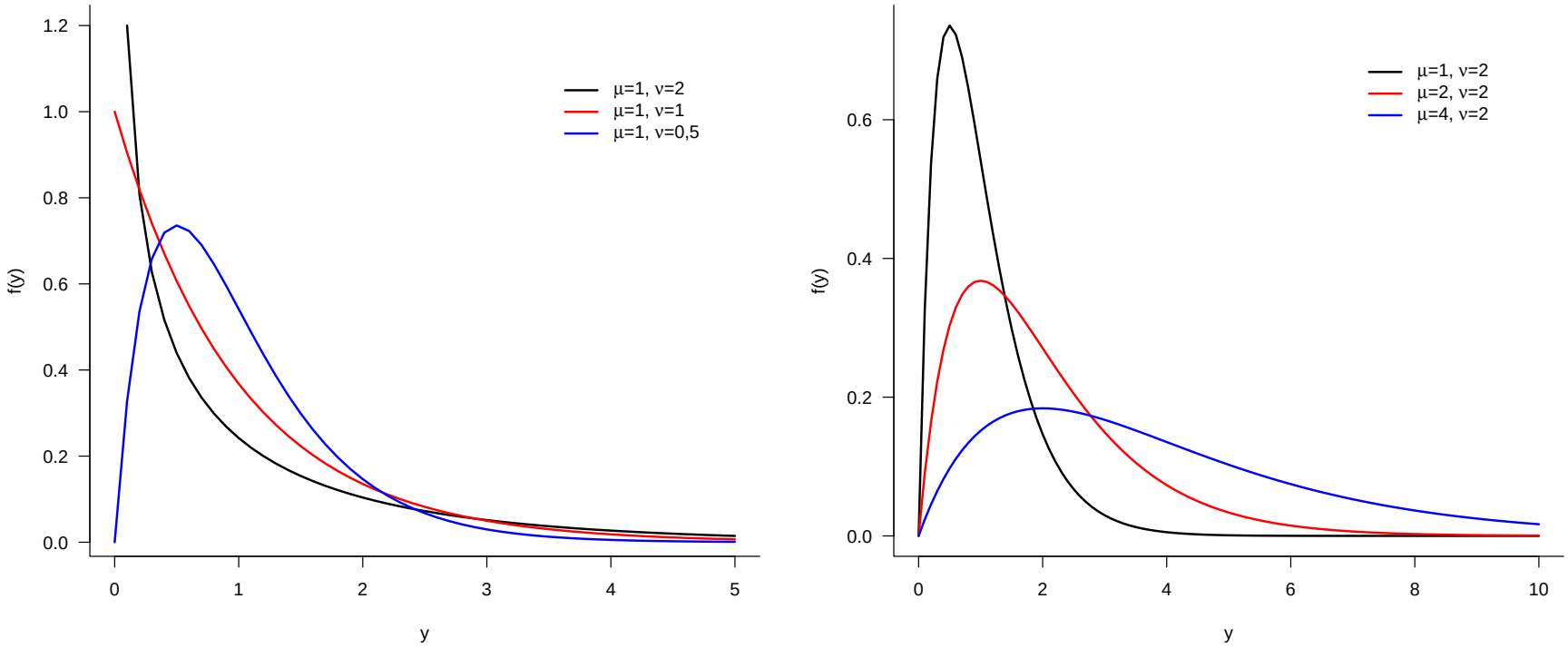


Figura 9: Ilustração - distribuição gama

Distribuição gama

- A distribuição gama é usada na análise de dados contínuos não negativos em que a variância aumenta conforme a média, particularmente no caso em que o coeficiente de variação é aproximadamente constante.

Distribuição normal inversa

- Uma variável aleatória contínua tem distribuição normal inversa se sua função densidade de probabilidade é dada por:

$$f(y_i;\mu_i,\lambda) = \sqrt{\frac{1}{2\pi\phi y_i^3}} exp\left\{-\frac{(y_i-\mu_i)^2}{2\mu^2\phi y_i}\right\}, \tag{10}$$

com $y_i > 0$, $\mu_i > 0$, $\phi > 0$.

- A distribuição normal inversa se aplica à análise de dados contínuos, não negativos com distribuição acentuadamente assimétrica.

Distribuição normal inversa

- Neste caso, temos que:
  - $\theta_i = -\frac{1}{2\mu_i^2}$, $b(\theta_i) = -(-2\theta_i)^{1/2}$;
  - $a(\phi) = \phi$, $c(y_i, \phi) = -\frac{1}{2}\left[log(2\pi\phi y_i^3) + \frac{1}{\phi y_i}\right]$
  - $E(y_i) = b'(\theta_i) = \mu_i$;
  - $V(\mu_i) = b''(\theta_i) = \mu_i^3$;
  - $Var(y_i) = a(\phi)V(\mu_i) = \phi\mu_i^3$.

Distribuição normal inversa

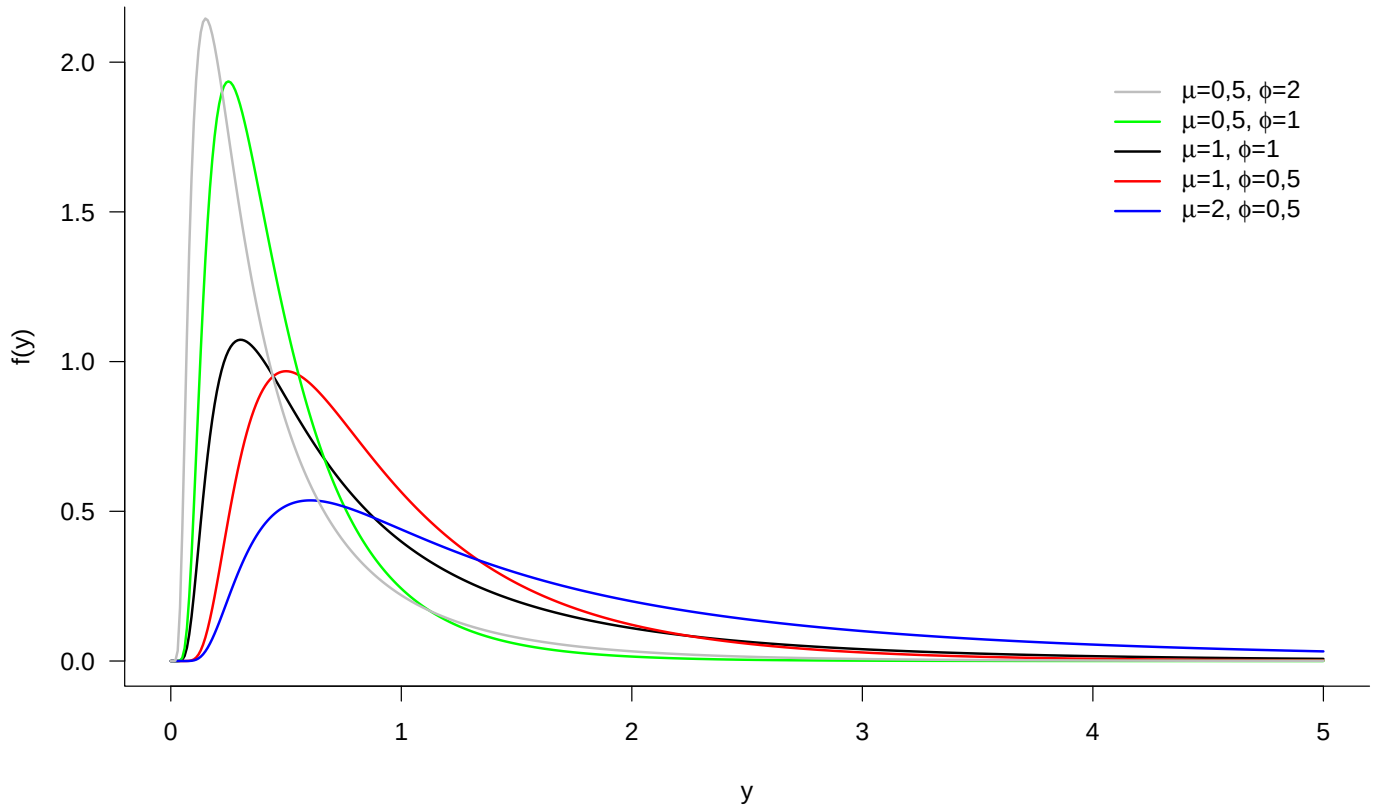


Figura 10: Ilustração - distribuição normal inversa

Distribuição binomial negativa

- Uma variável aleatória discreta Y tem distribuição binomial negativa se a sua função de probabilidades é dada por:

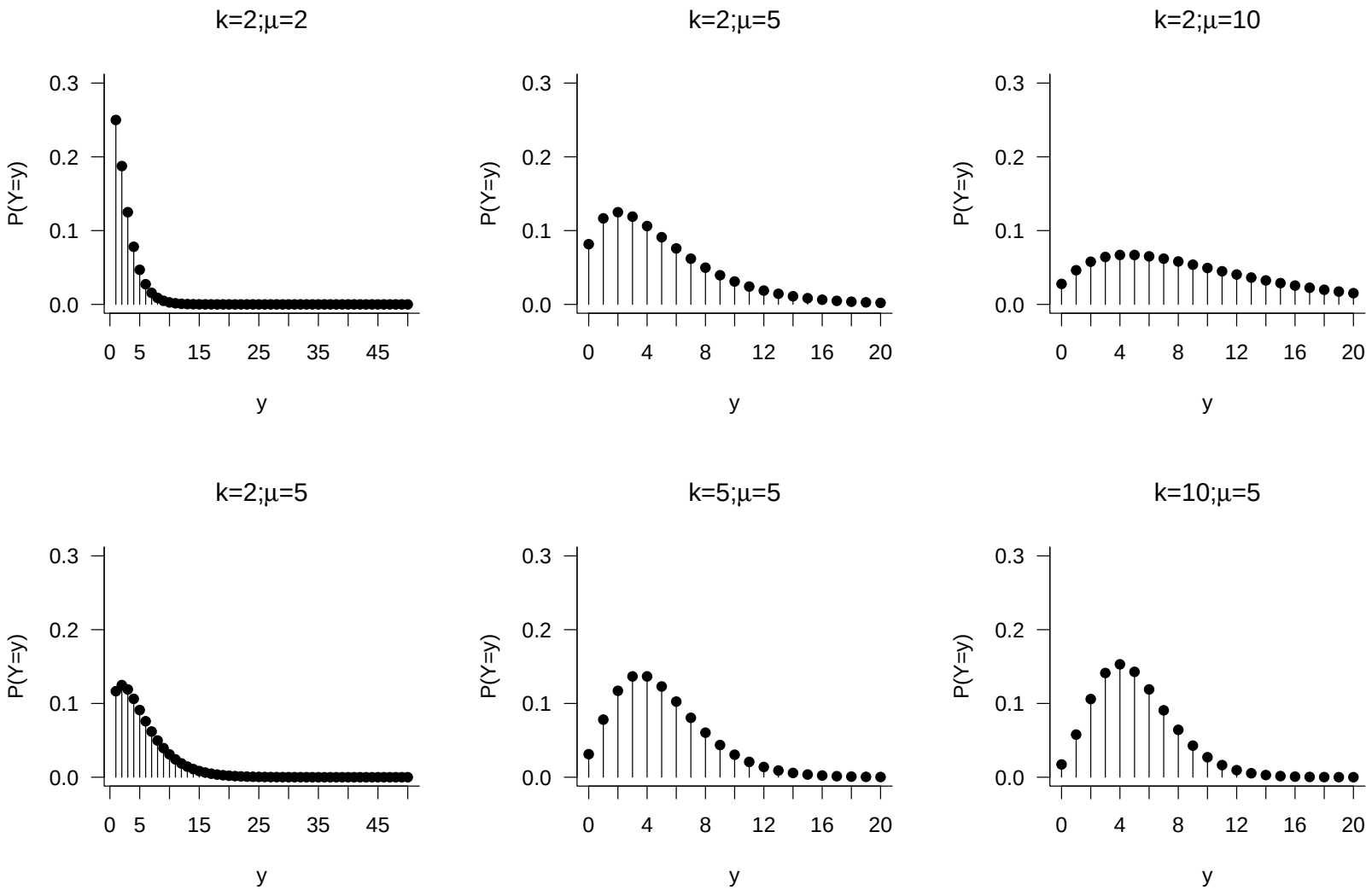
$$f(y_i; \mu_i, k) = \frac{\Gamma(k + y_i)}{\Gamma(k) y_i!} \frac{\mu_i^{y_i} k^k}{(\mu_i + k)^{k+y_i}}, \tag{11}$$

com $y_i = 0, 1, 2, ...$; $\mu_i > 0$; $k > 0$.

Distribuição binomial negativa

- Neste caso, temos que:
  - $\theta_i = log\left(\frac{\mu_i}{\mu_i+k}\right)$, $b(\theta_i) = -klog(1+e^{\theta_i})$;
  - $a(\phi) = 1$, $c(y_i, \phi) = log\left[\frac{\Gamma(k+y_i)}{\Gamma(k)y_i!}\right]$;
  - $\mu_i = b'(\theta_i) = \mu_i$;
  - $V(\mu_i) = b''(\theta_i) = \mu_i\left(\frac{\mu_i}{k}+1\right)$;
  - $Var(y_i) = \mu_i\left(\frac{\mu_i}{k}+1\right)$.

Distribuição binomial negativa



Distribuição binomial negativa

- O modelo binomial negativo é uma alternativa ao de Poisson em situações em que a variância dos dados aumenta mais rapidamente que a média;
- Para valores inteiros de k, usa-se também a denominação *modelo de Pascal*;
- Para $k = 1$, temos como caso particular a distribuição geométrica.

Parte 3- Modelo Linear Generalizado

Componentes de um modelo linear generalizado

- Um modelo linear generalizado é definido pela especificação de três componentes: o **componente aleatório**, o **componente sistemático** e uma **função de ligação**.

- **Componente aleatório:** Um conjunto de variáveis aleatórias independentes $y_1, y_2, ..., y_n$ com distribuição pertencente à família exponencial de dispersão:

$$f(y_i; \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)\right\}, \tag{12}$$

- Como vimos anteriormente, são membros dessa família as distribuições binomial, Poisson, normal, gama, normal inversa...

Componentes de um modelo linear generalizado

- **Componente sistemático:** preditor linear do modelo, em que são inseridas as covariáveis por meio de uma combinação linear de parâmetros.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}. \tag{13}$$

- **Função de ligação:** Função real, monótona e diferenciável, denotada por $c(\dot{})$, que conecta os componentes aleatório e sistemático do modelo.

Componentes de um modelo linear generalizado

- Seja $\mu_i = E(y_i|x_{i1}, x_{i2}, ..., x_{ip})$, para $i = 1, 2, ..., n$. Então:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}. \quad (14)$$

- Pelas propriedades de $g(\cdot)$, o modelo pode ser escrito de maneira equivalente por:

$$\mu_i = g^{-1}(\eta_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}. \quad (15)$$

Especificação do componente aleatório

- Requer a definição de uma distribuição de probabilidades para a variável resposta.
- A variável resposta é discreta ou contínua? Sua distribuição é simétrica? Qual o conjunto de valores com probabilidade não nula?
- Deve-se propor um modelo que tenha propriedades compatíveis à distribuição dos dados;
- Não se tendo convicção sobre uma particular escolha, pode-se testar diferentes alternativas ou usar alguma abordagem que não exija essa especificação.

Especificação do preditor linear

- Quais variáveis explicativas devem ser consideradas?
- Como essas variáveis serão incorporadas ao modelo? Avaliar a necessidade (conveniência) de escalonar, transformar, categorizar ou incluir potências de variáveis numéricas. . .
- Avaliar a necessidade de incluir efeitos de interação entre variáveis.

Especificação da função de ligação

- A função de ligação tem o papel de linearizar a relação entre os componentes aleatório e sistemático do modelo
- Deve produzir valores no espaço paramétrico (para μ_i) para qualquer valor produzido por $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$.
- Apresentar propriedades estatísticas e computacionais desejadas (trataremos disso adiante);
- Proporcionar interpretações práticas para os parâmetros de regressão $\beta's$.

Função de ligação canônica

- A função de ligação $g(\cdot)$ definida pelo parâmetro canônico é chamada *função de ligação canônica*:

$$g(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}. \tag{16}$$

- Como exemplos de função de ligação canônicas temos a ligação logarítmica para a distribuição de Poisson; a logito para a distribuição binomial, a identidade para a normal...

- A ligação canônica garante algumas simplificações no algoritmo de estimação e propriedades desejadas no processo de ajuste do modelo.

Especificação da função de ligação

Tabela 1: Exemplos de funções de ligação

Ligação	$\eta = g(\mu)$	$\mu = g^{-1}(\eta)$	Ligação canônica
Identidade	μ	η	Normal
Logarítmica	$ln(\mu)$	e^{η}	Poisson
Inversa	μ^{-1}	η^{-1}	Gama
Inversa-quadrada	μ^{-2}	$\eta^{-1/2}$	Normal inversa
Raiz quadrada	$\sqrt{\mu}$	η^2	
Logito	$\ln \frac{\mu}{1-\mu}$	$\frac{e^{\eta}}{1+e^{\eta}}$	Binomial
Probito	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	
Log-log	$-ln[ln(\mu)]$	$exp[-exp(-\eta)]$	
Clog-log	$ln[-ln(1-\mu)]$	$1 - exp[-exp(\eta)]$	

Parte 4 - Estimação

Estimação em modelos lineares generalizados

- Seja y_i uma única observação de uma distribuição na forma:

$$f(y_i;\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)\right\}. \quad (17)$$

- A log-verossimilhança correspondente a essa observação fica dada por:

$$l_i = l(\theta_i;\phi,y_i) = log\left[f(y_i;\theta_i,\phi)\right] = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi). \quad (18)$$

- Considerando n observações independentes $y_1, y_2, ..., y_n$, a log-verossimilhança fica dada por:

$$l(\boldsymbol{\theta};\phi,\boldsymbol{y}) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} l(\theta_i;\phi,y_i) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)\right]. \quad (19)$$

Estimação em modelos lineares generalizados

- Considere um modelo linear generalizado com função de ligação $g(\cdot)$ e preditor linear $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{20}$$

- A estimação de $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ por máxima verossimilhança baseia-se na determinação de $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)'$ tal que:

$$\begin{aligned}
&\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = 0 \\
&\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = 0 \\
&\vdots \\
&\frac{\partial l(\boldsymbol{\beta})}{\partial \beta}\Big|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p} = 0
\end{aligned} \tag{21}$$

Estimação em modelos lineares generalizados

- Como $y_1, y_2, ..., y_n$ são independentes:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \tag{22}$$

para todo j.

- Observe que estamos denotando a log-verossimilhança por $l(\boldsymbol{\beta})$ uma vez que $\mu_i = b'(\theta_i)$; $g(\mu_i) = \eta_i$ e $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$.
- Assim, pela regra da cadeia:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}, \tag{23}$$

para $j = 0, 1, 2, ..., p$.

Estimação em modelos lineares generalizados

- Usando a definição e as propriedades de modelos lineares generalizados:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{var(y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}. \tag{24}$$

- Somando para as n observações, as equações de log-verossimilhança ficam dadas por:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{var(y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}, \quad j = 0, 1, 2, ..., p, \tag{25}$$

onde $\eta_i = \sum_{j=0}^{p} \beta_j x_{ij} = g(\mu_i)$.

Estimação em modelos lineares generalizados

- As equações de log-verossimilhança podem ser escritas na forma matricial da seguinte maneira:

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})=\boldsymbol{0}, \tag{26}$$

em que $\boldsymbol{V}$ é a matriz diagonal das variâncias das observações; $\boldsymbol{X}$ é a matriz do modelo; $\boldsymbol{D}$ é a matriz diagonal com entradas $\partial\mu_i/\partial\eta_i$ e $\boldsymbol{y}$ e $\boldsymbol{\mu}$ são os vetores de observações e de médias, respectivamente.

- As equações de log-verossimilhança são funções não lineares dos $\beta's$.
- Assim, a determinação das estimativas de máxima verossimilhança requer o uso de métodos iterativos. Vamos discutir mais adiante dois desses métodos: o **método de Newton-Raphson** e o **método Score de Fisher**.

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Os estimadores de máxima verossimilhança dos parâmetros de um MLG atendem às propriedades gerais de estimadores de máxima verossimilhança.

- Assim, assintoticamente:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\jmath}^{-1}), \tag{27}$$

em que $\boldsymbol{\jmath}$ é a matriz informação de Fisher (ou matriz informação esperada), com entradas $-E(\partial^2 l(\boldsymbol{\beta})/\partial\beta_r\partial\beta_s)$.

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Usando o fato que, sob condições de regularidade atendidas pela família exponencial de distribuições:

$$E\left(-\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right) = E\left[\left(\frac{\partial l(\beta)}{\partial\beta_r}\right)\left(\frac{\partial l(\beta)}{\partial\beta_s}\right)\right] \tag{28}$$

chega-se a:

$$E\left(-\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right) = \sum_{i=1}^{n}\frac{x_{ir}x_{is}}{var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{29}$$

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Seja $\boldsymbol{W}$ a matriz diagonal com elementos:

$$\omega_i = \frac{(\partial \mu_i / \partial \eta_i)^2}{var(y_i)}. \tag{30}$$

- Então, generalizando para toda a matriz de informação, temos:

$$\jmath = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}, \tag{31}$$

em que $\boldsymbol{X}$ é a matriz do modelo. A matriz $\jmath$ depende da função de ligação, uma vez que $\partial \eta_i / \partial \mu_i = g'(\mu_i)$.

Distribuição assintótica dos estimadores dos parâmetros de um MLG

- Assim, a distribuição assintótica de $\hat{\beta}$ é dada por:

$$\hat{\beta} \sim N(\beta, (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}). \tag{32}$$

- A matriz de covariância assintótica é estimada por

$$\widehat{var}(\hat{\beta}) = (\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{33}$$

sendo que $\hat{\boldsymbol{W}}$ é $\boldsymbol{W}$ avaliado em $\hat{\beta}$.

Método de Newton-Raphson

- O método de Newton-Raphson é aplicado na solução de equações não lineares (no caso, na determinação do ponto em que a função assume seu máximo);
- O método inicia com um valor incial como primeira aproximação para a solução;
- Na sequência, uma segunda aproximação é obtida na vizinhança do valor inicial através de um polinômio de segundo grau, e encontrando o ponto de máximo do polinômio.
- Após a repetição de uma sequência de aproximações, o processo converge para a locação do máximo se a função é bem comportada e o valor inicial bem escolhido..

Método de Newton-Raphson

- Formalizando: desejamos determinar $\hat{\beta}$ que maximiza $L(\beta)$. Seja:

$$\boldsymbol{S} = \left(\frac{\partial l(\beta)}{\partial \beta_0}, \frac{\partial l(\beta)}{\partial \beta_1}, ..., \frac{\partial l(\beta)}{\partial \beta_p} \right)' . \tag{34}$$

- Seja $\boldsymbol{H}$ a matriz hessiana, definida pelas deriavdas parciais de segunda ordem:

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_p} \end{bmatrix}$$

Método de Newton-Raphson

- Sejam $\boldsymbol{S}^{(t)}$ e $\boldsymbol{H}^{(t)}$, respectivamente, $\boldsymbol{S}$ e $\boldsymbol{H}$ avaliados em $\beta^{(t)}$, a aproximação no passo t para $\hat{\beta}$.

- O método se Newton-Raphson aproxima $l(\beta)$ em torno de $\beta^{(t)}$ por meio de uma expansão em série de Taylor de segunda ordem:

$$l(\beta) \approx l(\beta^{(t)}) + \boldsymbol{S}^{(\boldsymbol{t})'}(\beta - \beta^{(\boldsymbol{t})}) + \left(\frac{1}{2}\right)(\beta - \beta^{(\boldsymbol{t})})'\boldsymbol{H}^{(t)}(\beta - \beta^{(\boldsymbol{t})}) \quad (35)$$

Método de Newton-Raphson

- Resolvendo $\partial L(\beta)/\partial\beta \approx \boldsymbol{S}^{(t)} + \boldsymbol{H}^{(t)}(\beta - \boldsymbol{\beta^{(t)}}) = \mathbf{0}$ para β, temos como aproximação para $\hat{\beta}$ no passo $t+1$:

$$\beta^{(t+1)} = \beta^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{S}^{(t)}, \tag{36}$$

assumindo que $\boldsymbol{H}^{(t)}$ é não singular.

- As iterações continuam até que a mudança em $\beta^{(t)}$ em passos sucessivos seja suficientemente pequena (convergência).

Método Score de Fisher

- A diferença do método Score de Fisher para o método de Newton Raphson é que o primeiro usa o valor esperado da matriz Hessiana, que é a matriz de *informação esperada*, enquanto o segundo usa a própria hessiana, que é a matriz de *informação observada*.

- Seja $\mathcal{J}^{(t)}$ a aproximação no passo t para a matriz de informação esperada, com entradas $-E(\partial^2 l(\beta)/\partial\beta_r\partial\beta_s)$ avaliado em $\beta^{(t)}$. O algoritmo score de Fisher fica dado por:

$$\beta^{(t+1)} = \beta^{(t)} - (\mathcal{J}^{(t)})^{-1}\boldsymbol{S}^{(t)}. \tag{37}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- Existe uma relação entre o algoritmo Score de Fisher, aplicado na estimação de máxima verossimilhança dos parâmetros de um MLG, e o método de mínimos quadrados ponderados.
- Após algumas passagens, as equações do algoritmo Score de Fisher podem ser expressas na seguinte forma:

$$(\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})\beta^{(t+1)} = \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}, \tag{38}$$

em que $\boldsymbol{z}^{(t)}$ é um vetor com elementos:

$$z_i^{(t)} = \sum_j x_{ij}\beta_j^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}. \tag{39}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- Isolando $\beta^{(t+1)}$:

$$\beta^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)} \tag{40}$$

- O vetor $\boldsymbol{z}^{(t)}$ é uma forma linearizada da função de ligação g avaliada em $\boldsymbol{y}$:

$$g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)})g'(\mu_i^{(t)}) = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}. \tag{41}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

- O algoritmo de mínimos quadrados ponderados iterativamente para o ajuste por máxima verossimilhança de modelos lineares generalizados fica definido da seguinte forma:

1. Defina valores iniciais para μ_i e calcule $\eta_i = g_{(\mu_i)}$, $i = 1, 2, ..., n$, denotados por $\mu_i^{(0)}$ e $\eta_i^{(0)}$. Uma escolha simples é $\mu_i^{(0)} = y_i$ e $\eta_i^{(0)} = g(y_i)$;
2. Calcule os elementos do vetor $\boldsymbol{z}$ e a matriz $\boldsymbol{W}$ com base nos valores de μ_i atribuídos no passo anterior:

$$z_i^{(0)} = \eta_i^{(0)} + (y_i - \mu_i^{(0)}) \times g'(\mu_i^{(0)}); \tag{42}$$

$$\omega_{ii} = \frac{1}{V(\mu_i^{(0)}) \times [g'(\mu_i^{(0)})]^2}; \tag{43}$$

Máxima verossimilhança e método de mínimos quadrados ponderados iterativamente

3. Calcular a aproximação de β no próximo passo por:

$$\beta^{(1)} = (\boldsymbol{X}'\boldsymbol{W}^{(0)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(0)}\boldsymbol{z}^{(0)} \tag{44}$$

4. Repetir os passos 2 e 3, com a aproximação atual de β, até verificar convergência.

- Um critério de convergência que pode ser considerado é o seguinte:

$$\sum_{j=0}^{p}\left(\frac{\beta_j^{(t)} - \beta_j^{(t-1)}}{\beta_j^{(t-1)}}\right)^2. \tag{45}$$

Estimação em modelos lineares generalizados

- Uma vez obtidos os estimadores dos parâmetros de um modelo linear generalizado, o ajuste pode ser apresentado na escala do preditor:

$$g(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p, \tag{46}$$

ou na escala da média (denominaremos de escala da resposta ao longo do curso):

$$\hat{\mu} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p). \tag{47}$$

Estimação do parâmetro de dispersão

- Nas situações em que ϕ é desconhecido, precisamos estimá-lo para avaliação dos erros das estimativas, construção de intervalos de confiança...
- Um estimador consistente para ϕ baseia-se na estatística X^2 de Pearson:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{48}$$

onde $V(\mu)$ é a função de variância, sendo definido por:

$$\hat{\phi} = \frac{X^2}{n-p}. \tag{49}$$

Robustez dos MLG's quanto à especificação incorreta do modelo

- Os estimadores dos parâmetros de modelos lineares generalizados são consistentes ainda que a distribuição especificada esteja incorreta, mas desde que a especificação do preditor linear e da função de ligação esteja correta;
- Entretanto, ao assumir uma distribuição incorreta, a função de variância também estará errada, de forma que $Var(\hat{\beta})$ (e os resultados subsequentes) estarão incorretos;
- Estudaremos adiante como contornar os problemas decorrentes da especificação incorreta da função de variância sem precisar, para isso, assumir um particular modelo probabilístico (abordagem de quasi-verossimilhança).

Parte 5- Inferência

Testes de hipóteses

- Vamos discutir neste momento testes para hipóteses do tipo:

$$H_0 : \beta = \beta_0 \quad H_1 : \beta \neq \beta_0 \tag{50}$$

- Nas hipóteses apresentadas, β representa um ou mais parâmetros do modelo ajustado, e β_0 valores postulados (fixados) para esses parâmetros na hipótese nula.

Testes de hipóteses

- Apenas para ilustração, considere o seguinte modelo:

$$ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{51}$$

- Exemplos de hipóteses:

$$H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad vs \quad H_1 : \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix};$$

Testes de hipóteses

$$H_0 : \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad vs \quad H_1 : \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

$$H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0;$$

$$H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix} \quad vs \quad H_1 : \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \neq \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix} ...$$

- Os principais testes de hipóteses em modelos lineares generalizados são:
  - Teste da razão de verossimilhanças;
  - Teste de Wald;
  - Teste escore.

Ilustração - Testes

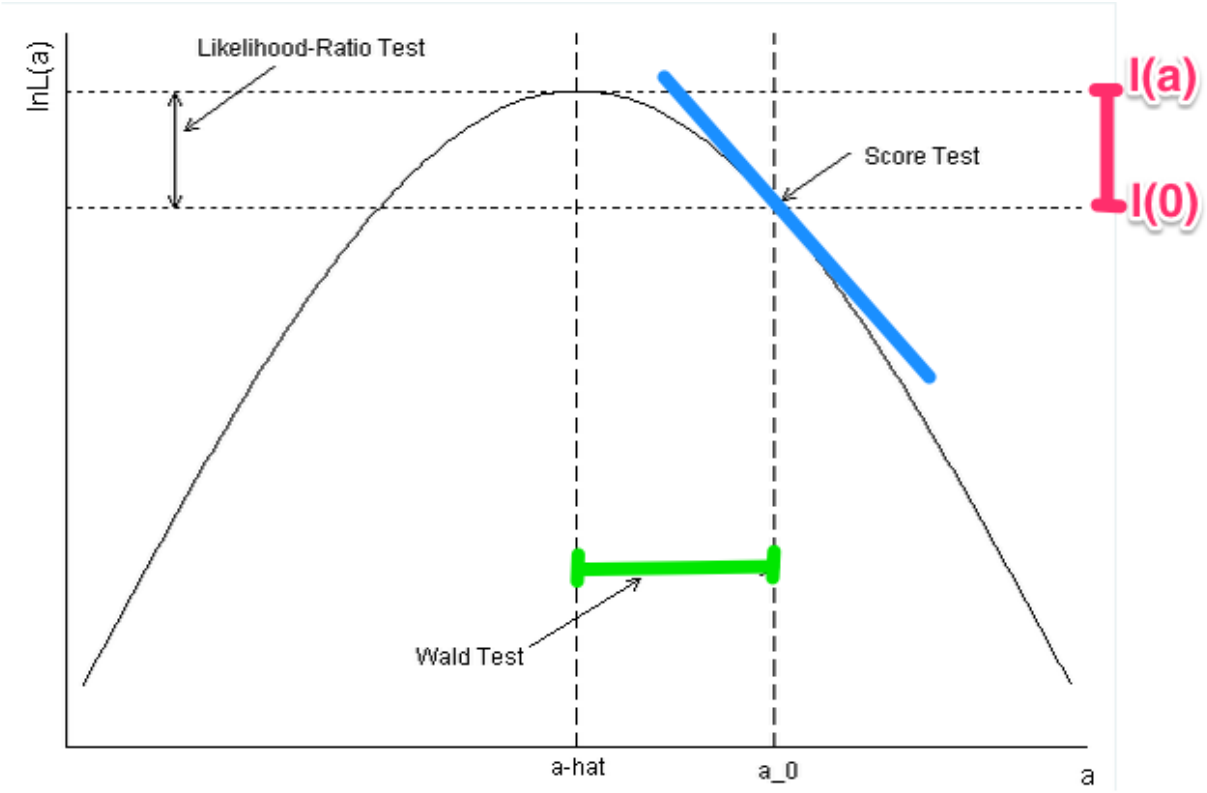


Figura 12: Log-verossimilhança e informação usada nos três testes para a hipótese $H_0 : \beta = a_0$.

Testes da razão de verossimilhanças

- Seja L_0 a verossimilhança maximizada sob a hipótese nula, e L_1 a verossimilhança maximizada de forma não restrita (permitindo que H_0 ou H_1 seja verdade).

- A razão $\Lambda = L_0/L_1 \leq 1$, uma vez que L_0 resulta da maximização sobre um conjunto restrito de valores para β.

Testes da razão de verossimilhanças

- A estatística do teste da razão de verossimilhança é definida por:

$$-2ln\Lambda = -2ln(L_0/L_1) = -2(l_0 - l_1), \tag{52}$$

sendo l_0 e l_1 as log-verossimilhanças maximizadas restrita e não restrita.

- Sob H_0 e ϕ conhecido, a estatística do teste tem distribuição assintótica ($n \to \infty$) χ^2 com q graus de liberdade, sendo q o número de parâmetros fixados em H_0.

Nota: se o parâmetro de dispersão é desconhecido (sendo estimado), o teste da razão de verossimilhança tem melhor aproximação pela distribuição F.

Testes de Wald

- Para o teste de um único parâmetro, com hipótese nula $H_0 : \beta_k = \beta_0$, a estatística do teste de Wald fica dada por:

$$z = \frac{\hat{\beta}_k - \beta_0}{\sqrt{\widehat{Var}(\hat{\beta}_k)}} \tag{53}$$

tendo, sob H_0, distribuição assintótica $N(0, 1)$ quando ϕ é conhecido.

Intervalos de confiança

- Intervalos de confiança para qualquer dos três métodos podem ser obtidos invertendo as respectivas estatísticas de teste.

- Por exemplo, um intervalo de confiança 95% para um único parâmetro β_k, é definido pelo conjunto de valores β_0 tais que $H_0 : \beta_k = \beta_0$ não é rejeitada ao nível de significância de 5%.

- Um intervalo de confiança assintótico $100(1-\alpha)\%$ para β_k, **baseado no teste de Wald**, tem limites:

$$IC(\beta_k; 1-\alpha) = \hat{\beta}_k \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_k)}, \tag{54}$$

em que $z_{\alpha/2}$ é o quantil $\alpha/2$ da distribuição normal padrão.

Intervalos de confiança

- Pode-se obter um intervalo de confiança para β_k baseado na **verossimilhança perfilada**.
- Seja $H_0 : \beta_k = \beta_0$ e Ψ representando o conjunto dos demais parâmetros do modelo.
- Ao inverter o teste da razão de verossimilhanças, para determinar o conjunto de valores β_0 que compõem o intervalo de confiança, a estimativa de máxima verossimilhança de Ψ varia para os diferentes valores de β_0.
- O intervalo de confiança baseado na verossimilhança perfilada de β_k é definido pelo conjunto de valores β_0 tais que:

$$-2[L(\beta_0, \hat{\Psi}(\beta_0)) - L(\hat{\beta}_k, \hat{\Psi})] < \chi^2_1(\alpha), \tag{55}$$

sendo $L(\beta_0, \hat{\Psi}(\beta_0))$ a verossimilhança maximizada para $\beta_k = \beta_0$ e $L(\hat{\beta}_k, \hat{\Psi})$ a verossimilhança maximizada de forma irrestrita.

Intervalos de confiança

- Além de intervalos de confiança para os parâmetros, é interessante também obter intervalos de confiança para $\mu_{\boldsymbol{x}} = E[y|\boldsymbol{x}]$, sendo $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$ um específico vetor de covariáveis.

- A estimativa pontual para $\mu_{\boldsymbol{x}}$ é dada por:

$$\hat{\mu}_{\boldsymbol{x}} = g^{-1}(\boldsymbol{x}'\hat{\beta}). \tag{56}$$

- Seja $\hat{\eta}_{\boldsymbol{x}} = \boldsymbol{x}'\hat{\beta}$. Como $\hat{\eta}_{\boldsymbol{x}}$ é uma combinação linear dos $\hat{\beta}'s$, decorre que, assintoticamente:

$$\hat{\eta}_{\boldsymbol{x}} \sim Normal(\boldsymbol{x}'\beta, \boldsymbol{x}'Var(\hat{\beta})\boldsymbol{x}) \tag{57}$$

Intervalos de confiança

- Um intervalo de confiança assintótico $100(1-\alpha)\%$ para $\eta_{\boldsymbol{x}} = \boldsymbol{x}'\boldsymbol{\beta}$ fica dado por:

$$IC(\eta_{\boldsymbol{x}}; 1-\alpha) = \boldsymbol{x}'\hat{\boldsymbol{\beta}} \pm z_{\alpha/2}\sqrt{\boldsymbol{x}'\widehat{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}}. \quad (58)$$

- Dessa forma, um intervalo de confiança assintótico $100(1-\alpha)\%$ para $\mu_{\boldsymbol{x}} = g^{-1}(\eta_{\boldsymbol{x}})$ fica dado por:

$$IC(\mu_{\boldsymbol{x}}; 1-\alpha) = (g^{-1}(LI); g^{-1}(LS)), \quad (59)$$

se $g(\cdot)$ é uma função estritamente crescente, e:

$$IC(\mu_{\boldsymbol{x}}; 1-\alpha) = (g^{-1}(LS); g^{-1}(LI)), \quad (60)$$

se $g(\cdot)$ é uma função estritamente decrescente, onde LI e LS denotam os

Parte 6- Diagnóstico do ajuste

Diagnóstico do ajuste

- A análise de diagnóstico (ou diagnóstico do ajuste) configura uma etapa fundamental no ajuste de modelos de regressão.

- O objetivo principal dessa etapa da análise é a avaliação do modelo ajustado. No caso de MLGs, baseia-se, dentre outros, na verificação dos seguintes itens:
  - Avaliação da distribuição proposta;
  - Avaliação da parte sistemática do modelo;
  - Adequação da função de ligação.
  - Identificação e avaliação do efeito de observações mal ajustadas;
  - Identificação de pontos influentes.

Diagnóstico do ajuste

- Boa parte dos métodos de diagnóstico em MLGs configuram extensões dos procedimentos utilizados em regressão linear.
- O uso de simulação no diagnóstico de MLGs (por exemplo na obtenção de qq-plots com envelopes simulados) é importante.
- O principal componente no diagnóstico de MLGs é, novamente, a análise de resíduos.

Resíduo de Pearson

- O resíduo de Pearson é definido por:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}. \tag{61}$$

- Para um MLG Poisson, o resíduo de Pearson fica definido por:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{62}$$

- Já para um MLG binomial:

$$e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}. \tag{63}$$

Resíduo de Pearson

- O resíduo de Pearson tem uma versão padronizada, com média 0 e variância aproximadamente 1, definida por:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V(\hat{\mu}_i)(1 - \hat{h}_{ii})}}, \tag{64}$$

em que $\hat{h}_{ii}$ é o i-ésimo elemento da diagonal da matriz

$$\hat{H} = W^{1/2} X (X' W X)^{-1} X' W^{1/2}, \tag{65}$$

que é a matriz de projeção do algoritmo de estimação dos MLGs.

Resíduo componente da deviance

- A **deviance escalonada** de um modelo proposto é definida como a estatística da razão de verossimilhança relativa ao modelo saturado (com n parâmetros):

$$S(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2[l(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - l(\boldsymbol{y}; \boldsymbol{y})]. \tag{66}$$

- Resgatando a log-verossimilhança para a família exponencial:

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi) \right], \tag{67}$$

temos:

$$S(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \frac{\left[y_i \tilde{\theta}_i - b(\tilde{\theta}_i) \right]}{a(\phi)} - 2 \sum_{i=1}^{n} \frac{\left[y_i \hat{\theta}_i - b(\hat{\theta}_i) \right]}{a(\phi)}, \tag{68}$$

em que $\tilde{\theta}_i$ e $\hat{\theta}_i$ são as estimativas de θ_i sob os modelos saturado e proposto, respectivamente.

Resíduo componente da deviance

- Caso mais geral, quando $a(\phi) = \phi/\omega_i$, temos:

$$S(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \frac{\omega_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right]}{\phi} = \frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{\phi}, \qquad (69)$$

em que $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ é a deviance do modelo proposto.

- Uma vez que $l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \leq l(\boldsymbol{y}, \boldsymbol{y})$, $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \geq 0$, de forma que, quanto pior o ajuste do modelo proposto, maior a deviance.

Resíduo componente da deviance

Tabela 2: Deviances para alguns modelos mais usuais

Distribuição	Deviance
Normal	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (\hat{\mu}_i - y_i)\right]$
Binomial	$2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{m_i\hat{\mu}_i}\right) + (m_i - y_i)\ln\left\{\frac{\left(1-\frac{y_i}{m_i}\right)}{(1-\hat{\mu}_i)}\right\}\right]$
Gama	$2\sum_{i=1}^{n}\left[\ln\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right]$
Normal inversa	$\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{y_i\hat{\mu}_i^2}$

Resíduo componente da deviance

- O resíduo componente da deviance fica definido pela contribuição de cada observação para a deviance do modelo:

$$d_i = sinal(y_i - \hat{\mu}_i) \times \sqrt{2\omega_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right]} \tag{70}$$

- Uma versão padronizada do resíduo componente da deviance é dada por:

$$d_i^* = \frac{d_i}{\sqrt{\hat{\phi}(1 - \hat{h}_{ii})}}. \tag{71}$$

Análise de resíduos

- Os resíduos de Pearson e componente da deviance geralmente não tem boa aproximação com a distribuição Normal, ainda que o modelo ajustado esteja correto.

- A avaliação da qualidade do ajuste baseada em gráficos probabilísticos normais (ou meio-normais), para esses tipos de resíduos, requer simulação (envelopes simulados). Veremos adiante.

- Um tipo de resíduo que, por construção, tem distribuição normal caso o modelo ajustado esteja correto, é o **resíduo quantílico aleatorizado**.

Resíduo quantílico aleatorizado (Dunn, 1997)

- O resíduo quantílico aleatorizado baseia-se no método método da transformação integral da probabilidade.
- Seja y_i uma variável aleatória contínua com FDA $F(y_i; \mu_i, \phi)$. O método da transformação integral da probabilidade baseia-se no seguinte resultado:

$$u_i = F(y_i; \mu_i, \phi) \sim U(0, 1). \tag{72}$$

- Adicionalmente, considerando que u_i tem distribuição uniforme entre 0 e 1, temos que:

$$\Phi^{-1}(F(y_i; \mu_i, \phi)) \sim N(0, 1), \tag{73}$$

resultado bastante utilizado para simular dados de uma distribuição Normal padrão.

Resíduo quantílico aleatorizado

- Assim, o resíduo quantílico aleatorizado fica definido por:

$$q_i = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi})), \tag{74}$$

tal que, se o modelo tiver corretamente especificado, tem distribuição Normal(0,1).

Resíduo quantílico aleatorizado

- Se a variável y_i for discreta, então $F(y_i; \mu_i, \phi)$ é uma função discreta, com 'saltos' em cada valor de y_i com probabilidade não nula.

- Neste caso, consideramos a seguinte adaptação:

$$F^*(y_i; \hat{\mu}_i, \hat{\phi}) = F(y_i-; \hat{\mu}_i, \hat{\phi}) + u_i f(y_i; \hat{\mu}_i, \hat{\phi}), \tag{75}$$

em que $F(y_i-; \hat{\mu}_i, \hat{\phi})$ é o limite de $F(y-; \hat{\mu}_i, \hat{\phi})$ pela esquerda, u_i é um valor aleatório da distribuição $U(0, 1)$ e $f(y_i; \hat{\mu}_i, \hat{\phi})$ é a massa de probabilidade em y_i.

Análise gráfica dos resíduos

- **Resíduos vs valores ajustados**- Para um modelo bem ajustado, deve-se observar a dispersão aleatória dos pontos, centrada em zero, com média e variância constantes e sem valores extremos.

Nota: É recomendável plotar os resíduos padronizados, e os valores ajustados na escala do preditor.

- **Resíduos vs variáveis incluídas no modelo:** Padrões não aleatórios indicam que a variável não está bem acomodada no modelo;
- **Resíduos vs variáveis não incluídas no modelo:** Padrões não aleatórios sinalizam a necessidade (e a forma) de inclusão da variável no modelo;
- **Gráfico de resíduos versus ordem de coleta dos dados** - Padrões não aleatórios indicam a dependência das observações gerada pela ordem de coleta (no tempo, no espaço,...).

Análise gráfica dos resíduos

- **Gráfico da variável ajustada versus preditor linear** - Plotando z_i vs $\hat{\eta}_i$ podemos avaliar a adequação da função de ligação.

- Uma forma alternativa de testar a função de ligação é a seguinte:

1. Ajusta-se o modelo extrai-se $\hat{\eta}_i$;
2. Ajusta-se novamente o modelo incorporando $\hat{\eta}_i^2$ como uma nova covariável;
3. Se o efeito de $\hat{\eta}_i^2$ for significativo, então a função de ligação não é adequada.

Gráficos meio normais com envelopes simulados

- Os gráficos meio-normais consistem na plotagem de alguma medida de diagnóstico (resíduos, distância de Cook, leverage) versus a esperança das estatísticas de ordem da distribuição meio-normal:

$$\Phi^{-1}\left(\frac{i+1-\frac{1}{8}}{2n+\frac{1}{2}}\right), i = 1, 2, ..., n. \tag{76}$$

- Em modelos lineares generalizados, a distribuição dos resíduos (Pearson, deviance) e das medidas de influência, dentre outros, em geral não é normal, o que pode prejudicar a avaliação dos gráficos meio-normais.
- A solução é usar simulação para poder avaliar adequadamente a disposição dos pontos em um gráfico meio-normal (envelopes simulados).

Obtenção dos envelopes simulados para gráficos meio-normais (Moral, 2013)

1. Obter $d_{(i)}$, os valores de uma quantidade diagnóstica em valor absoluto e em ordem crescente;
2. Simular 99 amostras do modelo ajustado com os mesmos valores para as variáveis explanatórias;
3. Fazer o ajuste do modelo para as 99 amostras e, para cada ajuste, obter a quantidade diagnóstica de interesse, $d^*_{j(i)}$, $j = 1, 2, ..., 99$, em valor absoluto e em ordem crescente;
4. Para cada i computar os percentis 5%, 50% e 95%;
5. Fazer o gráfico desses percentis e dos $d_{(i)}$'s observados contra as estatísticas de ordem da distribuição meio-normal.

- Se a maior parte dos valores observados estiver contida no envelope simulado, há indícios de que o modelo está bem ajustado aos dados.

Diagnóstico de influência

- Assim como no caso de modelos lineares, também para MLGs o diagnóstico de influência é útil para identificar pontos que exercem grande influência sobre o ajuste do modelo.

- A estratégia para diagnóstico de influência, novamente, é do tipo *leave-one-out*, em que se avalia o quanto resultados dos modelos (estimativas dos coeficientes, erros padrões,...) mudam ao desconsiderar uma particular observação i, $i = 1, 2, ..., n$;

- Assim como no caso dos modelos lineares, não há a necessidade de ajustar o mesmo modelo n vezes (uma para a deleção de cada observação), dispondo-se de aproximações adequadas para as medidas de influência.

Parte 7- Regressão para dados binários

Introdução

- Interesse em modelar fenômenos aleatórios com dois desfechos possíveis (*sucesso* ou *fracasso*) como função de covariáveis;
- A distribuição binomial (e, como caso particular, a distribuição Bernoulli) surge como principal alternativa para a modelagem de dados binários;
- Grande quantidade e variedade de potenciais aplicações.

Exemplos de motivação

- Prognóstico clínico de pacientes (ex: cura ou não cura) em função de variáveis clínicas, genéticas, comportamentais...
- Risco de crédito (pagamento ou não) por clientes de um banco em função de variáveis sócio-econômicas, referentes à modalidade de empréstimo, ao relacionamento do cliente com o banco...
- Resultado de partidas de basquete (vitória do mandante ou do visitante) em função do desempenho das equipes no campeonato, histórico de confrontos, circunstâncias da partida,...
- Presença ou não de certa espécie vegetal em regiões de uma floresta em função de variáveis ambientais e climáticas.

Distribuição de Bernoulli

- A distribuição de Bernoulli associa valores 0 e 1 a cada um dos dois desfechos.
- Uma variável aleatória y tem distribuição de Bernoulli com parâmetro π se sua função de probabilidades é dada por:

$$f(y;\pi) = \pi^y(1-\pi)^{1-y}, \quad y = 0,1; \quad 0 < \pi < 1, \tag{77}$$

ou seja, $f(0;\pi) = P(y = 0|\pi) = 1 - \pi$ e $f(1;\pi) = P(y = 1|\pi) = \pi$.

- Propriedades da distribuição Bernoulli:

$$E(y) = \mu = \pi; \quad Var(Y) = \mu(1-\mu). \tag{78}$$

- A distribuição Bernoulli pertence à família exponencial com $V(\mu) = \mu(1-\mu)$ e $\phi = 1$.

Distribuição binomial

- Considere n realizações independentes de um experimento de Bernoulli, todos com mesma probabilidade de sucesso π.

- Seja Y a fração de sucessos observada nas n realizações. Então:

$$f(y; n, \pi) = \binom{n}{ny} \pi^{ny} (1-\pi)^{n-(ny)}; y = 0, \frac{1}{n}, \frac{2}{n}, ..., 1; 0 < \pi < 1. \quad (79)$$

- A Bernoulli é um caso particular da distribuição binomial (quando $n = 1$).

Distribuição binomial

- A média e a variância de y são dadas por:

$$E(y) = \mu = \pi; \quad Var(Y) = \frac{\mu(1-\mu)}{n}. \tag{80}$$

- A distribuição binomial pertence à família exponencial com $V(\mu) = \mu(1-\mu)$ e $\phi = 1$.

Regressão para dados binários

- No contexto de modelos lineares generalizados, considere $y_1, y_2, ..., y_n$ variáveis aleatórias independentes com

$$y_i \sim binomial(m_i, \pi_i), \quad i = 1, 2, ..., n. \tag{81}$$

- Adicionalmente, sejam $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ os vetores de covariáveis associados a cada observação;

- Especificação do modelo linear generalizado:

$$y_i|\boldsymbol{x_i} \sim binomial(m_i, \pi_i);$$
$$g(\pi_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{82}$$

Escolha da função de ligação

- Dentre os requisitos para a escolha de uma função de ligação adequada, destacamos:
  - A função de ligação deve ser contínua, diferenciável e monótona;
  - Capaz de 'mapear' os valores de π no intervalo (0,1);
  - Capaz de linearizar a relação entre os componentes aleatório e sistemático do modelo;
  - Que proporcione interpretações simples.

Escolha da função de ligação

- Boa parte dos requisitos indicados são atendidos se adotarmos, como função de ligação, a inversa de F, a função distribuição acumulada (fda) de alguma variável aleatória contínua:

$$g(\pi_i) = F^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \quad (83)$$

ou, de forma equivalente,

$$\pi_i = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}). \quad (84)$$

- Três funções usuais para MLGs com dados binários são as ligações **logito**, **probito** e **complemento log-log**, discutidas na sequência.

Função de ligação logito

- A função de ligação logito baseia-se na fda da distribuição logística em sua forma padrão ($\mu = 0$ e $\sigma = 1$):

$$F(z) = \frac{e^z}{(1 + e^z)}. \tag{85}$$

- Assim como a distribuição normal padrão, a distribuição logística padrão é definida em todo o conjunto dos reais, com forma de sino centrada em zero.
- O MLG baseado na função de ligação logito fica definido por:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}}}, \tag{86}$$

ou, na escala do preditor:

$$\ln\left(\frac{\pi_i}{1-\pi}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}. \tag{87}$$

Função de ligação probito

- A função de ligação probito fica definida pela (inversa) da fda da distribuição normal padrão, definindo o seguinte MLG:

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{88}$$

ou, na escala da resposta:

$$\pi_i = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}), \tag{89}$$

em que $\Phi(\cdot)$ denota a fda da distribuição normal padrão.

- Na prática, as funções de ligação probito e logito têm comportamentos bastante semelhantes, sobretudo no intervalo (0.1,0.9).

Função de ligação complemento log-log

- A função de ligação complemento log-log baseia-se na (inversa) da fda da distribuição Gumbel, definindo o seguinte MLG:

$$\ln[-\ln(1-\pi_i)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{90}$$

ou, na escala da resposta,

$$\pi_i = 1 - exp\left[-exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}\right)\right]. \tag{91}$$

- Diferentemente das funções de ligação probito e logito, a função de ligação complemento log-log é assimétrica em relação a $\pi = 0.5$, o que pode ser conveniente em algumas aplicações.

Ligação Aranda-Ordaz

- A família de ligações Aranda-Ordaz é definida por:

$$g(\pi_i) = \ln\left[\frac{(1-\pi_i)^{-\alpha}-1}{\alpha}\right], \tag{92}$$

sendo α um parâmetro que pode ser estimado.

- Como caso particular, para $\alpha = 1$ temos a função de ligação logito;
- Para $\alpha \to 0$ temos a função de ligação complemento log-log.

Regressão logística

- O modelo de regressão logística fica definido pelo uso da ligação logito em um MLG binomial.

- Neste caso, considerando $y_i|\boldsymbol{x}_i \sim binomial(m_i, \pi_i)$, $i = 1, 2, ..., n$, independentes, então o modelo de regressão logística fica dado por:

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \quad (93)$$

Regressão logística

- De forma equivalente, na escala da *odds*:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}. \tag{94}$$

- Na escala da probabilidade (resposta):

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}+1}. \tag{95}$$

Regressão logística - interpretação dos parâmetros

- A interpretação dos parâmetros em um modelo de regressão logística baseia-se em razões de chances (*odds ratios*).
- Comecemos pelo caso da regressão logística simples, com apenas uma covariável (x). Assumindo que x seja uma variável numérica, então:

$$OR\{x+1, x\} = \frac{odds\{x+1\}}{odds\{x\}} = \frac{\pi_{x+1}/(1-\pi_{x+1})}{\pi_x/(1-\pi_x)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}, \tag{96}$$

em que $\pi_x = P(Y = 1|x)$.

- Assim, e^{β_1} corresponde ao acréscimo na chance de resposta ($y = 1$) para um aumento unitário em x.

Regressão logística - interpretação dos parâmetros

- Para um aumento de k unidades em x, verifica-se facilmente que a chance de resposta fica aumentada em $e^{k\beta_1}$.

- Para o caso de uma covariável dicotômica (com categorias A e B), inserida ao modelo por meio de uma variável indicadora de B, temos:

$$OR\{B, A\} = \frac{odds\{B\}}{odds\{A\}} = \frac{\pi_B/(1-\pi_B)}{\pi_A/(1-\pi_A)} = \frac{e^{\beta_0+\beta_1\times 1}}{e^{\beta_0+\beta_1\times 0}} = e^{\beta_1}, \qquad (97)$$

em que π_A é a probabilidade de resposta para um indivíduo da categoria A.

- Assim, e^{β_1} corresponde à razão das chances de resposta da categoria B para a categoria A da covariável.

Regressão logística - interpretação dos parâmetros

- Se houvesse uma terceira categoria (C), então seriam necessárias duas variáveis indicadoras (x_1, indicadora de B; x_2, indicadora de C). Assim, teríamos:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{98}$$

- As razões de chances ficariam dadas por:

$$OR\{B,A\} = \frac{odds\{B\}}{odds\{A\}} = \frac{e^{\beta_0+\beta_1\times 1+\beta_2\times 0}}{e^{\beta_0+\beta_1\times 0+\beta_2\times 0}} = e^{\beta_1}; \tag{99}$$

$$OR\{C,A\} = \frac{odds\{C\}}{odds\{A\}} = \frac{e^{\beta_0+\beta_1\times 0+\beta_2\times 1}}{e^{\beta_0+\beta_1\times 0+\beta_2\times 0}} = e^{\beta_2}; \tag{100}$$

$$OR\{C,B\} = \frac{odds\{C\}}{odds\{B\}} = \frac{e^{\beta_0+\beta_1\times 0+\beta_2\times 1}}{e^{\beta_0+\beta_1\times 1+\beta_2\times 0}} = e^{\beta_2-\beta_1}. \tag{101}$$

Regressão logística - interpretação dos parâmetros

- Caso o preditor linear contenha múltiplas variáveis, as interpretações são idênticas, devendo-se ressaltar, no entanto, que a interpretação da razão de chances calculada para uma particular variável, é válida fixando os valores das demais variáveis.

Parte 8- Modelos preditivos

Introdução

- Modelos de regressão para dados binários são bastante utilizados para predição, ou seja, classificar indivíduos conforme suas probabilidades estimadas.

- Alguns exemplos:
  - Predição (classificação) de clientes em bons ou maus pagadores;
  - Predição de e-mails em spams ou não spams;
  - Predição do resultado de um jogo de basquete (vitória do time mandante ou do time visitante);
  - Prognóstico de um paciente (cura ou não cura)...

Introdução

- É fortemente recomendável avaliar o poder preditivo do modelo ajustado com dados que não foram usados no ajuste.

- Ajustar o modelo e avaliar a qualidade preditiva usando os mesmos dados tende a produzir resultados excessivamente otimistas.

- Algumas possibilidades:
  - Separar aleatoriamente a amostra em duas partes (uma para ajuste, a outra para predição);
  - Usar validação cruzada (caso particular: *leave one out*).

Predição

- Sejam $\hat{\pi}_i$ as estimativas de $P(y_i = 1)$, $i = 1, 2, ..., n$.

- Considere uma regra do tipo:
  - Predizer $\hat{y}_i = 0$ se $\hat{\pi}_i \leq p_0$;
  - Predizer $\hat{y}_i = 1$ se $\hat{\pi}_i > p_0$,

para algum valor (ponto de corte) especificado p_0 e $i = 1, 2, ..., n$.

- É comum (mas não obrigatório) usar $p_0 = 0.5$, classificando pelo resultado com maior probabilidade.

- Diferentes valores de p_0 conduzem a diferentes regras de predição.

Tabelas de classificação

- Dadas as predições e os valores realmente observados de y, podemos construir uma tabela de classificação.

Tabela 3: Tabela de classificação

	$\hat{y}$	
y	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

- Duas medidas úteis para sumarizar o poder preditivo de um modelo são a sensibilidade e a especificidade.

Sumarizando o poder preditivo

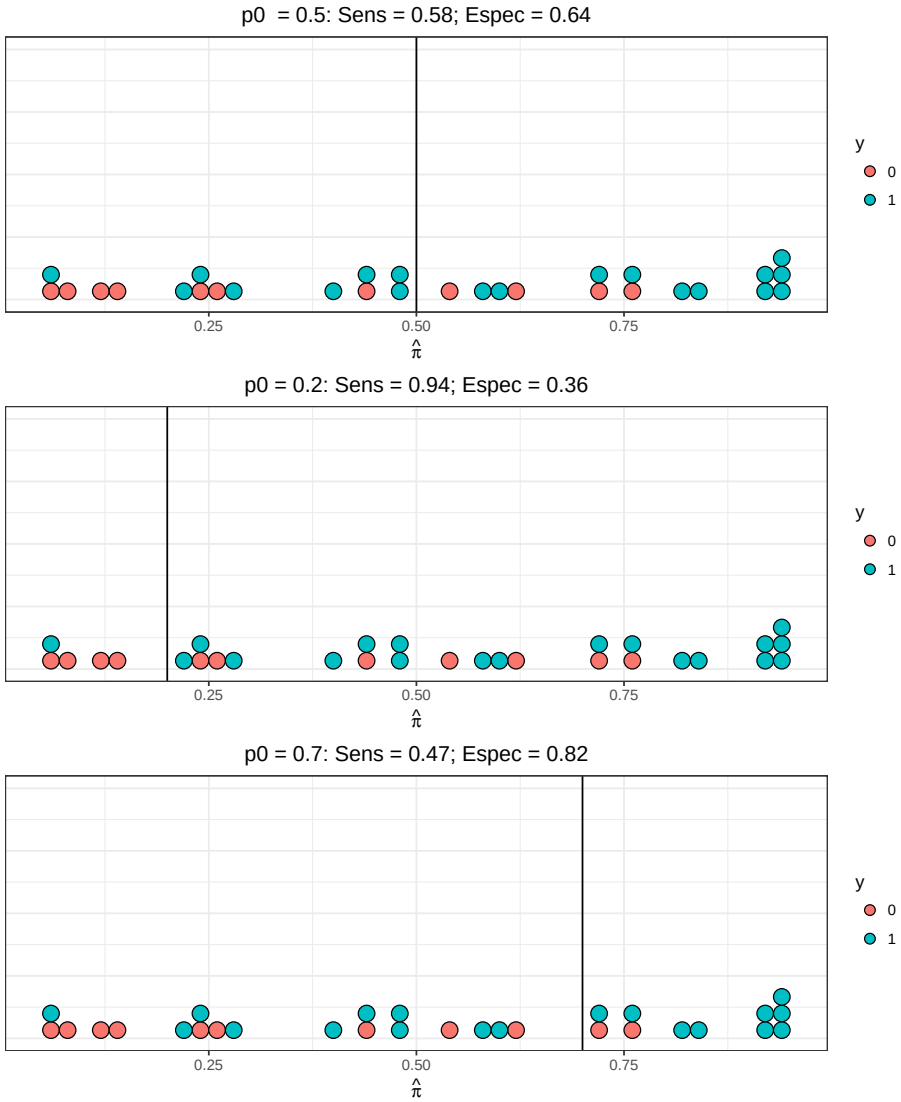
- A **sensibilidade** de um modelo (ou de uma regra de classificação) é definida por $P(\hat{y} = 1|y = 1)$;

- A **especificidade** de um modelo (ou de uma regra de classificação) é definida por $P(\hat{y} = 0|y = 0)$.

- Podemos estimar a sensibilidade e a especificidade com base nas frequências de uma tabela de classificação:

$$\widehat{Sens} = \frac{n_{11}}{n_{10} + n_{11}}; \quad \widehat{Esp} = \frac{n_{00}}{n_{00} + n_{01}}. \tag{102}$$

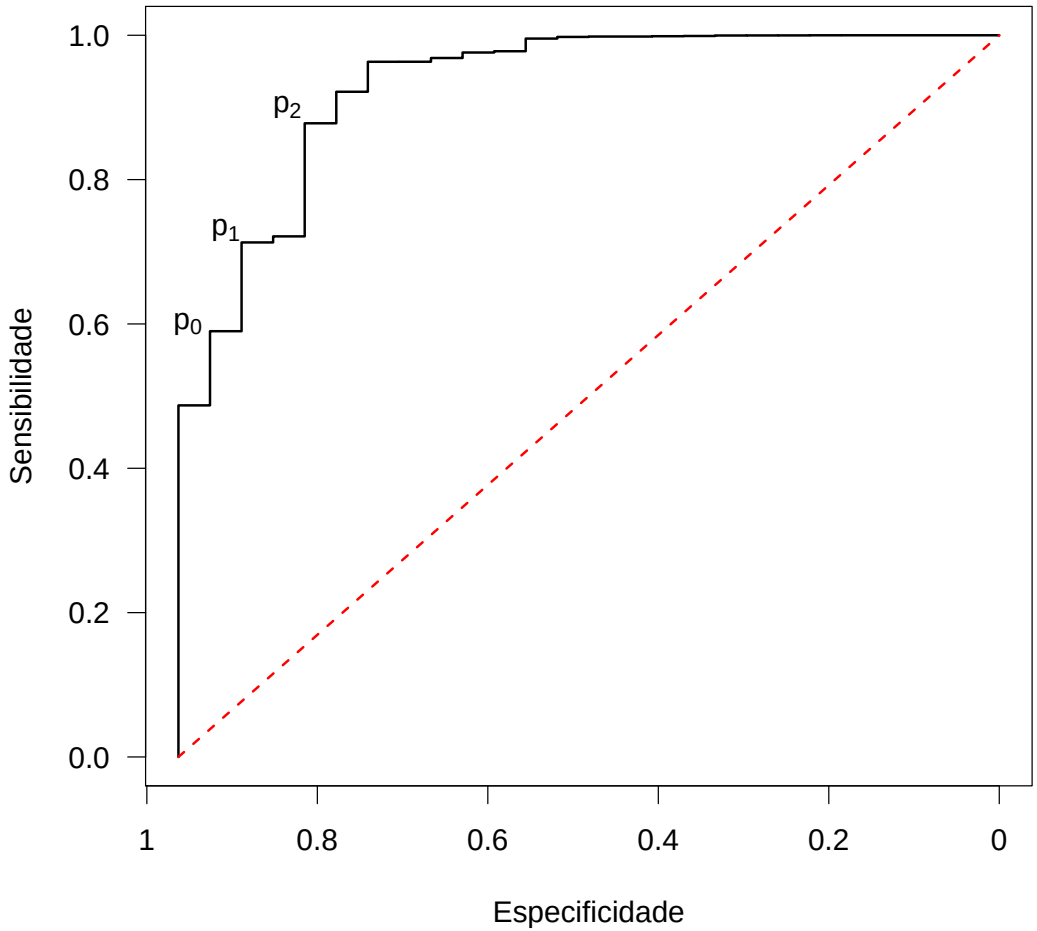
Sumarizando o poder preditivo



Curva ROC

- Uma forma de analisar o poder preditivo associado a diferentes regras de decisão (valores de p_0) é por meio da **curva ROC**.
- A curva ROC permite avaliar conjuntamente a sensibilidade e a especificidade para diferentes valores de p_0.
- Para valores $p_0 \approx 1$, temos sensibilidade próxima de zero e especificidade próxima de um;
- Para valores $p_0 \approx 0$, temos sensibilidade próxima de um e especificidade próxima de zero;
- Em geral, busca-se p_0 tal que se tenha, conjuntamente, elevadas sensibilidade e especificidade;
- A **área sob a curva ROC** é uma medida de poder preditivo do modelo.

Curva ROC


1.0
0.8
0.6
0.4
0.2
0.0
Sensibilidade
p2
p1
p0
1
0.8
0.6
0.4
0.2
0
Especificidade

Parte 9- Regressão para dados de contagens

Introdução

- Interesse em modelar a distribuição de uma variável referente a algum tipo de contagem em função de covariáveis. Normalmente, tem-se uma contagem de eventos em unidades de tempo ou espaço.

- O modelo de regressão mais comum nessas situações assume distribuição de Poisson com função de ligação logarítmica (canônica).

- Como para a distribuição de Poisson temos $\mu > 0$, a ligação logarítmica garantirá valores positivos para μ quaisquer que sejam os valores das covariáveis.

Exemplos de motivação

- Modelagem do número de sinistros de automóveis em função de características do motorista e do veículo;
- Número de casos de dengue em diferentes quadras de um município em função de variáveis geográficas e demográficas;
- Número de gols marcados por times de futebol nas partidas de um campeonato em função de variáveis referentes ao desempenho em campeonatos anteriores, ao investimento em jogadores, aos valores dos patrocínios, . . . ;
- Abundância de certa espécie animal em quadrantes de uma floresta em função de variáveis ambientais e climáticas.

Distribuição de Poisson

O modelo de Poisson configura uma distribuição de probabilidades discreta, obtida:

- Como aproximação à distribuição binomial quando $n \to \infty$ e $p \to 0$ de tal forma que np permaneça constante;
- Quando os eventos sob contagem ocorrem aleatoriamente ao longo do tempo (ou espaço), com probabilidade de ocorrência proporcional ao tamanho do intervalo e independente das contagens em outros intervalos (Processo de Poisson);
- Quando os tempos entre ocorrências de eventos são independentes e identicamente distribuídos segundo uma distribuição exponencial;
- A distribuição Poisson se caracteriza pela equidispersão ($E(y) = Var(y)$).

Processos de contagens

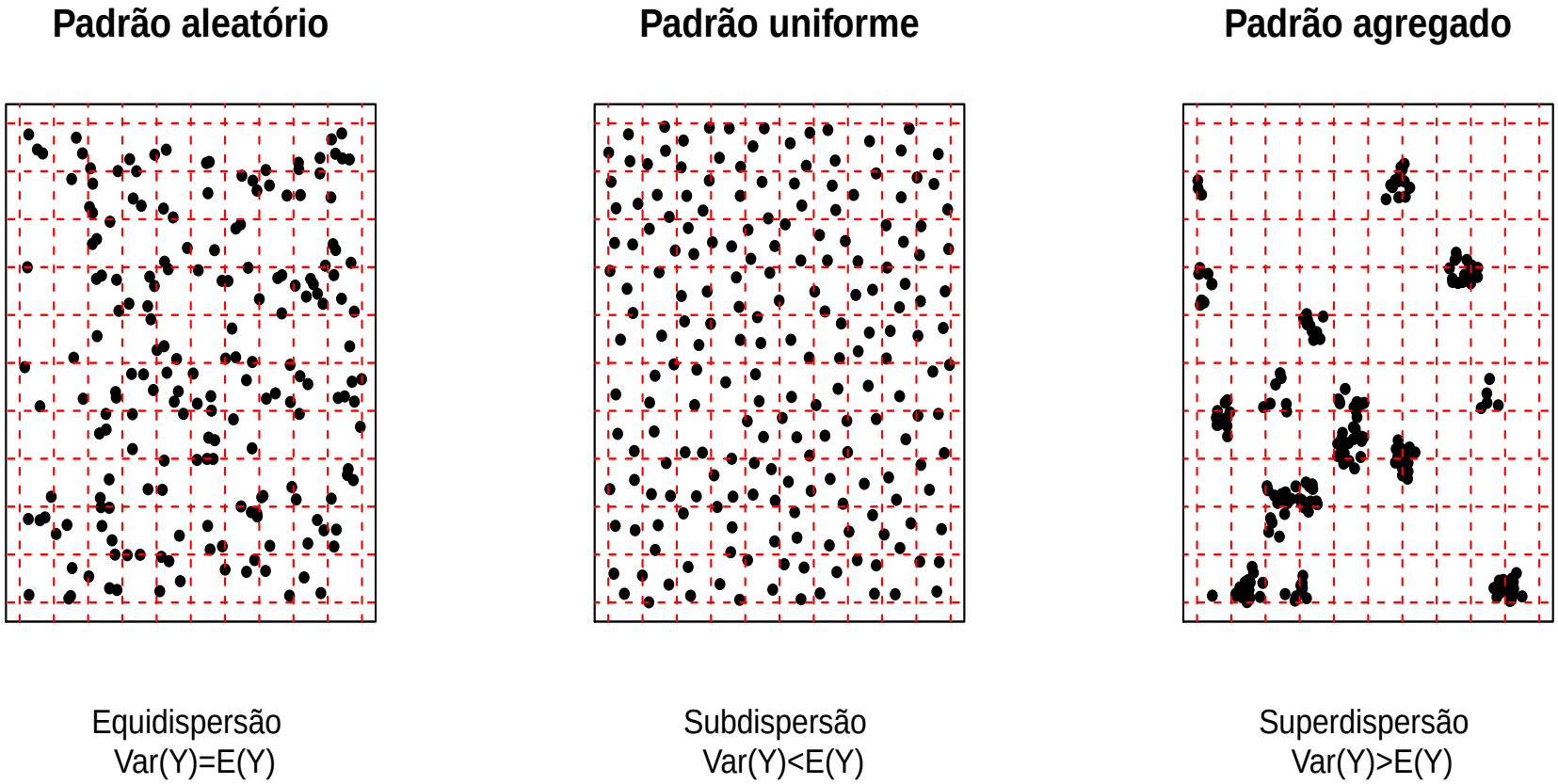


Figura 15: Ilustração de processos de contagens com diferentes padrões espaciais.

Modelo de Poisson

- Uma variável aleatória Y tem distribuição de Poisson se sua função de probabilidades for dada por:

$$f(y;\mu) = P(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, ...; \quad \mu > 0. \tag{103}$$

- Como ressaltado anteriormente, como propriedade da distribuição Poisson temos:

$$E(y) = Var(y) = \mu. \tag{104}$$

- A distribuição de Poisson pertence à família exponencial de distribuições, tendo função de variância $V(\mu) = \mu$ e parâmetro de dispersão $\phi = 1$.

- Adicionalmente, a medida que μ aumenta, a distribuição Poisson torna-se mais simétrica, aproximando-se da distribuição Normal quando $\mu \to \infty$

Modelo de Poisson

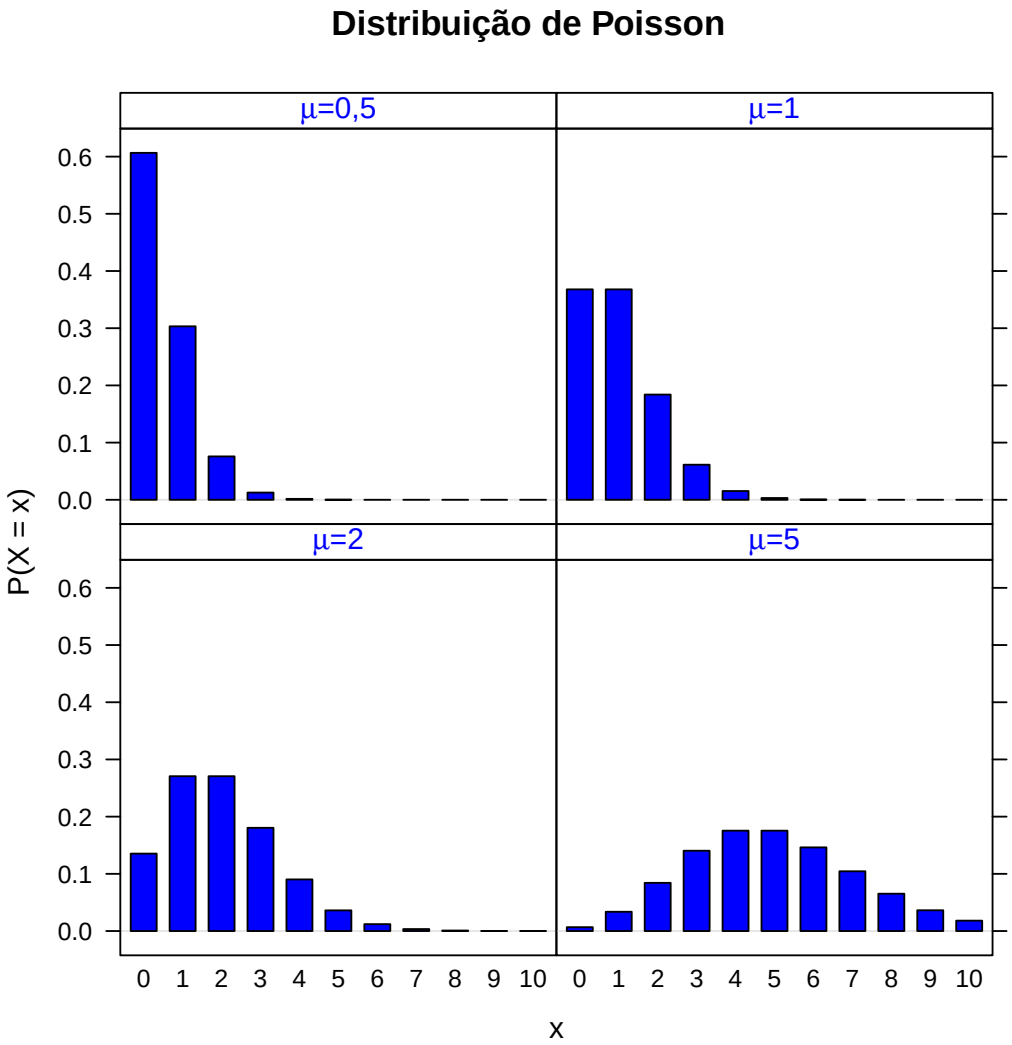


Figura 16: Distribuição de Poisson para diferentes valores de μ

Modelo de regressão log-linear

- No contexto de modelos lineares generalizados, vamos considerar $y_1, y_2, ..., y_n$ variáveis aleatórias independentes, com $y_i|\boldsymbol{x}_i \sim Poisson(\mu_i)$, $i = 1, 2, ..., n$.

- Adicionalmente, considere $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ os vetores de covariáveis correspondentes a cada observação na amostra.

- Considerando função de ligação logarítmica, o modelo log-linear de Poisson fica definido por:

$$y_i|\boldsymbol{x}_i \sim Poisson(\mu_i)$$
$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}.$$

Modelo de regressão log-linear

- Para o modelo log-linear de Poisson, as equações de estimação, baseadas na maximização da (log) verossimilhança, ficam dadas por:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 0, 1, ..., p \quad (\textit{verificar}!) \tag{105}$$

- O modelo log-linear fica expresso na escala da resposta (média) por:

$$\mu_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}. \tag{106}$$

- Verifica-se facilmente que os efeitos, para o modelo log-linear, são multiplicativos. Basta observar que:

$$\mu_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}} = e^{\beta_0}(e^{\beta_1})^{x_{i1}}(e^{\beta_2})^{x_{i2}}...(e^{\beta_p})^{x_{ip}}. \tag{107}$$

Modelo log-linear - interpretação dos parâmetros

- Considere o modelo log-linear com apenas uma covariável (x). Assumindo que x seja uma variável numérica, então:

$$\frac{\mu_{x+1}}{\mu_x} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}, \tag{108}$$

- Assim, o aumento de uma unidade em x tem efeito multiplicativo na média igual a e^{β_1}.

Modelo log-linear - interpretação dos parâmetros

- Para um aumento de k unidades em x, verifica-se facilmente que o efeito multiplicativo na média fica dado por $e^{k\beta_1}$.

- Para o caso de uma covariável dicotômica (com categorias A e B), inserida no modelo por meio de uma variável indicadora de B, temos:

$$\frac{\mu_B}{\mu_A} = \frac{e^{\beta_0+\beta_1\times 1}}{e^{\beta_0+\beta_1\times 0}} = e^{\beta_1}. \tag{109}$$

- Assim, e^{β_1} corresponde à razão das médias para as categorias B e A (efeito multiplicativo de B).

Modelo log-linear - interpretação dos parâmetros

- Se houvesse uma terceira categoria (C), então seriam necessárias duas variáveis indicadoras (x_1, indicadora de B; x_2, indicadora de C). Assim, teríamos:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{110}$$

- As razões de médias ficariam dadas por:

$$\frac{\mu_B}{\mu_A} = \frac{e^{\beta_0+\beta_1\times 1+\beta_2\times 0}}{e^{\beta_0+\beta_1\times 0+\beta_2\times 0}} = e^{\beta_1}; \tag{111}$$

$$\frac{\mu_C}{\mu_A} = \frac{e^{\beta_0+\beta_1\times 0+\beta_2\times 1}}{e^{\beta_0+\beta_1\times 0+\beta_2\times 0}} = e^{\beta_2}; \tag{112}$$

$$\frac{\mu_C}{\mu_B} = \frac{e^{\beta_0+\beta_1\times 0+\beta_2\times 1}}{e^{\beta_0+\beta_1\times 1+\beta_2\times 0}} = e^{\beta_2-\beta_1}. \tag{113}$$

Modelo log-linear - interpretação dos parâmetros

- Caso o preditor linear contenha múltiplas variáveis, as interpretações são idênticas, devendo-se ressaltar, no entanto, que a interpretação do efeito para uma particular variável é válida fixando os valores das demais variáveis.

Modelagem de taxas

- Na análise de dados de contagens, é comum que os indivíduos na amostra apresentem diferentes *níveis de exposição* (pacientes acompanhados por períodos de tempo distintos; contagens de peixes em trechos de um rio com diferentes volumes de água, ...);

- Em situações desse tipo, é necessário incorporar o nível de exposição ao modelo, de maneira a modelar a taxa de ocorrência por *unidade de exposição*.

- Seja y_i a contagem correspondente ao indivíduo i, com exposição t_i (ex: y_i : número de animais de certa espécie em $t_i = 100m^2$ de uma reserva).

- Neste caso, temos uma taxa de y_i/t_i de animais por *unidade de área*, com valor esperado $\lambda_i = \mu_i/t_i$.

Modelagem de taxas

- O modelo log-linear, aplicado à modelagem de taxas, fica dado por:

$$log(\lambda_i) = log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \quad (114)$$

tal que:

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + log(t_i). \quad (115)$$

- Neste caso, a variável $log(t_i)$ deve ser incluída como covariável no preditor, forçando seu coeficiente a ser 1. (Chamamos $log(t_i)$ de termo *offset*).
- Escrevendo o modelo na escala da resposta (média), temos:

$$\mu_i = t_i e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}. \quad (116)$$

Qualidade do ajuste

- A deviance para o MLG Poisson fica dada por:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left[y_i \, log \left(\frac{y_i}{\hat{\mu}_i} \right) - y_i + \hat{\mu}_i \right]. \tag{117}$$

- Para a distribuição de Poisson, $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}$ quando $\mu_i \to \infty$ para todo i.
- A estatística X^2 de Pearson, definida por:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{118}$$

também tem distribuição χ^2_{n-p} quando $\mu_i \to \infty$ para todo i.

- Nessas condições, tanto a deviance quanto a estatística X^2 podem ser usadas para testar a hipótese nula de que o modelo está bem ajustado.