Extensões ao modelo normal

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 10 de junho de 2024

Roteiro

Roteiro

Modelo de regressão linear

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por um conjunto de variáveis preditoras ($\boldsymbol{x}$), da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

Modelo de regressão linear

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por um conjunto de variáveis preditoras ($\boldsymbol{x}$), da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

em que $\mu = \boldsymbol{x}^\top \beta$ e σ^2 são a média e variância, respectivamente, da distribuição de Y, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ é um vetor conhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ e σ^2 são parâmetros desconhecidos.

Modelo de regressão linear

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por um conjunto de variáveis preditoras ($\boldsymbol{x}$), da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

em que $\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$ e σ^2 são a média e variância, respectivamente, da distribuição de Y, $\boldsymbol{x} = (x_1, x_2, \ldots, x_{\boldsymbol{p}})^\top$ é um vetor conhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{\boldsymbol{p}})^\top$ e σ^2 são parâmetros desconhecidos.

Modelo de regressão linear

É conveniente escrever a relação entre as variáveis resposta e preditora da seguinte forma:

$$Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon,$$

em que $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, sendo ε denominado de erro.

Modelo de regressão linear

É conveniente escrever a relação entre as variáveis resposta e preditora da seguinte forma:

$$Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon,$$

em que $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, sendo ε denominado de erro.

Modelo de regressão linear

A fim de estimar $\boldsymbol{\beta}$ e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$,

Modelo de regressão linear

A fim de estimar β e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

em que $\mu_\ell = \boldsymbol{x}_\ell^\top \beta$, $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão linear

A fim de estimar $\boldsymbol{\beta}$ e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

em que $\mu_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$, $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão linear

De forma alternativa,

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

com $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão linear

De forma alternativa,

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

com $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$. Em outras palavras, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2.

Modelo de regressão linear

De forma alternativa,

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

com $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$. Em outras palavras, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2.

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Modelo de regressão normal linear

Para obter estimativas dos parâmetros, somente é necessário usar os estimadores oriundos do **método de mínimos quadrados**. Porém, para uso de procedimentos inferenciais, como testes de hipóteses ou intervalos de confiança,

Modelo de regressão normal linear

Para obter estimativas dos parâmetros, somente é necessário usar os estimadores oriundos do **método de mínimos quadrados**. Porém, para uso de procedimentos inferenciais, como testes de hipóteses ou intervalos de confiança, a seguinte suposição precisa ser adicionada:

$$Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2),$$

Modelo de regressão normal linear

Para obter estimativas dos parâmetros, somente é necessário usar os estimadores oriundos do **método de mínimos quadrados**. Porém, para uso de procedimentos inferenciais, como testes de hipóteses ou intervalos de confiança, a seguinte suposição precisa ser adicionada:

$$Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2),$$

de forma equivalente, $Y_\ell = \boldsymbol{x}_\ell^\top \beta + \varepsilon_\ell$, com $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão normal linear

Para obter estimativas dos parâmetros, somente é necessário usar os estimadores oriundos do **método de mínimos quadrados**. Porém, para uso de procedimentos inferenciais, como testes de hipóteses ou intervalos de confiança, a seguinte suposição precisa ser adicionada:

$$Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2),$$

de forma equivalente, $Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell$, com $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão normal linear

Isto é, nós precisamos assumir que a natureza dos dados segue uma distribuição normal. Como ela é uma distribuição simétrica, com suporte na reta real,

Modelo de regressão normal linear

Isto é, nós precisamos assumir que a natureza dos dados segue uma distribuição normal. Como ela é uma distribuição simétrica, com suporte na reta real, supor normalidade em situações em que os dados são assimétricos, estritamente positivos ou de contagem pode não ser razoável.

Modelo de regressão normal linear

Isto é, nós precisamos assumir que a natureza dos dados segue uma distribuição normal. Como ela é uma distribuição simétrica, com suporte na reta real, supor normalidade em situações em que os dados são assimétricos, estritamente positivos ou de contagem pode não ser razoável.

Roteiro

Família exponencial de distribuições

Uma alternativa para ajustar os dados é supor que a sua natureza segue alguma distribuição da **família exponencial**, i.e.,

$$Y \sim \mathcal{FE}(\theta, \phi).$$

Família exponencial de distribuições

Uma alternativa para ajustar os dados é supor que a sua natureza segue alguma distribuição da **família exponencial**, i.e.,

$$Y \sim \mathcal{F}\mathcal{E}(\theta, \phi).$$

Família exponencial de distribuições

Se $Y \sim \mathcal{FE}(\theta, \phi)$, sua função densidade de probabilidade pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi[y\theta - b(\theta) + c(y)] + d(y, \phi)\right\},$$

Família exponencial de distribuições

Se $Y \sim \mathcal{FE}(\theta, \phi)$, sua função densidade de probabilidade pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi[y\theta - b(\theta) + c(y)] + d(y, \phi)\right\},$$

em que $b(\cdot)$, $c(\cdot)$ e $d(\cdot,\cdot)$ são funções conhecidas,

Família exponencial de distribuições

Se $Y \sim \mathcal{FE}(\theta, \phi)$, sua função densidade de probabilidade pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi[y\theta - b(\theta) + c(y)] + d(y, \phi)\right\},$$

em que $b(\cdot)$, $c(\cdot)$ e $d(\cdot, \cdot)$ são funções conhecidas, θ e ϕ, são, respectivamente, os parâmetros canônico e de precisão

Família exponencial de distribuições

Se $Y \sim \mathcal{FE}(\theta, \phi)$, sua função densidade de probabilidade pode ser escrita da seguinte forma

$$f(y) = \exp\{\phi[y\theta - b(\theta) + c(y)] + d(y, \phi)\},$$

em que $b(\cdot)$, $c(\cdot)$ e $d(\cdot, \cdot)$ são funções conhecidas, θ e ϕ, são, respectivamente, os parâmetros canônico e de precisão (o inverso, ϕ^{-1}, é o parâmetro de dispersão).

Família exponencial de distribuições

Se $Y \sim \mathcal{F}\mathcal{E}(\theta, \phi)$, sua função densidade de probabilidade pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi[y\theta - b(\theta) + c(y)] + d(y, \phi)\right\},$$

em que $b(\cdot)$, $c(\cdot)$ e $d(\cdot, \cdot)$ são funções conhecidas, θ e ϕ, são, respectivamente, os parâmetros canônico e de precisão (o inverso, ϕ^{-1}, é o parâmetro de dispersão).

Família exponencial de distribuições

Além disso, se $Y \sim \mathcal{FE}(\theta, \phi)$, então

$$\mathbb{E}(Y) = \frac{db(\theta)}{d\theta} = \mu \text{ e } \mathrm{Var}(Y) = \phi^{-1}\frac{d^2 b(\theta)}{d\theta^2} = \phi^{-1} V(\mu),$$

Família exponencial de distribuições

Além disso, se $Y \sim \mathcal{FE}(\theta, \phi)$, então

$$\mathbb{E}(Y) = \frac{db(\theta)}{d\theta} = \mu \text{ e } \mathrm{Var}(Y) = \phi^{-1}\frac{d^2b(\theta)}{d\theta^2} = \phi^{-1}V(\mu),$$

em que $V = V(\mu)$ é denominada de função de variância e

Família exponencial de distribuições

Além disso, se $Y \sim \mathcal{FE}(\theta, \phi)$, então

$$\mathbb{E}(Y) = \frac{db(\theta)}{d\theta} = \mu \text{ e } \mathrm{Var}(Y) = \phi^{-1}\frac{d^2 b(\theta)}{d\theta^2} = \phi^{-1}V(\mu),$$

em que $V = V(\mu)$ é denominada de função de variância e $\theta = \int V^{-1}\mathrm{d}\mu = q(\mu)$,

Família exponencial de distribuições

Além disso, se $Y \sim \mathcal{FE}(\theta, \phi)$, então

$$\mathbb{E}(Y) = \frac{db(\theta)}{d\theta} = \mu \text{ e } \mathrm{Var}(Y) = \phi^{-1}\frac{d^2b(\theta)}{d\theta^2} = \phi^{-1}V(\mu),$$

em que $V = V(\mu)$ é denominada de função de variância e $\theta = \int V^{-1}\mathrm{d}\mu = q(\mu)$, sendo $q(\mu)$ uma função conhecida um-a-um de μ.

Família exponencial de distribuições

Além disso, se $Y \sim \mathcal{FE}(\theta, \phi)$, então

$$\mathbb{E}(Y) = \frac{db(\theta)}{d\theta} = \mu \text{ e } \mathrm{Var}(Y) = \phi^{-1}\frac{d^2 b(\theta)}{d\theta^2} = \phi^{-1}V(\mu),$$

em que $V = V(\mu)$ é denominada de função de variância e $\theta = \int V^{-1}\mathrm{d}\mu = q(\mu)$, sendo $q(\mu)$ uma função conhecida um-a-um de μ. As Tabelas 1 e 2 apresentam as quantidades mencionadas até aqui para cinco casos.

Família exponencial de distribuições

Além disso, se $Y \sim \mathcal{FE}(\theta, \phi)$, então

$$\mathbb{E}(Y) = \frac{db(\theta)}{d\theta} = \mu \text{ e } \mathrm{Var}(Y) = \phi^{-1}\frac{d^2 b(\theta)}{d\theta^2} = \phi^{-1}V(\mu),$$

em que $V = V(\mu)$ é denominada de função de variância e $\theta = \int V^{-1}\mathrm{d}\mu = q(\mu)$, sendo $q(\mu)$ uma função conhecida um-a-um de μ. As Tabelas 1 e 2 apresentam as quantidades mencionadas até aqui para cinco casos.

Tabela 1: Distribuições discretas pertencentes à $\mathcal{FE}$.

Quantidade	Binomial	Poisson
θ	$\log \dfrac{\mu}{1-\mu}$	$\log \mu$
ϕ	n	1
$b(\theta)$	$\log(1+e^{\theta})$	e^{θ}
$c(y)$	0	0
$d(y,\phi)$	$\log \dbinom{\phi}{\phi y}$	$-\log y!$
$V(\mu)$	$\mu(1-\mu)$	μ

Tabela 2: Distribuições contínuas pertencentes à $\mathcal{FE}$.

Quantidade	Gama	Normal	Normal inversa
θ	$-\frac{1}{\mu}$	μ	$-\frac{1}{2\mu^2}$
ϕ	$\frac{1}{CV^2}$	$\frac{1}{\sigma^2}$	ϕ
$b(\theta)$	$-\log(-\theta)$	$\frac{\theta^2}{2}$	$-\sqrt{-2\theta}$
$c(y)$	$\log y$	$-\frac{y^2}{2}$	$-\frac{1}{2y}$
$d(y,\phi)$	$-\log y + \phi \log \phi - \log \Gamma(\phi)$	$\frac{1}{2}\log\frac{\phi}{2\pi}$	$\frac{1}{2}\log\frac{\phi}{2\pi y^3}$
$V(\mu)$	μ^2	1	μ^3

Família exponencial de distribuições

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e β.

Família exponencial de distribuições

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e $\boldsymbol{\beta}$.

Isso pode ser problemático, pois a média de uma Bernoulli, por exemplo, está contida no intervalo entre 0 e 1

Família exponencial de distribuições

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e $\boldsymbol{\beta}$.

Isso pode ser problemático, pois a média de uma Bernoulli, por exemplo, está contida no intervalo entre 0 e 1 e o resultado de $\boldsymbol{x}^\top \boldsymbol{\beta}$ poderá estar fora desse intervalo.

Família exponencial de distribuições

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e $\boldsymbol{\beta}$.

Isso pode ser problemático, pois a média de uma Bernoulli, por exemplo, está contida no intervalo entre 0 e 1 e o resultado de $\boldsymbol{x}^\top \boldsymbol{\beta}$ poderá estar fora desse intervalo.

Família exponencial de distribuições

Para contornar esse problema, ao invés de nós ajustarmos a média, nós podemos ajustar uma função dela, i.e.,

$$g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

Família exponencial de distribuições

Para contornar esse problema, ao invés de nós ajustarmos a média, nós podemos ajustar uma função dela, i.e.,

$$g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

em que $g(\cdot)$ é uma função invertível e conhecida, sendo ela denominada de **função de ligação**.

Para contornar esse problema, ao invés de nós ajustarmos a média, nós podemos ajustar uma função dela, i.e.,

$$g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

em que $g(\cdot)$ é uma função invertível e conhecida, sendo ela denominada de **função de ligação**.

Modelos lineares generalizados

Propostos por Nelder e Wedderburn (1972), os modelos lineares generalizados (MLG) supõem um fenômeno Y,

Modelos lineares generalizados

Propostos por Nelder e Wedderburn (1972), os modelos lineares generalizados (MLG) supõem um fenômeno Y, tal que $Y \sim \mathcal{FE}(\theta, \phi)$, com $g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}$.

Modelos lineares generalizados

Propostos por Nelder e Wedderburn (1972), os modelos lineares generalizados (MLG) supõem um fenômeno Y, tal que $Y \sim \mathcal{FE}(\theta, \phi)$, com $g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}$.

Modelos lineares generalizados

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$,

Modelos lineares generalizados

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$,

Modelos lineares generalizados

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{FE}(\theta_\ell, \phi),$$

Modelos lineares generalizados

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{FE}(\theta_\ell, \phi),$$

em que $g(\mu_\ell) = \boldsymbol{x}_\ell^\top \beta$, $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$, $\ell = 1, 2, \ldots, n$.

Modelos lineares generalizados

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{FE}(\theta_\ell, \phi),$$

em que $g(\mu_\ell) = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$, $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$, $\ell = 1, 2, \ldots, n$.

Modelos lineares generalizados

Os estimadores do vetor $\boldsymbol{\beta}$ e escalar ϕ são obtidos pelo método da máxima verossimilhança.

Roteiro

Família de dispersão

Uma segunda alternativa para ajustar um conjunto de dados é supor que a sua natureza segue alguma distribuição da **família de dispersão**, i.e.,

$$Y \sim \mathcal{FD}(\theta, \phi).$$

Família de dispersão

Uma segunda alternativa para ajustar um conjunto de dados é supor que a sua natureza segue alguma distribuição da **família de dispersão**, i.e.,

$$Y \sim \mathcal{FD}(\theta, \phi).$$

Família de dispersão

Se $Y \sim \mathcal{FD}(\theta, \phi)$, sua função densidade de probabilidade (ou função de probabilidade) pode ser escrita da seguinte forma

Família de dispersão

Se $Y \sim \mathcal{FD}(\theta, \phi)$, sua função densidade de probabilidade (ou função de probabilidade) pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi t(y, \theta) + a(y, \phi)\right\},$$

Família de dispersão

Se $Y \sim \mathcal{FD}(\theta, \phi)$, sua função densidade de probabilidade (ou função de probabilidade) pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi t(y,\theta) + a(y,\phi)\right\},$$

em que $t(\cdot,\cdot)$ e $a(\cdot)$ são funções conhecidas, $\theta \in \mathbb{R}$ e $\phi > 0$.

Família de dispersão

Se $Y \sim \mathcal{FD}(\theta, \phi)$, sua função densidade de probabilidade (ou função de probabilidade) pode ser escrita da seguinte forma

$$f(y) = \exp\left\{\phi t(y, \theta) + a(y, \phi)\right\},$$

em que $t(\cdot, \cdot)$ e $a(\cdot)$ são funções conhecidas, $\theta \in \mathbb{R}$ e $\phi > 0$.

Tabela 3: Distribuições da família de dispersão.

Distribuição	Suporte
Binomial	$\{0, 1, \ldots, n\}$
Gama	$(0, \infty)$
Normal inversa	$(0, \infty)$
Normal	$(-\infty, \infty)$
Poisson	$\{0, 1, \ldots\}$
Secante hiperbólica generalizada	$(-\infty, \infty)$
Binomial negativa	$\{0, 1, \ldots\}$
Família exponencial	$\subset \mathbb{R}$
Modelos Morris	$\subset \mathbb{R}$
Classe Tweedie	$\subset \mathbb{R}$

Tabela 4: Distribuições da família de dispersão.

Distribuição	Suporte
Hipérbole	$(0, \infty)$
Hiperbólica	$(-\infty, \infty)$
Leipnik	$(-1, 1)$
Log-gama	$(0, \infty)$
Normal inversa recíproca	$(0, \infty)$
Gama inversa	$(0, \infty)$
Leipnik transformada	$(0, 1)$
Simplex	$(0, 1)$
von-Mises	$(0, 2\pi)$

Família de dispersão

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\theta = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e β.

Família de dispersão

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\theta = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e $\boldsymbol{\beta}$.

Isso pode ser problemático, pois se θ está contido no intervalo entre 0 e 1, o resultado de $\boldsymbol{x}^\top \beta$ precisará estar dentro desse intervalo.

Família de dispersão

Partindo da mesma ideia de um modelo de regressão linear, nós podemos supor que $\theta = \boldsymbol{x}^\top \boldsymbol{\beta}$. Todavia, notem que, não há nenhuma restrição sobre os vetores $\boldsymbol{x}$ e $\boldsymbol{\beta}$.

Isso pode ser problemático, pois se θ está contido no intervalo entre 0 e 1, o resultado de $\boldsymbol{x}^\top \boldsymbol{\beta}$ precisará estar dentro desse intervalo.

Família de dispersão

Para contornar esse problema, ao invés de nós ajustarmos θ, nós podemos ajustar uma função dele,

Família de dispersão

Para contornar esse problema, ao invés de nós ajustarmos θ, nós podemos ajustar uma função dele, i.e.,

$$g(\theta) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

Família de dispersão

Para contornar esse problema, ao invés de nós ajustarmos θ, nós podemos ajustar uma função dele, i.e.,

$$g(\theta) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

em que $g(\cdot)$ é uma função invertível e conhecida, sendo ela denominada de **função de ligação**.

Família de dispersão

Para contornar esse problema, ao invés de nós ajustarmos θ, nós podemos ajustar uma função dele, i.e.,

$$g(\theta) = \boldsymbol{x}^\top \boldsymbol{\beta},$$

em que $g(\cdot)$ é uma função invertível e conhecida, sendo ela denominada de **função de ligação**.

Modelos de dispersão

Propostos por Jørgensen (1997), os modelos de dispersão supõem um fenômeno Y,

Modelos de dispersão

Propostos por Jørgensen (1997), os modelos de dispersão supõem um fenômeno Y, tal que $Y \sim \mathcal{FD}(\theta, \phi)$, com $g(\theta) = \boldsymbol{x}^\top \beta$.

Modelos de dispersão

Propostos por Jørgensen (1997), os modelos de dispersão supõem um fenômeno Y, tal que $Y \sim \mathcal{FD}(\theta, \phi)$, com $g(\theta) = \boldsymbol{x}^\top \boldsymbol{\beta}$.

Modelos de dispersão

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$,

Modelos de dispersão

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$,

Modelos de dispersão

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{FD}(\theta_\ell, \phi),$$

Modelos de dispersão

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{FD}(\theta_\ell, \phi),$$

em que $g(\theta_\ell) = \boldsymbol{x}_\ell^\top \beta$, $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$, $\ell = 1, 2, \ldots, n$.

Modelos de dispersão

Da mesma forma o que ocorre no modelo de regressão linear, para explicitar a relação entre as variáveis, é necessário estimar os parâmetros do modelo, para isso, é necessário retirar uma amostra independente de tamanho n do vetor $(Y, \boldsymbol{x}^\top)^\top$, i.e., $(Y_1, \boldsymbol{x}_1^\top)^\top, (Y_2, \boldsymbol{x}_2^\top)^\top, \ldots, (Y_n, \boldsymbol{x}_n^\top)^\top$, com

$$Y_\ell \sim \mathcal{FD}(\theta_\ell, \phi),$$

em que $g(\theta_\ell) = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$, $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$, $\ell = 1, 2, \ldots, n$.

Modelos de dispersão

Os estimadores do vetor $\boldsymbol{\beta}$ e escalar ϕ são obtidos pelo método da máxima verossimilhança.

Roteiro

Estudo de simulação

Para avaliar o desempenho dos estimadores de máxima verossimilhança, nós utilizaremos o método de Monte Carlo. Nossa suposição será de que $Y \sim \mathcal{FE}(\theta, \phi)$, com $g(\mu) = \beta_1 + \beta_2 x$.

Estudo de simulação

Para avaliar o desempenho dos estimadores de máxima verossimilhança, nós utilizaremos o método de Monte Carlo. Nossa suposição será de que $Y \sim \mathcal{FE}(\theta, \phi)$, com $g(\mu) = \beta_1 + \beta_2 x$.

Serão 5.000 réplicas, para tamanho de amostras $n = 10, 20, 40, 80, 160$, para as distribuição normal (ligação identidade), Bernoulli (ligação logito) e gama (ligação log), $\beta = (1, 1)^\top$ e $\phi = 10$, para a Bernoulli, $\phi = 1$.

Estudo de simulação

Para avaliar o desempenho dos estimadores de máxima verossimilhança, nós utilizaremos o método de Monte Carlo. Nossa suposição será de que $Y \sim \mathcal{FE}(\theta, \phi)$, com $g(\mu) = \beta_1 + \beta_2 x$.

Serão 5.000 réplicas, para tamanho de amostras $n = 10, 20, 40, 80, 160$, para as distribuição normal (ligação identidade), Bernoulli (ligação logito) e gama (ligação log), $\boldsymbol{\beta} = (1, 1)^\top$ e $\phi = 10$, para a Bernoulli, $\phi = 1$.

Estudo de simulação

O desempenho será avaliado através do viés absoluto e do erro quadrático médio (EQM). Sem perda generalidade, o viés absoluto e o EQM do estimador de um parâmetro δ são dados, respectivamente por

Estudo de simulação

O desempenho será avaliado através do viés absoluto e do erro quadrático médio (EQM). Sem perda generalidade, o viés absoluto e o EQM do estimador de um parâmetro δ são dados, respectivamente por

$$\text{Viés absoluto} = \left|\hat{\delta}_{\mathrm{MC}} - \delta\right| \text{ e EQM} = \sum_{i=1}^{5.000} \frac{(\hat{\delta}_i - \delta)^2}{5.000},$$

em que $\hat{\delta}_{\mathrm{MC}} = 5.000^{-1} \sum_{i=1}^{5.000} \hat{\delta}_i$.

Estudo de simulação

O desempenho será avaliado através do viés absoluto e do erro quadrático médio (EQM). Sem perda generalidade, o viés absoluto e o EQM do estimador de um parâmetro δ são dados, respectivamente por

$$\text{Viés absoluto} = \left|\hat{\delta}_{\mathrm{MC}} - \delta\right| \text{ e EQM} = \sum_{i=1}^{5.000} \frac{(\hat{\delta}_i - \delta)^2}{5.000},$$

em que $\hat{\delta}_{\mathrm{MC}} = 5.000^{-1} \sum_{i=1}^{5.000} \hat{\delta}_i$.

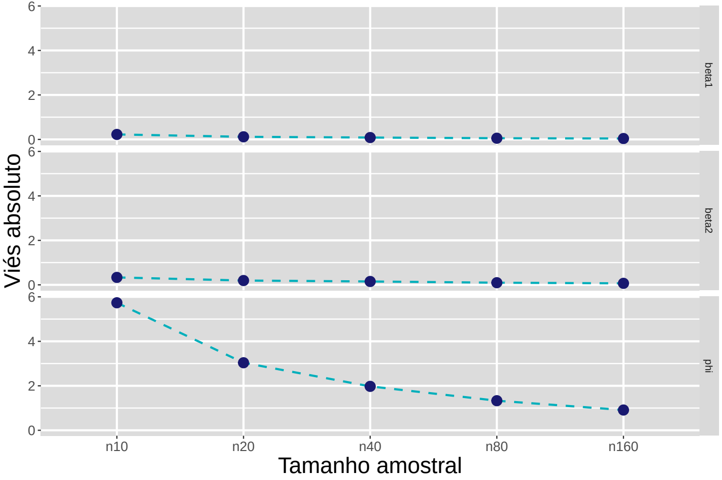

Figura 1: Viés absoluto dos parâmetros - normal.

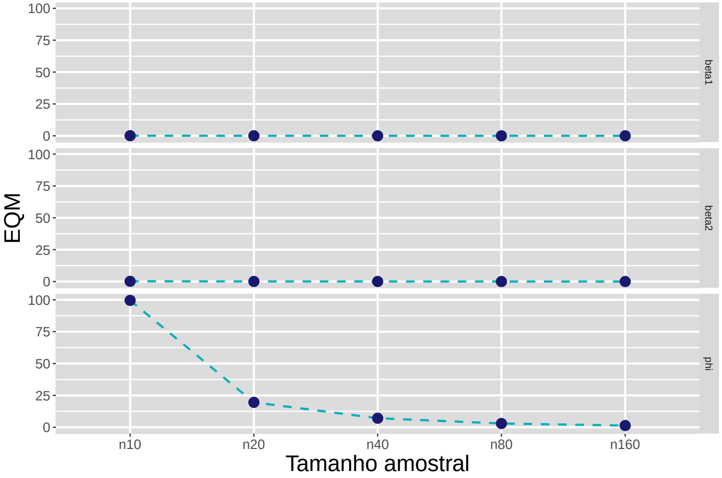

Figura 2: EQM dos parâmetros - normal.

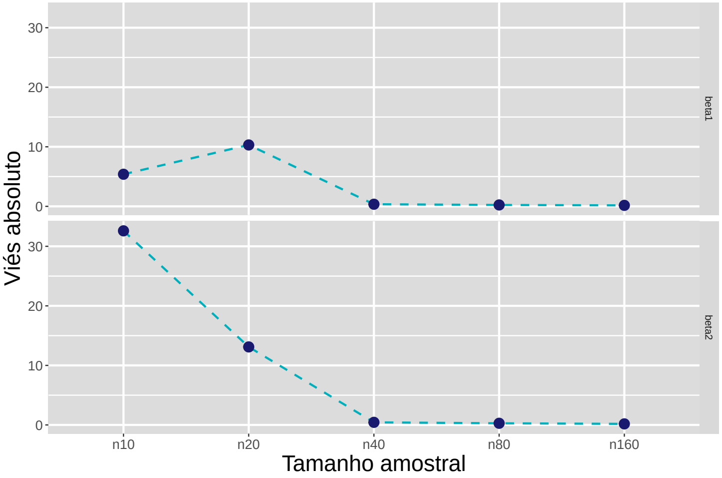

Figura 3: Viés absoluto dos parâmetros - Bernoulli.

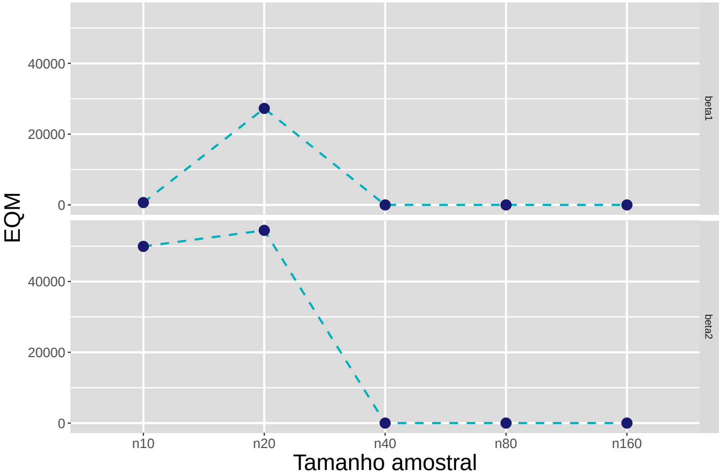

Figura 4: EQM dos parâmetros - Bernoulli.

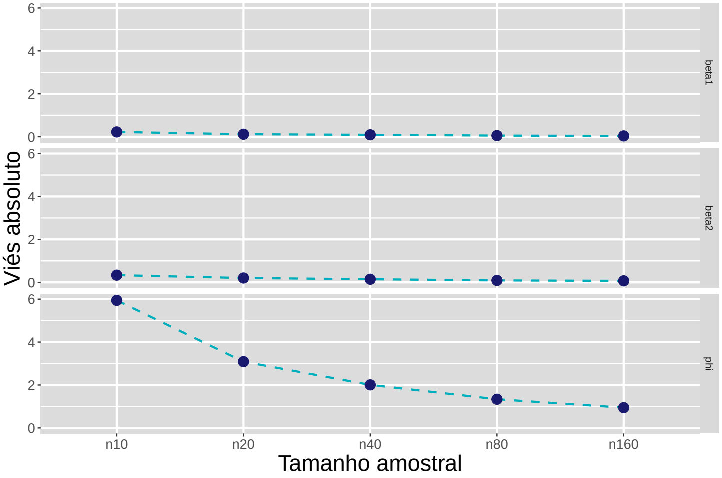

Figura 5: Viés absoluto dos parâmetros - gama.

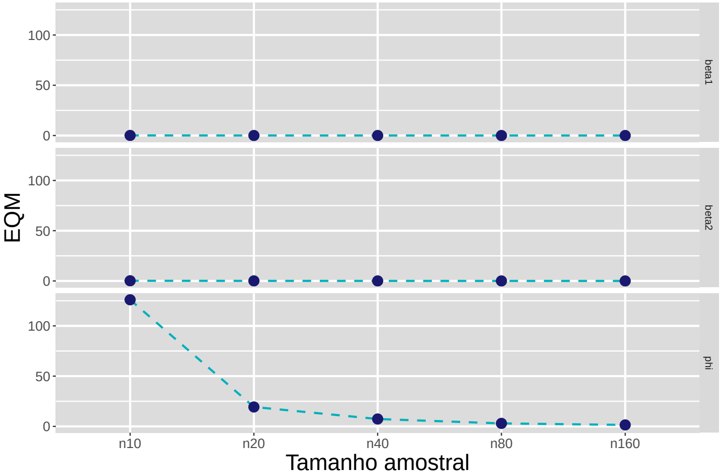

Figura 6: EQM dos parâmetros - gama.

Estudo de simulação

De maneira geral, para tamanhos de amostras grandes (80 e 160), os estimadores para β e ϕ têm viés absoluto e EQM próximos a zero.

Estudo de simulação

De maneira geral, para tamanhos de amostras grandes (80 e 160), os estimadores para $\boldsymbol{\beta}$ e ϕ têm viés absoluto e EQM próximos a zero.

Para amostras pequenas (10 e 20), o estimador de ϕ parece não ser confiável.

Estudo de simulação

De maneira geral, para tamanhos de amostras grandes (80 e 160), os estimadores para $\boldsymbol{\beta}$ e ϕ têm viés absoluto e EQM próximos a zero.

Para amostras pequenas (10 e 20), o estimador de ϕ parece não ser confiável. No caso da Bernoulli, o estimador de $\boldsymbol{\beta}$ também retornou estimativas longes dos valores verdadeiros.

Estudo de simulação

De maneira geral, para tamanhos de amostras grandes (80 e 160), os estimadores para $\boldsymbol{\beta}$ e ϕ têm viés absoluto e EQM próximos a zero.

Para amostras pequenas (10 e 20), o estimador de ϕ parece não ser confiável. No caso da Bernoulli, o estimador de $\boldsymbol{\beta}$ também retornou estimativas longes dos valores verdadeiros.

Roteiro

Referências bibliográficas I

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Chapman & Hall, London.

Nelder, J. A. e Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A (General)* **135**(3), 370–384.

Obrigado!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Departamento de Estatística, Sala 319