UNIVERSIDADE FEDERAL DE SERGIPE
DECAT
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

Modelos lineares generalizados

Disciplina ofertada pelo DECAT/UFS

Código: ESTAT0092
Nível: Graduação
Carga horária: 60h
Período: 2020.2

Professor responsável e ministrante: Luiz Henrique Dore

Tópico 2: família exponencial de distribuições

Sumário

Metas

1. Apresentar uma definição para a família exponencia de distribuições, algumas de suas propriedades e alguns membros dessa família de distribuições.

Objetivos

- Após estudar esse tópico, o aluno ou aluna será capaz de:
  1. caracterizar uma família exponencial de distribuições;
  2. identificar se uma dada distribuição de probabilidade é membro da família exponencial de distribuições;

Pré-requisitos

1. Não há pré-requisitos.

Modelagem de regressão e o modelo linear normal

- A estatística é uma ciência que lida com desenvolvimento de teorias e métodos matemáticos e computacionais com o objetivo de compreender o comportamento de uma entidade, a partir de observações dessa entidade.
- A entidade é representada por características mensuráveis, as quais são denominadas variáveis.
- Cada vez que a entidade é observada, as variáveis são medidas e seus valores são registrados.
- Em geral, os valores referentes a diferentes observações da entidade são diferentes entre si.
- Dessa forma, o conjunto dos valores registrados exibe uma variabilidade.

Modelagem de regressão e o modelo linear normal

- Essa variabilidade é analisada a partir da frequência com a qual os valores ocorrem. O objetivo é identificar padrões de frequência no conjunto dos valores registrados que permitam responder questões a respeito do comportamento da entidade.
- Métodos matemáticos e computacionais, com fundamentos na teoria da probabilidade, são utilizados para esse fim.
- Tais métodos são denominados métodos estatísticos.
- A premissa básica para o funcionamento de métodos estatísticos é a de que a variabilidade observada no conjunto de valores registrados é regida por algum mecanismo aleatório.
- Frequentemente, o objetivo da análise é verificar a existência de uma **associação** entre as variáveis.

Modelagem de regressão e o modelo linear normal

Exemplo 1

O conjunto de dados **waterPolution.dat**, que está na pasta **Dados**, contém informações sobre a qualidade da água em vinte bacias hidrográficas, no estado de Nova York, e sobre o uso da terra no entorno dessas bacias. Esses dados foram coletados com o objetivo de investigar como o uso da terra contribui para a poluição das águas [1, p. 6]. Para cada bacia, foram registrados os valores da porcentagem da área em uso agrícola, da porcentagem de floresta na área, da porcentagem da área em uso residencial, da porcentagem da área em uso comercial/industrial e da concentração média de nitrogênio.

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 1

Esse conjunto de dados encontra-se disponível, sob o nome de **P010.txt**, na seguinte página:

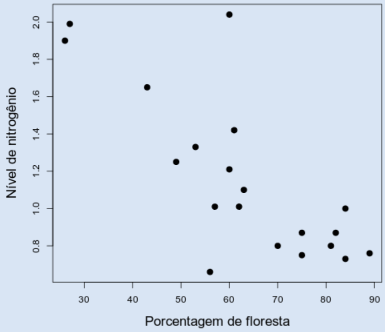
http://www1.aucegypt.edu/faculty/hadi/RABE4/.

O gráfico de dispersão é a ferramenta mais básica para detecção de associação entre duas variáveis. Um padrão de variação **sistemática** no gráfico, lembrando o gráfico de uma função, indica que há associação entre as variáveis. A figura 1 contém o gráfico de dispersão do nível de nitrogênio versus a porcentagem de floresta na área. Pode-se ver que o nível de nitrogênio **tende** a diminuir quando a porcentagem de floresta aumenta, indicando uma associação entre essas duas variáveis.

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 1

Figura 1: Nível de nitrogênio vs Porcentagem de floresta.



Modelagem de regressão e o modelo linear normal

Exemplo 2

O conjunto de dados **store.dat**, na pasta **Dados**, contém informações sobre os clientes de uma determinada loja, oriundos de 110 áreas de uma cidade [7, p. 299]. Para cada área, foram registrados os valores das seguintes variáveis: número de clientes da loja, número de domicílios (em mil), renda média anual (em mil USD), idade média dos domicílios (em anos), distância ao concorrente mais próximo (em milhas) e distância à loja (em milhas). Pretende-se investigar se o número de clientes da loja está relacionado às demais variáveis e como se dá essa relação.

Modelagem de regressão e o modelo linear normal

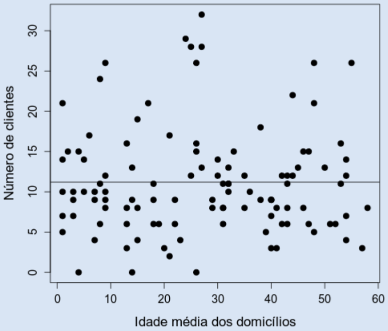
Continuação do exemplo 2

As figuras 2 e 3 contém os gráficos de dispersão do número de clientes vs a idade média dos domicílios e do número de clientes vs a distância ao concorrente mais próximo. Os pontos no gráfico na figura 2 se distribuem em torno de uma reta constante. Ou seja, quando a idade média dos domicílios aumenta (ou diminui), o número de clientes permanece, mais ou menos, no mesmo nível. Isso indica que não há associação entre o número de clientes e a idade média do domicílio. Já o gráfico da figura 3 indica uma associação entre o número de clientes e a distância ao concorrente mais próximo: quando a distância aumenta, o número de clientes **tende** a aumentar.

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 2

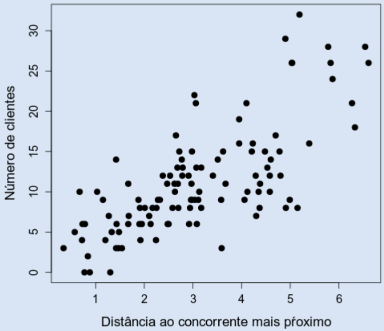
Figura 2: Número de clientes vs idade média dos domicílios.



Modelagem de regressão e o modelo linear normal

Continuação do exemplo 2

Figura 3: Número de clientes vs distância ao concorrente.



Modelagem de regressão e o modelo linear normal

Exemplo 3

O conjunto de dados **icu.csv**, na pasta **Dados**, contém informações a respeito de 200 pessoas que foram parte de um estudo sobre a sobrevivência de pacientes internados numa UTI de um hospital [5, p. 22]. O objetivo desse estudo foi elaborar um modelo matemático para **predizer** a probabilidade de que um paciente sobreviva ao período de internação numa UTI, com base em algumas características que descrevem a condição do paciente no momento da internação. Para cada paciente, foram registrados os valores de 21 variáveis. As descrições dessas variáveis encontram-se nas páginas 13 e 14 do arquivo **aplore3.pdf**, na pasta **Dados**.

Modelagem de regressão e o modelo linear normal

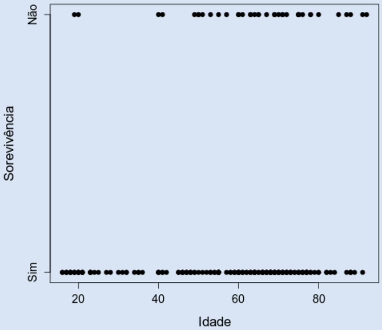
Continuação do exemplo 3

A variável **stat** é uma variável binária que indica se o paciente sobreviveu ao período na UTI (Lived) ou não (Died). A variável **age** é a idade do paciente em anos. A figura 4 contém o gráfico de dispersão da variável **stat** versus a variável **age**. Pode-se ver que há uma concentração dos pacientes que não sobrevivem em faixas etárias maiores. Porém, há também vários pacientes sobreviventes nessas faixas etárias maiores. O gráfico não exibe um padrão claro de variação sistemática, sendo, portanto, pouco informativo. Isso se deve à natureza dicotômica da variável **stat** e ao fato de haver ao menos um sobrevivente e um não sobrevivente de quase todas as idades.

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 3

Figura 4: Sobrevivência vs Idade.



Modelagem de regressão e o modelo linear normal

Continuação do exemplo 3

Um gráfico mais útil pode ser elaborado agrupando-se os pacientes por faixa etária e calculando-se a proporção de sobreviventes em cada faixa etária. A tabela 1 contém as faixas etárias e as proporções de sobreviventes. A figura 5 contém o gráfico de dispersão da proporção de sobreviventes versus a faixa etária. Pode-se ver que a proporção de sobrevivência **tende** a diminuir com o aumento da faixa etária. Isso indica que há associação entre a sobrevivência do paciente e a idade do paciente. Nota-se que a proporção não pode aumentar ou diminuir indefinidamente, sendo, no máximo 1, e podendo diminuir com a idade até o mínimo 0.

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 3

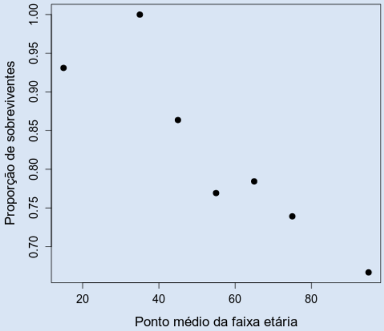
Tabela 1: Proporções de sobreviventes por faixa etária

Faixa etária	Proporção de sobreviventes
0-29	0,93
30-39	1
40-49	0,86
50-59	0,77
60-69	0,78
70-79	0,74
80-109	0,67

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 3

Figura 5: Proporção de sobreviventes vs Faixa etária.



Modelagem de regressão e o modelo linear normal

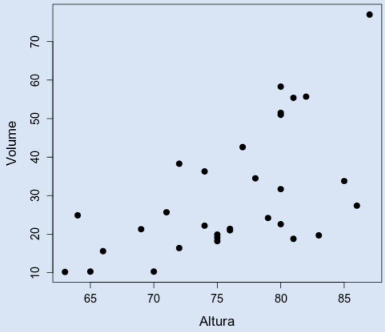
Exemplo 4

O conjunto de dados **trees.dat**, na pasta **Dados**, contém informações sobre 31 árvores cerejeiras, situadas no estado da Pênsilvania, EUA [7, p. 111]. Para cada árvore, registrou-se os valores do diâmetro, da altura e do volume. Esses dados foram coletados com objetivo de investigar a relação do volume com a altura e o diâmetro. As figuras 6 e 7 contém os gráficos de dispersão do volume vs a altura e do volume vs o diâmetro. Os gráficos revelam que, quando o diâmetro ou a altura da árvore aumentam, o seu volume **tende** a aumentar, o que indica uma associação do volume com a altura e com o diâmetro.

Modelagem de regressão e o modelo linear normal

Continuação do exemplo 4

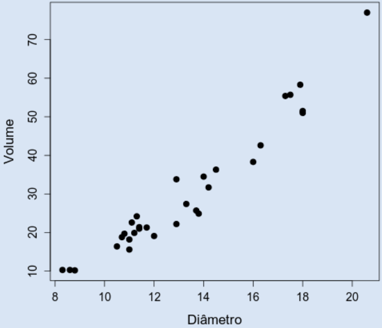
Figura 6: Volume vs altura.



Modelagem de regressão e o modelo linear normal

Continuação do exemplo 4

Figura 7: Volume vs diâmetro.



Modelagem de regressão e o modelo linear normal

- A associação entre variáveis pode ser utilizada de diversas formas e uma delas é a **predição**.
- O conhecimento sobre a forma como duas ou mais variáveis são associadas permite desenvolver métodos para **predizer** o valor de uma das variáveis a partir dos valores das demais.

Exemplo 5

No exemplo 3, viu-se que há uma associação entre a probabilidade de sobrevivência de um paciente internado numa UTI e sua idade. Essa informação pode ser usada para desenvolver procedimentos que permitam predizer a probabilidade de sobrevivência do paciente a partir da sua idade.

Modelagem de regressão e o modelo linear normal

- A associação entre variáveis também pode ser utilizada para **explicar** como uma das variáveis varia em termos das demais.

Exemplo 6

No exemplo 1, viu-se que, quando há um aumento na porcentagem de área com floresta no entorno de uma bacia, o nível de nitrogênio nas águas dessa bacia tende a diminuir. Essa informação pode ser usada para desenvolver procedimentos que permitam explicar como se dá a redução no nível de nitrogênio com o aumento na porcentagem de área com floresta. Por exemplo, pode-se descrever a redução no nível de nitrogênio que se obtém para cada 1% a mais de área com floresta.

Modelagem de regressão e o modelo linear normal

- A capacidade de predizer e/ou de explicar levam à capacidade de **controlar** o valor de uma variável, controlando os valores das variáveis associadas a ela.

Exemplo 7

A associação entre o nível de nitrogênio nas águas de uma bacia e a porcentagem de floresta no entorno dessa bacia, no exemplo 1, pode ser aplicada no desenvolvimento de métodos que permitam controlar o nível de nitrogênio, controlando-se a porcentagem de floresta. A partir de tais métodos, pode-se obter a porcentagem de floreta necessária para atingir um nível aceitável de nitrogênio e direcionar políticas de manejo floresta com base nessa informação.

Modelagem de regressão e o modelo linear normal

- Várias são as razões para se querer predizer uma variável a partir de outras variáveis associadas a ela [4, p. 80].
  - Quando **custa caro obter o valor de uma variável**, mas a obtenção dos valores de outras variáveis associadas a ela é relativamente barata. Nesse caso, os valores “baratos” são utilizados para predizer os valores “caros”. Na fazenda onde se cultiva as árvores mencionadas no exemplo 4, deseja-se saber se o volume de uma árvore é grande o suficiente para que ela seja cortada e sua madeira seja vendida. A princípio, para se medir o volume da árvore, seria necessário cortá-la. Porém, isso poderia custar caro se, após a medição do volume, fosse constatado que a árvore não estava pronta para ser cortada. A associação do volume com a altura e o diâmetro pode ser utilizada para predizer o volume a partir da altura e do diâmetro, evitando assim que a árvore seja cortada fora de tempo.

Modelagem de regressão e o modelo linear normal

- Quando **é impossível medir o valor de uma variável, visto que tal valor é o resultado da ocorrência de eventos futuros**, porém é desejável conhecer o valor da variável no presente, como auxílio num processo de tomada de decisão. No exemplo 3, o conhecimento da probabilidade de sobrevivência de um paciente a partir da idade do paciente pode ajudar na tomada de decisão a respeito da alocação de pacientes em leitos de UTI, em situações de extrema emergência nas quais tal decisão é necessária.
- Quando **se deseja apenas obter a relação de uma variável com outras**. Nesses casos, o foco não está, necessariamente, na predição. Provavelmente, é barato e fácil de se medir o nível de nitrogênio nas águas das bacias hidrográficas mencionadas no exemplo 1. Entretanto, talvez o objetivo não seja predizer o nível de nitrogênio, e sim controlá-lo, conforme mencionado no exemplo 7

Modelagem de regressão e o modelo linear normal

- A **modelagem de regressão** consiste na aplicação de métodos estatísticos para 1) verificar a existência de associação entre duas ou mais variáveis e 2) usar essa associação para representar uma das variáveis, denominada **variável resposta**, como função das demais, denominadas **variáveis preditoras**.
- A representação da variável resposta como uma função das variáveis preditoras é feita por modelos matemáticos conhecidos como **modelos de regressão**.
- A modelagem de regressão é mais comumente utilizada quando se deseja obter o valor **desconhecido** de uma variável resposta a partir dos valores **conhecidos** das variáveis preditoras.
- Portanto, nesse curso, **os modelos de regressão assumem que os valores das variáveis preditoras são conhecidos**.

Modelagem de regressão e o modelo linear normal

- Um modelo de regressão é composto por equações, as quais descrevem a relação funcional entre a resposta e as variáveis preditoras, e, possivelmente, um conjunto de suposições sobre os termos dessas equações.
- Um dos modelos de regressão mais populares é o **modelo de regressão linear normal**.
- Seja Y a variável resposta e sejam $x_1, \cdots, x_p$ os valores das variáveis preditoras.
- O modelo de regressão linear normal assume que

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

onde β_0, $\beta_1, \cdots$, β_{p-1} e β_p são constantes reais e ϵ é uma variável aleatória tal que $\epsilon \sim N(0, \sigma^2)$.

Modelagem de regressão e o modelo linear normal

- Os coeficientes $\beta_0, \cdots, \beta_p$ e a variância σ^2 são os **parâmetros** do modelo.
- Em geral, os parâmetros do modelo são **desconhecidos** e devem ser **estimados** com base numa amostra de $x_1, \cdots, x_p$ e Y.
- A variável aleatória ϵ é denominada **erro aleatório**.
- O modelo linear normal possui as seguintes propriedades:
  1. $E[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
  2. $Var[Y] = \sigma^2$;
  3. Y segue uma distribuição normal.
- A propriedade 1 diz que **a média de Y varia linearmente com cada uma das variáveis preditoras**.
- A propriedade 2 diz que a variância de Y é constante. Isto é, a variável resposta é **homoscedástica**.

Modelagem de regressão e o modelo linear normal

- Logo, **o modelo linear normal assume homoscedasticidade da variável resposta**
- O oposto de homoscedasticidade é **heteroscedasticidade**.
- Quando a variância da variável resposta não é constante, a variável resposta é dita ser **heteroscedástica**.
- Essas três propriedades permitem concluir que Y é uma variável aleatória tal que

$$Y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2).$$

- Ou seja, **o modelo linear normal assume que a variável resposta é uma variável aleatória que segue distribuição normal com média** $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ **e variância** σ^2.

Modelagem de regressão e o modelo linear normal

- Na modelagem de regressão, o principal objetivo é obter um modelo que descreva, de forma "satisfatória", a variação da resposta em termos da variação das variáveis preditoras.
- O processo de obtenção de um modelo de regressão é constituído por quatro etapas básicas [3, p. 21]:
  1. especificação do modelo;
  2. estimação dos parâmetros do modelo;
  3. verificação da adequacidade do modelo;
  4. aplicação do modelo.
- Na etapa 1, apresenta-se as variáveis resposta e preditoras, as equações que compõem o modelo e as suposições adotadas.
- Na etapa 2, os parâmetros são estimados com base numa amostra de das variáveis resposta e preditoras.

Modelagem de regressão e o modelo linear normal

- Na etapa 3, procura-se analisar a qualidade das predições da resposta fornecidas pelo modelo, bem como avaliar, por meio de técnicas de diagnóstico, possíveis violações nas suposições adotadas e a existência de observações não explicadas pelo modelo, isto é, observações atípicas (*outliers*).
- Uma vez que o modelo tenha sido validado na etapa 3, ele é aplicado na etapa 4. As aplicações incluem a realização de inferência sobre as associações entre a variável resposta e as variáveis preditoras, a predição de valores desconhecidos da resposta e a descrição do comportamento da resposta em termos do comportamento das preditoras.
- Espera-se, ao final desse processo, obter um modelo capaz de reproduzir o comportamento observado da variável resposta.

Modelagem de regressão e o modelo linear normal

- Conforme mencionado acima, é na etapa 3 que se busca por violações nas suposições do modelo de regressão.
- No caso do modelo linear normal, procura-se avaliar se a variável resposta 1) apresenta uma relação linear com cada uma das variáveis preditoras; 2) é homoscedástica e 3) segue distribuição normal, fixados os valores das variáveis preditoras.
- A não confirmação de algum desses itens pode significar que o modelo linear normal não é adequado aos dados em mãos.
- Situações como essa podem ser contornadas aplicando-se uma transformação à variável resposta e/ou às variáveis preditoras.
- A aplicação das transformações de Box-Cox à resposta costuma linearizar a relação entre a resposta e as preditoras, bem como tornar a resposta homoscedástica e normalmente distribuída.

Modelagem de regressão e o modelo linear normal

- Há uma outra maneira, conceitualmente mais correta, de lidar com a violação nas suposições do modelo linear normal.
- O modelo linear normal, o qual estabelece a relação entre a variável resposta Y e as variáveis preditoras $x_1, \cdots, x_p$, admite a seguinte formulação:
  1. $Y \sim N(\mu, \sigma^2)$;
  2. $\eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
  3. $\mu = \eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p)$.
- Dessa forma, o modelo linear normal pode ser visto como um modelo constituído por três componentes: 1) uma distribuição de probabilidade para a resposta Y; 2) uma função η linear nas preditoras e 3) uma relação funcional entre a média de Y e η.
- Diferentes modelos, mais adequados aos dados em mãos, podem ser obtidos apenas modificando-se esses componentes.

Regressão Poisson

- Na seção anterior, viu-se que um modelo de regressão pode ser formulado especificando-se três termos: 1) uma distribuição de probabilidade para a resposta Y e 2) uma função η linear nas preditoras e 3) uma relação funcional entre a média de Y e η.
- A distribuição da variável resposta é, em geral, determinada pela natureza da própria variável resposta.
- No exemplo 2, a variável resposta Y é o número de clientes de uma loja, que residem numa determinada área.
- Nesse caso, Y é uma variável quantitativa discreta, a qual pode assume valores no conjunto $\{0, 1, 2, 3, \cdots\}$.
- Uma distribuição de probabilidade, que é adequada a variáveis aleatórias com essas características, é a distribuição de Poisson.

Regressão Poisson

- Se Y é uma variável aleatória discreta, a qual se distribui de acordo com uma distribuição de Poisson com parâmetro μ, então a função de probabilidade Y é dada por

$$P(Y=y)=\frac{e^{-\mu}\mu^{y}}{y!}, \quad y \in \{0,1,2,3,\cdots\}.$$

- Além disso, a média e a variância de Y são iguais a μ.
- Portanto, um modelo de regressão pode ser formulado como
  1. $Y \sim \text{Poisson}(\mu)$;
  2. $\eta(x_1,\cdots,x_p;\beta_0,\cdots,\beta_p)=\beta_0+\beta_1x_1+\cdots+\beta_px_p$;
  3. $\mu=\eta(x_1,\cdots,x_p;\beta_0,\cdots,\beta_p)$.
- Um problema com esse modelo é que a função η pode assumir valores negativos. Entretanto, a média μ de Y é positiva.

Regressão Poisson

- Pode-se resolver esse problema considerando-se um modelo no qual uma transformação $g : [0, \infty) \to \mathbb{R}$ é aplicada à μ.
- A transformação mais comumente empregada, nesses casos, é a transformação logarítmica $g(\mu) = \ln(\mu)$.
- Um modelo de regressão pode, então, ser formulado como
  1. $Y \sim \text{Poisson}(\mu)$;
  2. $\eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
  3. $\ln(\mu) = \eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p)$.
- Esse modelo é conhecido como **modelo de regressão Poisson**.
- De acordo com o modelo de regressão Poisson, a média da variável resposta é

$$\mu = e^{\eta} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p},$$

garantindo que $\mu > 0$.

Regressão Poisson

- Essa relação funcional entre a média de Y e as preditoras permite concluir que o modelo de regressão Poisson postula uma associação não linear entre a resposta e as preditoras.
- Além disso, como $Var(Y) = \mu$, tem-se que

$$Var(Y) = e^{\eta} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}.$$

- Ou seja, de acordo com o modelo de regressão Poisson, a variável resposta é heteroscedástica.
- Dessa forma, o modelo de regressão Poisson é um modelo que pode ser útil em situações nas quais 1) a variável resposta é uma contagem; 2) há uma associação não linear entre a resposta e as preditoras; 3) a resposta é heteroscedástica; 4) a distribuição da resposta não pode ser considerada uma normal.

Regressão logística

- No exemplo 3, a resposta Y é uma variável binária que indica se o paciente sobrevive ($Y = 1$) ou não ($Y = 0$) ao período de internação numa UTI.
- Por conta de sua natureza dicotômica, uma distribuição adequada à variável resposta Y é a distribuição de Bernoulli.
- Se Y é uma variável aleatória discreta, distribuída de acordo com uma distribuição de Bernoulli com parâmetro μ, então a função de probabilidade de Y é dada por

$$P(Y = y) = \begin{cases} 1 - \mu & \text{se } y = 0; \\ \mu & \text{se } y = 1. \end{cases}$$

- O parâmetro μ é denominado **probabilidade de sucesso**.

Regressão logística

- Nesse exemplo, o parâmetro μ é interpretado como sendo a probabilidade sobrevivência do paciente.
- Sabe-se que $E(Y) = \mu$ e $Var(Y) = \mu(1-\mu)$.
- Portanto, um modelo de regressão pode ser formulado como
    1. $Y \sim \text{Bernoulli}(\mu)$;
    2. $\eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
    3. $\mu = \eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p)$.
- Um problema com esse modelo é que a função η pode assumir quaisquer valores reais, enquanto μ, por ser uma probabilidade, assume valores apenas no intervalo $[0, 1]$.
- Pode-se resolver esse problema considerando-se um modelo no qual uma transformação $g : [0, 1] \to \mathbb{R}$ é aplicada à μ.

Regressão logística

- A transformação mais comumente empregada, nesses casos, é a transformação **logit**:
$$g(\mu) = \text{logit}(\mu) = \ln\left(\frac{\mu}{1-\mu}\right).$$
- Um modelo de regressão pode, então, ser formulado como
  1. $Y \sim \text{Bernoulli}(\mu)$;
  2. $\eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
  3. $\text{logit}(\mu) = \eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p)$.
- Esse modelo é chamado **modelo de regressão logística** [5].
- De acordo com o modelo de regressão logística, a probabilidade de sucesso é dada por
$$\mu = \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}},$$
garantindo que $0 \leq \mu \leq 1$.

Regressão logística

- Portanto, o modelo de regressão logística postula uma associação não linear entre probabilidade de sucesso μ e as preditoras.
- Além disso, como $Var(Y) = \mu(1-\mu)$, tem-se que

$$Var(Y) = \frac{e^{\eta}}{(1+e^{\eta})^2} \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{(1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p})^2}$$

- Ou seja, de acordo com o modelo de regressão logística, a variável resposta é heteroscedástica.
- Dessa forma, o modelo de regressão logística é um modelo que pode ser útil em situações nas quais 1) a variável resposta é binária; 2) há uma associação não linear entre a probabilidade de sucesso e as preditoras.

Modelos lineares generalizados

- Conforme visto nas seções anteriores, os modelos de regressão linear normal, Poisson e logística possuem a mesma estrutura.
- Essa estrutura é constituída por três componentes:
  1. uma distribuição de probabilidade para a variável resposta Y;
  2. uma função η linear nas variáveis preditoras;
  3. uma relação funcional entre a média μ de Y e η.
- Esses três modelos compartilham uma outra característica: as distribuições da resposta são membros da **família exponencial**.
- A distribuição de uma variável aleatória Y é membro da família exponencial se sua função de densidade de probabilidade, ou função de probabilidade, pode ser escrita como

$$f(y;\theta,\phi) = \exp\left\{\phi^{-1}\left[\theta y - b(\theta)\right] + c(y,\phi)\right\},$$

onde θ e ϕ são parâmetros e $b(\cdot)$ e $c(\cdot,\cdot)$ são funções conhecidas.

Modelos lineares generalizados

- Além dessas três distribuições, várias outras, tais como a normal inversa, a gama, a binomial e a binomial negativa, pertencem à família exponencial.
- Portanto, uma forma de generalizar os modelos de regressão linear normal, Poisson e logística é considerar que a distribuição da variável resposta é membro da família exponencial.
- Os **modelos lineares generalizados** (MLGs), para uma variável resposta Y e variáveis preditoras $x_1, \cdots, x_p$, assumem que Y é uma variável aleatória tal que
  1. a distribuição de Y pertence à família exponencial;
  2. $\eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$,
  3. $g(\mu) = \eta(x_1, \cdots, x_p; \beta_0, \cdots, \beta_p)$;

  onde μ é a média de Y e g é uma função monótona diferenciável.

Modelos lineares generalizados

- Os MLGs, propostos por Nelder e Wedderburn [6], estendem os modelos de regressão linear, Poisson e logística de duas formas.
- Por um lado, a distribuição da variável resposta deixa de ser uma distribuição específica e passa a ser qualquer membro da família exponencial.
- Por outro, a transformação aplicada à média também deixa de ser uma transformação específica e passa a ser qualquer função monótona e diferenciável g.
- Essas duas generalizações tornam bastante flexíveis os MLGs.
- Além da flexibilidade, os MLGs permitem que teorias e métodos, envolvidos nos processos de estimação, inferência e diagnóstico para diversos modelos de regressão, sejam desenvolvidos de forma unificada.

Modelos lineares generalizados

- Os MLGs são formulados especificando-se três componentes.
- O primeiro componente é a distribuição de probabilidade da variável resposta. Esse é o **componente aleatório**.
- O segundo componente é a função η, a qual é linear nas variáveis preditoras. Esse é o **componente sistemático**.
- O terceiro componente é a **função de ligação** g.
- A função η é denominada **preditor linear**.
- Os MLGs estabelecem uma relação funcional entre a resposta e as variáveis preditoras assumindo que há uma relação funcional entre a média da resposta e as variáveis preditoras.
- A função de ligação lineariza a relação funcional entre a média e as variáveis preditoras.

Modelos lineares generalizados

- Em geral, a função de ligação é escolhida de forma a garantir que $g^{-1}(\eta)$ seja um valor possível para μ.
- Pode-se ver que a família exponencial de distribuições ocupa posição central na formulação dos MLGs.
- O principal objetivo do presente tópico é apresentar a família exponencial de distribuições, algumas de suas propriedades e exemplos de distribuições que pertencem a essa família.
- Isso é feito na seção a seguir.

Definição e propriedades

- Seja Y uma variável aleatória contínua (discreta) com função de densidade (função de probabilidade) f.
- A distribuição de probabilidade de Y é um membro da **família exponencial** de distribuições se f pode ser escrita como
$$f(y;\theta,\phi) = \exp\left\{\phi^{-1}\left[\theta y - b(\theta)\right] + c(y,\phi)\right\}, \qquad (1)$$
onde θ e ϕ são parâmetros e $b(\cdot)$ e $c(\cdot,\cdot)$ são funções conhecidas.
- θ é o **parâmetro canônico** e ϕ é o **parâmetro de dispersão**.
- Várias distribuições são membros da família exponencial.

Exemplo 8

Se $Y \sim \text{Bernoulli}(\mu)$, então
$$f(y;\mu) = P(Y=y) = \begin{cases} 1-\mu & \text{se } y = 0; \\ \mu & \text{se } y = 1. \end{cases}$$

Definição e propriedades

Continuação do exemplo 8

A função de probabilidade de Y pode ser escrita como

$$f(y;\mu) = \mu^y(1-\mu)^{1-y}.$$

Daí, obtém-se

$$\begin{aligned} f(y;\mu) &= \exp\{\ln[f(y;\mu)]\} = \exp\left\{\ln[\mu^y(1-\mu)^{1-y}]\right\} = \\ &= \exp\{y\ln(\mu) + (1-y)\ln(1-\mu)\} = \\ &= \exp\{y\ln(\mu) - y\ln(1-\mu) + \ln(1-\mu)\} = \\ &= \exp\left\{y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu)\right\}. \end{aligned}$$

Fazendo $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$, obtém-se $\mu = \frac{e^\theta}{1+e^\theta}$.

Definição e propriedades

Continuação do exemplo 8

Logo, $\ln(1-\mu) = \ln\left(1 - \frac{e^\theta}{1+e^\theta}\right) = \ln\left(\frac{1}{1+e^\theta}\right) = -\ln\left(1+e^\theta\right)$.

Segue-se que $f(y;\mu)$ pode ser escrita como

$$
\begin{aligned}
f(y;\mu) &= \exp\left\{y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu)\right\} = \\
&= \exp\left\{y\theta - \ln\left(1+e^\theta\right)\right\} = \\
&= \exp\left\{1\cdot\left[y\theta - \ln\left(1+e^\theta\right)\right] + 0\right\} = \\
&= \exp\left\{\phi^{-1}[y\theta - b(\theta)] + c(y;\phi)\right\} = f(y;\theta,\phi),
\end{aligned}
$$

onde $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$, $b(\theta) = \ln\left(1+e^\theta\right)$, $\phi = 1$ e $c(y;\phi) = 0$.

Definição e propriedades

Exemplo 9

Se $Y \sim \text{Poisson}(\mu)$, então

$$f(y;\mu) = P(Y=y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y \in \{0,1,2,3,\cdots\}.$$

Logo,

$$\begin{aligned} f(y;\mu) &= \exp\{\ln[f(y;\mu)]\} = \exp\left\{\ln\left[\frac{\mu^y e^{-\mu}}{y!}\right]\right\} = \\ &= \exp\left\{\ln(\mu^y) + \ln(e^{-\mu}) - \ln(y!)\right\} = \\ &= \exp\{y\ln(\mu) - \mu - \ln(y!)\}. \end{aligned}$$

Fazendo $\theta = \ln(\mu)$, obtém-se $\mu = e^{\theta}$.

Definição e propriedades

Continuação do exemplo 9

Segue-se que $f(y;\mu)$ pode ser escrita como

$$
\begin{aligned}
f(y;\mu) &= \exp\{y\ln(\mu) - \mu - \ln(y!)\} = \\
&= \exp\left\{y\theta - e^{\theta} - \ln(y!)\right\} = \\
&= \exp\left\{1 \cdot \left[y\theta - e^{\theta}\right] - \ln(y!)\right\} = \\
&= \exp\left\{\phi^{-1} \cdot [y\theta - b(\theta)] + c(y,\phi)\right\} = f(y;\theta,\phi),
\end{aligned}
$$

onde $\theta = \ln(\mu)$, $b(\theta) = e^{\theta}$, $\phi = 1$ e $c(y,\phi) = -\ln(y!)$.

Definição e propriedades

Exemplo 10

Se $Y \sim N(\mu, \sigma^2)$, então

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad y \in \mathbb{R}.$$

Logo,

$$\begin{aligned} f(y; \mu, \sigma^2) &= \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2} - \ln\left(\sqrt{2\pi\sigma^2}\right)\right\} = \\ &= \exp\left\{\frac{-(y^2 - 2y\mu + \mu^2)}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right\} = \\ &= \exp\left\{\frac{1}{\sigma^2} y\mu - \frac{1}{\sigma^2}\frac{\mu^2}{2} - \frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right\}. \end{aligned}$$

Definição e propriedades

Continuação do exemplo 10

Fazendo $\theta = \mu$, obtém-se

$$\begin{aligned} f(y;\mu,\sigma^2) &= \exp\left\{ \frac{1}{\sigma^2} y\mu - \frac{1}{\sigma^2}\frac{\mu^2}{2} - \frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \right\} = \\ &= \exp\left\{ \frac{1}{\sigma^2} y\theta - \frac{1}{\sigma^2}\frac{\theta^2}{2} - \frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \right\} = \\ &= \exp\left\{ \frac{1}{\sigma^2}\left[y\theta - \frac{\theta^2}{2} \right] - \frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \right\} = \\ &= \exp\left\{ \phi^{-1}\left[y\theta - b(\theta) \right] + c(y,\phi) \right\} \end{aligned}$$

onde $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^2$ e $c(y,\phi) = -\frac{1}{2}\frac{y^2}{\phi} - \frac{1}{2}\ln\left(2\pi\phi\right)$.

Definição e propriedades

- Os métodos de estimação e inferência para a família exponencial são baseados na teoria da verossimilhança.
- Para que os resultados teoria da verossimilhança possam ser aplicados, é preciso assumir que a família exponencial satisfaz certas **condições de regularidade** [2, p. 97].
- Uma dessas condições diz que o suporte da família exponencial não depende dos parâmetros.
- Essa condição garante que, quando aplicadas a $f(y;\theta,\phi)$, as operações de diferenciação com relação a θ e integração com relação a y são permutáveis [2, p. 10]
- Isso significa, por exemplo, que a seguinte igualdade é válida:

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(y;\theta,\phi)dy = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta} f(y;\theta,\phi)dy. \tag{2}$$

Definição e propriedades

- Com base na igualdade (2), pode-se mostrar que

$$E(Y) = \frac{\partial}{\partial\theta} b(\theta) = b'(\theta).$$

- De fato, como f é uma função de densidade de probabilidade, pode-se concluir que $\int_{-\infty}^{\infty} f(y;\theta,\phi)dy = 1$ é constante.
- Portanto,

$$\frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f(y;\theta,\phi)dy = 0. \tag{3}$$

- Substituindo (3) na igualdade (2), obtém-se

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(y;\theta,\phi)dy = 0. \tag{4}$$

Definição e propriedades

- A derivada de f com relação a θ é dada por

$$\frac{\partial}{\partial\theta}f(y;\theta,\phi) = f(y;\theta,\phi)\phi^{-1}[y - b'(\theta)]. \qquad (5)$$

Substituindo (5) em (4), obtém-se

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(y;\theta,\phi)dy = \int_{-\infty}^{\infty}f(y;\theta,\phi)\phi^{-1}[y - b'(\theta)]dy =$$
$$= \phi^{-1}\int_{-\infty}^{\infty}yf(y;\theta,\phi)dy - \phi^{-1}b'(\theta)\int_{-\infty}^{\infty}f(y;\theta,\phi)dy =$$
$$= \phi^{-1}E(Y) - \phi^{-1}b'(\theta) = 0,$$

de onde se conclui que $E(Y) = b'(\theta)$.

Definição e propriedades

- Aplicando um raciocínio análogo, pode-se mostrar que

$$Var(Y) = \phi b''(\theta).$$

- Portanto, **se Y é uma variável aleatória, cuja distribuição de probabilidade pertence à família exponencial (1), então a média e a variância de Y são, respectivamente**,

$$E(Y) = b'(\theta) \tag{6}$$

e

$$Var(Y) = \phi b''(\theta). \tag{7}$$

Definição e propriedades

Exemplo 11

Conforme visto no exemplo 8, se $Y \sim \text{Bernoulli}(\mu)$, então

$$\theta = \ln\left(\frac{\mu}{1-\mu}\right), \quad b(\theta) = \ln\left(1+e^{\theta}\right) \quad \text{e} \quad \phi = 1$$

Logo,

$$E(Y) = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \frac{\mu/(1-\mu)}{1+\mu/(1-\mu)} = \mu$$

e

$$Var(Y) = \phi b''(\theta) = \frac{e^{\theta}(1+e^{\theta}) - \left(e^{\theta}\right)^2}{(1+e^{\theta})^2} = \frac{e^{\theta}}{(1+e^{\theta})^2} =$$

$$= \frac{\mu/(1-\mu)}{1/(1-\mu)^2} = \mu(1-\mu).$$

Definição e propriedades

Exemplo 12

Conforme visto no exemplo 9, se $Y \sim \text{Poisson}(\mu)$, então

$$\theta = \ln(\mu), \quad b(\theta) = e^{\theta} \quad \text{e} \quad \phi = 1$$

Logo,

$$E(Y) = b'(\theta) = e^{\theta} = \mu \quad \text{e} \quad Var(Y) = \phi b''(\theta) = e^{\theta} = \mu.$$

Conforme visto no exemplo 10, se $Y \sim N(\mu, \sigma^2)$, então

$$\theta = \mu, \quad b(\theta) = \frac{\theta^2}{2} \quad \text{e} \quad \phi = \sigma^2$$

Logo,

$$E(Y) = b'(\theta) = \theta = \mu \quad \text{e} \quad Var(Y) = \phi b''(\theta) = \sigma^2.$$

Outros casos particulares

- Nessa seção, são apresentados exemplos de outros membros da família exponencial.

Exemplo 13

Se $Y \sim \text{binomial}(m, \pi)$, **com** m **conhecido**, então

$$f(y; \pi, m) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \;\; y \in \{0, 1, 2, \cdots, m\}.$$

Logo,

$$\begin{aligned} f(y; \pi, m) &= \exp\left\{ \ln\left[\binom{m}{y} \pi^y (1-\pi)^{m-y} \right] \right\} = \\ &= \exp\left\{ y \ln\left(\frac{\pi}{1-\pi} \right) + m \ln(1-\pi) + \ln\binom{m}{y} \right\}. \end{aligned}$$

Outros casos particulares

Continuação do exemplo 13

Fazendo $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$, obtém-se

$$\pi = \frac{e^{\theta}}{1+e^{\theta}} \quad \text{e} \quad \ln(1-\pi) = -\ln\left(1+e^{\theta}\right).$$

Logo, $f(y;\pi,m)$ pode ser escrita como

$$f(y;\pi,m) = \exp\left\{y\theta - m\ln\left(1+e^{\theta}\right) + \ln\binom{m}{y}\right\} =$$
$$= \exp\left\{\phi^{-1}[y\theta - b(\theta)] + c(y,\phi)\right\} = f(y;\theta,\phi),$$

onde $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = m\ln\left(1+e^{\theta}\right)$, $\phi = 1$ e $c(y;\phi) = \ln\binom{m}{y}$.

Outros casos particulares

Continuação do exemplo 13

Pelas equações (6) e (7), tem-se que

$$E(Y) = b'(\theta) = m \cdot \frac{e^{\theta}}{1 + e^{\theta}} = m \cdot \frac{\pi/(1-\pi)}{1 + \pi/(1-\pi)} = m\pi$$

e

$$Var(Y) = \phi b''(\theta) = m \cdot \frac{e^{\theta}(1 + e^{\theta}) - \left(e^{\theta}\right)^2}{\left(1 + e^{\theta}\right)^2} =$$

$$= m \cdot \frac{e^{\theta}}{\left(1 + e^{\theta}\right)^2} = m \cdot \frac{\pi/(1-\pi)}{1/(1-\mu)^2} = m\pi(1-\pi).$$

Outros casos particulares

Exemplo 14

Se $Y \sim \text{Gama}(a, b)$, então

$$f(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}, \quad y > 0.$$

Logo,

$$\begin{aligned} f(y; a, b) &= \exp\left\{\ln\left[\frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}\right]\right\} = \\ &= \exp\left\{\ln\left(b^a\right) - \ln\Gamma(a) + (a-1)\ln(y) - by\right\} = \\ &= \exp\left\{-by + a\ln(b) - \ln\Gamma(a) + (a-1)\ln(y)\right\} = \\ &= \exp\left\{a\left[-\frac{b}{a}y + \ln(b)\right] - \ln\Gamma(a) + (a-1)\ln(y)\right\} \end{aligned}$$

Outros casos particulares

Continuação do exemplo 14

Fazendo $\theta = -\frac{b}{a}$, obtém-se $b = -a\theta$. Logo,

$$\begin{aligned}
f(y; a, b) &= \exp\left\{a\left[-\frac{b}{a}y + \ln(b)\right] - \ln\Gamma(a) + (a-1)\ln(y)\right\} = \\
&= \exp\{a[\theta y + \ln(-a\theta)] - \ln\Gamma(a) + (a-1)\ln(y)\} = \\
&= \exp\{a[\theta y + \ln(-\theta)] + a\ln(a) - \\
&\qquad - \ln\Gamma(a) + (a-1)\ln(y)\} = \\
&= \exp\left\{\phi^{-1}[\theta y - b(\theta)] + c(y,\phi)\right\} = f(y; \theta, \phi),
\end{aligned}$$

onde $\theta = -\frac{b}{a}$, $b(\theta) = -\ln(-\theta)$, $\phi = a^{-1}$ e

$$c(y,\phi) = -\phi\ln(\phi) - \ln\Gamma(\phi) + (\phi - 1)\ln(y).$$

Outros casos particulares

Continuação do exemplo 14

Pelas equações (6) e (7), tem-se que

$$E(Y) = b'(\theta) = -\frac{(-1)}{(-\theta)} = -\frac{1}{\theta} = \frac{a}{b}$$

e

$$Var(Y) = \phi b''(\theta) = a^{-1} \cdot \frac{1}{\theta^2} = \frac{1}{a} \cdot \frac{a^2}{b^2} = \frac{a}{b^2}.$$

Referências I

[1] S. Chatterjee and A. S. Hadi, *Regression analysis by example*, 4 ed., John Wiley & Sons, Hoboken, NJ, EUA, 2006.

[2] G. M. Cordeiro, *Introdução à teoria assintótica*, 1999, disponível em https://www.ime.usp.br/~abe/lista/pdftCtIOIA62A.pdf.

[3] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4 ed., CRC Press, Boca Raton, FL, EUA, 2018.

[4] F. A. Graybill and H. K. Iyer, *Regression analysis: Concepts and applications*, Duxbury Press, Belmont, CA, EUA, 1994.

Referências II

[5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3 ed., John Wiley & Sons, Hoboken, NJ, EUA, 2013.

[6] J. A. Nelder and R. W. M. Wedderburn, *Generalized linear models*, Journal of the Royal Statistical Society. Series A (General) **135** (1972), no. 3, 370–384.

[7] G. A. Paula, *Modelos lineares generalizados com apoio computacional*, 2013, disponível em https://www.ime.usp.br/~giapaula/texto_2013.pdf.