UNIVERSIDADE FEDERAL DE SERGIPE
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

Aula 12: testando a significância global do modelo

Sumário

Metas

1. Apresentar o teste da razão de verossimilhanças para a avaliação da significância global do modelo linear generalizado.

Objetivos

- Após estudar essa aula, o aluno ou aluna será capaz de:
  1. desenvolver o teste da razão de verossimilhanças para avaliar a significância global do modelo linear generalizado;
  2. desenvolver esse teste para os principais casos particulares do modelo linear generalizado;
  3. interpretar os resultados desse teste;
  4. executar esse teste utilizando a plataforma computational **R**.

Pré-requisitos

1. Unidade 1.

Introdução

- Na aula 2, viu-se que o processo de modelagem de regressão é constituído por quatro etapas básicas: 1) formulação do modelo; 2) estimação dos parâmetros do modelo; 3) verificação da adequacidade do modelo; 4) aplicação do modelo.
- As etapas 1 e 2 desse processo foram abordadas na unidade 1.
- A unidade 2 trata da etapa 4, a aplicação do modelo.
- Conforme mencionado na aula 2, dentre as possíveis aplicações do modelo, encontram-se a realização de inferência sobre as associações entre a variável resposta e as variáveis preditoras.
- Uma maneira de realizar inferência sobre as associações entre a resposta e as preditoras é avaliar a significância global dos coeficientes de regressão do modelo.

Introdução

- O MLG, que relaciona a resposta Y às preditoras $x_1, \cdots, x_p$, é formulado especificando-se
    - uma distribuição na família exponencial para Y;
    - um preditor linear $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
    - e uma função de ligação $g(\mu) = \eta$,

  onde μ é a média da variável resposta Y.
- O teste de significância global do MLG corresponde ao teste de hipóteses cujas hipóteses nula e alternativa são

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0; \\ H_1 &: \beta_j \neq 0, \text{ para ao menos um } j \in \{1, 2, \cdots, p\}. \end{aligned}$$

- Assumindo que H_0 é verdadeira, obtém-se

$$\eta = \beta_0 \quad \text{e} \quad \mu = g^{-1}(\beta_0).$$

Introdução

- Isto é, a média da variável resposta é constante com relação às variáveis preditoras.
- Portanto, sob a hipótese H_0, a média da variável resposta não depende das variáveis preditoras.
- No caso do MLG, significa dizer que não há associação entre a variável resposta e as variáveis preditoras.
- A hipótese H_1 corresponde à negação de H_0.
- O seja, sob H_1, existe uma associação entre a variável resposta e ao menos uma das variáveis preditoras.
- Dessa forma, o teste de significância global procura avaliar se é plausível assumir que não há associação entre a variável resposta e qualquer uma das variáveis preditoras.

Introdução

- Os principais testes de significância global, no contexto dos MLG's, são o teste da razão de verossimilhanças, o teste de Wald e o teste de escore [1, p. 116].
- Há também o teste F, o qual não depende do parâmetro de dipsersão ϕ, sendo, portanto, útil em situações nas quais ϕ é desconhecido [2, p. 30 e 79].
- Na presente aula, o teste da razão de verossimilhanças para avaliação da significância global do MLG é desenvolvido.
- Versões desse teste são apresentadas para alguns dos principais casos particulares do MLG.
- Por fim, mostra-se como o teste pode ser executado utilizando a plataforma **R**.

Tópico 24: teste da razão de verossimilhanças

- Seja Y a variável resposta e sejam $x_1, x_2, \cdots, x_p$ as variáveis preditoras.
- Supõe-se que Y e as variáveis preditoras possuem uma relação governada por um MLG.
- Isto é, Y é uma variável aleatória tal que
  1. a distribuição de Y pertence à família exponencial;
  2. $\eta(\boldsymbol{x};\boldsymbol{\beta}) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$;
  3. $g(\mu) = \eta(\boldsymbol{x};\boldsymbol{\beta})$;

  onde μ é a média de Y, g é a função de ligação,
  $$\boldsymbol{x} = (1, x_1, \cdots, x_p)^T$$
  é o vetor de variáveis preditoras e
  $$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$$
  é o vetor dos coeficientes de regressão.

Tópico 24: teste da razão de verossimilhanças

- Seja $\{Y_1, \cdots, Y_n\}$ uma amostra aleatória simples de Y.
- Para cada $i \in \{1, \cdots, n\}$, sejam $x_{i1}, \cdots, x_{ip}$ os valores das preditoras $x_1, \cdots, x_p$, respectivamente, correspondentes a Y_i.
- Nesse caso, supõe-se que
  1. a distribuição de Y_i pertence à família exponencial;
  2. $\eta(\boldsymbol{x}_i; \boldsymbol{\beta}) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$;
  3. $g(\mu_i) = \eta(\boldsymbol{x}_i; \boldsymbol{\beta}) = \eta_i$;

  onde μ_i é a média de Y_i, g é a função de ligação,

$$\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{ip})^T$$

  é o vetor da i-ésima observação das variáveis preditoras e

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$$

  é o vetor dos coeficientes de regressão.

Tópico 24: teste da razão de verossimilhanças

- As hipóteses do teste da razão de verossimilhanças para avaliar a significância global do modelo são as seguintes:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0;$$
$$H_1 : \beta_j \neq 0, \text{ para ao menos um } j \in \{1, 2, \cdots, p\}.$$

- No modelo sob H_0, as preditoras são excluídas e, portanto, há apenas dois parâmetros a serem estimados: o coeficiente β_0 e o parâmetro de dispersão ϕ.
- O teste da razão de verossimilhanças avalia a plausibilidade da hipótese nula comparando o desempenho do modelo sob hipótese nula, isto é, do modelo sem variáveis preditoras, com o desempenho do modelo contendo todas as variáveis preditoras, o qual, aqui, é denominado modelo completo.

Tópico 24: teste da razão de verossimilhanças

- A medida de desempenho utilizada pelo teste é o valor máximo da função de log-verossimilhança.
- Sejam $\widehat{\boldsymbol{\beta}}$ e $\hat{\phi}$ os estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e de ϕ, respectivamente, no modelo completo.
- Sejam $\hat{\beta}_0^0$ e $\hat{\phi}_0^0$ os estimadores de máxima verossimilhança de β_0 e de ϕ, respectivamente, no modelo sob H_0.
- O valor máximo da função de log-verossimilhança, referente ao modelo completo, é $l(\widehat{\boldsymbol{\beta}}, \hat{\phi})$.
- O valor máximo da função de log-verossimilhança, referente ao modelo sob H_0 é $l(\hat{\beta}_0^0, \hat{\phi}_0^0)$.
- A estatística do teste de razão de verossimilhanças para avaliar a significância global do modelo é dada por

$$T_{RV} = 2\{l(\widehat{\boldsymbol{\beta}}, \hat{\phi}) - l(\hat{\beta}_0^0, \hat{\phi}_0^0)\}. \tag{1}$$

Tópico 24: teste da razão de verossimilhanças

- Quanto maior o valor de T_{RV}, melhor o desempenho do modelo completo em relação ao modelo sob H_0 e, portanto, maior a evidência contra H_0.
- Nota-se que o modelo completo contém $p + 2$ parâmetros e o modelo sob H_0 contém 2 parâmetros.
- Isto é, o modelo completo contém p parâmetros a mais em relação ao modelo sob H_0.
- Portanto, se n é suficientemente grande, então a distribuição da estatística T_{RV}, sob H_0, pode ser aproximada pela distribuição qui-quadrado com p graus de liberdade.
- A regra de decisão do teste é: rejeita-se a hipótese nula, ao nível de significância α, se o valor observado de T_{RV} é maior do que o quantil de ordem $1 - \alpha$ da distribuição qui-quadrado.

Tópico 24: teste da razão de verossimilhanças

- Ou seja, denotando por $\chi^2_p(1-\alpha)$ o quantil de ordem $1-\alpha$ da distribuição qui-quadrado com p graus de liberdade e por t_{RV} o valor observado de T_{RV}, tem-se que, H_0 é rejeitada, ao nível de significância α, se $t_{RV} > \chi^2_p(1-\alpha)$.
- O p-valor do teste é dado por

$$\text{p-valor} = P(T_{RV} > t_{RV} | H_0).$$

- Esse p-valor pode ser calculado utilizando a aproximação a distribuição de T_{RV} pela distribuição qui-quadrado com p graus de liberdade, conforme mencionado anteriormente.
- Um teste com nível de significância α é obtido adotando-se a seguinte regra de decisão: rejeita-se H_0 se p-valor $< \alpha$.

Tópico 24: teste da razão de verossimilhanças

- Se ϕ é conhecido, então não precisa ser estimado.
- Nesse caso, o modelo completo passa a ter $p + 1$ parâmetros, os coeficientes $\beta_0, \beta_1, \cdots, \beta_p$, e o modelo sob H_0 passa a ter um único parâmetro, o coeficiente β_0.
- A função de log-verossimilhança passa a ser vista como função apenas dos coeficientes de regressão e, portanto, T_{RV} pode ser escrita como

$$T_{RV} = 2\{l(\widehat{\boldsymbol{\beta}}, \phi) - l(\hat{\beta}_0^0, \phi)\} = 2\{l(\widehat{\boldsymbol{\beta}}) - l(\hat{\beta}_0^0)\}. \qquad (2)$$

- A aproximação qui-quadrado, o cálculo do p-valor e as regras de decisão, mencionadas no caso em que ϕ é desconhecido, são aplicadas exatamente da mesma forma.

Tópico 24: teste da razão de verossimilhanças

- Na plataforma **R**, uma função que pode ser utilizada para realizar o teste da razão de verossimilhanças para avaliação global do modelo é a função `lr.test`, do pacote `mdscore`.

Tópico 24: teste da razão de verossimilhanças

Exemplo 1

Considere o conjunto de dados **store.dat**, cuja descrição foi feita no exemplo 2 da aula 2. O objetivo é ajustar um modelo de regressão Poisson, no qual a variável resposta é o número de clientes e as variáveis preditoras são a distância ao concorrente mais próximo e a distância à loja, e executar o teste da razão de verossimilhanças para avaliar a significância global do modelo. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_1.ipynb** ou no arquivo **exemplo_1.html**, localizados na pasta da aula 12.

Tópico 24: teste da razão de verossimilhanças

Exemplo 2

Considere o conjunto de dados **icu.csv**, que se encontra descrito no exemplo 3 da aula 2. O objetivo é ajustar um modelo de regressão logística, tendo a variável **sta** (Sobrevivência) como resposta e a variável **age** (Idade) como preditora, e executar o teste da razão de verossimilhanças para avaliar a significância global do modelo. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_2.ipynb** ou no arquivo **exemplo_2.html**, localizados na pasta da aula 12.

Tópico 24: teste da razão de verossimilhanças

Exemplo 3

Considere o conjunto de dados **trees.dat**, que se encontra descrito no exemplo 4 da aula 2. O objetivo é ajustar um modelo de regressão gama, com ligação logarítmica, tendo a variável volume como variável resposta e as variáveis altura e diâmetro como variáveis preditoras, e executar o teste da razão de verossimilhanças para avaliar a significância global do modelo. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_3.ipynb** ou no arquivo **exemplo_3.html**, ambos na pasta da aula 12.

Referências I

[1] G. M. Cordeiro e C. G. B. Demétrio, *Modelos lineares generalizados e extensões*, 2013, disponível em https://docs.ufpr.br/~taconeli/CE22518/LivClarice.pdf.

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