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

Aula 13

Tópicos 30 e 31: predição da resposta média

Sumário

Metas

1. Apresentar métodos para, com base no MLG, predizer valores desconhecidos da variável resposta.

Objetivos

- Após estudar essa aula, o aluno ou aluna será capaz de:
  1. aplicar o MLG para predizer valores desconhecidos da resposta, utilizando o estimador da resposta média como preditor;
  2. calcular intervalos de predição;
  3. executar os métodos de predição apresentados na plataforma computacional **R**.

Pré-requisitos

1. Aula 12.
2. Parte 1 da aula 13.

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.
- A inferência sobre as associações entre a resposta e as variáveis preditoras foi tratada na aula 12 e na parte 1 da aula 13.

Introdução

- Na aula 12, foi apresentada uma versão do teste da razão de verossimilhanças que permite avaliar a significância global do MLG, isto é, se ao menos um dos coeficientes de regressão é significativamente diferente de zero.
- Uma vez que se tenha decidido que o modelo é globalmente significativo, é necessário identificar quais dos coeficientes de regressão são, de fato, significativamente diferentes de zero.
- Isso pode ser feito por meio do teste de Wald para avaliação da significância individual dos coeficientes de regressão, que foi apresentado na parte 1 da aula 13.
- Uma outra aplicação do modelo, mencionada na aula 2, é a predição de valores desconhecidos da variável reposta.

Introdução

- Em geral, um dos principais objetivos, com a modelagem de regressão, é obter um modelo matemático que permita predizer o valor desconhecido de uma variável resposta, a partir de valores conhecidos das variáveis preditoras associadas a ela.
- Várias são as situações que justificam a necessidade de predição de uma resposta a partir de variáveis preditoras.
- Algumas dessas situações encontram-se descritas na aula 2.
- Na presente aula, são apresentados métodos para predição do valor médio da variável resposta num MLG.
- Dois tipos de métodos são considerados: um método de predição pontual e um de predição intervalar, ambos utilizados para predizer o valor da resposta média.

Tópico 30: predição da variável resposta

- 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 30: predição da variável resposta

- 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 30: predição da variável resposta

- Deseja-se predizer o valor da resposta média μ, correspondente a um dado vetor de variáveis preditoras $\boldsymbol{x} = (1, x_1, \cdots, x_p)^T$, com base no MLG.
- Sejam $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^T$ e $\hat{\phi}$ os estimadores de máxima verossimilhança de $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ e ϕ, respectivamente.
- O estimador de máxima verossimilhança de η é dado por

$$\hat{\eta} = x^T\widehat{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{1}$$

- O estimador utilizado para predizer o valor de μ é o estimador de máxima verossimilhança de μ, o qual é dado por

$$\hat{\mu} = g^{-1}(\hat{\eta}) = g^{-1}(x^T\widehat{\boldsymbol{\beta}}) = g^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p). \tag{2}$$

Tópico 30: predição da variável resposta

- Dessa forma, para predizer o valor de μ, calcula-se a estimativa de máxima verossimilhança de μ pela equação (2) e utiliza-se o valor calculado como estimativa de μ.
- Na plataforma **R**, uma função que pode ser usada para calcular intervalo de confiança é a função `predict`.

Tópico 31: intervalos de predição

- Conforme mencionado na parte 1 da aula 13, se a amostra é suficientemente grande, então a distribuição de $\widehat{\boldsymbol{\beta}}$ pode ser aproximada pela normal multivariada, com média $\boldsymbol{\beta}$ e matriz de covariância $\boldsymbol{K}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}$.
- Isto é, se n é grande, então

$$\widehat{\boldsymbol{\beta}} \overset{\text{aprox.}}{\sim} \boldsymbol{N}_{p+1}\left(\boldsymbol{\beta}, \boldsymbol{K}^{-1}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}\right). \tag{3}$$

- Consequentemente, se n é grande, então

$$\hat{\eta} = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} \overset{\text{aprox.}}{\sim} N\left(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{x}^T\boldsymbol{K}^{-1}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}\boldsymbol{x}\right) = N\left(\eta, \hat{\sigma}^2_{\hat{\eta}}\right), \tag{4}$$

onde $\hat{\sigma}^2_{\hat{\eta}} = \boldsymbol{x}^T\boldsymbol{K}^{-1}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}\boldsymbol{x}$.

Tópico 31: intervalos de predção

- O resultado (4) pode ser utilizado para construir um intervalo de confiança para η.
- Se a amostra é suficientemente grande, então (4) implica que

$$\frac{\hat{\eta}-\eta}{\sqrt{\hat{\sigma}^2_{\hat{\eta}}}} \overset{\text{aprox.}}{\sim} N(0,1).$$

- Nesse caso, denotando por $z_{(1-\frac{\alpha}{2})}$ o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição $N(0,1)$, tem-se que

$$P\left(\frac{|\hat{\eta}-\eta|}{\sqrt{\hat{\sigma}^2_{\hat{\eta}}}} \leq z_{(1-\frac{\alpha}{2})}\right) \approx 1-\alpha.$$

- Equivalentemente,

$$P\left(\hat{\eta} - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2_{\hat{\eta}}} < \eta < \hat{\eta} + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2_{\hat{\eta}}}\right) \approx 1-\alpha.$$

Tópico 31: intervalos de predição

- Portanto, se n é grande, o intervalo

$$\left(\hat{\eta} - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2_{\hat{\eta}}};\ \hat{\eta} + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2_{\hat{\eta}}}\right) \tag{5}$$

  é um intervalo de confiança para η, com nível de confiança aproximadamente igual a $1-\alpha$.

- Como $\mu = g^{-1}(\eta)$, pode-se concluir que o intervalo

$$\left(g^{-1}\left(\hat{\eta} - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2_{\hat{\eta}}}\right);\ g^{-1}\left(\hat{\eta} + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{\sigma}^2_{\hat{\eta}}}\right)\right) \tag{6}$$

  é um intervalo de confiança para μ, com nível de confiança aproximadamente igual a $1-\alpha$.

- O intervalo de predição de μ é definido como sendo o intervalo de confiança de μ, dado pela equação (6).

Tópico 31: intervalos de predição

- Enquanto $\hat{\mu}$ é uma estimativa pontual de μ, o intervalo de predição de μ é uma estimativa intervalar de μ.
- Na plataforma **R**, uma função que pode ser usada para calcular intervalo de confiança é a função `add_ci`, do pacote `ciTools`.

Tópico 31: intervalos de predição

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 realizar as predições dos números de clientes correspondentes às distâncias da ao concorrente mais próximo e às distâncias à loja dadas na tabela abaixo. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_1.ipynb** ou no arquivo **exemplo_1.html**, localizados na pasta da aula 13.

distConc	1	3	5
distLoja	5	7	10

Tópico 31: intervalos de predição

Exemplo 2

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 como variável resposta a variável volume e como variáveis preditoras as variáveis altura e diâmetro e realizar as predições dos volumes correspondentes às alturas e aos diâmetros dados na tabela abaixo. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_2.ipynb** ou no arquivo **exemplo_2.html**, localizados na pasta da aula 13.

Altura	70	75	83
Diâmetro	11	12	18

Tópico 31: intervalos de predição

Exemplo 3

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 realizar as predições das probabilidades de sobrevivência correspondentes às idades na tabela abaixo. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_3.ipynb** ou no arquivo **exemplo_3.html**, localizados na pasta da aula 13.

Idade	35	60	78