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 28 e 29: avaliação da significância individual dos coeficientes de regressão

Sumário

1. Informações sobre a aula
   - Metas
   - Objetivos
   - Pré-requisitos
2. Introdução
3. Aula 13
   - Tópico 28: teste de Wald para a significância individual
   - Tópico 29: intervalos de confiança para os coeficientes
4. Referências

Metas

1. Apresentar métodos, baseados na estatística de Wald, para avaliar a significância individual dos coeficientes de regressão.

Objetivos

- Após estudar essa aula, o aluno ou aluna será capaz de:
  1. desenvolver o teste de Wald para avaliar a significância individual dos coeficientes de regressão do MLG;
  2. calcular intervalos de confiança, com base na estatística de Wald, para os coeficientes de regressão do MLG;
  3. interpretar os resultados do teste e os intervalos de confiança;
  4. executar o teste e calcular os intervalos de confiança utilizando a plataforma computacional **R**.

Pré-requisitos

1. Aula 12.

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.
- Conforme visto na aula 12, o teste de significância global do MLG é o teste cujas hipóteses nula e alternativa são

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

- Isto é, o teste de significância global procura avaliar se ao menos um coeficiente de regressão, associado a alguma variável preditora, é significativamente diferente de zero.

Introdução

- Em outras palavras, o teste de significância global do modelo procura avaliar se há uma associação, significativa do ponto de vista estatístico, entre a variável resposta e ao menos uma das variáveis preditoras.
- Uma vez que a hipótese nula, no teste de significância global, tenha sido rejeitada, pode-se concluir que ao menos um dos coeficientes de regressão, associados às variáveis preditoras, é significativamente diferente de zero.
- Nesse ponto, a pergunta que se faz é a seguinte: quais são os coeficientes de regressão que podem ser considerados diferentes de zero?
- Essa pergunta é respondida avaliando-se a significância individual de cada um dos coeficientes de regressão.

Introdução

- A avaliação da significância individual de um coeficiente pode ser feita por meio de um teste de hipóteses.
- O teste de hipóteses, para avaliar significância individual do coeficiente β_j, é o teste cujas hipóteses são

$$H_0 : \beta_j = 0;$$
$$H_1 : \beta_j \neq 0.$$

- Se a hipótese nula não é rejeitada, então o teste indica que o coeficiente de regressão β_j não é significativamente diferente de zero e que, portanto, não há uma associação estatisticamente significativa entre a resposta e a preditora x_j. Nesse caso, pelo menos a princípio, a variável x_j poderia ser excluída do modelo.
- A significância individual do coeficiente β_j também pode ser avaliada por meio de um intervalo de confiança para β_j.

Introdução

- Na presente aula, são desenvolvidos um teste de hipóteses para a significância individual de um coeficiente e um intervalo de confiança para o coeficiente.
- Tanto o teste quanto intervalo de confiança baseiam-se na estatística de Wald [2, p. 117] e na normalidade assintótica dos estimadores de máxima verossimilhança [2, p. 96].
- Versões desses procedimentos são apresentadas para alguns dos principais casos particulares do MLG.
- Por fim, mostra-se como executar o teste e calcular o intervalo utilizando a plataforma **R**.

Tópico 28: teste de Wald para a significância individual

- 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 28: teste de Wald para a significância individual

- 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 28: teste de Wald para a significância individual

- 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.
- Sejam $\boldsymbol{\psi} = (\beta_0, \beta_1, \cdots, \beta_p, \phi)^T = \left(\boldsymbol{\beta}^T, \phi\right)^T$ o vetor contendo os parâmetros do MLG e $\widehat{\boldsymbol{\psi}} = \left(\widehat{\boldsymbol{\beta}}^T, \hat{\phi}\right)^T$ o estimador de máxima verossimilhança de $\boldsymbol{\psi}$.
- Se n é suficientemente grande, então a distribuição de $\widehat{\boldsymbol{\psi}}$ pode ser aproximada por uma normal multivariada, com média $\boldsymbol{\psi}$ e matriz de covariância $\boldsymbol{K}^{-1}(\boldsymbol{\beta}, \phi)$ [3, p. 28].
- Isto é, para grandes amostras, tem-se, aproximadamente,

$$\widehat{\boldsymbol{\psi}} \sim \boldsymbol{N}_{p+1}\left(\boldsymbol{\psi}, \boldsymbol{K}^{-1}(\boldsymbol{\beta}, \phi)\right). \tag{1}$$

- Esse resultado é consequência da normalidade assintótica dos estimadores de máxima verossimilhança [1, p. 102].

Tópico 28: teste de Wald para a significância individual

- Conforme visto no tópico 12,

$$\boldsymbol{K}(\boldsymbol{\beta},\phi)=\begin{pmatrix}\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \underset{(p+1)\times 1}{\mathbf{0}} \\ \underset{1\times(p+1)}{\mathbf{0}} & K_{\phi\phi}\end{pmatrix} \tag{2}$$

- Logo,

$$\boldsymbol{K}^{-1}(\boldsymbol{\beta},\phi)=\begin{pmatrix}\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} & \underset{(p+1)\times 1}{\mathbf{0}} \\ \underset{1\times(p+1)}{\mathbf{0}} & K_{\phi\phi}^{-1}\end{pmatrix}. \tag{3}$$

- De (3) e (1), tem-se que, se n é grande, então

$$\widehat{\boldsymbol{\beta}} \overset{\text{aprox.}}{\sim} \boldsymbol{N}_{p+1}\left(\boldsymbol{\beta},\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\right) \quad \text{e} \quad \hat{\phi} \overset{\text{aprox.}}{\sim} N\left(\phi, K_{\phi\phi}^{-1}\right). \tag{4}$$

Tópico 28: teste de Wald para a significância individual

- Isto é, se a amostra é grande, então $\widehat{\boldsymbol{\beta}}$ segue, aproximadamente, distribuição normal multivariada, com média $\boldsymbol{\beta}$ e matriz de covariância $\boldsymbol{K}^{-1}_{\boldsymbol{\beta},\boldsymbol{\beta}}$.
- Consequentemente, se a amostra é grande, então $\hat{\beta}_j$ segue, aproximadamente, distribuição normal com média β_j e variância dada pelo elemento $j+1$ da diagonal de $\boldsymbol{K}^{-1}_{\boldsymbol{\beta},\boldsymbol{\beta}}$.
- Ou seja, para cada $j \in \{0, 1, \cdots, p\}$, se n é grande, então

$$\hat{\beta}_j \overset{\text{aprox.}}{\sim} N\left(\beta_j, Var(\hat{\beta}_j)\right), \tag{5}$$

  onde $Var(\hat{\beta}_j)$ é o elemento $j+1$ da diagonal de $\boldsymbol{K}^{-1}_{\boldsymbol{\beta},\boldsymbol{\beta}}$.
- O teste de Wald, para avaliação da significância individual do coeficiente de regressão β_j, é baseado no resultado (5).

Tópico 28: teste de Wald para a significância individual

- As hipóteses do teste são

$$H_0 : \beta_j = 0;$$
$$H_1 : \beta_j \neq 0.$$

- A estatística do teste é

$$T_{W_j} = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var(\hat{\beta}_j)}}}, \tag{6}$$

onde $\widehat{Var(\hat{\beta}_j)}$ é o estimador da variância de $\hat{\beta}_j$, isto é, $\widehat{Var(\hat{\beta}_j)}$ é o elemento $j+1$ da diagonal de $\boldsymbol{K}^{-1}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}}$.

- Pelo resultado (5), tem-se que, sob H_0, se n é grande, então

$$T_{W_j} \overset{\text{aprox.}}{\sim} N(0,1). \tag{7}$$

Tópico 28: teste de Wald para a significância individual

- Quanto mais distante de zero o valor de T_{W_j}, maior o grau de evidência contra H_0.
- Com base nessa constatação e no resultado (7), elabora-se a seguinte regra de decisão: rejeita-se a hipótese nula, ao nível de significância α, se o valor observado de $|T_{W_j}|$ é maior do que o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição $N(0,1)$.
- Sejam t_{W_j} o valor observado de T_{W_j} e $z_{(1-\frac{\alpha}{2})}$ o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição $N(0,1)$. De acordo com essa regra de decisão, a hipótese H_0 é rejeitada se $t_{W_j} > z_{(1-\frac{\alpha}{2})}$.
- O p-valor do teste é dado por

$$P(|T_{W_j}| > |t_{W_j}|) = 2[1 - P(Z > |t_{W_j}|)], \tag{8}$$

onde $Z \sim N(0,1)$.

Tópico 28: teste de Wald para a significância individual

- Um teste com nível de significância α é obtido adotando-se a seguinte regra de decisão: rejeita-se H_0 se p-valor $< \alpha$.
- Na plataforma **R**, uma função que pode ser utilizada para realizar o teste de Wald, para avaliar a significância individual dos coeficientes de regressão, é a função `summary`.

Tópico 29: intervalos de confiança para os coeficientes

- O resultado (5) pode ser utilizado para construir um intervalo de confiança para β_j, para cada $j \in \{0, 1, \cdots, p\}$.
- Se a amostra é suficientemente grande, então (5) implica que

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var(\hat{\beta}_j)}}} \overset{\text{aprox.}}{\sim} N(0, 1).$$

- 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

$$P\left(\frac{|\hat{\beta}_j - \beta_j|}{\sqrt{\widehat{Var(\hat{\beta}_j)}}} \leq z_{(1-\frac{\alpha}{2})}\right) \approx 1 - \alpha.$$

Tópico 29: intervalos de confiança para os coeficientes

- Equivalentemente,

$$P\left(\hat{\beta}_j - z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{Var(\hat{\beta}_j)}} < \beta_j < \hat{\beta}_j + z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{Var(\hat{\beta}_j)}}\right) \approx 1-\alpha.$$

- Portanto, o intervalo

$$\left(\hat{\beta}_j - z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{Var(\hat{\beta}_j)}};\ \hat{\beta}_j + z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{Var(\hat{\beta}_j)}}\right) \qquad (9)$$

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

- Na plataforma **R**, uma função que pode ser usada para calcular intervalo de confiança (9) é a função `confint.default`.

Tópico 29: intervalos de confiança para os coeficientes

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, executar o teste de Wald, ao nível de 5% de significância, para avaliar a significância individual dos coeficientes de regressão e calcular o intervalo de confiança, com 95% de confiança, para cada um dos coeficientes de regressão. 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.

Tópico 29: intervalos de confiança para os coeficientes

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, executar o teste de Wald, ao nível de 5% de significância, para avaliar a significância individual dos coeficientes de regressão e calcular o intervalo de confiança, com 95% de confiança, para cada um dos coeficientes de regressão. Os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_2.ipynb** ou no arquivo **exemplo_2.html**, ambos na pasta da aula 13.

Tópico 29: intervalos de confiança para os coeficientes

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, executar o teste de Wald, ao nível de 1% de significância, para avaliar a significância individual dos coeficientes de regressão e calcular o intervalo de confiança, com 99% de confiança, para cada um dos coeficientes de regressão. 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.

Referências I

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

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

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