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 15: seleção das variáveis

Sumário

Metas

1. Apresentar métodos para selecionar as variáveis preditoras a serem inclusas no MLG.

Objetivos

- Após estudar essa aula, o aluno ou aluna será capaz de:
  1. descrever os método forward, backward e stepwise para seleção das variáveis preditoras a serem inclusas no MLG, com base no critério de Akaike;
  2. executar, na plataforma computacional **R**, os método forward, backward e stepwise, com base no critério de Akaike.

Pré-requisitos

1. Aula 13.

Introdução

- 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.

Introdução

- 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 1que
  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.

Introdução

- Especificar o segundo componente do MLG, isto é, o preditor linear, equivale a especificar as p variáveis preditoras que farão parte do modelo.
- Em algumas áreas do conhecimento, a teoria sendo utilizada para compreender o fenômeno em questão pode guiar a seleção das variáveis preditoras [7, p. 417].
- Isso ocorre no exemplo 4 da aula 2.
- Naquele exemplo, deseja-se obter um modelo no qual a variável resposta é o volume da árvore e as variáveis preditoras são a altura e o diâmetro da árvore.
- A escolha da altura e do diâmetro como variáveis preditoras se justifica pelo fato de a forma de uma árvore poder ser aproximada pelas formas de um cilindro ou de um cone.

Introdução

- O volume de um cilindro, com altura h e diâmetro d, é
$$V = \frac{\pi h d^2}{4}.$$
- O volume de um cone, com altura h e base com diâmetro d, é
$$V = \frac{\pi h d^2}{12}.$$
- Em ambos os casos, tem-se que $\ln(V) = \beta_0 + \beta_1 \ln(h) + \beta_2 \ln(d)$.
- Isso leva ao modelo proposto no exemplo 11 da aula 6, segundo o qual o volume médio da árvore é dado por
$$\ln(\mu) = \beta_0 + \beta_1 \ln(h) + \beta_2 \ln(d).$$
- Ou seja, as variáveis preditoras são os logaritmos do diâmetro e da altura e a função de ligação é a logarítmica.

Introdução

- Em situações nas quais não há suporte teórico, a seleção das variáveis preditoras ganha bastante relevância [1, p. 281].
- Nesses casos, os investigadores são frequentemente forçados a prospectar variáveis preditoras que, possivelmente, apresentam relação com a variável resposta em questão [7, p. 418].
- Procura-se evitar a escolha de variáveis preditoras cujos valores possam apresentar erros consideráveis de mensuração ou cujos valores são difíceis de se medir.
- Evita-se também a escolha de variáveis preditoras que não são fundamentalmente ligadas ao fenômeno sendo estudado.
- Tendo escolhido as p variáveis preditoras e tendo coletado os dados, passa-se à modelagem de regressão, seguindo as etapas básicas mencionadas na aula 2.

Introdução

- O modelo a ser utilizado, a princípio, contém as p variáveis preditoras disponíveis.
- Entretanto, em algumas situações, é desejável obter um modelo contendo apenas algumas dessas p variáveis preditoras.
- Tal modelo é dito ser um **submodelo** do **modelo completo**, o qual contém as p preditoras.
- Razões gerais que justificam a opção por um submodelo são:
  - o custo de obtenção da informação para monitorar e atualizar modelos com menos variáveis é menor [2, p. 327];
  - encontrar submodelos que se adéquem bem aos dados e que contenham variáveis preditoras mais fáceis e/ou mais baratas de serem medidas [5, p. 502];
  - modelos com menos variáveis são menos propensos a problemas computacionais.

Introdução

- Além disso, as técnicas apresentadas nesse curso assumem que o número de variáveis preditoras é menor ou igual ao número de observações. Portanto, se $p > n$, então o investigador tem que optar, necessariamente, por um submodelo.
- Quando o interesse é obter um submodelo, o que se deseja é um modelo, formado apenas por um subconjunto das p variáveis preditoras, mas que, em algum sentido pré-definido, se adéque bem aos dados.
- Como identificar tal subconjunto das p variáveis preditoras?
- A estratégia básica para responder essa questão consiste em realizar uma **busca exaustiva**, isto é, procurar o submodelo desejado dentre todos os submodelos que podem ser formados com as p variáveis preditoras disponíveis.

Introdução

- Com p variáveis, pode-se formar 2^p submodelos distintos.
- Se $p = 2$, então há $2^2 = 4$ submodelos distintos:

$$(1)\ \eta = \beta_0; \qquad (2)\ \eta = \beta_0 + \beta_1 x_1;$$
$$(3)\ \eta = \beta_0 + \beta_2 x_2; \qquad (4)\ \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

- Se $p = 3$, então há $2^3 = 8$ submodelos distintos:

$$(1)\ \eta = \beta_0; \qquad (2)\ \eta = \beta_0 + \beta_1 x_1;$$
$$(3)\ \eta = \beta_0 + \beta_2 x_2; \qquad (4)\ \eta = \beta_0 + \beta_3 x_3;$$
$$(5)\ \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2; \qquad (6)\ \eta = \beta_0 + \beta_1 x_1 + \beta_3 x_3;$$
$$(7)\ \eta = \beta_0 + \beta_2 x_2 + \beta_3 x_3; \qquad (8)\ \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

- Sendo assim, o submodelo pode ser obtido comparando-se os desempenhos dos 2^p submodelos, segundo algum critério de avaliação de desempenho pré-definido, e escolhendo-se aquele com o melhor desempenho.

Introdução

- A escolha do critério a ser utilizado para avaliar e comparar os desempenhos dos submodelos, depende do uso que será feito do modelo obtido ao final do processo de modelagem.
- Há três usos básicos para um modelo de regressão [1, p. 284]:
  - descrição/explicação do comportamento da variável resposta em termos das variáveis preditoras;
  - predição do valor da variável resposta, dados os valores das variáveis preditoras;
  - controle do valor da variável resposta, controlando os valores das variáveis preditoras.
- Uma breve discussão sobre esses usos, com a apresentação de alguns exemplos, pode ser encontrada na aula 2.
- Cada um desses usos possui suas próprias peculiaridades no que diz respeito à seleção das variáveis preditoras.

Introdução

- Quando o objetivo é descrever o comportamento da variável resposta em termos das variáveis preditoras, procura-se
  - evitar a inclusão no modelo, ao mesmo tempo, de variáveis preditoras correlacionadas umas com as outras, o que poderia caracterizar **multicolinearidade**; a multicolinearidade prejudica a interpretação dos coeficientes;
  - dar preferência aos modelos com poucas variáveis preditoras, por serem mais facilmente interpretados e compreendidos; esse é o princípio da parcimônia [1, p. 69, 284];
  - obter estimativas precisas dos coeficientes de regressão, isto é, com erros padrões pequenos, pois os coeficientes de regressão exercem papel fundamental na interpretação do modelo.
- No caso em que do modelo será usado para controlar a variável resposta, procura-se obter submodelos cujos coeficientes de regressão são estimados com precisão [1, p. 284].

Introdução

- No caso em que o objetivo é predição, o foco é na obtenção de submodelos que forneçam predições com um alto nível de precisão, por exemplo, predições com erros quadráticos médios pequenos [1, p. 284]. Nesse caso, nem a quantidade de variáveis preditoras e nem a multicolinearidade são problemas, embora uma redução no número de variáveis prditoras possa ajudar a reduzir a variância amostral [8]. Portanto, a preferência não precisa, necessariamente, ser dada aos submodelos com poucas variáveis preditoras, podendo-se escolher submodelos com um número maior de variáveis, desde que apresentem boa capacidade preditiva.
- Na presente aula, o foco é obter um submodelo com o objetivo de descrever o comportamento da variável resposta em termos das variáveis preditoras.

Introdução

- Conforme já visto, para esse uso, deve-se obter um submodelo que (1) explique bem a variabilidade nos dados (2) com o menor número de variáveis preditoras possível e (3) possibilite estimar com precisão os coeficientes de regressão.
- O requisito (1) concorre com os requisitos (2) e (3).
- De fato, modelos com um número maior de variáveis preditoras têm maior capacidade de capturar a variabilidade nos dados.
- Por outro lado, as estimativas dos coeficientes de regressão em modelos com mais variáveis preditoras são menos precisas.
- É preciso equilibrar a capacidade de capturar a variabilidade e o número de variáveis preditoras.
- Um critério que pode ser utilizado para esse fim é o critério de informação de Akaike (AIC) [3, p. 124], [4, p. 341].

Introdução

- 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 critério de informação de Akaike é definido como

$$AIC = -2l(\widehat{\boldsymbol{\beta}}, \hat{\phi}) + 2p, \tag{1}$$

  onde $l(\cdot\ , \cdot)$ é a função de log-verossimilhança de um modelo com p variáveis preditoras (ver tópico 10).
- Sabe-se que a função de log-verossimilhança pode ser usada para medir a qualidade do modelo ajustado: quanto maior o valor de $l(\widehat{\boldsymbol{\beta}}, \hat{\phi})$, melhor a qualidade do ajuste.
- Sabe-se também que modelos com mais parâmetros têm maiores valores de $l(\widehat{\boldsymbol{\beta}}, \hat{\phi})$.

Introdução

- Portanto, se a qualidade dos modelos for avaliada apenas pela função de log-verossimilhança, o melhor modelo será aquele com mais variáveis preditoras.
- O critério de Akaike penaliza a inclusão de variáveis preditoras no modelo, de forma que modelos com mais variáveis preditoras nem sempre sejam os melhores.
- A inclusão de variáveis preditoras no modelo aumenta o valor de $l(\widehat{\boldsymbol{\beta}}, \hat{\phi})$, causando uma redução no valor do AIC.
- Por outro lado, ao incluir variáveis preditoras no modelo, o valor de p aumenta, causando um aumento no valor do AIC.
- O melhor modelo, pelo critério de informação de Akaike, é aquele que apresenta menor valor do AIC.

Introdução

- Portanto, ao utilizar o critério de informação de Akaike para comparar as qualidades dos modelos, nem sempre aquele com mais variáveis preditoras é avaliado como sendo o melhor, o que torna possível a escolha de modelos parcimoniosos.
- Sendo assim, a seleção de um submodelo, com base no critério de informação de Akaike, por busca exaustiva, pode ser descrita da seguinte forma:
  1. para cada um dos 2^p modelos, calcula-se o valor de AIC;
  2. seleciona-se o submodelo com menor valor de AIC.
- A busca exaustiva é o procedimento ideal para a seleção do submodelo, pois considera todos os possíveis submodelos.
- O problema com a busca exaustiva é que a quantidade de submodelos aumenta rapidamente com o número de preditoras.

Introdução

- Quando há quatorze variáveis preditoras disponíveis, o número de submodelos possíveis é $2^{14} = 16.384$. Com dezoito variáveis preditoras, esse número salta para $2^{18} = 262.144$.
- Diversos algoritmos têm sido propostos, visando otimizar a seleção do submodelo por busca exaustiva [6, p. 49].
- Porém, dependendo do número de variáveis preditoras a serem consideradas, pode ser impraticável selecionar o submodelo por busca exaustiva, mesmo usando esses algoritmos.
- Segundo [6, p. 48], a avaliação de todos os possíveis submodelos é viável quando há, no máximo, algo em torno de vinte variáveis preditoras disponíveis.
- Para esses casos, há métodos de seleção alternativos, os quais não envolvem a avaliação de todos os possíveis submodelos.

Introdução

- Dentre tais métodos, os mais conhecidos são os algoritmos **forward**, **backward** e **stepwise** [9, p. 221], [4, p. 150].
- Esses algoritmos realizam ações de inclusão no modelo e/ou exclusão do modelo de uma variável por vez, de forma a obter um submodelo satisfatório sem ter que, necessariamente, avaliar todos os submodelos possíveis.
- A presente aula trata da seleção do submodelo por meio desses três algoritmos e da busca exaustiva, adotando-se o AIC como critério para avaliação e comparação dos modelos
- Além do critério de Akaike, há outros critérios de informação que podem ser utilizados como, por exemplo, o critério de informação Bayesiana (BIC) [3, p. 124], [4, p. 342].

Tópico 35: algoritmo forward

- Partindo do modelo sem variáveis preditoras, em cada passo do algoritmo forward, uma nova variável preditora é inclusa no modelo.
- A variável preditora inclusa é aquela que promove uma maior melhora no desempenho do modelo.
- Esse processo de inclusão é realizado até que a inclusão de qualquer nova variável preditora não melhore o modelo ou até que todas as variáveis preditoras tenham sido inclusas.
- Os passos do algoritmo forward são os seguintes:
  1. selecione, como submodelo atual, o submodelo sem variáveis preditoras e calcule o seu AIC;
  2. ajuste os submodelos que podem ser obtidos incluindo-se uma nova variável preditora ao submodelo atual e calcule seus AICs;

Tópico 35: algoritmo forward

3. dentre os submodelos ajustados no passo 2, selecione aquele com menor AIC;
4. se o AIC do submodelo selecionado no passo 3 for menor do que o AIC do submodelo atual, selecione, como submodelo atual, o submodelo selecionado no passo 3 e vá para o passo 5; caso contrário, pare o algoritmo e selecione, como resultado final do algoritmo, o submodelo atual;
5. se o submodelo atual tiver todas as p variáveis preditoras, pare o algoritmo e selecione, como resultado final do algoritmo, o submodelo atual; caso contrário, vá para o passo 2.

Tópico 36: algoritmo backward

- Partindo do modelo com todas as variáveis preditoras, em cada passo do algoritmo backward, uma variável preditora é excluída do modelo.
- A variável preditora excluída é aquela cuja exclusão promove uma maior melhora no desempenho do modelo.
- Esse processo de exclusão é realizado até que a exclusão de qualquer variável preditora não melhore o modelo ou até que todas as variáveis preditoras tenham sido excluídas.
- Os passos do algoritmo backward são os seguintes:
  1. selecione, como submodelo atual, o submodelo sem variáveis preditoras e calcule o seu AIC;
  2. ajuste todos os submodelos que podem ser obtidos excluindo-se uma variável preditora do submodelo atual e calcule seus AICs;

Tópico 36: algoritmo backward

3. dentre os submodelos ajustados no passo 2, selecione aquele com menor AIC;
4. se o AIC do submodelo selecionado no passo 3 for menor do que o AIC do submodelo atual, selecione, como submodelo atual, o submodelo selecionado no passo 3 e vá para o passo 5; caso contrário, pare o algoritmo e selecione, como resultado final do algoritmo, o submodelo atual;
5. se o submodelo atual não contiver variáveis preditoras, pare o algoritmo e selecione, como resultado final do algoritmo, o submodelo atual; caso contrário, vá para o passo 2.

Tópico 37: método stepwise

- O algoritmo stepwise combina as ações dos algoritmos backward e forward.
- Partindo do modelo sem variáveis preditoras, num dado passo do algoritmo stepwise, uma nova variável preditora é inclusa.
- A variável preditora inclusa é aquela que promove uma maior melhora no desempenho do modelo.
- Em seguida, uma variável preditora é excluída do modelo obtido.
- A variável preditora excluída é aquela cuja exclusão promove uma maior melhora no desempenho do modelo.
- Esse processo é realizado até que as ações de inclusão e exclusão não melhorem o modelo, ou que todas as variáveis tenham sido incluídas, ou que todas as variáveis tenham sido excluídas.

Tópico 37: algoritmo stepwise

- Os passos do algoritmo stepwise são os seguintes:
  1. selecione, como submodelo atual, o submodelo com todas as variáveis preditoras e calcule o seu AIC;
  2. ajuste os submodelos que podem ser obtidos incluindo-se uma nova variável preditora ao submodelo atual e calcule seus AICs;
  3. ajuste todos os submodelos que podem ser obtidos excluindo-se uma variável preditora do submodelo atual e calcule seus AICs;
  4. dentre os submodelos ajustados nos passos 2 e 3, selecione aquele com menor AIC;
  5. se o AIC do submodelo selecionado no passo 4 for menor do que o AIC do submodelo atual, selecione, como submodelo atual, o submodelo selecionado no passo 4 e vá para o passo 6; caso contrário, pare o algoritmo e selecione, como resultado final do algoritmo, o submodelo atual;

Tópico 37: algoritmo stepwise

6. se o submodelo atual tiver todas as p variáveis preditoras, pare o algoritmo e selecione, como resultado final do algoritmo, o submodelo atual; caso contrário, vá para o passo 2.

- Na plataforma **R**, a função `step`, do pacote `stats`, pode ser usada para selecionar variáveis preditoras através dos algoritmos forward, backward e stepwise. A busca exaustiva pode ser realizada utilizando-se a função `glmulti` do pacote `glmulti`.

Tópico 37: método stepwise

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 **possíveis** variáveis preditoras são o número de domicílios, a renda média anual, a idade média dos domicílios, a distância ao concorrente mais próximo e a distância à loja. A busca exaustiva e o algoritmo stepwise são aplicados para determinar quais dessas variáveis preditoras devem fazer parte 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 15.

Tópico 37: método stepwise

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, com a variável **sta** (Sobrevivência) como variável resposta e tendo como **possíveis** variáveis preditoras as outras dezenove variáveis no conjunto de dados. O método da busca exaustiva e o algoritmo stepwise devem ser aplicados para determinar quais dessas variáveis preditoras devem fazer parte do modelo. Todos os detalhes desse exemplo podem ser vistos no jupyter notebook **exemplo_2.ipynb** ou no arquivo **exemplo_2.html**, localizados na pasta da aula 15.

Referências I

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

[2] N. R. Draper and H. Smith, *Applied regression analysis*, 3 ed., John Wiley & Sons, New York, NY, EUA, 1998.

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

[4] E. W. Frees, *Regression modeling with actuarial and financial applications*, Cambridge University Press, New York, NY, EUA, 2009.

Referências II

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

[6] A. J. Miller, *Subset selection in regression*, 2 ed., Chapman & Hall/CRC, Boca Raton, FL, EUA.

[7] J. Neter, W. Wasserman, and M. H. Kutner, *Applied linear regression models*, R. D. Irwin, Homewood, IL, EUA, 1983.

[8] Galit Shmueli, *To Explain or to Predict?*, Statistical Science **25** (2010), no. 3, 289 – 310.

[9] S. Weisberg, *Applied linear regression*, John Wiley & Sons, Hoboken, NJ, EUA, 2005.