

CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

23 de agosto, 2017

Aula 2 - Uma breve revisão sobre modelos lineares

Modelos lineares

- Modelos de regressão são utilizados para modelar a relação entre uma variável aleatória y e um conjunto de variáveis explicativas x_1, x_2, \dots, x_p .
- As variáveis explicativas são incorporadas ao modelo juntamente com um conjunto de parâmetros desconhecidos, que são estimados com base nos dados disponíveis.
- Uma classe de modelos de regressão são os modelos lineares, que podem ser expresso na seguinte forma geral:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon, \quad (1)$$

em que $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ são os parâmetros do modelo e ϵ é o erro, aleatório e não observável, ao qual assumimos $E[\epsilon] = 0$ e $Var[\epsilon] = \sigma^2$.

Modelos lineares

- Vamos denotar o modelo linear por:

$$y = f(\beta; \mathbf{x}) + \epsilon = \mathbf{x}'\beta + \epsilon, \quad (2)$$

em que $\mathbf{x} = (1, x_1, \dots, x_p)'$ é o vetor de variáveis explicativas e $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ é o vetor de parâmetros.

- Importante notar que o termo *linear* se refere à forma como os parâmetros (e não as variáveis explicativas) são inseridos no modelo.
- Assim, um modelo é linear se cada derivada parcial do tipo

$$\frac{\partial f(\beta; \mathbf{x})}{\partial \beta_j} \quad (3)$$

não depender de β_j , $j = 0, 1, 2, \dots, p$.

- Os seguintes preditores definem modelos lineares:

$$f(\boldsymbol{\beta}; \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2; \quad (4)$$

$$f(\boldsymbol{\beta}; \mathbf{x}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3; \quad (5)$$

$$f(\boldsymbol{\beta}; \mathbf{x}) = \beta_0 + \beta_1 \ln x_1 + \beta_2 \left(\frac{1}{x_2} \right); \quad (6)$$

$$f(\boldsymbol{\beta}; \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 + x_2 + \beta_3 x_1 x_2. \quad (7)$$

- Os seguintes preditores definem modelos não lineares:

$$f(\boldsymbol{\beta}; \mathbf{x}) = \beta_0 + \beta_1 \exp\{\beta_1 x_1\}; \quad (8)$$

$$f(\boldsymbol{\beta}; \mathbf{x}) = \frac{\beta_0}{1 + \exp\{\beta_1 x\}}; \quad (9)$$

$$f(\boldsymbol{\beta}; \mathbf{x}) = \beta_0 + \beta_1 x_1 + \sin(\beta_2 + \beta_3 x_2). \quad (10)$$

Representação matricial de modelos lineares

- Considere um conjunto de n observações do tipo $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)$, $\mathbf{x}'_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$.
- A representação matricial de um modelo linear fica dada por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (11)$$

em que N_n representa a distribuição Normal n -variada, \mathbf{I}_n a matriz identidade $n \times n$ e

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \quad (12)$$

Representação matricial de modelos lineares

- Uma representação alternativa de modelos lineares pode ser feita em duas etapas.
- Considere $\mu_1, \mu_2, \dots, \mu_n$, em que $E(y_i|\mathbf{x}_i) = \mu_i$, $i = 1, 2, \dots, n$. Então:

$$\begin{aligned}y_i|\mathbf{x}_i &\sim N(\mu_i, \sigma^2); \\ \mu_i &= \mathbf{x}'_i\boldsymbol{\beta} = \beta_0 + \beta_1x_1 + \dots + \beta_px_p.\end{aligned}\tag{13}$$

- Esta representação é mais flexível e será adotada ao longo da disciplina.

Ajuste do modelo linear pelo método de mínimos quadrados

- O ajuste de um modelo linear via mínimos quadrados baseia-se na determinação de $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ que minimizam a **soma de quadrados dos erros**:

$$SQE(\beta) = \|(\mathbf{y} - \boldsymbol{\mu})\|^2 = \sum_i (y_i - \mu_i)^2 = \sum_{i=1}^n \left(y_i - \sum_j \beta_j x_{ij} \right)^2. \quad (14)$$

- Por se tratar de uma soma de quadrados, a minimização de $SQE(\beta)$ fica determinada pela solução do seguinte conjunto de equações de estimação:

$$\frac{\partial SQE(\beta)}{\partial \beta_j} = 0, \quad j = 0, 1, 2, \dots, p. \quad (15)$$

Ajuste do modelo linear pelo método de mínimos quadrados

- Uma vez que as equações de estimação são lineares com relação aos parâmetros, é possível obter os estimadores de mínimos quadrados de maneira analítica (sem recorrer a métodos numéricos).
- Após alguma álgebra matricial, o estimador de mínimos quadrados de β , denotado por $\hat{\beta}$, fica dado por:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (16)$$

Propriedades dos estimadores de mínimos quadrados em modelos lineares

- $E(\hat{\beta}) = \beta$;
- $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$;
- Na classe de estimadores lineares não viciados, $\hat{\beta}$ tem variância mínima (eficiência);
- Se assumirmos erros com distribuição Normal, então $\hat{\beta}$ tem distribuição Normal:

$$\hat{\beta} \sim N_{p+1} \left(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right), \quad (17)$$

tal que:

$$\hat{\beta}_j \sim N \left(\beta_j, \sigma^2 (\mathbf{x}'_j \mathbf{x}_j)^{-1} \right). \quad (18)$$

Propriedades dos estimadores de mínimos quadrados em modelos lineares

- Dada uma combinação linear dos parâmetros:

$$\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \dots + c_p\beta_p, \quad (19)$$

em que $\mathbf{c}' = (c_0, c_1, \dots, c_p)$ é um vetor de constantes, então o estimador de mínimos quadrados para $\mathbf{c}'\boldsymbol{\beta}$ é $\mathbf{c}'\hat{\boldsymbol{\beta}}$, com:

$$E(\mathbf{c}'\hat{\boldsymbol{\beta}}) = \mathbf{c}'\boldsymbol{\beta}; \quad \text{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2 \quad (20)$$

e, sob a suposição de normalidade,

$$\mathbf{c}'\hat{\boldsymbol{\beta}} \sim N(\mathbf{c}'\boldsymbol{\beta}, \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2). \quad (21)$$

Resultados adicionais sobre a estimação por mínimos quadrados

- A matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é o projetor ortogonal de \mathbf{Y} no espaço coluna de \mathbf{X} , sendo chamada “matriz chapéu” (*hat matrix*)
- Pelo teorema de Pitágoras:

$$\|\mathbf{Y}\|^2 = \|\hat{\mathbf{Y}}\|^2 + \|\hat{\boldsymbol{\epsilon}}\|^2, \quad (22)$$

ou seja, o vetor de observações pode ser decomposto na soma de dois vetores ortogonais: o vetor $\hat{\mathbf{Y}}$ do vetor estimação e o vetor $\hat{\boldsymbol{\epsilon}}$ do espaço resíduo.

Resultados adicionais sobre a estimação por mínimos quadrados

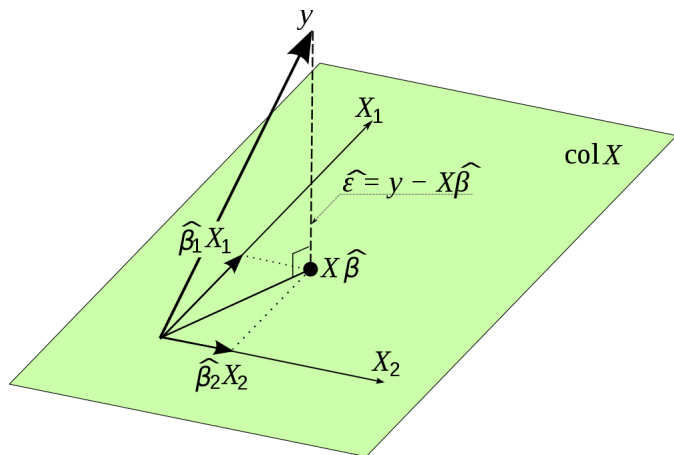


Figura 1: Projeção ortogonal de Y no espaço coluna de X

Resultados adicionais sobre os estimadores de mínimos quadrados em modelos lineares

- Novamente sob a suposição de normalidade dos erros, os estimadores de mínimos quadrados são equivalentes aos de máxima verossimilhança;
- Se a matriz do modelo (\mathbf{X}) não tem posto completo, então $(\mathbf{X}'\mathbf{X})^{-1}$ também não tem;
- Conseqüentemente, o sistema de equações de estimação admite infinitas soluções, não existindo estimador de mínimos quadrados ($\hat{\beta}$).
- A solução nesse caso é considerar uma matriz inversa generalizada para $(\mathbf{X}'\mathbf{X})^{-1}$, o que implica no uso de restrições para os parâmetros.

Inferência estatística em modelos lineares

- A inferência estatística em modelos lineares tem como principais objetivos estimar e testar hipóteses sobre os parâmetros, bem como obter predições.
- Inicialmente, vamos tratar da inferência para um particular parâmetro β_j do modelo.
- Testes de hipóteses e intervalos de confiança para os parâmetros do modelo podem ser obtidos a partir do seguinte resultado:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{var}(\hat{\beta}_j)}} \sim t_{n-p-1}, \quad (23)$$

para $j = 0, 1, 2, \dots, p$, em que t_ν representa a distribuição t -Student com ν graus de liberdade.

Inferência estatística em modelos lineares

- Assim, um intervalo de confiança $100(1 - \alpha)\%$ para β_j fica dado por:

$$\hat{\beta}_j \pm t_{n-p-1; \alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta}_j)} = \hat{\beta}_j \pm t_{n-p-1; \alpha/2} \sqrt{\hat{\sigma}^2 (\mathbf{x}'_j \mathbf{x}_j)^{-1}}, \quad (24)$$

em que $\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-p-1}$.

- De maneira similar o teste de $H_0 : \beta_j = \beta_{j0}$ versus $H_1 : \beta_j \neq \beta_{j0}$, sendo β_{j0} um valor postulado para β_j , baseia-se na estatística:

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{\sigma}^2 (\mathbf{x}'_j \mathbf{x}_j)^{-1}}}, \quad (25)$$

rejeitando-se H_0 , ao nível de significância α , se $|t| > t_{n-p-1; 1-\alpha/2}$.

- Vamos considerar agora o teste da hipótese de nulidade conjunta dos parâmetros do modelo:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0. \quad (26)$$

- Esse teste baseia-se na partição da variabilidade total dos dados, conforme pode ser apresentado num quadro de análise de variância.

Tabela 1: Análise de variância

Fonte	Soma de Quadrados	gl	Quadrado médio	F
Regressão	$SQT - SQRes$	p	$\frac{SQReg}{p}$	$\frac{QMReg}{QMRes}$
Resíduos	$\sum_i (y_i - \hat{y}_i)^2$	$n-p-1$	$\frac{SQRes}{n-p-1}$	
Total	$\sum_i (y_i - \bar{y})^2$	$n-1$		

- Sob a hipótese nula, a estatística F tem distribuição F-Snedecor com parâmetros p e $n - p - 1$.
- Para um nível de significância α , H_0 deve ser rejeitada se $F > F_{p, n-p-1; 1-\alpha/2}$.

- Outra possibilidade é o teste da nulidade conjunta de um subconjunto de parâmetros:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_q = 0, \quad 1 \leq q \leq p. \quad (27)$$

- O teste de H_0 baseia-se nos resultados dos ajustes “completo” (com $p+1$ parâmetros) e “reduzido” (com $q+1$ parâmetros):

$$\begin{aligned} \text{Modelo reduzido: } & y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_q x_{iq} \\ \text{Modelo completo: } & y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \end{aligned} \quad (28)$$

Inferência estatística em modelos lineares

- Sejam $SQReg_0$ e $SQReg_1$ as somas de quadrados de regressão dos modelos reduzido e completo, respectivamente.
- Sob a hipótese nula (de nulidade conjunta do subconjunto de parâmetros), a estatística F :

$$F = \frac{(SQReg_1 - SQReg_0)/q}{SQReg_1/(n - p - 1)} \quad (29)$$

tem distribuição F-Snedecor com q e $n - p - 1$ graus de liberdade, fundamentando o teste da hipótese.

Inferência estatística em modelos lineares

- Intervalos de confiança e testes de hipóteses para uma combinação linear $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \dots + c_p\beta_p$ podem ser feitos com base na distribuição t_{n-p-1} . Começando pelo IC $100(1 - \alpha)\%$:

$$\mathbf{c}'\hat{\boldsymbol{\beta}} \pm t_{n-p-1;1-\alpha/2} \sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\hat{\sigma}^2}, \quad (30)$$

e, para o teste bilateral de $H_0 : \mathbf{c}'\boldsymbol{\beta} = 0$, a hipótese é rejeitada se

$$t = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\hat{\sigma}^2}} > t_{n-p-1;1-\alpha/2}, \quad (31)$$

para um nível de significância α .

Diagnóstico do ajuste

- O diagnóstico do ajuste de um modelo de regressão é uma etapa fundamental da análise, tendo como objetivos:
 - Avaliar se o modelo proposto, de maneira geral, se ajusta bem aos dados;
 - Checar se as pressuposições do modelo são atendidas;
 - Identificar quais as causas de possível falta de ajuste e medidas corretivas apropriadas;
 - Identificar *outliers* e pontos influentes. Estudar o impacto desses pontos no ajuste do modelo.

Diagnóstico do ajuste

- Dentre as principais ferramentas para diagnóstico do ajuste, destacam-se:
 - Métodos gráficos;
 - Medidas de qualidade de ajuste;
 - Testes de hipóteses;
 - Medidas de qualidade preditiva.

Resíduo

Medida da diferença entre valores observados de uma variável e os correspondentes valores ajustados por um modelo.

- Resíduo ordinário:

$$r_i = y_i - \hat{y}_i, \quad (32)$$

sendo \hat{y}_i o valor ajustado pelo modelo para a i -ésima observação, $i = 1, 2, \dots, n$.

Nota: Os resíduos ordinários não têm variância constante, comprometendo sua utilização no diagnóstico do ajuste. Versões padronizadas são recomendadas.

Diagnóstico do ajuste - Análise de resíduos

- Resíduo padronizado:

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma} \sqrt{1 - h_i}}, \quad i = 1, 2, \dots, n, \quad (33)$$

sendo h_i o i -ésimo elemento da diagonal de $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

- Resíduo studentizado:

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_i}}, \quad i = 1, 2, \dots, n, \quad (34)$$

em que $\hat{\sigma}_{(-i)}$ é a estimativa de σ obtida sem considerar a i -ésima observação (*leave one out*).

Diagnóstico do ajuste - Alguns gráficos para análise de resíduos

- **Gráfico quantil-quantil normal:** permite avaliar a pressuposição de normalidade, avaliar a forma da distribuição em caso de não normalidade, identificar *outliers*;
- **Resíduos vs valores ajustados:** investigar padrões não aleatórios, variância não homogênea, presença de *outliers* e potenciais pontos influentes;
- **Resíduos vs ordem de coleta (no tempo, no espaço, . . .):** avaliar possível dependência relacionada à ordem de coleta;
- **Resíduos versus variáveis explicativas:** detectar possível falta de ajuste em relação às variáveis explicativas inseridas no modelo;
- **Resíduos versus variáveis não incluídas no modelo:** verificar se há variáveis não incluídas no ajuste que deveriam ser consideradas. . .

Objetivo

Medir o impacto de cada observação no ajuste global (ou em componentes) do modelo.

- **Leverage** h_i : Medida de distância da i -ésima observação, no espaço das variáveis explicativas, ao centróide das demais observações;
- **Distância de Cook**: Medida de diferença das estimativas dos parâmetros do modelo ao considerar e ao desconsiderar uma particular observação no ajuste;

Medidas de influência

- **DFFITS:** Medida de diferença dos valores ajustados para uma particular observação ao considerar e ao desconsiderar essa observação no ajuste;
- **DFBETAS:** Medida de diferença das estimativas dos parâmetros do modelo (avaliados um a um) ao considerar e ao desconsiderar uma particular observação no ajuste;
- **COVRATIO:** Medida de alteração na precisão das estimativas dos parâmetros do modelo ao considerar e ao desconsiderar uma particular observação no ajuste.

Nota: Observe que as medidas de influência usam a estratégia *leave one out*. Para a análise, pode-se construir gráficos dos valores de uma particular medida vs o índice da observação.

- O pacote **car** disponibiliza diversas funções para diagnóstico do ajuste, com diferentes gráficos para resíduos e medidas de influência;
- Os pacotes **effects** e **lsmeans** dispõem recursos para explorar os efeitos das variáveis usadas no ajuste do modelo e produção de inferências;
- Usaremos pacotes adicionais, nas aulas práticas, para seleção de covariáveis e teste da qualidade do ajuste, dentre outros.

- Vamos trabalhar com três exemplos, com scripts disponíveis na página da disciplina:
 - 1 Análise da viscosidade de um polímero segundo a temperatura e a taxa de alimentação do catalisador em uma reação química;
 - 2 Vendas de um produto sob quatro diferentes tipos de embalagens;
 - 3 Total em vendas de representantes de uma marca de cosméticos segundo a idade, tempo de escolaridade, anos de experiência e tamanho da população atendida.