

Introdução aos Modelos Lineares Generalizados

- Introdução

Generalização dos modelos lineares, incluem entre outros:

- { de regressão linear
- de análise de variância
- logit e probit
- log-lineares e resposta multinomial
- usados em análise de sobrevivência

- Propriedades comuns permitem tratá-los como uma única classe
→ MLG.

História

- Início do século XIX : Gauss e Legendre

Dados sobre Astronomia (contínuos).

Modelos Lineares Clássicos e Mínimos Quadrados

Erros normalmente distribuídos para descrever variabilidade.

Muitas propriedades dos estimadores não dependem de normalidade, mas sim de variância constante e independência (propriedades de segunda ordem).

Uma propriedade similar aplica-se a todos os MLG's.

História e algumas aplicações

- Século XVIII

Cálculo de probabilidades para várias configurações de jogos de cartas e dados \Rightarrow métodos para lidar com dados na forma de contagem de eventos.

Desenvolvimento de métodos para tratar eventos discretos em vez de quantidades contínuas.

No contexto de eventos raros \Rightarrow distribuição de Poisson.

Outros modelos relacionados \Rightarrow Bernoulli e binomial (dados na forma de proporções ou razões de contagens).

- Experimentos clínicos o interesse principal não está na incidência de uma doença, mas sim como ela é afetada por fatores tais como idade, classe social, condições de moradia, exposição a poluentes e a quaisquer tratamentos sob estudo.

Algumas vantagens

- Medidas contínuas não-normais podem ser incluídas para análise na classe dos MLG's.
- Quantidades positivas com distribuição assimétrica à direita (ex. estudos de tempos de sobrevivência) podem ser modeladas, por exemplo, pela distribuição exponencial ou gama.

O Problema de analisar os dados

- DADOS \Rightarrow suponha que tenhamos uma série de medidas associadas com algumas informações contextuais

Ordem na qual os dados foram coletados

Instrumentos de medida que foram utilizados e outras diferenças nas condições sob as quais as medidas individuais foram feitas.

- INTERPRETAÇÃO

\Rightarrow Busca-se por um padrão de comportamento (por exemplo, um instrumento pode ter produzido sistematicamente leituras maiores do que outro).

Estes padrões são chamados efeitos sistemáticos. Tais efeitos podem ser mascarados por outra variação de natureza mais aleatória. Esta variação é geralmente descrita em termos estatísticos.

Modelos Estatísticos

- Efeitos sistemáticos e efeitos aleatórios
- Valor do modelo \Rightarrow sumário simples dos dados em termos dos efeitos sistemáticos junto com um sumário da natureza e magnitude da variação não explicada.
- Analisar inteligentemente os dados

Formulação de padrões capazes de descrever sucintamente não somente a variação sistemática nos dados sob consideração, como também os padrões em dados similares que podem ser coletados por outro investigador em outra época e em outro lugar.

Teoria como padrão

- Teorias como geradoras de observações numéricas

Podem "substituir" os dados

Podem ser descritas em termos de um número menor de quantidades (*parâmetros*).

Fornecendo-se os valores dos parâmetros, padrões específicos podem ser gerados.

- Exemplo: Seja o modelo:

$$y = \alpha + \beta x$$

Relaciona duas quantidades y e x por meio do par de parâmetros (α, β)

Define uma relação "afim" entre y e x .

Modelo

- Suponha que exista uma relação casual entre x e y na qual x está sob controle (fixo) e afeta y que pode ser medido (idealmente) sem erro.

Então, se x assumir os valores

$$x_1, x_2, \dots, x_n$$

y toma os valores

$$\alpha + \beta x_1, \alpha + \beta x_2, \dots, \alpha + \beta x_n$$

para os valores designados à α e β .

Se conhecemos α e β , podemos reconstruir os valores de y exatamente a partir dos valores dados de x .

\Rightarrow o par (α, β) é um sumário exato de y_1, \dots, y_n .

Modelo

- Na prática nunca conseguiremos medir os y 's exatamente, tal que a relação entre y e x é apenas aproximadamente linear.
- Apesar desta falta de exatidão, ainda podemos escolher valores de α e β , a e b , que de alguma forma melhor representem a relação aproximadamente linear entre y e x .
- As quantidades $a + bx_1, \dots, a + bx_n$, que denotaremos por $\hat{y}_1, \dots, \hat{y}_n$ ou $\hat{\mu}_1, \dots, \hat{\mu}_n$ são os valores ajustados pelo modelo e os dados.
- Estas quantidades não reproduzem y_1, \dots, y_n exatamente.
- O padrão que elas representam aproxima os valores dos dados e podem ser sumariados pelo par (a, b) .

O ajuste do modelo

- Escolha no conjunto de todos os pares possíveis de valores dos parâmetros, um particular par (a, b) que torna o conjunto $\hat{y}_1, \dots, \hat{y}_n$ o mais "próximo" possível dos dados observados.
- Precisamos de uma medida de "proximidade" ou, alternativamente, de distância ou discrepância entre os y 's observados e os \hat{y} 's ajustados.
- Exemplos de tais funções de discrepância incluem
 - (1) Norma L_1 : $S_1(y, \hat{y}) = \sum_i |y_i - \hat{y}_i|$
 - (2) Norma L_∞ : $S_\infty(y, \hat{y}) = \max_i |y_i - \hat{y}_i|$
 - (3) Norma L_2 : $S_2(y, \hat{y}) = \sum_i (y_i - \hat{y}_i)^2$

Fórmulas de Discrepâncias

- Implicações

1) Soma de desvios individuais tais como $|y_i - \hat{y}_i|$ ou $(y_i - \hat{y}_i)^2$, onde cada termo depende de somente uma observação, implica que as observações são todas feitas sobre a mesma escala física e sugere que as observações são independentes, ou pelo menos permutáveis.

2) O uso de diferenças aritméticas $y_i - \hat{y}_i$ implica que um dado desvio carrega o mesmo peso qualquer que seja o valor de \hat{y} .

- Na terminologia estatística, a adequação das normas L_p como medidas de discrepância depende da independência estocástica e também da suposição de que a variância de cada observação seja independente da sua média.

- Estas suposições embora comuns e frequentemente razoáveis na prática não são universalmente aplicáveis.

Fórmulas de Discrepâncias

- As funções de discrepância podem ser justificadas em termos puramente estatísticos.
- Exemplo: o critério de mínimos quadrados surge se considerarmos os valores x como fixos ou não estocásticos e os valores y são supostos ter distribuição normal com média μ na qual

$$f(y|\mu) \propto \exp\left\{-\frac{1}{2} \left(\frac{y-\mu}{\sigma}\right)^2\right\} \quad (1)$$

onde μ é linearmente relacionado com x por meio dos coeficientes α e β .

- O fator de escala σ , que é o desvio padrão de y , descreve o "comprimento" dos erros quando medidos sobre o valor médio.

Critério dos Mínimos Quadrados

- Podemos interpretar a função (1) de duas formas.
 - (1) Como uma função de y para μ fixado, ela especifica a densidade de probabilidade das observações.
 - (2) Para um dado y , podemos interpretá-la como uma função de μ para o particular valor de y observado.
- Esta segunda interpretação, é conhecida como a função de verossimilhança (Fisher).
- Note $-2l$ é igual a

$$\frac{1}{\sigma^2} \sum_i (y_i - \mu_i)^2$$

- Em outras palavras, exceto pelo termo σ^2 , aqui suposto conhecido, $-2l$ é idêntica ao critério da soma de quadrados. A medida que μ varia, $-2l$ assume seu valor mínimo em $\mu = \bar{y}$.

Função de Verossimilhança

- Para um modelo mais complicado, no qual μ varia de forma sistemática de observação para observação, definimos $\hat{\mu}$ como o conjunto de valores que maximizam a verossimilhança ou, equivalentemente, que minimizam $-2l$.
- Mais geralmente, podemos estender o interesse além do único ponto que minimiza $-2l$ para a forma da superfície da verossimilhança na vizinhança do mínimo.
- Esta forma nos diz, na terminologia usada por Fisher, quanta informação referente aos parâmetros existe nos dados.
- Voltando ao exemplo de uma relação linear, podemos fazer um gráfico com eixos α e β , dos contornos de igual discrepância $-2l$ para os dados y .

Função Verossimilhança

- Neste exemplo particular, $-2l$ é uma função quadrática de (α, β) e os contornos são elipses similares na forma e na orientação, com a estimativa de máxima verossimilhança (a, b) correspondendo ao centro.
- A informação nos dados sobre os parâmetros é fornecida pela matriz de curvatura ou matriz Hessiana da função quadrática.
- Se os eixos das elipses não estão alinhados com os eixos (α, β) , então as estimativas são correlacionadas.
- A informação é maior na direção para a qual a curvatura é maior. Em certas circunstâncias, a forma da superfície de informação pode ser determinada antes do experimento ser realizado.

Adequação de um modelo

- Seja um modelo escolhido que satisfaça algum critério de "otimalidade", por exemplo minimizar a soma de quadrados $\sum_i (y_i - \hat{\mu}_i)^2$
- Pode parecer que um bom modelo é aquele que ajusta os dados observados da melhor maneira possível, isto é, que torna $\hat{\mu}$ o mais próximo possível de y .
- Incluindo um número suficiente de parâmetros no nosso modelo, podemos tornar o ajuste cada vez melhor e, até chegar a um ajuste perfeito.
- Porém, fazendo isto, não alcançamos nenhuma redução na complexidade – não produzimos nenhum padrão teórico simples para os dados.

Adequação de um modelo

- Simplicidade: representada pela parcimônia – característica desejável de qualquer modelo.
- Uma outra característica importante de um modelo é a sua "abrangência", isto é, o conjunto de condições sobre as quais ele fornece boas previsões.
- Abrangência e parcimônia são de alguma forma relacionadas.
- Modelo construído de forma a se ajustar muito bem para um particular conjunto de dados pode não ser capaz de incorporar mudanças inevitáveis que podem ser necessárias quando outro conjunto de dados relacionado ao mesmo fenômeno for coletado.

Modelagem: Princípios

- (I) Todos os modelos estão errados, alguns porém, são mais úteis do que outros e devemos buscar por estes modelos.
- (II) Não se deve "apaixonar" por um modelo excluindo modelos alternativos. Os dados, com frequência, apontarão com igual ênfase para vários modelos possíveis.
- (III) Deve-se realizar verificações completas sobre o ajuste de um modelo para os dados, por exemplo, usando resíduos e outras estatísticas derivadas do ajuste para procurar por *outliers* e desvios das suposições do modelo.

Modelos Lineares Clássicos

- Teve início com a análise de dados sobre Astronomia
- Modelos Lineares Clássicos: $EY = X\beta$
- Legendre(1805) \Rightarrow Mínimos quadrados
- Gauss (1809) $\Rightarrow N(0, \sigma^2)$ para os erros
- Gauss (1823) retirou a suposição de normalidade mantendo apenas a suposição de variância constante \Rightarrow estimadores de β pelo critério de mínimos quadrados têm variância mínima dentro da classe de estimatidores não tendenciosos.
- A extensão desta suposição mais fraca para MLG's foi feita por Wedderburn (1974) \Rightarrow conceito de quasi-verossimilhança.

Modelos Fatoriais Ensaios Biológicos

- Modelos Fatoriais - Fisher (1919)

A diferença está na matriz do modelo X que aqui é composta por zeros e uns.

A influência de Fisher foi além dos modelos fatoriais e inclui modelos especiais para a análise de certos tipos de contagens e proporções.

- Ensaios de diluição (“dilution assay”) – Fisher (1922)

Uma solução contendo um organismo infectado é progressivamente diluída. Uma estimativa da concentração de organismos infectados na solução original é feita. Supondo que as diluições são feitas em potências de 2, depois de x diluições, o número de organismos infectados ρ_x por unidade de volume é

$$\rho_x = \frac{\rho_0}{2^x}, \quad x = 0, 1, 2, \dots$$

Ensaio de diluição

- Onde ρ_0 é a densidade de organismos infectados na solução original.
- O número esperado de organismos infectados em cada diluição de volume ν é $\rho_x \nu$.
- Sob condições apropriadas, o número de organismos infectados segue uma distribuição de Poisson com parâmetro $\rho_x \nu$.
- A probabilidade de uma diluição estar infectada é

$$\pi_x = 1 - \exp\{-\rho_x \nu\}$$

Segue que na diluição x : $\log(\log(1 - \pi_x)) = \log(\rho_x \nu) = \log(\nu) + \log(\rho_0) - x \log(2)$.

$\rightarrow \alpha = \log(\nu) + \log(\rho_0)$ e $\beta = -\log(2)$ (conhecido).

Ensaio de diluição

- Se para uma diluição no nível x tivermos y organismos infectados, num total de m , a proporção y/m pode ser considerado como sendo uma variável aleatória Y satisfazendo:

$$E(Y \mid x) = \pi_x$$

- Note que a transformação $\eta = \log(-\log(1 - \pi_x)) = \alpha + \beta x$ é uma função linear de x ao invés de $E(Y \mid x)$

Modelos “Probito”

- Bliss (1935)

Animais são divididos em grupos. Cada grupo j recebe dose x_j .
 y_j dos n_j do j -ésimo grupo sobrevivem.

Estimar π_{x_j} probabilidade de sobrevivência na dose x_j .

Modelo: $\pi_x = \phi(\alpha + \beta x)$ onde $\phi(\cdot)$ representa a fda. da $N(0, 1)$.

Outros Exemplos

- Modelos “Logit” para proporções:

$$\log \frac{\pi}{1-\pi} = \text{logit}\pi = \alpha + \beta_1 x_1 + \dots + \beta_p x_p.$$

- Modelos Log-lineares para contagens
- Polinômios Inversos

$$\frac{x}{y} = \alpha + \beta x$$

- Modelos para Dados de sobrevivência

Estudo onde procura-se relacionar tempo de falha de unidades observacionais com covariáveis.

Processos no Ajuste de Modelos

[Seleção] → [Estimação] → [Crítica] → [Previsão]

- Seleção: Características dos MLG's:
 - (i) Independência das observações (autocorrelação não é permitida).
 - (ii) Existência de um único erro para cada observação.
- A escolha da escala para análise é um um aspecto importante da seleção do modelo. “O que caracteriza uma boa escala?”.
- Escolha das covariáveis. (Parcimônia)
- Estimação: A escolha dos valores dos parâmetros desconhecidos é feita segundo algum critério de “otimalidade” (de bondade do ajuste - *goodness-of-fit*). Por exemplo, podemos escolher os estimadores de máxima-verossimilhança.

Formulação de um Modelo Linear

- Modelos lineares clássicos:
 - (i) Componente aleatória de Y : normais e independentes com $EY = \mu$ tal que $\mu = X\beta$ e variância constante.
 - (ii) Componente sistemática: covariáveis x_1, \dots, x_p produzem um preditor linear η dado por $\eta = \sum_j x_j \beta_j$.
 - (iii) A ligação entre as componentes sistemática e aleatória $\eta = \mu$.

Componentes de um MLG

- MLG's permitem duas extensões:

A distribuição na comp. (i) pode ser da família exponencial.

A função de ligação na comp. (iii) pode ser qualquer função monotonicamente diferenciável.

- Se escrevemos $\eta_i = g(\mu_i)$ então $g(\cdot)$ é chamada função de ligação.
- Nesta formulação, modelos lineares clássicos têm uma distribuição normal para a comp. (i) e a função de ligação para a comp. (iii) é a função identidade.

Programa da disciplina e cronograma das aulas

- Na primeira metade do curso abordaremos apenas os casos clássicos de modelos lineares sob o ponto de vista da álgebra linear: modelo de regressão linear simples, a teoria geral para modelos lineares, modelos de análise de variância a um fator e a uma amostra, modelos de regressão múltipla (incluindo regressão polinomial e regressão ponderada). Finalmente falaremos um pouco sobre a análise de resíduos.
- A referência principal é o texto
Jørgensen, Bent. (1993). *The theory of Linear Models*. Chapman & Hall.

Programa da disciplina

- Na segunda metade do curso abordaremos os casos não clássicos de MLG's incluindo modelos para dados binários (Cap.4), modelos para dados politômicos (Cap.5), modelos log-lineares (Cap.6), modelos para dados com coeficiente de variação constante (Cap.8). Voltaremos também à questão de verificação de modelos (Cap.12).
- A referência principal é o texto

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Second Edition. Chapman & Hall.

Ao longo do curso outras referências complementares serão indicadas.

MODELOS LINEARES GENERALIZADOS

Referências principais

1. Jørgensen, B., (1993). *The Theory of Linear Models*. Chapman and Hall.
2. McCullagh, P. and Nelder, J. A., (1989). *Generalized Linear Models*. Second edition, Chapman and Hall.

Avaliação

1. Provas
2. Listas de exercícios.

Introdução ao Modelo de Regressão Linear Simples

- Modelos lineares \Rightarrow estudo da explicação de uma dada variável em termos de uma combinação linear de um conjunto de variáveis explicativas dadas.
- Caso particular: uma variável explicativa.

Considere um experimento no qual fazemos medições simultâneas de duas variáveis x e y . Se n medições são feitas, sejam

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$

os n pares de observações correspondentes.

Modelo de Regressão Simples: Nomenclatura

- Modelo estatístico para a situação onde a relação entre x e y possa ser pensada como linear ou, aproximadamente linear.

$\Rightarrow x$ representa as condições experimentais, e y representa o resultado do experimento.

y é chamada variável resposta ou variável dependente.

$\Rightarrow y_1, y_2, \dots, y_n$ são realizações independentes das variáveis aleatórias Y_1, Y_2, \dots, Y_n .

$\Rightarrow x_1, x_2, \dots, x_n$ são considerados constantes (não-aleatórias), e x é chamada variável explicativa ou variável independente.

A distinção entre as variáveis resposta e explicativa deve ser clara em análise de regressão.

Modelo de Regressão Simples: Nomenclatura

- Se x_1, x_2, \dots, x_n são realizações das variáveis aleatórias X_1, X_2, \dots, X_n , podemos pensar x_1, x_2, \dots, x_n como valores fixados, no sentido de considerar a distribuição condicional de $Y_1, Y_2, \dots, Y_n \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$.
- O primeiro passo na análise é construir o diagrama de dispersão de y versus x .
- Se o gráfico mostra uma relação aproximadamente linear entre as duas variáveis, podemos propor um modelo estatístico adequado para esta relação.

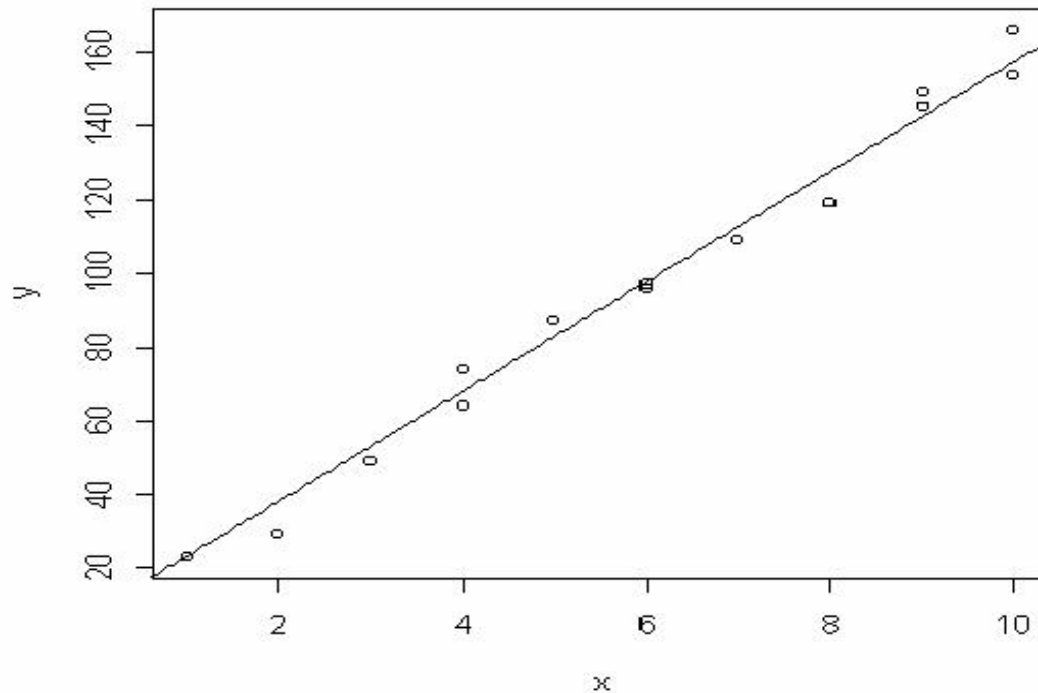
O Modelo Linear Clássico

- Sejam Y_1, Y_2, \dots, Y_n variáveis aleatórias independentes, tal que y_1, y_2, \dots, y_n sejam realizações de Y_1, Y_2, \dots, Y_n , respectivamente.
- Suponha que x_1, x_2, \dots, x_n sejam constantes e que $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, n$, onde as médias $\mu_i = EY_i$ seguem o modelo

$$\mu_i = \beta_1 + \beta_2 x_i, \quad i = 1, \dots, n.$$

- Os parâmetros do modelo são $(\beta_1, \beta_2, \sigma^2)$, com domínio $\mathbb{R}^2 \times \mathbb{R}_+$.
- Resumindo
 - i. A média de Y_i é uma função linear de x_i .
 - ii. As variáveis Y_1, Y_2, \dots, Y_n , são independentes.
 - iii. A variância de Y_i é constante.
 - iv. A distribuição de Y_i é normal.

Figura 1: *Diagrama de Dispersão de Y vs. X*



Inspeção das hipóteses do modelo

- 1) Diagrama de dispersão de y versus $x \Rightarrow$ julgar (i) linearidade e (iii) homogeneidade das variâncias. A homogeneidade das variâncias implica que a dispersão vertical dos pontos é a mesma para qualquer valor de x .
- 2) Normal plot dos resíduos \Rightarrow para avaliar a normalidade.

A suposição de independência (ii) é mais difícil de ser verificada na prática e é quase sempre suposta em função da informação do experimento.

Condições que podem levar a violação desta suposição são, por exemplo, a existência de grupos de observações feitas sobre circunstâncias muito similares, diferentes de grupo para grupo, ou outras fontes de correlação, como por exemplo, correlações seriais ou espaciais.

Estimação dos parâmetros

- Estimadores de máxima-verossimilhança de β_1 , β_2 e σ^2 .
- Temos que Y_1, Y_2, \dots, Y_n são variáveis aleatórias independentes com distribuição $N(\beta_1 + \beta_2 x_i, \sigma^2)$.
- Função de verossimilhança $L(\beta_1, \beta_2, \sigma^2)$:

$$\begin{aligned} L(\beta_1, \beta_2, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2\right\} \end{aligned} \quad (1)$$

Maximizar a função de verossimilhança é equivalente a maximizar o log da verossimilhança l :

$$\begin{aligned} l(\beta_1, \beta_2, \sigma^2) &= \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2 \end{aligned}$$

Estimador de máxima verossimilhança

- Os estimadores de máxima verossimilhança $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$ é o valor de $(\beta_1, \beta_2, \sigma^2)$ que maximiza l .
- Este valor pode ser obtido como a solução das equações de verossimilhança:

$$\frac{\delta l}{\delta \beta_1} = 0, \quad \frac{\delta l}{\delta \beta_2} = 0, \quad \frac{\delta l}{\delta \sigma^2} = 0$$

- Porém, vamos dividir este problema de maximização em subproblemas.

Observe, da equação (1) que o log. da verossimilhança tem a forma

$$l(\beta_1, \beta_2, \sigma^2) = c(\sigma^2) - \frac{1}{2\sigma^2} D(\beta_1, \beta_2)$$

- Onde:

$$c(\sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) \text{ e,}$$

$D(\beta_1, \beta_2) = \sum_i (y_i - \beta_1 - \beta_2 x_i)^2$ é a soma dos quadrados dos desvios que chamaremos “deviance”.

- Para um dado valor de σ^2 , o valor de (β_1, β_2) que maximiza l também minimiza $D(\beta_1, \beta_2)$.
- Em particular, o valor mínimo é o mesmo qualquer que seja o valor de σ^2 .
- Seja $t_i = x_i - \bar{x}_+$, $i = 1, \dots, n$ e defina um novo parâmetro α por $\alpha = \beta_1 + \beta_2 \bar{x}_+$, onde
 $x_+ = x_1 + \dots + x_n$ e $\bar{x}_+ = x_+/n$.
- Usaremos a mesma notação para y tal que $y_+ = y_1 + \dots + y_n$ e $\bar{y}_+ = y_+/n$.

Estimador de máxima verossimilhança

- O modelo pode então ser escrito como

$$\mu_i = \alpha + \beta_2 t_i \quad (2)$$

- Observe que $t_+ = t_1 + \dots + t_n = 0$.
- Defina $\tilde{D}(\alpha, \beta_2) = D(\beta_1, \beta_2)$.
- Minimizar D é equivalente a minimizar \tilde{D} .

$\tilde{D}(\alpha, \beta_2) = \sum_{i=1}^n (y_i - \alpha - \beta_2 t_i)^2$ tal que

$$\frac{\delta \tilde{D}}{\delta \alpha} = -2 \sum_{i=1}^n (y_i - \alpha - \beta_2 t_i) = -2(y_+ - n\alpha) \text{ e}$$

$$\frac{\delta \tilde{D}}{\delta \beta_2} = -2 \sum_{i=1}^n t_i (y_i - \alpha - \beta_2 t_i) = -2(S_{ty} - \beta_2 S_t),$$

onde $S_{ty} = \sum_i y_i t_i$, $S_t = \sum_i t_i^2$.

Estimador de máxima verossimilhança

- Segue que a solução das equações

$\frac{\delta \tilde{D}}{\delta \alpha} = 0$ e $\frac{\delta \tilde{D}}{\delta \beta_1} = 0$ é dada por

$$\hat{\alpha} = \bar{y}_+ \quad \text{e} \quad \hat{\beta}_2 = \frac{S_{ty}}{S_t} \quad (3)$$

- A matriz das derivadas parciais de ordem 2 de \tilde{D} é $2 \begin{pmatrix} n & 0 \\ 0 & S_t \end{pmatrix}$ que é positiva definida se $S_t > 0$.
- Assim, se $S_t > 0$, a equação (3) nos fornece o único valor de (α, β_2) que minimiza $\tilde{D}(\alpha, \beta_2)$.
- Como $\beta_1 = \alpha - \beta_2 \bar{x}_+$, segue que o EMV de β_1 é $\hat{\beta}_1 = \bar{y}_+ - \hat{\beta}_2 \bar{x}_+$.

Estimador de máxima verossimilhança

- Destes resultados, segue que l é maximizada com respeito à (β_1, β_2) para um dado valor de σ^2 com
$$\tilde{l}(\sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{D(\hat{\beta}_1, \hat{\beta}_2)}{2\sigma^2}.$$
- Esta equação é chamada perfil do log. da verossimilhança para σ^2 pois pode ser visualizada como o perfil do gráfico da função $l(\beta_1, \beta_2, \sigma^2)$ quando analisada ao longo do eixo σ^2 .
- Assim, para encontrar o EMV de σ^2 resolvemos a equação $\tilde{l}'(\sigma^2) = 0$.
- Como candidato a EMV temos então $\hat{\sigma}^2 = \frac{1}{n}D(\hat{\beta}_1, \hat{\beta}_2)$ e, como $\tilde{l}''(\sigma^2) < 0$, segue que $\hat{\sigma}^2$ é o EMV de σ^2 .

Propriedades do EMV

- Os EMV's de β_1 , β_2 e σ^2 são

$$\hat{\beta}_2 = \frac{s_{ty}}{s_t}, \hat{\beta}_1 = \bar{y}_+ - \hat{\beta}_2 \bar{x}_+ \text{ e}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2.$$

- Suficiência

Se (Y_1, \dots, Y_n) segue um modelo estatístico caracterizado por um vetor de parâmetros θ , uma estatística $T = t(Y_1, \dots, Y_n)$ é dita ser suficiente para θ , se a distribuição condicional de Y dado T não envolve θ .

Pelo critério de fatoração de Neyman-Fisher, T é suficiente para θ se, e somente se, a densidade de probabilidade de Y pode ser escrita na forma $f_\theta(y) = g_\theta\{t(y)\}h(y)$, onde g_θ e h são funções apropriadas tal que h não depende de θ e $g_\theta\{t(y)\}$ depende de y somente através de $t(y)$.

Propriedades do EMV

- Isto é equivalente a dizer que o log. da verossimilhança tem a forma

$$l(\theta) = \log f_{\theta}(y) = a_{\theta}(t(y)) + b(y)$$

para funções a_{θ} e b adequadas.

- O log. da verossimilhança para α , β_2 e σ^2 no modelos de regressão linear simples é $l(\alpha, \beta_2, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta_2 t_i)^2$

- Assim, escrevendo μ_i como $\alpha + \beta_2 t_i$, tem-se $l(\alpha, \beta_2, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^2 + \mu_i^2 - 2y_i \mu_i) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{S_y}{2\sigma^2} - \frac{S_{\mu}}{2\sigma^2} + \frac{\alpha}{\sigma^2} y_+ + \frac{\beta_2}{\sigma^2} S_{ty}$.

- Assim, encontramos que a estatística

$T(S_y, y_+, S_{ty})$ é suficiente para $(\alpha, \beta_2, \sigma^2)$.

Propriedades do EMV

- Os EMV's dos parâmetros são sempre funções da estatística suficiente T . Neste caso,

$$\hat{\alpha} = \bar{Y}_+ \text{ e } \hat{\beta}_2 = \frac{S_{ty}}{S_t} \text{ e } \hat{\sigma}^2 = \frac{1}{n}(S_y - n\hat{\alpha}^2 - S_t\hat{\beta}_2^2).$$

- Como há uma relação 1 a 1 entre T e o vetor de EMV's $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$, este último também é uma estatística suficiente.
- Distribuição dos coeficientes de regressão

Tanto $\hat{\alpha}$ como $\hat{\beta}_2$ são combinações lineares das variáveis Y_1, \dots, Y_n e, portanto, são conjuntamente normalmente distribuídos.

Tem-se $E[\hat{\alpha}] = E[\bar{Y}_+] = \alpha$ e

$$Var(\hat{\alpha}) = \frac{1}{n^2}Var(\sum Y_i) = \sigma^2/n.$$

Logo, $\hat{\alpha} \sim N(\alpha, \sigma^2/n)$.

Propriedades do EMV

- Para uma sequência de x_i 's onde \bar{x}_+ é fixado, $\hat{\alpha}$ converge em probabilidade para α tal que $\hat{\alpha}$ é um estimador consistente de α pela desigualdade de Chebyshev (resultado esperado pois $\hat{\alpha}$ é um EMV).

Para $\hat{\beta}_2$ temos que $E[\hat{\beta}_2] = \beta_2$ e $Var(\hat{\beta}_2) = \frac{\sigma^2}{S_t}$, e, assim, temos, $\hat{\beta}_2 \sim N(\beta_2, \sigma^2/S_t)$

- Segue que $\hat{\beta}_2$ é um estimador consistente se e só se $S_t \rightarrow \infty$ quando $n \rightarrow \infty$.

Esta condição é satisfeita, por exemplo, se cada valor de x é replicado muitas vezes, ou se os valores de x dispersam-se cada vez mais quando n cresce. Isto conforma-se com a intuição, de que para se obter uma estimativa precisa da inclinação da reta de regressão, precisamos ou de estimativas mais precisas para pelo menos dois pontos dados, ou ter observações sobre um amplo campo de valores de x .

Propriedades do EMV

- As propriedades do estimador $\hat{\beta}_1$ podem ser obtidas a partir das propriedades de $\hat{\alpha}$ e $\hat{\beta}_2$. Como $\hat{\beta}_1$ é uma combinação linear de dois estimadores conjuntamente normalmente distribuídos, sua distribuição também é normal.

$$E[\hat{\beta}_1] = E[\hat{\alpha} - \bar{x}_+ \hat{\beta}_2] = \alpha - \beta_2 \bar{x}_+ = \beta_1$$

- Para calcular a variância de $\hat{\beta}_1$, precisaremos da covariância entre $\hat{\alpha}$ e $\hat{\beta}_2$:

$$\text{cov}(\hat{\alpha}, \hat{\beta}_2) = \text{cov}\left(\frac{1}{n} \sum_i Y_i, \frac{1}{S_t} \sum_i t_i Y_i\right) = \frac{\sigma^2}{n S_t} t_+ = 0$$

- Como $\hat{\alpha}$ e $\hat{\beta}_2$ são normalmente distribuídos, segue que $\hat{\alpha}$ e $\hat{\beta}_2$ são independentes e, assim,

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\hat{\alpha}) + \bar{x}_+^2 \text{Var}(\hat{\beta}_2) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_+^2}{S_t} \right)$$

Propriedades do EMV

- Distribuições Amostrais

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\frac{1}{n} + \frac{\bar{x}_+^2}{S_t}))$$

- n e S_t fixados a variância de $\hat{\beta}_1$ cresce com o quadrado de \bar{x}_+ .
- Se os valores para x estão afastada da origem, o intercepto β_1 será determinado de forma imprecisa.
- É preferível trabalhar com α e β_2 em vez de β_1 e β_2 pois as propriedades de $\hat{\alpha}$ e $\hat{\beta}_2$ são mais simples e α e β_2 são parâmetros mais estáveis.
- A instabilidade dos parâmetros β_1 e β_2 também é refletida na correlação entre $\hat{\beta}_1$ e $\hat{\beta}_2$.

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \text{cov}(\hat{\alpha} - \hat{\beta}_2 \bar{x}_+, \hat{\beta}_2) = -\frac{\bar{x}_+ \sigma^2}{S_t}$$

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{x}_+}{\sqrt{S_t/n + \bar{x}_+^2}} = -\frac{\bar{x}_+}{\sqrt{\sum_i x_i^2/n}}$$

Propriedades do EMV

- Variância (σ^2)

O i -ésimo resíduo é definido por

$$r_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i = y_i - \hat{\alpha} - \hat{\beta}_2 t_i, \quad i = 1, \dots, n$$

A v.a. correspondente será denotada por R_i .

É fácil mostrar que $\sum_i r_i = \sum_i r_i t_i = 0$. Obtemos então, $\sum_i r_i^2 = \sum_i y_i^2 - \hat{\alpha} y_+ - \hat{\beta}_2 S_{ty}$.

$$E[R_i] = \alpha + \beta_2 t_i$$

Para calcular a variância de R_i mostraremos primeiro que R_i e $(\hat{\alpha}, \hat{\beta}_2)$ são independentes.

Propriedades do EMV

- Covariâncias

$$\text{cov}(R_i, \hat{\alpha}) = \text{cov}(Y_i - \hat{\alpha} - \hat{\beta}_2 t_i, \hat{\alpha}) = 0.$$

$$\text{cov}(R_i, \hat{\beta}_2) = \text{cov}(Y_i - \hat{\alpha} - \hat{\beta}_2 t_i, \hat{\beta}_2) = 0.$$

- Como as variáveis tem distribuição conjunta normal, a correlação zero implica em independência. Assim, para cada $i = 1, \dots, n$; R_i e $(\hat{\alpha}, \hat{\beta}_2)$ são independentes.
- Por outro lado: $Y_i = R_i + \hat{\alpha} + \hat{\beta}_2 t_i$
- Pela independência temos

$$\text{Var}(Y_i) = \text{Var}(R_i) + \text{Var}(\hat{\alpha}) + t_i^2 \text{Var}(\hat{\beta}_2)$$

- Logo

$$\text{Var}(R_i) = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{t_i^2}{S_t} \right) \right]$$

Propriedades do EMV

- Assim, para t_i fixado, se $n \rightarrow \infty$ e $S_t \rightarrow \infty$, $Var(R_i) \rightarrow \sigma^2$. Observe que $Var(R_i) < \sigma^2$.
- Como R_i é uma combinação linear das observações temos $R_i \sim N(0, \sigma^2(1 - \frac{1}{n} - \frac{t_i^2}{S_t}))$
- Considere agora $\hat{\sigma}^2 = \frac{1}{n} \sum_i R_i^2$.
- Como (R_1, \dots, R_n) é independente de $(\hat{\alpha}, \hat{\beta}_2)$, e $\hat{\sigma}^2$ é função de R_1, \dots, R_n então $\hat{\sigma}^2$ é independente de $(\hat{\alpha}, \hat{\beta}_2)$. Assim, os estimadores $\hat{\alpha}$, $\hat{\beta}_2$ e $\hat{\sigma}^2$ são mutuamente independentes.
- Similarmente, $\hat{\sigma}^2$ e $\hat{\beta}_1$ são independentes.
- A distribuição de $\hat{\sigma}^2$ será obtida mais a frente. Agora apenas enunciaremos o resultado:

$D(\hat{\beta}_1, \hat{\beta}_2) = \sum_i R_i^2 \sim \sigma^2 \chi_{(n-2)}^2$, onde $\chi_{(f)}^2$ denota uma distribuição de Qui-quadrado com f graus de liberdade.

Propriedades da Variância Residual

- A média e a variância de uma $\chi^2_{(f)}$ são f e $2f$, respectivamente.

Assim,

$$E[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2 \text{ e } Var(\hat{\sigma}^2) = \frac{2(n-2)}{n^2}\sigma^4.$$

- Assim, vemos que $\hat{\sigma}^2$ é assintoticamente não tendencioso e consistente quando $n \rightarrow \infty$.
- A consistência de $\hat{\sigma}^2$ não requer quaisquer condições sobre os x 's exceto $S_t > 0$ tal que para S_t pequeno, $Var(\hat{\sigma}^2)$ é pequena para n grande.
- Um estimador não tendencioso para σ^2 é dado por

$$\tilde{\sigma}^2 = \frac{D(\hat{\beta}_1, \hat{\beta}_2)}{n-2}.$$

Propriedades dos Resíduos

- O resíduo não padronizado é definido por $r_i = y_i - \hat{\mu}_i$ e $\hat{\mu}_i$ é o valor ajustado pelo modelo definido por
$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta}_2 t_i.$$
- Vimos que os resíduos têm média zero e variância $\sigma^2(1 - \frac{1}{n} - \frac{t_i^2}{S_t})$ enquanto que os valores ajustados têm média μ_i e variância $\sigma^2(\frac{1}{n} + \frac{t_i^2}{S_t})$.
- Assim, para grandes amostras tal que S_t também seja grande, as variâncias dos resíduos tendem para σ^2 enquanto que as variâncias dos valores ajustados tendem para zero.
- Os resíduos são úteis para verificar algumas das suposições do modelo.

Uso dos resíduos

- Verificação de normalidade dos Y_i 's.

Se o modelo estiver correto

$$R_i \sim N(0, \sigma^2(1 - \frac{1}{n} - \frac{t_i^2}{S_t})).$$

- Assim, podemos construir o 'Normal-plot' dos resíduos, 'plotando' u_i versus $r_{(i)}$, onde

$u_i = \phi^{-1}((i - 1/2)/n)$ e $r_{(i)}$ são os resíduos ordenados. ϕ^{-1} representa a inversa da distribuição acumulada normal padrão.

- A idéia por trás do 'Normal-plot' é a seguinte: suponha Z_1, \dots, Z_n iid's $N(\mu, \sigma^2)$. Então, $Z_i = \mu + \sigma\epsilon_i$, onde $\epsilon_i \sim N(0, 1)$ e a mesma relação vale para os valores ordenados, $Z_{(i)} = \mu + \sigma\epsilon_{(i)}$, $i = 1, \dots, n$.
- A esperança de $\epsilon_{(i)}$ pode ser aproximada pelo escore normal u_i sugerindo uma relação aproximadamente linear entre u_i e $z_{(i)}$.

- Se os R_i 's são normais, aproximadamente independentes e identicamente distribuídos, o gráfico deve mostrar uma tendência linear, a linha passando através da origem (pois os resíduos têm média zero) e inclinação σ^{-1} .
- Desvios sistemáticos da linearidade no 'Normal-plot' indicam evidência de que os resíduos e, conseqüentemente as observações Y_i não são normais.
- Cuidados com a interpretação:
 - (i) Os resíduos são combinações lineares das observações e, assim, um efeito limite central tenderá a produzir resíduos normais mesmo se a distribuição dos Y_i 's não for normal.
 - (ii) Os resíduos não têm todos a mesma variância, como foi evidenciado. Isto pode ser corrigido trabalhando-se com os resíduos padronizados s_i .

Uso dos resíduos

- Resíduos Padronizados

$$s_i = \frac{r_i}{\sqrt{1 - \frac{1}{n} - \frac{t_i^2}{S_t}}},$$

onde todos têm a mesma variância (σ^2).

(iii) Os resíduos são correlacionados.

- Estes fatos implicam que grandes amostras são necessárias para mostrar evidência contra normalidade, e que o normal plot pode mostrar linearidade mesmo se a distribuição de Y não for normal.
- Porém, como veremos adiante, os resíduos são muito úteis para verificar outros tipos de desvios de um modelo linear tal como não-linearidade ou variância não constante.

Inferência sobre os parâmetros

- Primeiro, considere a estimação de σ^2 . Como vimos, o EMV de σ^2 é tendencioso com média $\sigma^2(1 - 2/n)$. Uma explicação para este fato é que $\hat{\sigma}^2$ é baseado sobre o valor minimizado da soma de quadrados $D(\beta_1, \beta_2)$. Assim, observe que no verdadeiro ponto do parâmetro, $D(\beta_1, \beta_2) \sim \chi_{(n)}^2$.
- Em particular, $E[D(\beta_1, \beta_2)/n] = \sigma^2$ tal que se β_1, β_2 são conhecidos o estimador $\hat{\sigma}^2$ será não tendencioso.
- Assim, o fato de termos que estimar β_1, β_2 juntos com σ^2 induz a uma distorção no EMV de σ^2 .
- Um melhor estimador será $\tilde{\sigma}^2 = \frac{D(\hat{\beta}_1, \hat{\beta}_2)}{n-2}$ pois será não-tendencioso e sua variância será

$$Var(\tilde{\sigma}^2) = \frac{2}{n-2}\sigma^4.$$

Inferência sobre os parâmetros

- Assim, a variância de $\tilde{\sigma}^2$ é maior que a de $\hat{\sigma}^2$. Porém, mostraremos, pelo teorema de Lehmann-Scheffé que o estimador $\tilde{\sigma}^2$ é o único estimador não-tendencioso de variância mínima de σ^2 .
- Os argumentos usados anteriormente mostram que $\tilde{\sigma}^2$ é independente de $(\hat{\beta}_1, \hat{\beta}_2)$.
- Consideraremos agora testes e intervalos de confiança para β_1 e β_2 baseados na distribuição t .
- Inferência sobre β_2 .

Vimos que

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{S_t})$$

O erro padrão de $\hat{\beta}_2$, denotado por $s.e.(\hat{\beta}_2)$, é dado por $\tilde{\sigma}/\sqrt{S_t}$.

Inferência sobre os parâmetros

- A razão t para β_2 é $t(Y) = \frac{\hat{\beta}_2 - \beta_2}{s.e.(\hat{\beta}_2)}$
com $t(Y) \sim t_{(n-2)}$, e $t_{(f)}$ representando uma distribuição t com f graus de liberdade.

- Teste de Hipóteses

$$H_0 : \beta_2 = \beta_2^{(0)} \text{ contra } H_A : \beta_2 \neq \beta_2^{(0)},$$

rejeitamos H_0 , ao nível de significância α , se

$$|t(y)| > t_{1-\alpha/2(n-2)},$$

onde $t(y) = (\hat{\beta}_2 - \beta_2^{(0)})/s.e.(\hat{\beta}_2)$ e $t_{q(f)}$ é tal que

$$P(t < t_{q(f)}) = q, \text{ com } t \sim t_{(f)}.$$

Equivalentemente, a equação

$$|t(y)| = t_{1-\frac{p}{2}(n-2)}$$

define o p-valor do teste.

Inferência sobre os parâmetros

- Podemos também usar a razão t para definir um intervalo de confiança para β_2

$$IC(\beta_2, \gamma) : \hat{\beta}_2 \pm t_{\frac{1+\gamma}{2}(n-2)} s.e.(\hat{\beta}_2)$$

onde $\gamma = 1 - \alpha$ é o coeficiente de confiança do intervalo.

- Inferência sobre α e β_1 .

Os resultados aqui são similares aos obtidos para β_2 . Lembre que

$$\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$$

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\frac{1}{n} + \frac{\bar{x}_+^2}{S_t}))$$

Aqui, $s.e.(\hat{\beta}_1) = \tilde{\sigma}(\frac{1}{n} + \frac{\bar{x}_+^2}{S_t})^{1/2}$ e $s.e.(\hat{\alpha}) = \frac{\tilde{\sigma}}{\sqrt{n}}$.

Inferência sobre os parâmetros

- Inferências sobre σ^2

Também podemos construir testes ou intervalos de confiança para o parâmetro σ^2 . Neste caso, usamos o fato de que

$$\frac{\tilde{\sigma}^2}{\sigma^2} \sim \frac{\chi_{(n-2)}^2}{n-2}$$

- Teste de Hipóteses

Assim, no teste bilateral de $H_0 : \sigma^2 = \sigma^{2(0)}$ rejeitamos H_0 se

$$\frac{\tilde{\sigma}^2}{\sigma^{2(0)}} > \frac{\chi_{1-\alpha/2(n-2)}^2}{n-2} \text{ ou } \frac{\tilde{\sigma}^2}{\sigma^{2(0)}} < \frac{\chi_{\alpha/2(n-2)}^2}{n-2}$$

onde $\chi_{q(f)}^2$ é o 100q-ésimo quantil da distribuição $\chi_{(f)}^2$.

- Um intervalo de confiança $1 - \alpha$ é

$$IC(\sigma^2, 1 - \alpha) : \left(\frac{(n-2)\tilde{\sigma}^2}{\chi_{1-\alpha/2(n-2)}^2}, \frac{(n-2)\tilde{\sigma}^2}{\chi_{\alpha/2(n-2)}^2} \right)$$

Inferência sobre os parâmetros

Deve-se sempre apresentar as estimativas de uma análise de regressão na forma da seguinte tabela.

parâmetro	estimativa	erro padrão
β_1	$\hat{\beta}_1$	$s.e.(\hat{\beta}_1)$
β_2	$\hat{\beta}_2$	$s.e.(\hat{\beta}_2)$
	$\tilde{\sigma}^2$	g.l.=n-2

Um relatório contendo a análise estatística de dados deve conter (i) apresentação do problema e dos dados; (ii) análise estatística e (iii) conclusões. Estas partes podem ser subdivididas ou não.