Modelos Lineares

Professora Ariane Ferreira

Instituto Politécnico
Campus Regional da UERJ
Nova Friburgo - RJ

Conteúdos das Aulas

1. Apresentação da disciplina e introdução aos Modelos Lineares (ML)
2. Distribuição Normal Univariada
3. Testes de hipóteses – aderência à lei Normal
4. Regressão Linear Simples - RLS
5. Inferência e testes de hipóteses para RLS
6. Análise de resíduos para RLS
7. Análise de outliers para RLS
8. Introdução à análise multivariada de dados
   - pré-tratamentos de uma tabela de dados numéricos
9. Distribuição Normal Multivariada
10. Testes de hipóteses- aderência à lei normal multivariada

Conteúdos das Aulas

11. Regressão Linear Multipla - RLM
12. Inferência e testes de hipóteses para RLS
13. Análise de resíduos para RLM
14. Análise de outliers para RLM
15. Validação de um modelo de regressão
16. Multicolinearidade
17. Seleção de variáveis para modelos lineares
18. Modelos Polinomiais

CONTEÚDOS ADICIONAIS:

19. Introdução ao Modelo Linear Generalizado (GENMOD/GLM)
20. Conceitos básicos de Regressão Logística

Aulas

Total de Aulas Teóricas: 20

Total de Aulas de Laboratório: 4 Aulas

Datas das Aulas de Laboratório:

- **22/11/2012; 11/12/2012; 10/01/2013; 31/01/2013**

Não havera aulas dias 18/12/2012 e 20/12/2012;

Aulas de laboratório : alunos agrupados em duplas para a realização dos exercícios;

Entrega de relatório das aulas de laboratório da dupla na semana seguinte à realização do laboratório.

Software utilizado nas aulas de laboratorio: Scilab com toolbox Cascilib

Avaliação

40% Nota : Entrega de relatórios com os exercícios de cada aula de laboratório (dupla).

60% Nota : Prova única com questões multipla escolha e exercícios sobre o conteudo total (exceto itens 19 e 20).

Datas:

Prova Geral: 26/02/2013

Recuperação PG: 05/03/2013

Substitui obrigatoriamente a nota da Prova Geral.

Bibliografia

[1] Neter, J., Wasserman, W., Kutner, M.H. Applied linear statistical models.
[2] Draper, N.R., Smith,H. Applied regression analysis. 2ed. New York: Wiley, 1981.
[3] Montgomery, D.; Runger,D.C. Applied Statistics and Probability for Engineers, 4 ed.J.Wiley
[4] Montgomery and Peck, Introduction to Linear Regression Analysis;
[5] Charnet, R. Analise de modelos de Regressão linear com aplicações. Ed. da Unicamp, 1999.
[6] Tenenhaus, M. Statistique Méthodes pour décrire, expliquer et prévoir, ed. Dunod.
[7] Seber, G.A.F., Linear Regression Analysis.
[8] Myers and Montgomery, Generalized Linear Models.

Web site da disciplina: http://wiki.nosdigitais.teia.org.br/Modelos_Lineares
Lista google: UERJ_EC_Modelos_Lineares

Modelos

Os modelos podem ser:

- **Físicos:** protótipos e plantas-piloto.
- **Matemáticos:** representação abstrata da realidade por equações.

O que é um modelo matemático?

"É uma representação dos aspectos essenciais de um sistema, que apresenta conhecimento desse sistema em uma forma utilizável.“ (Eykhoff, 1974)

"É um sistema de equações, cuja solução, dado um conjunto de dados de entrada, é representativa da resposta do processo.“

(Denn, 1986)

"Um modelo nada mais é do que uma abstração matemática ele um processo real.“ (Seborg et al, 2004)

Modelo Matemático: Equação

A equação ou conjunto de equações que compõe o modelo é uma aproximação do processo real.

Dessa forma, o modelo não pode incorporar todas as características, tanto macroscópicas como microscópicas, do processo real.

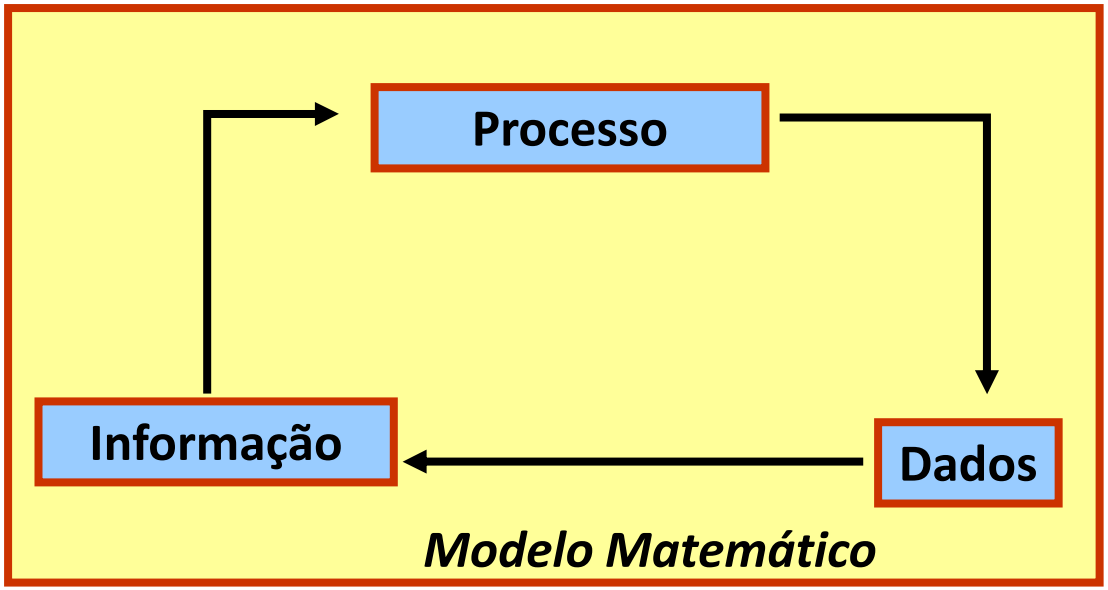
Deve-se normalmente buscar um compromisso entre o custo de se ter o modelo, isto é:

- o tempo e o esforço requeridos para obtê-la e verificá-lo,
- e o nível de detalhes no mesmo,
- bem como os benefícios esperados de sua aplicação.

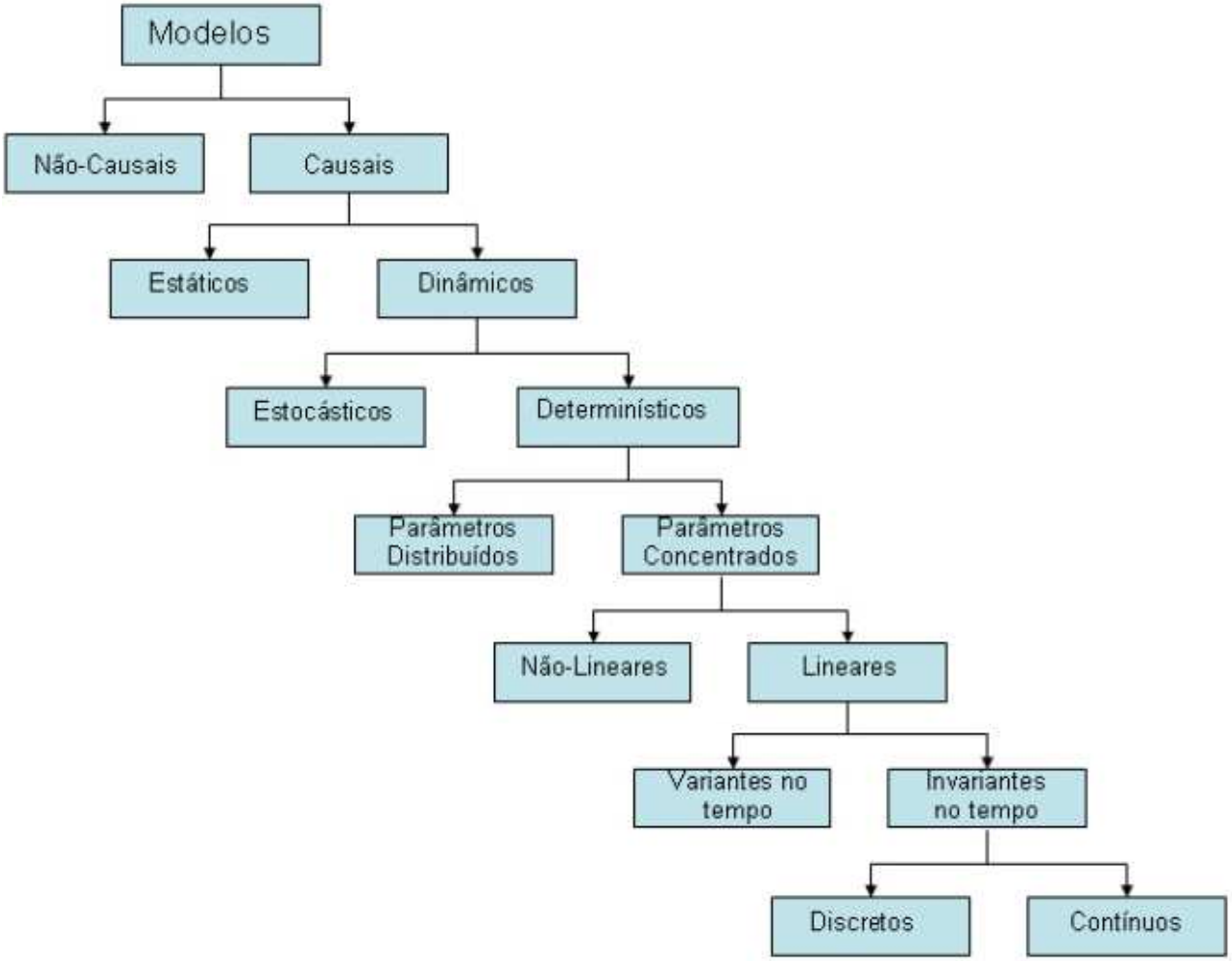
Modelo Matemático: Equação

O propósito do modelo determina, em última análise, sua precisão.

Um processo pode ser físico, químico, biológico, social, econômico etc.



Classificação dos Modelos Matemáticos



Modelos Matemáticos: Causal x Não-Causal

Um sistema causal depende somente

- de condições presentes ou passadas,
- e não dependem de estados futuros.

Sistemas físicos são todos sistemas causais.

Modelos Matemáticos: Estático x Dinâmico

Estático: processo cujo valor das variáveis permanece constante no tempo (se as entradas permanecem as mesmas, as saídas ficam inalteradas).

- Este tipo de modelo não possui **"memória",** daí o efeito de uma variável de entrada ser apenas instantâneo.
- O modelo é um sistema de equações algébricas.
- Não depende de estados passados.

Modelos Matemáticos: Estático x Dinâmico

Dinâmico: as variáveis variam no tempo, que é a variável independente.

- A solução completa consiste dos regimes permanente e transitório.
- O efeito de um sinal de entrada irá influenciar o comportamento do sistema nosinstantes subsequentes.
- O modelo é um sistema de equações diferenciais ou de diferenças.

Depende de estados passados e presentes.

Modelos Matemáticos: Determinísticos x Estocásticos

Determinístico: Em um modelo determinístico a saída pode ser calculada de forma exata tão logo se conheça o sinal de entrada e as condições iniciais.

Estocástico: Em contraste, um modelo estocástico contém termos aleatórios que tornam impossível um cálculo exato da saída.

- Os termos aleatórios do modelo podem ser encarados como uma descrição das perturbações.

Normalmente, o modelo determinístico engloba apenas o processo, enquanto o estocástico considera também as perturbações e ruídos.

Modelos Matemáticos: Parâmetros Concentrados x Parâmetros Distribuídos

Modelos à parâmetros concentrados: as variações espaciais são desprezadas:

- propriedades (estados) do sistema são considerados homogêneos em todo o volume de controle.
- Eles são descritos por um número finito de equações diferenciais ou de diferenças ordinárias.

Modelos a parâmetros distribuídos: variações espaciais são consideradas no comportamento das variáveis.

- Eles são descritos por um número infinito de equações ordinárias ou por equações diferenciais parciais.

Todo sistema real é distribuído.

Modelos Matemáticos: Parâmetros Concentrados x Parâmetros Distribuídos

Se as variações espaciais são pequenas, pode-se aproximar o comportamento do sistema por um modelo a parâmetros concentrados.

No caso de modelos a parâmetros concentrados, assume-se que as variáveis de interesse sofram alterações como função de apenas uma variável independente (tempo, posição etc) dentro do volume de controle.

Exemplo: modelar a temperatura dentro de uma sala, pode-se supor que essa variável seja homogênea em toda a sala e que apenas varie com o tempo.

Neste caso se tem um modelo a parâmetros concentrados e avariação de temperatura pode ser representada como: $\frac{dT}{dt}$

Modelos Matemáticos: Parâmetros Concentrados x Parâmetros Distribuídos

Por outro lado, caso se deseje considerar que a temperatura na sala não seja homogênea e que

- pode haver, por exemplo, uma variação da temperatura em função do tempo e da cota ‘z’ da sala, t
- em-se agora um modelo a parâmetros distribuídos e, neste caso, pode-se representar asvariações de temperatura como

$$\frac{\partial T}{\partial z \partial t}$$

Modelos Matemáticos: Linear x Não-Linear

Um modelo é linear se a(s) saída(s) depende(m) linearmente da(s) entrada(s) e possíveis perturbações, caso contrário ele é não-linear.

Equações (e portanto modelos) são lineares se variáveis dependentes ou suas derivadas aparecem apenas no 1º. grau.

Considere um sistema cujas variáveis tenham condições iniciais nulas.

- Se sua resposta a uma entrada $u_1(t)$ é $y_1(t)$ e
- Se sua resposta a uma entrada $u_2(t)$ é $y_2(t)$
- ele é linear se sua resposta a

$$\alpha \cdot u_1(t) + \beta \cdot u_2(t)$$ é igual à $$\alpha \cdot y_1(t) + \beta \cdot y_2(t)$$

Constantes quaisquer

Modelos Matemáticos: Linear x Não-Linear

Uma forma simples de se verificar a linearidade de uma função é aplicar o seguinte teste:

$$f(x_1 + x_2) = f(x_1) + f(x_2)$$

$$f(Kx) = K \cdot f(x)$$ Onde K é uma Constante qualquer

A linearidade implica no Princípio da Superposição:

- o que significa que se pode calcular a saída de um sistema excitado por qualquer tipo de entrada
- dividindo-se a entrada em componentes simples e adicionando-se as respostas de cada componente.

Modelos Matemáticos: Linear x Não-Linear

Dinâmicas não-lineares fazem com que a resposta a qualquer variável de entrada seja afetada

- pelo comportamento das outras entradas,
- de forma que é necessário identificar as relações entre todas as entradas e saídas simultaneamente.

As relações entrada/saída podem ser identificadas uma por vez em um sistema linear, considerando-se:

- somente uma das variáveis de entrada como fonte de variações na saída. (Norton,1986).

Modelos Matemáticos: Invariantes no tempo x Variantes no tempo

Invariantes no tempo: seus parâmetros não variam ao longo do tempo.

- Modelos invariantes no tempo são os mais comuns.

Exemplos de propriedade invariante no tempo:

- a resistência elétrica de um motor,
- a área de um tanque,
- o atrito de uma mesa posicionadora.

NOTE A DIFERENÇA: a corrente na resistência, a altura de líquido do tanque, a posição da mesa podem variar pois não correspondem a parâmetros, e sim são sinais abstratos ou variáveis a serem medidas no sistema.

Modelos Matemáticos: Invariantes no tempo x Variantes no tempo

Variantes no tempo: seus parâmetros variam ao longo do tempo.

Exemplos de propriedade variante no tempo:

- um trocador de calor do tipo casco-tubo em que ocorre incrustação de material nas paredes dos tubos.
  - Neste caso, o coeficiente de transferência térmica entre o casco e os tubos sofre uma variação ao longo do tempo, alterando as características funcionais do trocador de calor.
- Um foguete é outro exemplo de um sistema variante no tempo,
  - pois sua massa vai diminuindo a medida que seu combustível é consumido ao longo do tempo.

Modelos Matemáticos: Tempo Contínuo x Tempo Discreto

Modelos à Tempo Discreto: Modelos em tempo discreto descrevem a relação entre entradas e saídas em pontos de tempo discreto.

Assume-se que esses pontos sejam eqüidistantes e o tempo entre dois pontos consecutivos seja usado como unidade tempo, de forma que o tempo t assuma valores inteiros

$$(t \rightarrow Z^{+} , 1, 2, 3...).$$

Normalmente os modelos em tempo discreto são descritos por equações de diferença, ao passo que os modelos em tempo contínuo são descritos por equações diferenciais.

Modelos Matemáticos: Representação de Modelos

Para trabalhar com sistemas lineares invariantes no tempo, necessitamos representá-los de alguma forma.

Formas de representação de modelos:

- Funções de Transferência;
- Espaço de Estados;
- **Regressão Linear.**

Modelos Matemáticos: Modelos de Regressão

Modelos de Regressão: Os métodos estatísticos (modelos lineares) são amplamente usados como parte do processo de aprendizagem do método científico.

Na biologia, física e ciências sociais, como também nos negócios e engenharia, os modelos lineares são úteis nos estágios de planejamento da pesquisa e na análise dos dados resultantes.

A teoria da Regressão permite que se estabeleçam relações entre variáveis que se interrelacionam cujas informações estão disponíveis (dados pré-coletados), relações às quais associam-se os modelos de regressão.

Modelos Matemáticos: Teoria da Regressão

A teoria da Regressão permite que se estabeleçam relações entre variáveis que se interrelacionam cujas informações estão disponíveis (dados pré-coletados), relações às quais associam-se os modelos de regressão.

Uma vez estabelecida essa relação pelo modelo de regressão, é preciso avaliar a confiança que nela se pode colocar, realizando testes estatísticos.

Modelos de Regressão: Tipos de Informação

Temos dois tipos básicos de informação a considerar:

- Informação descrevendo as mudanças assumidas por uma variável através do tempo (dados de séries temporais);
- Informação descrevendo as atividades de pessoas, firmas etc. num dado instante de tempo (dados de corte transversal).

Para esses dois tipos de informação é possível estabelecer relações que descrevem as situações observadas por meio de modelos de regressão.

Modelos de Regressão: Tipos de Informação

Temos dois tipos básicos de informação a considerar:

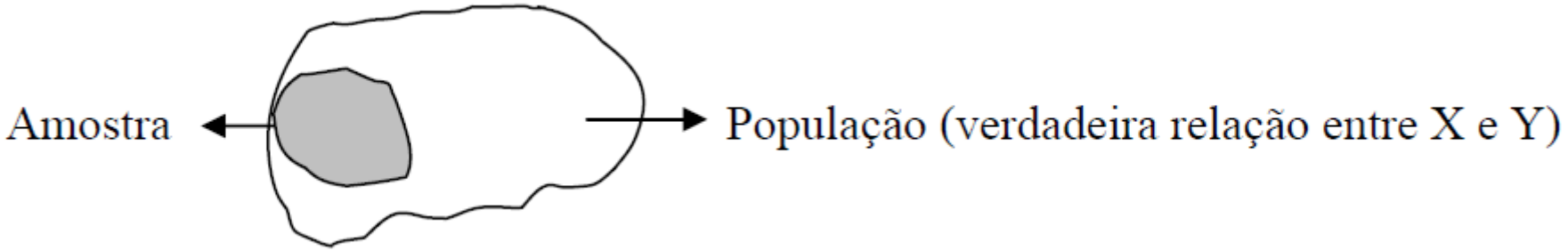
- Informação descrevendo as mudanças assumidas por uma variável através do tempo (dados de séries temporais);
- Informação descrevendo as atividades de pessoas, firmas etc. num dado instante de tempo (dados de corte transversal).

Para esses dois tipos de informação é possível estabelecer relações que descrevem as situações observadas por meio de modelos de regressão.

Modelos de Regressão: Tipos de Informação

Ou seja, dado um conjunto finito de observações *X* e *Y*, por meio do modelo de regressão é buscado estabelecer relações entre *X* e *Y*.

Esse conjunto finito de observações corresponde a uma amostra representativa do universo de informações ou população, a qual permitiria estabelecer a verdadeira relação entre *X* e *Y*



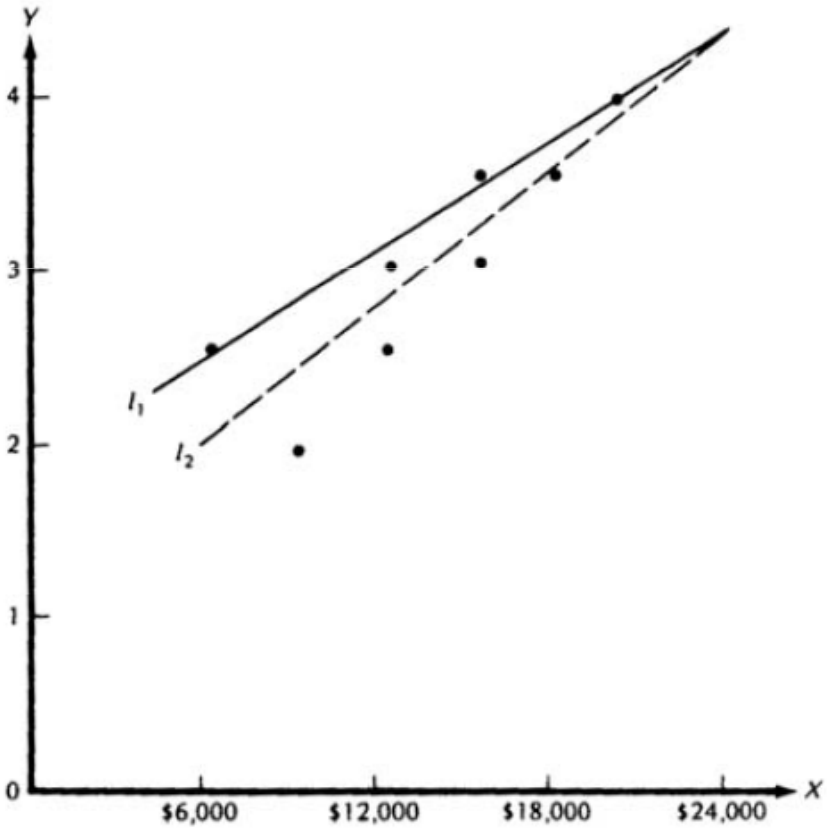
Relação entre a amostra e a população ou universo de informações

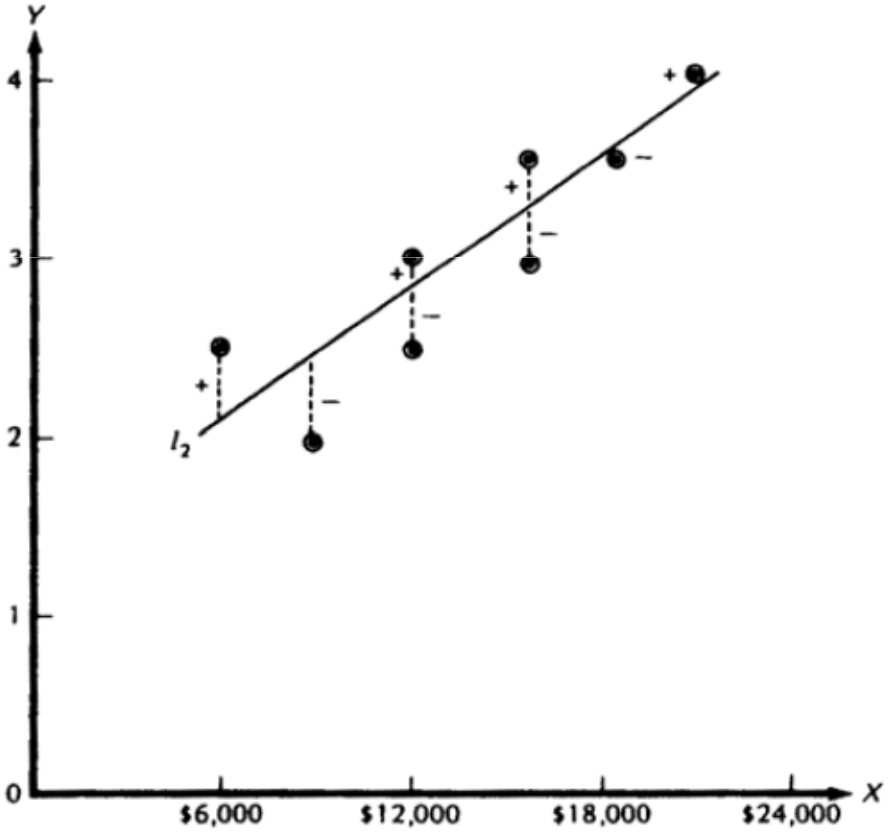
Modelos de Regressão: Ajuste da reta

Tome-se por hipótese que exista a relação linear l_i entre X e Y. No diagrama de dispersão da são representadas

- as linhas l_1 e l_2 que se procurou ajustar ao conjunto de pares ordenados (X, Y) do conjunto amostral,
- assim como os desvios (positivos e negativos) em relação a l_2

Diagrama de dispersão e desvios em relação à linha ajustada





Modelos de Regressão: Ajuste da reta

Definem-se desvios como os valores, segundo *Y*, das diferenças entre os valores observados e os valores sobre a linha l_i ajustada ao conjunto de pares (*X, Y*).

Como regra estabelece-se que a melhor linha l_i corresponde àquela cujo somatório dos desvios tende a zero (é minimizado).

A melhor linha ajustada define o modelo de regressão e pode ser obtida pela derivação de mínimos quadrados ordinários.

Modelo de Regressão Linear Simples

RLS

Na regressão linear simples, nós nos preocupamos em modelar a relação entre duas variáveis,

Exemplos:

- rendimento e número de anos de educação,
- altura e peso de pessoas, comprimento e largura de envelopes,
- altitude e temperatura de ebulição da água,
- dose de uma droga e resposta,
- quantidade de adubo e produção de gramíneas.

Para uma relação linear, nós usamos um modelo da forma:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad E(Y_i) = \alpha + \beta x_i$$

onde *Y* é a variável dependente ou variável resposta e *X* é a variável independente ou variável preditora.

A variável aleatória ε_i é o termo de erro no modelo.

Modelo de Regressão Linear Simples

Nesse contexto,o erro não significa engano ou equívoco, mas sim um termo estatístico que representa flutuações aleatórias, erros de medidas ou o efeito de fatores não controlados.

A linearidade do modelo em si é uma suposição.

Geralmente, nós adicionamos outras suposições sobre a distribuição do erro, independência dos valores observados de Y, assim por diante.

Usando valores observados de X e Y, nós estimamos $\beta 0$ e $\beta 1$ e fazemos inferências tais como intervalos de confiança e testes de hipóteses sobre $\beta 0$ e $\beta 1$.

Nós também podemos usar o modelo estimado para prever ou predizer o valor de Y para um particular valor de X.

Estimação e procedimentos inferenciais para o modelo de regressão linear simples são desenvolvidos nesta disciplina.

Modelo de Regressão Linear Simples

RLS

- X e Y duas variáveis contínuas.
- Os valores de x_i de X são amostrados sem erros de medida.
- Observamos os valores correspondentes $y_1, \ldots, y_n$ de Y.

Exemplos

- X pode ser o tempo e Y um valor medido em diferentes datas.
- Y pode ser a diferença de potencial medida nas extremidades de uma resistência para diferentes valores de intensidade de corrente.

Hypótese

- X e Y não são independentes e lo conhecimento X permete adquirir o conhecimento de Y.
- O valor médio E(Y|X=x), é a esperança condicional de Y sabendo que X = x.

Função Linear

$$\mathrm{E}(\mathrm{Y}_\mathrm{i}) = \alpha + \beta x_i$$

$$\mathrm{Y}_\mathrm{i} = \alpha + \beta x_i + \varepsilon_i$$

com $\mathrm{E}(\varepsilon_i)=0$, para todo $i=1,\ldots,n$

n = n° d'observações; ε_i =resíduos das obs i

Os Dados

Y = Variável à explicar

- numérica
- (ou dependente)

X_1, …, X_p = Variáveis explicativas

- numéricas ou binárias
- (ou independentes)

O Quadro de Dados

	X_1	…	X_p	Y
1	x_{11}	…	X_{1p}	y_1
⋮				
i	x_{1i}	…	X_{ip}	y_i
⋮				
n	x_{1n}	…	X_{np}	y_n

Onde os x_{ji} são fixos e ε_i é um termo aleatório de lei N(0,σ);
Os ε_i são indépendentes entre si.

O Modelo de Regressão Multipla

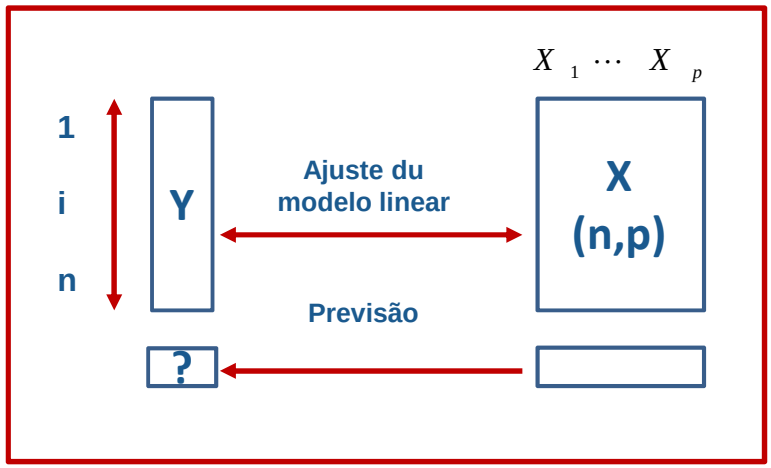
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Suposição : independência linear des X_i.

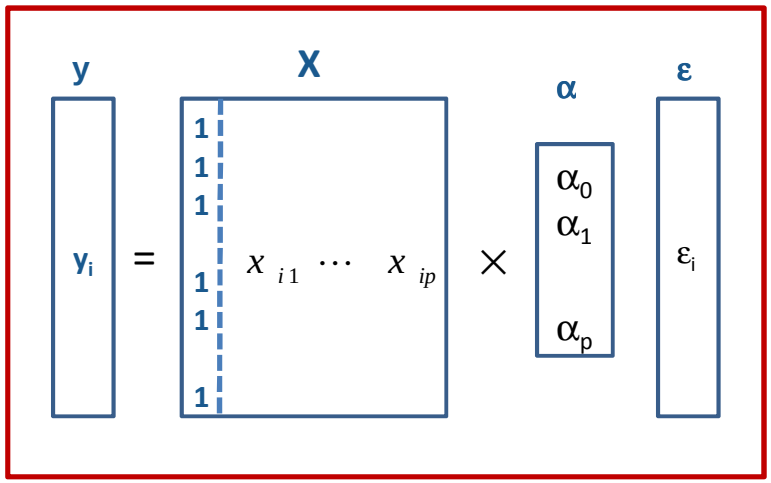
Modelo na forma matricial

$$y = \mathbf{X}\boldsymbol{\alpha} + \varepsilon$$

(n,1) (n,p+1) (p+1,1) (n,1)



Previsão Linear



Schema du modelo Linear

Hipóteses do modelo linear

Resíduos

- A variância dos resíduos é a mesma para todos os valores de X
  - **Homoscedasticidade** : $V(\varepsilon_i) = \sigma^2$
- Os resíduos são linearmente independentes: $cov(\varepsilon_i,\varepsilon_j) = 0 \; \forall \; i \neq j$
- Os resíduos são normalmente distribuídos : $\varepsilon_i \sim N(0,\sigma^2)$

A existência da componente estocástica (ε_i) corresponde ao fato de que :

- **variação sincrônica :** indivíduos com mesmo valor de x_i podem ter respostas Y diferentes;
- **Variação discrônica** : um mesmo indivíduo medido varias vezes para um mesmo valor de x_i pode ter respostas Y diferentes.

- **Temos equivalência de** $\varepsilon_i \sim N(0,\sigma^2)$ et $Y/X=x_i \sim N(\alpha + \beta x_i,\sigma^2)$

Os problemes

1. **Estimação dos coeficientes da regressão**

   $\beta_0, \beta_1, \ldots, \beta_p$.
2. **Estimação do desvio-padrão σ do termo residual ε_i**
3. **Analise dos residuos**
4. **Medir a força de ligação entre Y e as variaveis $X_1,\ldots,X_p$**

   Coeficientes : de correlação multipla (R), de determinação (R^2)
5. **A ligação global entre Y e $X_1,\ldots,X_p$ é significativa ?**
6. **A contribuição marginal de cada variavel X_j (adicionada às outras) na explicação de Y é significativa ?**
7. **Seleção automatica das « boas » variaveis X_j.**
8. **Comparação de modelos.**
9. **Intervalo de prédição à 95% de y.**
10. **Intervalo de confiança à 95% de E(Y).**

Notações

- As variáveis são designadas por letras maiúsculas em itálico (*Y, X_k*), o mesmo se passando com as suas médias $(\overline{Y},\ \overline{X}_k)$
- Os vetores são representados em letra minúscula em negrito $(\mathbf{y},\ \mathbf{x_k})$
- e os seus elementos em itálico e letra minúscula $(y_i,\ x_{ik})$
- Usam-se letras maiúsculas em negrito para as matrizes (**X, M, W**), sendo os seus elementos representados como os dos vetores.
- O acento circunflexo é usado para indicar que se trata de valores estimados $(\hat{\sigma},\ \hat{\mathbf{y}})$
- A transposição de vetores e matrizes é indicada por $(\mathbf{y}', \mathbf{M}')$
- Constantes e outros valores matemáticos são apresentados em letra minúscula e itálico.
- As referências a distribuições estatísticas são feitas do modo usual, indicando-se entre parêntesis ou em índice os parâmetros necessários.

Notações

- distribuição normal com média μ e variância σ^2 : N(μ, σ^2);
- distribuição *t* de Student com *n-p-1* graus de liberdade e nível de significância γ (teste bilateral): $t_{1-\gamma/2}(n-p-1)$

$\mathbf{y} = [y_1\, y_2 \,..\, y_n]'$	vector das leituras
$\mathbf{x_k} = [x_{1k}\, x_{2k} \,..\, x_{nk}]'$	vector das observações de cada variável X_k
$\mathbf{X} = [\mathbf{x_1}\, \mathbf{x_2} \,..\, \mathbf{x_p}]$	matriz de observações (elementos x_{ik}, $i=1..n$, $k=1..p$)
$\mathbf{b} = [a\ \mathbf{b_0}]' = [a\ b_1\ b_2 \,...\, b_p]'$	vector dos parâmetros
$\mathbf{e} = [e_1\, e_2 \,..\, e_n]'$	vector dos erros
$\mathbf{1} = [1 \,...\, 1]'$	vector unitário de dimensão n
$\mathbf{X_a} = [\mathbf{1}\, \mathbf{x_1}\, \mathbf{x_2} \,..\, \mathbf{x_p}]$	matriz aumentada de observações ($i=0..n$, $k=0..p$)

$\mathbf{y} = a.\mathbf{1} + \mathbf{X}.\mathbf{b_0} + \mathbf{e}$ Equação linear na forma compacta

Médias e variâncias

- Define-se a média de uma variável através de:

$$\bar{X}_k = \frac{\sum_{i=1}^{n} x_{ik}}{n} \text{ (observações)} \quad \text{ou} \quad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ (leituras)}$$

- as expressões acima referem-se à média amostral, que é uma estimativa não tendenciosa da média das variáveis.

- Defini-se soma dos quadrados dos desvios em relação à média de X_k

$$d_k^2 = \sum_{i=1}^{n} (x_{ik} - \bar{X}_k)^2$$

- a estimativa não tendenciosa da variância de X_k é dada por:

- O mesmo para a estimativa da variância de Y.

$$s_k^2 = \frac{d_k^2}{n-1} = \frac{\sum_{i=1}^{n} (x_{ik} - \bar{X}_k)^2}{n-1}$$

Variáveis não-numéricas

- A inclusão de categorias no modelo da regressão deve fazer-se recorrendo a variáveis binárias.
  - Para duas categorias (A e B), cria-se um variável X_1, com dois valores possíveis, correspondendo cada um a uma das categorias.
  - Os dois valores costumam ser 0 e 1, mas pode ser usado qualquer par de números.

Se houver *c* categorias, deverão criar-se as variáveis binárias necessárias para definir todas as categorias.

O número de variáveis a criar é o inteiro imediatamente superior (ou igual) a $log_2 c$

Variáveis não-numéricas

Por exemplo, para 3 categorias *A, B* e *C*, poderão criar-se 2 variáveis X_1 e X_2,

	A	B	C
X_1	1	0	0
X_2	0	1	0

Um erro frequente consiste em usar variáveis com mais de dois valores, o que institui uma ordem a priori e uma relação fixa entre classes.

No caso do exemplo, seria portanto errado usar apenas uma variável que tomasse os valores (0, 1, 2) para as três classes.

Transformações dos dados

Em alternativa ao uso das variáveis originais ("raw"), podem ser usadas variáveis:

- centradas ("centered"),
- estandardizadas ("standardized")
- ou com norma unitária ("unit lenght"),

Todos estes procedimentos visam compatibilizar, de algum modo, variáveis que podem ter escalas e dispersões muito diferentes.

A comparação da influência relativa das diversas variáveis, com base nos parâmetros estimados, só faz sentido se as variáveis forem normalizadas.

Centragem

Uma transformação simples consiste em centrar cada variável em relação à sua média.

A variável transformada M_k obtém-se de X_k através de:

$$m_{ik} = x_{ik} - \overline{X}_k$$

Semelhantemente ao que se fez para **X**, também aqui se define $\mathbf{M}=[\mathbf{m}_1\ \mathbf{m}_2\ ..\ \mathbf{m}_p]$.

Estandardização

A estandardização corresponde a uma transformação para média nula e desvio padrão unitário de cada variável original X_k.

A nova variável Z_k é obtida através de:

$$z_{ik} = \frac{x_{ik} - \overline{X}_k}{s_k}$$

também aqui se define $\mathbf{Z}=[\mathbf{z}_1\ \mathbf{z}_2\ ..\ \mathbf{z}_p]$.

Norma unitária

Esta transformação substitui os valores de cada variável X_k por uma nova variável W_k, obtida pela seguinte regra:

$$w_{ik} = \frac{x_{ik} - \overline{X}_k}{d_k}$$

também aqui se define $\mathbf{W}=[\mathbf{w}_1\ \mathbf{w}_2\ ..\ \mathbf{w}_p]$,

verifica-se que a matriz **W'.W** apresenta diagonal unitária (daí o nome da transformação).

Os restantes elementos $\mathbf{(W'.W)}_{uv}$ correspondem à correlação entre X_u e X_v.

Note-se ainda que **Z'.Z**=(*n-1*) **W'.W**.