Modelos de regressão aplicados à epidemiologia

Maria do Rosario Dias de Oliveira Latorre

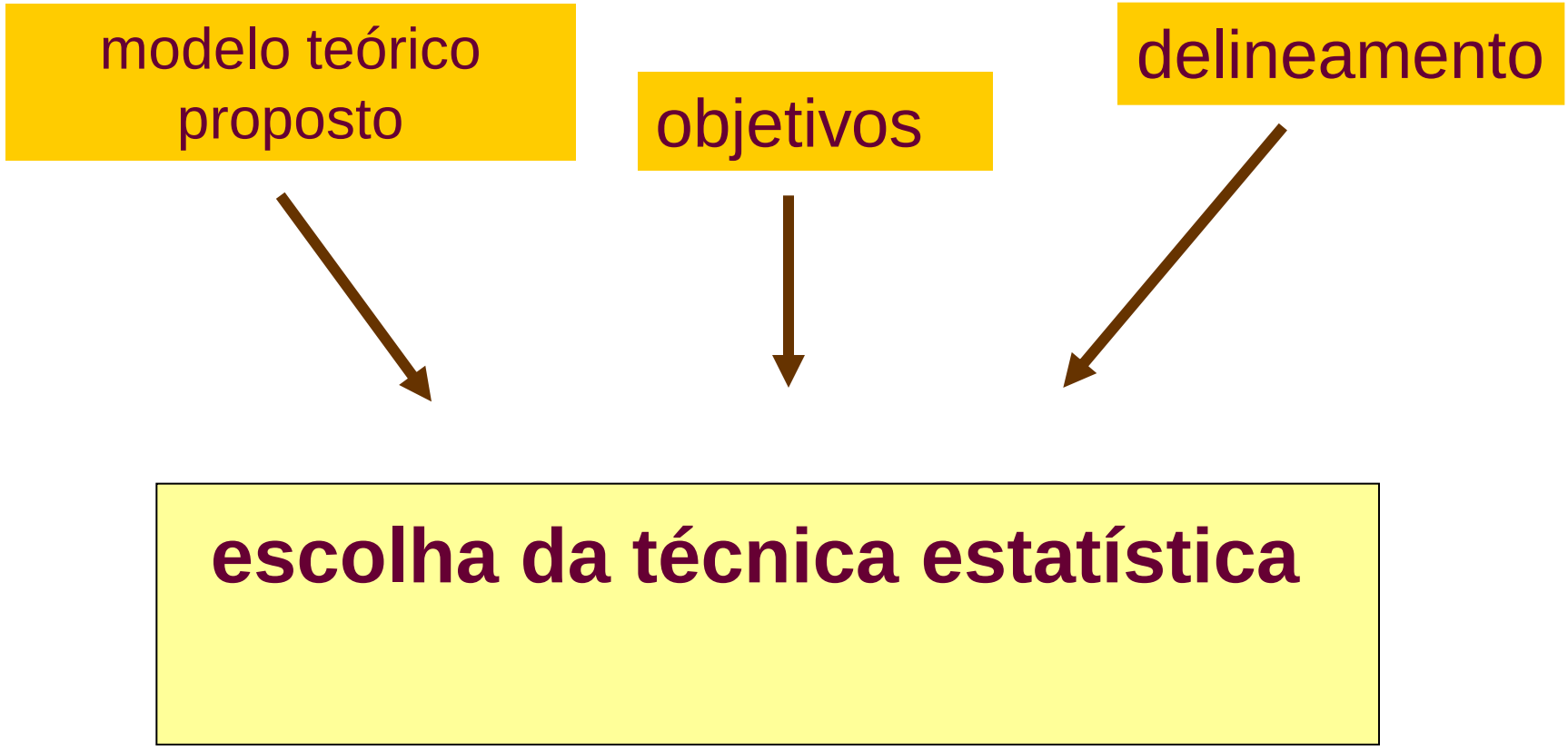
Departamento de Epidemiologia

Faculdade de Saúde Pública da USP

2021

- ✓ elementos que auxiliam na escolha da análise estatística
- ✓ primórdios dos modelos de regressão
- ✓ modelos de regressão utilizados em:
  - estudos ecológicos
  - estudos transversais
  - estudos caso-controle
  - estudos de coorte

qual a técnica estatística a ser utilizada?


modelo teórico proposto
objetivos
delineamento
escolha da técnica estatística

modelo teórico proposto

existe uma única variável dependente?

sugestões de análises estatísticas

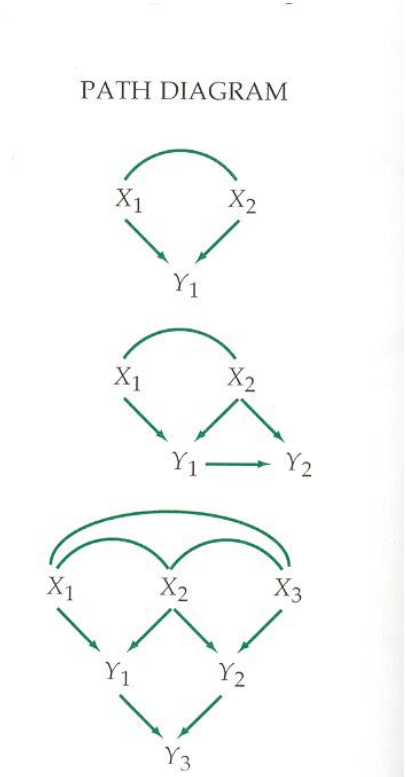
- ✓ não há variável dependente
- ✓ análise fatorial, de componentes principais, análise de cluster, análise espacial

sugestões de análises estatísticas

- ✓ não há variável dependente

- ✓ análise fatorial, de componentes principais, análise de cluster, análise espacial.

- ✓ há várias var dependentes



análise de correlação canônica, modelos de equação estrutural.

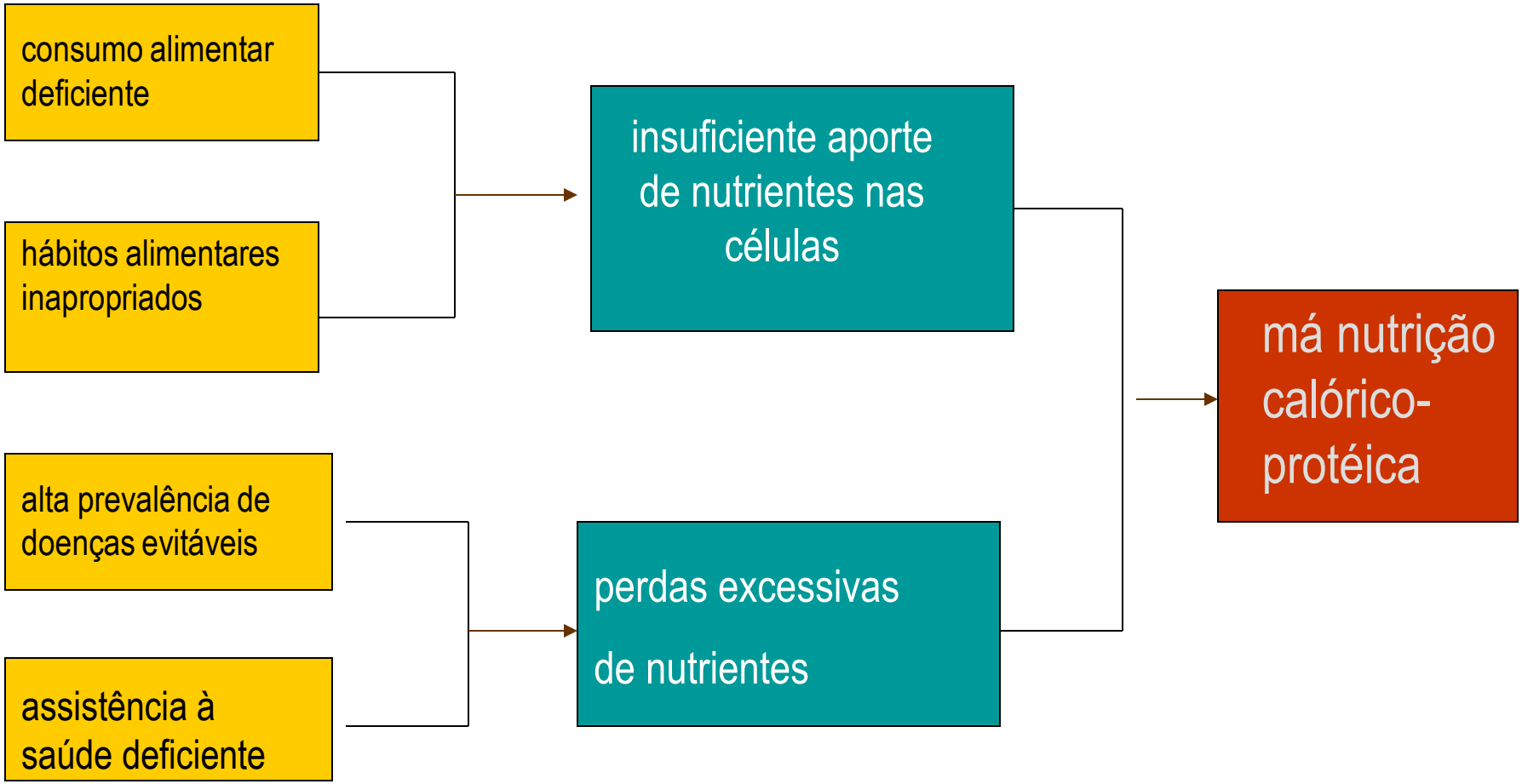
Análise discriminante

- Objetivo é, a partir de um conjunto de variáveis aleatórias, definir uma combinação linear das mesmas, de tal forma que ela possa “discriminar” grupos de pessoas.

sugestões de análises estatísticas

✓ não há variável dependente	✓ análise fatorial, de componentes principais, análise de cluster, análise espacial
✓ há várias variáveis dependentes	✓ análise de correlação canônica, modelos de equação estrutural
✓ há uma única variável dependente	✓ análise de regressão, análise de variância, análise de covariância

modelo teórico



(Pereira, 1999-p43)

Análise múltipla

- cada indivíduo da amostra é avaliado em relação a k variáveis. Sendo assim, cada indivíduo representa um vetor.

$$(v_1, v_2, \ldots\ldots, v_k)$$

- em uma amostra de n indivíduos, o banco de dados é uma matriz de *n* linhas e *k* colunas.

$$\begin{bmatrix} v_{1,1}, v_{1,2}, \ldots\ldots v_{1,k} \\ v_{2,1}, v_{2,2}, \ldots\ldots v_{2,k} \\ \ldots\ldots \\ v_{n,1}, v_{n,2}, \ldots\ldots v_{n,k} \end{bmatrix}$$

$v_{i,j} \rightarrow$ sendo i o indivíduo e j a variável

OBJETIVOS

Escolha da análise estatística

quando utilizar um modelo de regressão ?

- ✓ caracterizar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (X_i);
- ✓ controlar o efeito de outras variáveis (C_i);
- ✓ testar o efeito interativo de 2 ou mais variáveis independentes;
- ✓ comparar múltiplos relacionamentos derivados da análise de regressão;

Modelos de Regressão

Y: variável de interesse, resposta ou dependente

X_i: variáveis independentes, co-variáveis

?

$$Y=f(X_i)$$

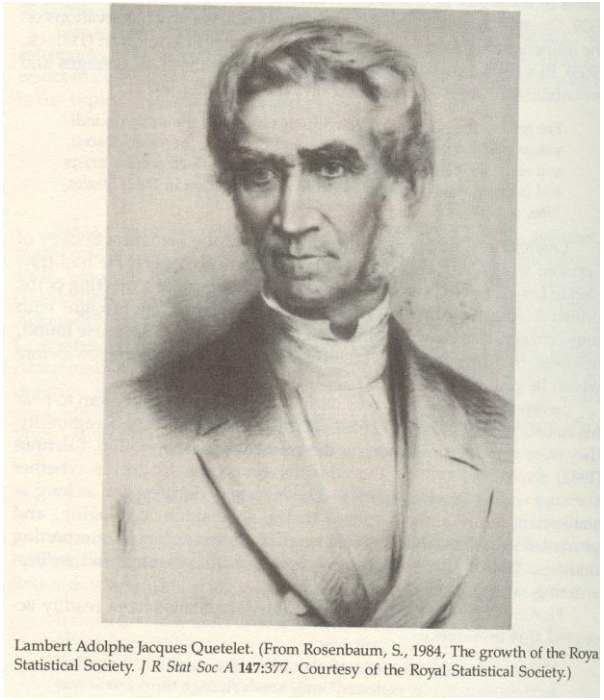
delineamento

tipo de variável dependente

Escolha do modelo de regressão

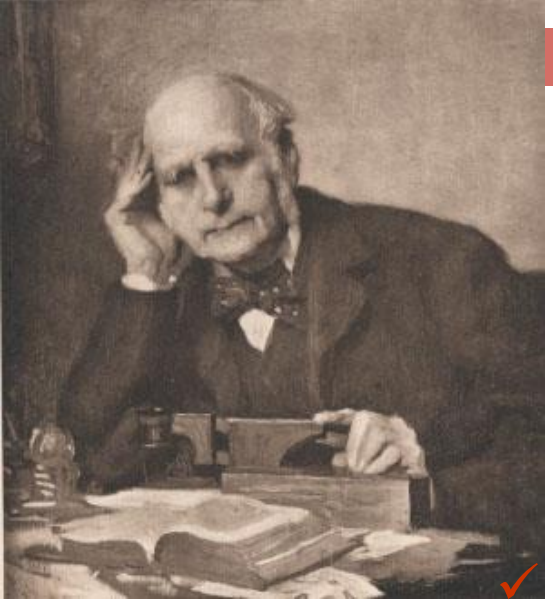
Primórdios dos modelos de regressão

Dr. Lambert Adolphe Jacques Quetelet (1796-1874)



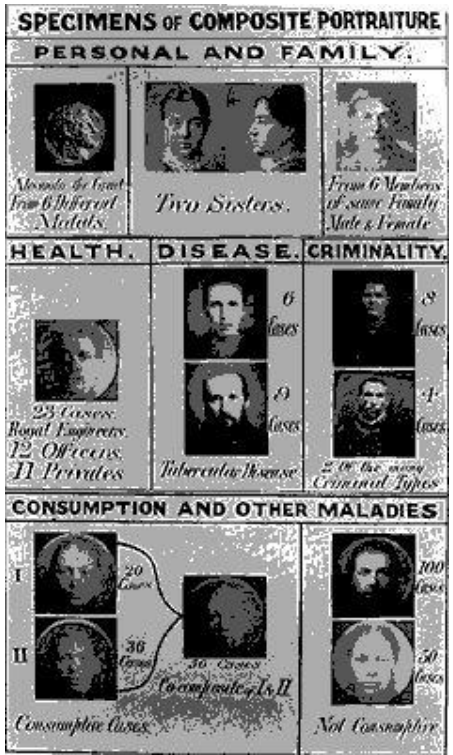
Lambert Adolphe Jacques Quetelet. (From Rosenbaum, S., 1984, The growth of the Roya Statistical Society. *J R Stat Soc A* **147**:377. Courtesy of the Royal Statistical Society.)

- ✓ 1835: defendeu a idéia de que as medidas antropométricas obedeciam a uma distribuição Normal
- ✓ 1853: organizou a 1ª. conferência internacional de estatística



Sir Francis Galton
(1822-1911)

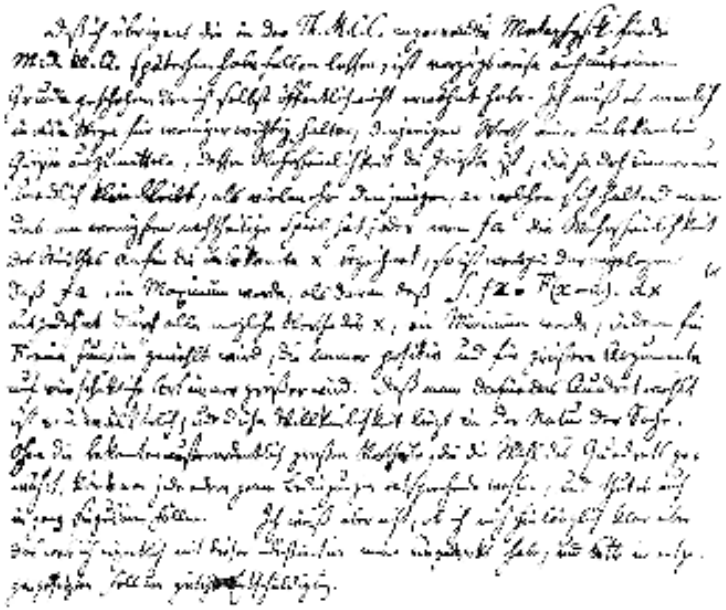
- 1869: "The law of deviation from an average"



- 1877: "Typical laws of heredity in man"

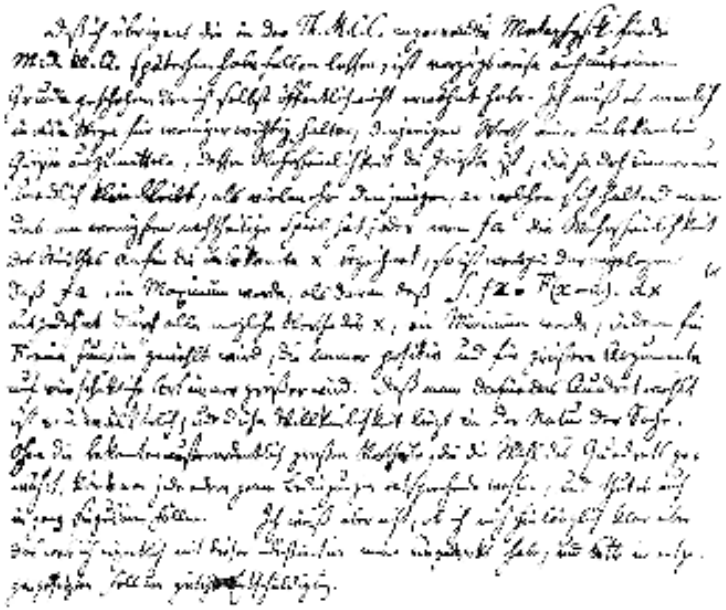
r: reversão (reversion), regressão (regression)

In a letter to Bessel dated Feb 28 1839, Gauss explains why he decided to drop the metaphysics of maximum likelihood in favour of the method of least squares...



and apologizes for using square loss function, which is used only for its obvious conveniences. He says he would take other choices where appropriate....

In a letter to Bessel dated Feb 28 1839, Gauss explains why he decided to drop the metaphysics of maximum likelihood in favour of the method of least squares...

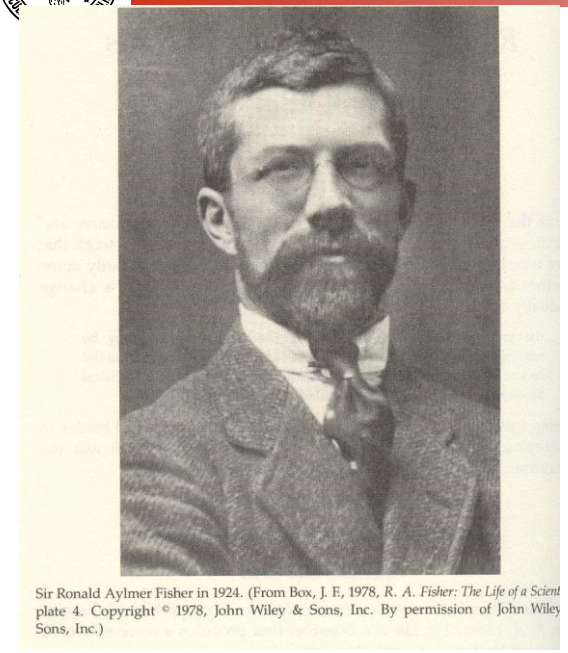


and apologizes for using square loss function, which is used only for its obvious conveniences. He says he would take other choices where appropriate....

- *New Scotland Yard : 12 medidas antropométricas (Bertillion)*

- Edgeworth (Galton) propôs 3 equações:
  - F1=0,16 estatura+0,51 antebraço+0,39 comprimento da perna
  - F2= -0,17 estatura+0,69 antebraço-0,09 comprimento da perna
  - F3= -0,15 estatura-0,25 antebraço+0,52 comprimento da perna

- WR Macdonell (K. Pearson): 7 variáveis de 3000 criminosos



Sir Ronald Aylmer Fisher in 1924. (From Box, J. E, 1978, *R. A. Fisher: The Life of a Scient* plate 4. Copyright © 1978, John Wiley & Sons, Inc. By permission of John Wiley Sons, Inc.)

Sir Ronald A.Fisher (1890-1962)

- 1922: teoria da máxima verossimilhança

Função de verossimilhança

$$L(\theta) = \text{Prob}(i_1) \,.\, \text{Prob}(i_2) \,.\, \ldots \,.\, \text{Prob}(i_n)$$

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$$

$$\ln(L(\theta)) = \sum_{i=1}^{n} f(y_i;\theta)$$

ESTUDOS ECOLÓGICOS

- Análise de tendências:
  - Modelos de regressão polinomial.

Análise de Séries Temporais

- Y: coeficientes de mortalidade, de morbidade ou o número de casos.
- X: dia, mês, ano.
- transformar X em (X-ponto médio do período)

$$Y = \beta_0 + \beta_1 \left(X_i - \overline{X} \right)$$

neste caso: $\beta_0 = \overline{Y}$ e

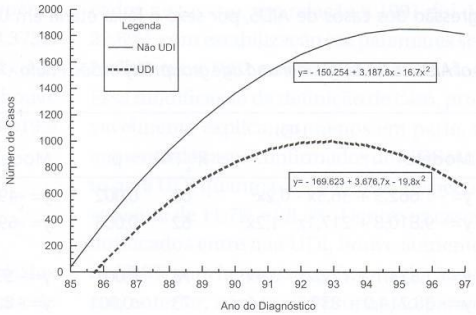
$\beta_1 =$ incremento médio do período

Exemplo

- Objetivo: Analisar a tendência da epidemia de Aids no Município de São Paulo, de 1985 a 1997, segundo uso de drogas.
  - Y= número de casos notificados
  - X= ano do diagnóstico-1991
  - Análise estratificada por sub-grupos

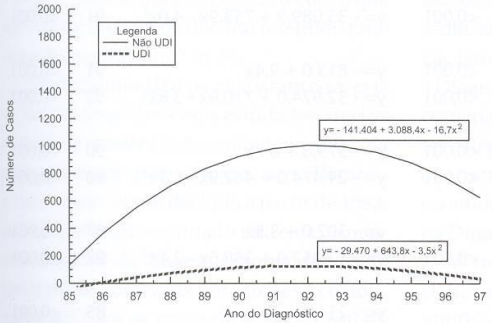
Bastos, Latorre, Waldman, 2001

Total



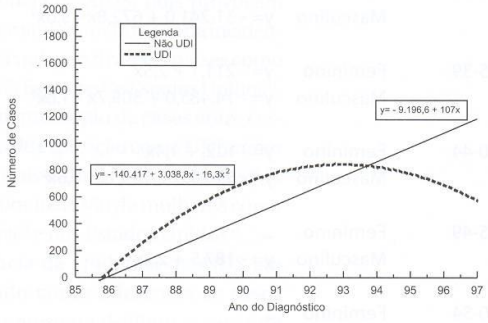
(A)

HSH



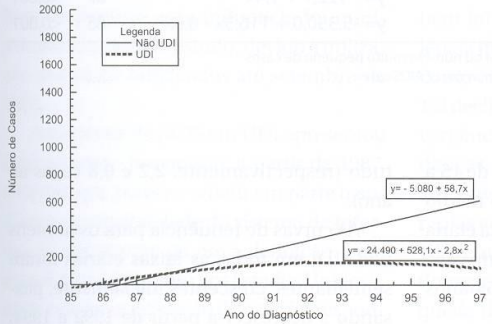
(B)

pessoas com práticas heterossexuais



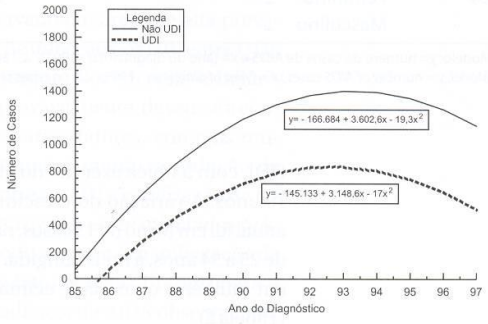
(C)

mulheres



(D)

homens



(E)

Figura 1 - Tendência dos casos de AIDS em pessoas UDI e não UDI (A); entre homens que fazem sexo com homens (HSH) (B); pessoas com práticas heterossexuais (C); mulheres (D); e homens (E) - Município de São Paulo - 1985 a 1997.

- Análise de tendências:
  - Modelos de regressão polinomial.
  - Modelos idade-período-coorte (age-period-cohort).

- O efeito da idade, período e coorte são estimados através de um modelo de Poisson, o qual assume que o número de casos ou óbitos segue uma distribuição de Poisson:

$$\log(d_{ij} / p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k$$

- d_{ij}: número de casos/óbitos na faixa etária i no período j;
- p_{ij} :população da faixa etária i no periodo j;
- μ: é o coeficiente médio (intercepto);
- α_i : é o efeito da faixa etária i;
- β_j : é o efeito do período j;
- γ_k : é o efeito da coorte de anscimento k.

Exemplo:

Trends in colon cancer incidence and mortality in São Paulo city - Brazil.

Marcolin; Latorre, Lisboa; Michels; Mirra.

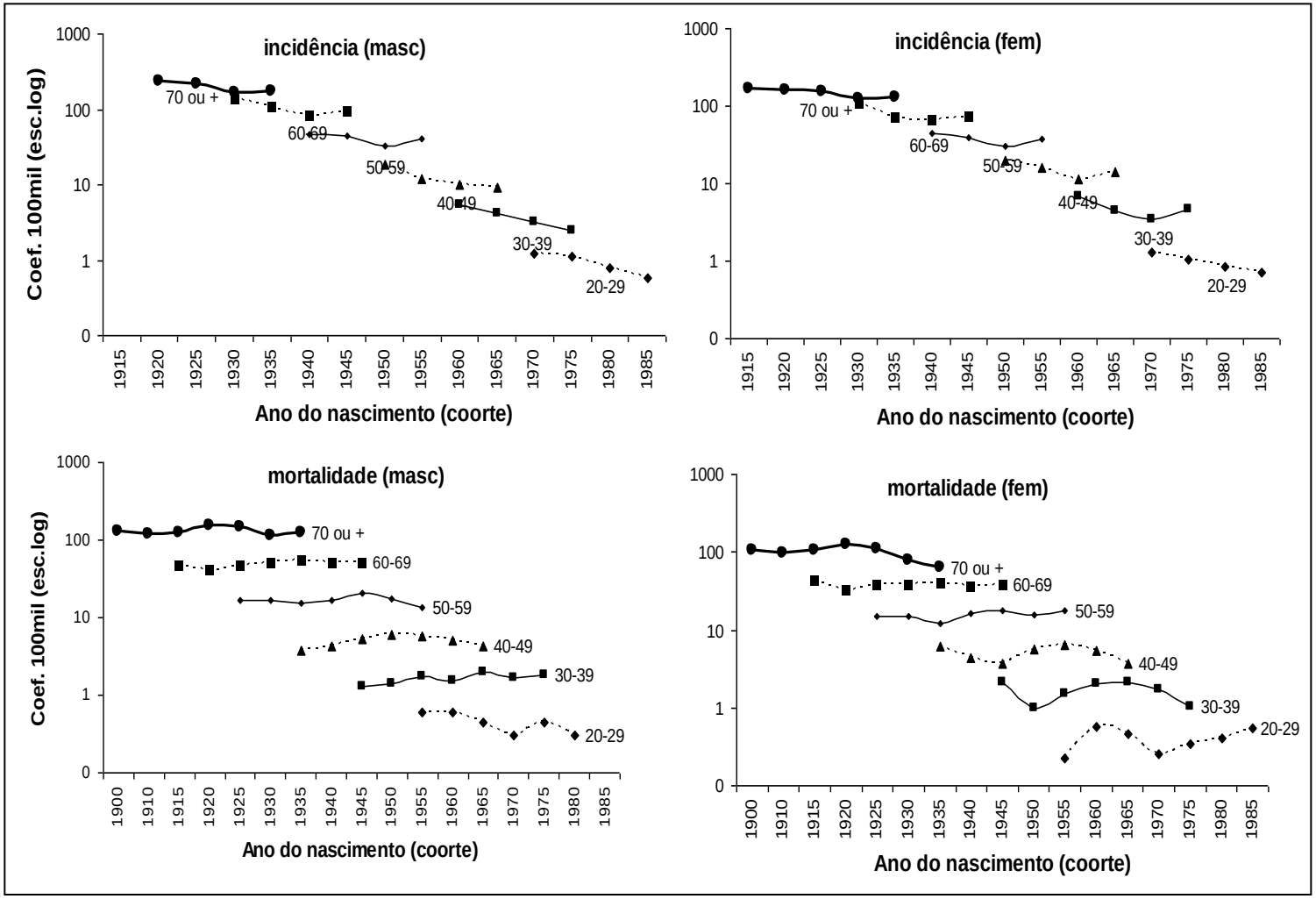


Table 1: Comparison and assessment of age-period-cohort models and period-cohort for colon cancer incidence and mortality, according to sex, 1982-2005

Efeitos	Masculino			Feminino		
Incidência (1997-2005)	*Deviance**	gl	p	*Deviance**	gl	p
Idade	95,0032	26		89,4111	26	
Idade-período	52,7094	24	0,1516	38,4997	24	0,7570
Idade-coorte	64,3378	19	< 0,0001	56,8835	19	< 0,0001
Idade-período-coorte	41,325	17		34,3045	17	
Mortalidade (1982-2005)						
Idade	358,2510	91		384,1349	91	
Idade-período	106,2979	84	0,1342	131,7111	84	< 0,0001
Idade-coorte	158,1389	83	< 0,0001	192,063	83	< 0,0001
Idade-período-coorte	93,8978	76		97,1815	76	

gl, graus de liberdade

*Deviance do modelo de Poisson.

Os níveis descritivos (p) são baseados nos testes de Qui-quadrado referentes a comparação entre os modelos com dois efeitos em relação ao modelo completo

- Análise de tendências:
  - Modelos de regressão polinomial.
  - Modelos idade-período-coorte (age-period-cohort).

- Análise espacial:
  - Modelos de regressão espacial.

Estudos transversais

Estudos transversais

	doente	não doente	TOTAL
EXPOSTO			
NÃO EXPOSTO			
TOTAL			N

O tamanho total da amostra é fixo (N).

modelos lineares generalizados (MLG)

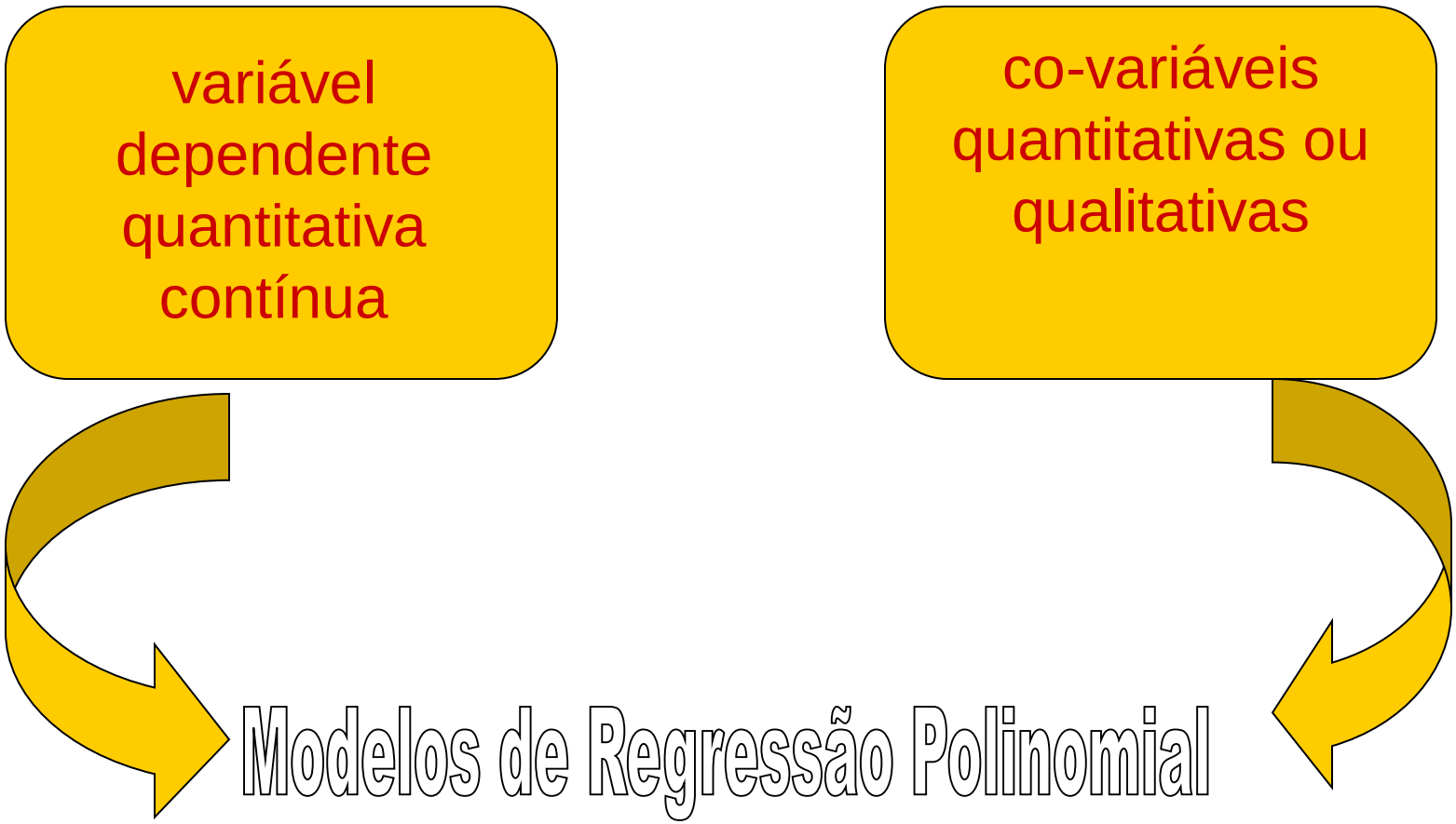
- são baseados na família exponencial de distribuição de probabilidades
  - Normal
  - Binomial
  - Poisson

- amostra com observações independentes

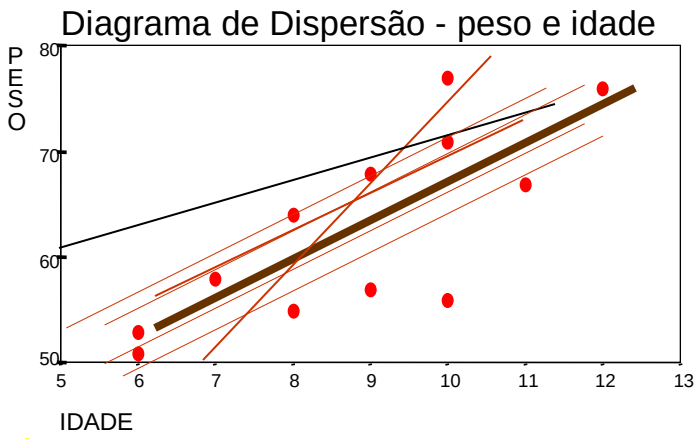
Estudos transversais

- ✓ modelos de regressão polinomial
  - Y é variável quantitativa contínua
- ✓ modelos de regressão de Poisson
  - Y é variável quantitativa discreta - processos de contagem
- ✓ modelos de regressão quantílica
  - Y é variável quantitativa e o interesse não é na estimativa da média.
- ✓ modelos de regressão logística binomial
  - Y é uma variável categórica binária
- ✓ modelos de regressão logística ordinal ou multinomial
  - Y é uma variável categórica com mais que 3 categorias.
- ✓ modelos log-lineares
  - processos de contagem - tabelas de contingência

Exemplos

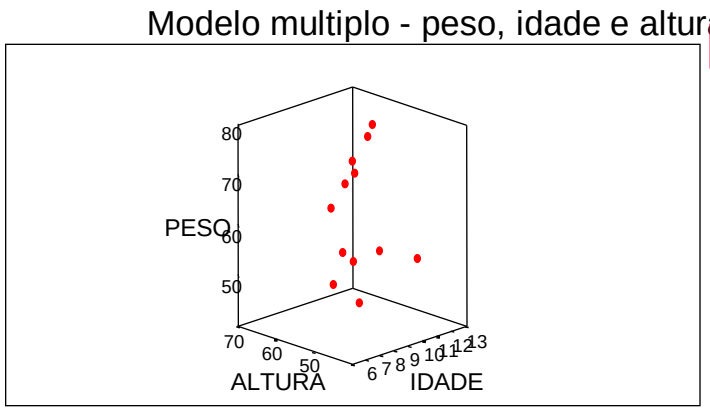

variável dependente quantitativa contínua
co-variáveis quantitativas ou qualitativas
Modelos de Regressão Polinomial

MODELOS DE REGRESSÃO LINEAR



modelo de regressão linear simples

$$Y = \beta_0 + \beta_1 X_1$$



modelo de regressão linear múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

SUPOSIÇÕES

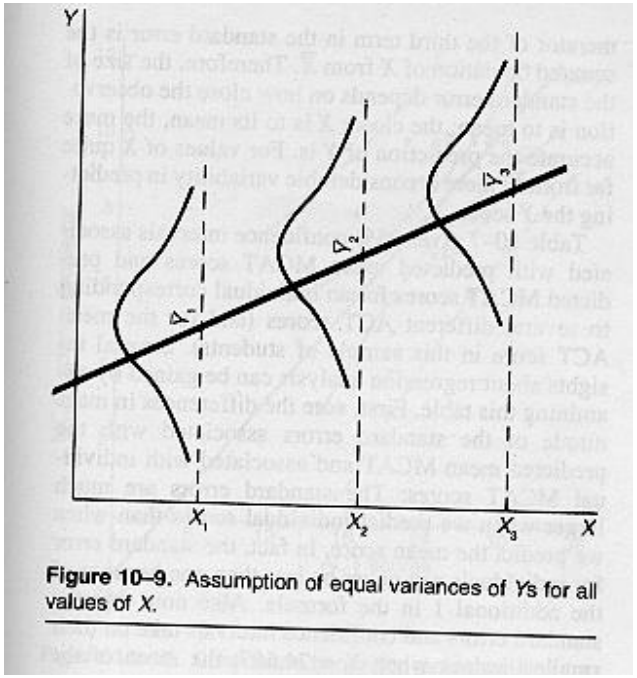
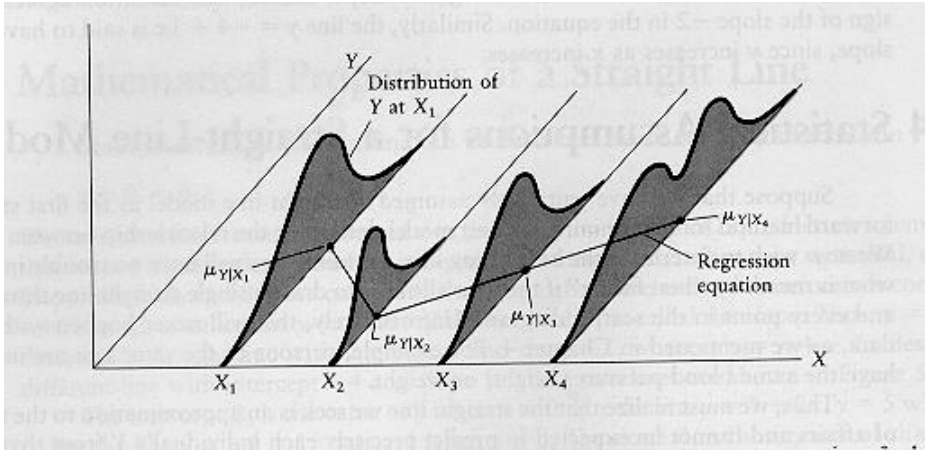
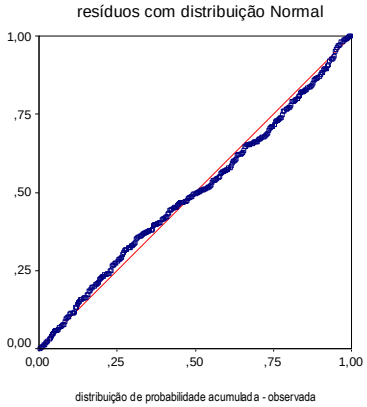
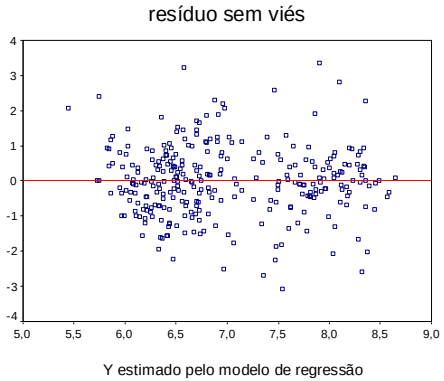


Figure 10–9. Assumption of equal variances of *Y*s for all values of *X*.



amostra independente





Exemplo

- Objetivo: determinar os fatores associados ao tempo entre o início dos sintomas e a procura de serviço especializado em tumores na infância.
  - Y: tempo (meses)
  - X_i: características das crianças, dos pais, do atendimento, queixas
  - análise estratificada segundo diagnóstico

Rodrigues, Camargo,Ribeiro, Latorre, 2003.

Resultados da análise múltipla

- ✓ LLA (ajustado por idade da criança e renda)
  - Presença de anemia (β=1,6 ; p=0,014)
  - Número de médicos (β=0,50 ; p=0,009)
  - Escolaridade materna (β= -0,5 ; p=0,050)
- ✓ Retinoblastoma (ajustado por idade e renda)
  - Estrabismo (β=10,4 ; p=0,002)
  - Tratamento oftalmológico (β=2,10 ; p=0,004)
- ✓ L. de Hodgkin
  - Nenhuma das características da criança, da mãe, do tratamento prévio, da queixa foram significativos.

como testar interação?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 . X_1 . X_2$$

Modelos com muitas variáveis qualitativas

Análise de covariância

Regressão de Poisson:

Y é uma variável quantitativa discreta (processo de contagem)

$$\mathrm{Prob}(Y;\mu) = \frac{\mu^{Y}\, e^{-\mu}}{Y!} \quad Y = 0,1,2,....\infty$$

$$RR_i = \exp(\beta_i)$$

Exemplo

- ✓ Objetivo: verificar a correlação entre os poluentes atmosféricos e a ocorrência de doença pulmonar obstrutiva crônica em idosos (DPOC).

  - Y: número de casos diários de idosos internados por doença respiratória obstrutiva crônica
  - X_i: os poluentes
  - C_i: variáveis climáticas e rodízio de veículos

Martins, Latorre, Saldiva, Braga, 2002

Poluente	Modelo 1	Modelo 2
	Coef. (ep)	Coef (ep)
O_3 (mm dia 4)	0,0036 (0,0013)	0,0030 (0,0014)
SO_2 (mm dia 6)	0,0140 (0,0056)	0,0104 (0,0059)
MP_{10} (mm dia 6)	0,0024 (0,0023)	
CO (mm dia 2)	0,0489 (0,0274)	
NO_2 (mm dia 3)	0,0009 (0,0011)	

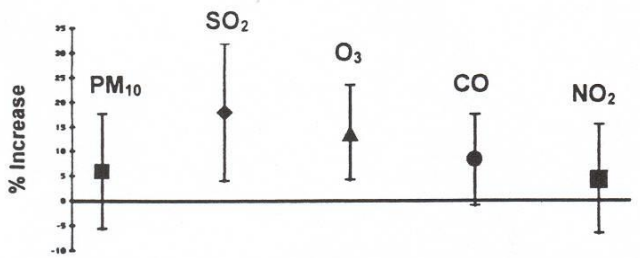
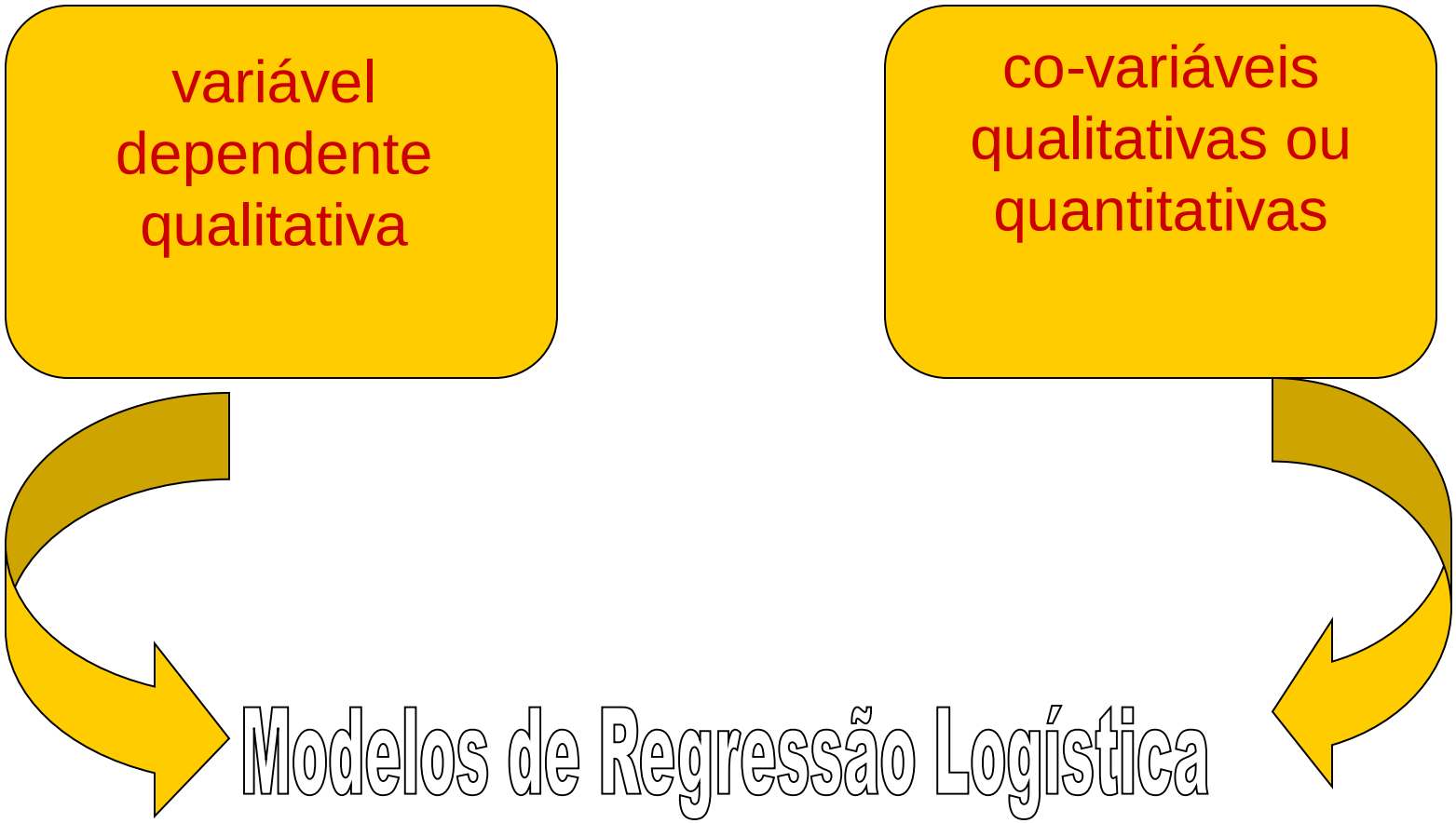
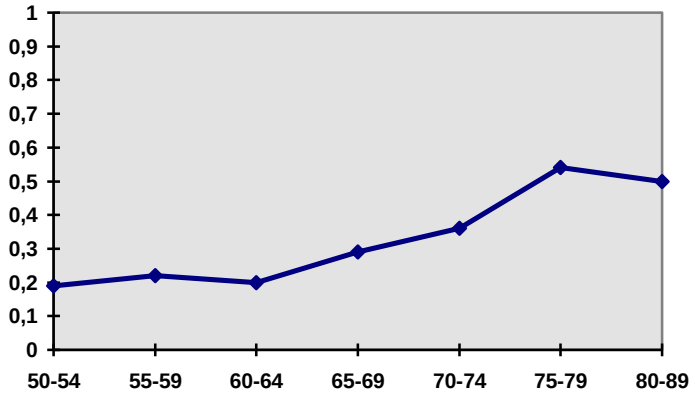


Fig. 2. Percent increase and 95% CI in CLRD emergency room visits due to an interquartile range increase in the 6-day moving average of PM_{10} (24.35 $\mu g/m^3$), 6-day moving average of SO_2 (11.82 $\mu g/m^3$), 4-day moving average of O_3 (35.87 $\mu g/m^3$), 2-day moving average of CO (1.63 ppm), and 3-day moving average of NO_2 (47.7 $\mu g/m^3$).


variável dependente qualitativa
co-variáveis qualitativas ou quantitativas
Modelos de Regressão Logística

$$Prob(Y=1) = p = \frac{1}{1+e^{-f(x)}}$$



Quando a f(x) é uma função linear, tem-se que

$$Prob(Y=1) = p = \frac{1}{1+e^{-\left(\beta_0+\beta_1 X\right)}}$$

e

$$\text{Prob}(Y \neq 1) = \text{Prob}(Y = 0) = 1 - p = 1 - \frac{1}{1+e^{-\left(\beta_0+\beta_1 X\right)}} = \frac{e^{-\left(\beta_0+\beta_1 X\right)}}{1+e^{-\left(\beta_0+\beta_1 X\right)}}$$

MEDIDA DE RISCO

$$OR(X_i) = exp(\beta_i)$$

MODELO PREDITIVO

$$Prob(Y = 1) = f(X_i)$$

regressão logística:

$$p = \text{Prob}(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

onde
$$\begin{cases} OR(X_1) = \exp(\beta_1) \\ OR(X_2) = \exp(\beta_2) \\ \ldots \\ OR(X_k) = \exp(\beta_k) \end{cases}$$

Exemplo

- ✓ Objetivo: Estudar os fatores associados à presença de osteoporose em homens com 50 anos ou mais.

  - Y: osteoporose (sim=1; não=0)
  - X_i: características sócio-demográficas, de estilo de vida, uso de medicamentos, história de fratura familiar

Tanaka, Latorre, Jaime, Florindo, Pippa, Zerbini, 2001.

resultados da análise múltipla

- ✓ características do indivíduo
  - IMC <25 (OR=13,3)
  - 70 e + anos (OR=3,5)
  - cor da pele branca (OR=4,5)

- ✓ outras co-variáveis
  - tabagismo
    - <20 cig. passado (OR=4,6)
    - 20+ cig. passado (OR=4,3)
    - atual (OR=6,4)
  - escore de EF/laser 12m
    - 1o. tercil (OR=15,4)
    - 2o. tercil (OR=5,9)
    - 3o. tercil (OR=1,0)
  - uso de diurético tiazídico (OR=0,28)
  - história materna de fratura não traumática (OR=6,3)

Exemplo

- Objetivo: definir grupos de risco para a organização de serviços de saúde materno-infantil em Curitiba (Programa “Nascer em Curitiba Vale a Vida”).
  - Y: óbito infantil tardio (sim e não).
  - X_i: variáveis constantes da DN.

Luhm, Cesar, Latorre, 2002.

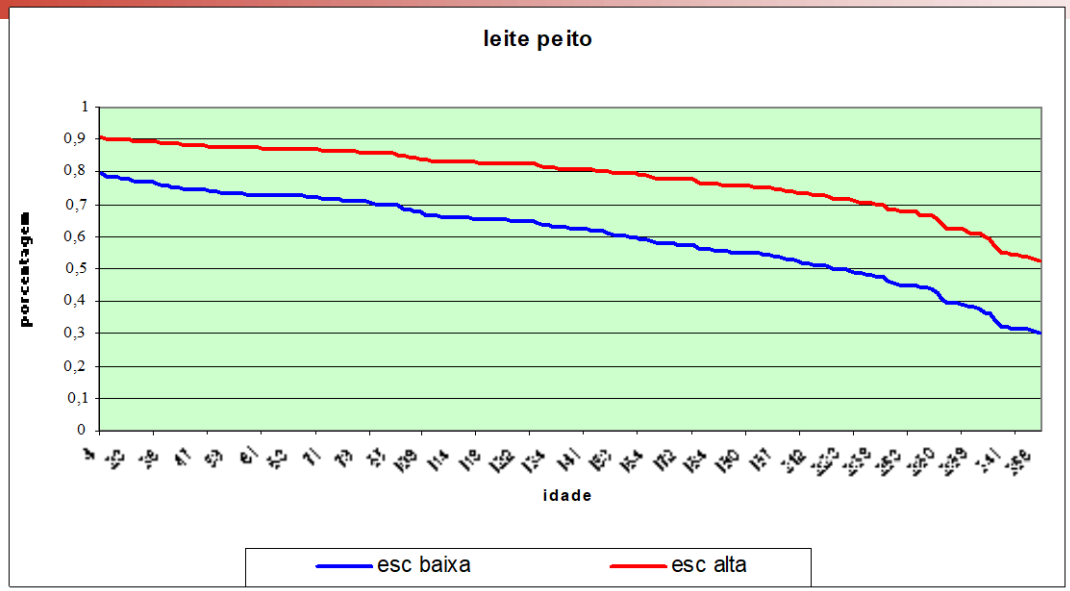
modelo	sensibilidade (%)	especificidade (%)	% de NV
PNCVV	60	73	28
M1	67	68	33
M2	59	77	24
M3	55	81	18

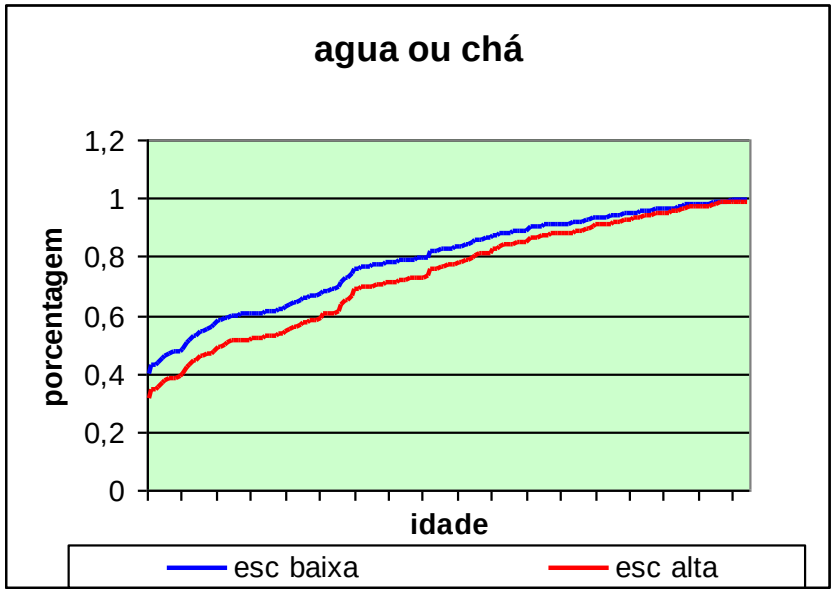
- PNCVV
  - Apgar 5’<7
  - Peso< 2500 g
  - IG<37 semanas
  - Parto fora do ambiente hospitalar
- M3
  - Apgar 5’<7
  - Peso<2500 g

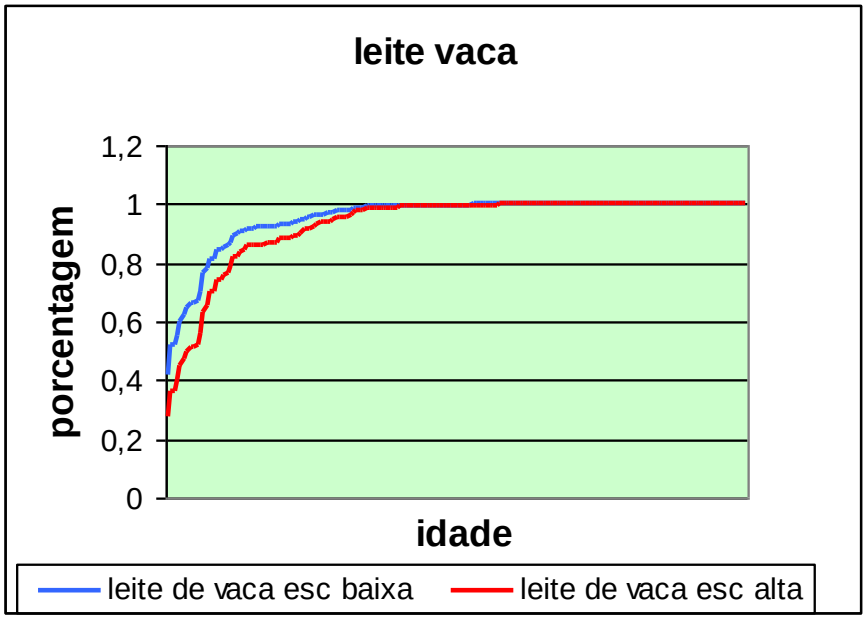
Exemplo

- Objetivo: verificar os fatores associados a diversas práticas alimentares em crianças menores de 1 ano.
  - Y: leite de peito (sim=1; não=0)
    água e chá (sim=1; não=0), etc
  - Xi: características maternas e da criança, sendo a idade da criança variável contínua.

MARCHIONI , LATORRE, SZARFAC, SOUZA, 2001


leite peito
porcentagem
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
idade
esc baixa
esc alta


agua ou chá
1,2
1
0,8
0,6
0,4
0,2
0
porcentagem
idade
esc baixa
esc alta


leite vaca
1,2
1
0,8
0,6
0,4
0,2
0
porcentagem
idade
leite de vaca esc baixa
leite de vaca esc alta

Exemplo

- Objetivo: estudar os fatores associados ao uso de mamadeira em lactentes até 6 meses de vida.
  - Y=uso de mamadeira (sim=1; não=0)
  - X_i: características maternas, da criança, do pré-natal, do atendimento no momento do parto e da puericultura.

AUDI, CORREA, LATORRE, 2003.

Como testar interação?

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1.X_2$$

Tabela 3. Análise múltipla dos fatores associados ao uso de mamadeira em lactentes até 6 meses de vida. Itapira, 1999.

variável independente	categoria	OR_{bruta}	$OR_{ajustada}$	$IC_{95\%}(OR_a)$	p
primípara	sim	1,64	1,66	1,06-2,59	0,027
	não	1,00	1,00		
hospital e puericultura	HM+puericultura	1,00	1,00		
	HM+sem puer.	0,39	0,36	0,07-1,73	0,201
	SC+sem puer.	2,75	2,71	1,34-5,48	0,006
	SC+puericultura	1,01	0,96	0,49-1,87	0,907
teste de	*Hosmer-*	*Lemeshow:*	*p=0,918*		

Estimativas de Probabilidade

$$Prob(Y=1)=\frac{1}{1+exp^{-(-0,626+0,505(primípara)-1,031(HM+sp)+0,995(SC+sp)-0,040(SC+p))}}$$

$$Prob(Y=1/\ não\ pri+HM+puer)=0,35$$

$$Prob(Y=1/\ não\ pri+SC+s/\ puer)=0,59$$

$$Prob(Y=1/\ pri+SC+s/\ puer)=0,71$$

Exemplo

- ✓ Objetivo: Analisar os fatores demográficos, de atividade física, de ingestão de energia e de macronutrientes associados à ocorrência de sobrepeso e obesidade, segundo sexo.

  - Y: estado nutricional:eutrófico, sobrepeso e obeso. (variável qualitativa ordinal)
  - Xi: características demográficas, atividade física e ingestão de energia e macronutrientes.

Costa, Latorre, Slater, 2008

Resultados da análise múltipla de regressão logística multinomial para o sexo masculino. Piracicaba, 2005.

Variável	categoria	sobrepeso	obeso
		OR* (IC95%)	OR* (IC95%)
consumo de energia (kcal/dia) (tercil)	1200,00 a 2500,99	1,37 (0,40; 4,65)	3,62 (0,85; 15,39)
	2501,00 a 3800,99	0,96 (0,20; 4,63)	6,74 (1,50; 30,20)
	3801,00 a 7000,00	1.0	1.0
idade (anos)	10,0 a 12,9	1,00 (0,33; 3,03)	2,01 (0,70; 5,72)
	13,0 a 14,9	1.0	1.0
tempo min/dia pratica AF/exercic	< 60 min/dia	0,25 (0,03; 2,03)	0,60 (0,14; 2,49)
	≥ 60 min/dia	1.0	1.0

OR*: eutrófico é categoria de referência.

Tabela 20. Modelo múltiplo final de regressão logística para os fatores associados à perda auditiva pela classificação BIAP.

Variável	Categoria	OR	p*
Otite média supurada	Não	1,0	
	Sim	5,7	0,001
Lamivudina (3TC)	Não	1,0	
	Sim	5,8	0,028

*p (teste Hosmer-Lemeshow) = 0,781

Tabela 21. Número e porcentagem de pacientes, segundo uso de lamivudina (3TC) e/ou ocorrência de otite média supurada e perda auditiva pela classificação BIAP. ICr, 2010.

Variável	Normal		Perda		Total		p
	N°	%	N°	%	N°	%	
Não recebeu 3TC+ teve ou não OM supurada*	23	88,5	3	11,5	26	100,0	<0,001
Só 3TC sem OM supurada	38	67,9	18	32,1	56	100,0	
3TC + OM supurada	7	29,2	17	70,8	24	100,0	
TOTAL	**68**	**64,2**	**38**	**35,8**	**106**	**100,0**	

*somente um paciente teve otite média supurada.

Exemplo

- Objetivos:O sistema financeiro influencia o crescimento econômico devido às funções que este desempenha, tais como: a) mobilização de recursos; b) alocação dos recursos no espaço e no tempo; c) administração do risco; d) seleção e monitoração de empresas; e e)produção e divulgação de informação. Aplicou-se a técnica de regressão quantílica para analisar esses aspectos para dados de 77 países, o que permitiu um mapeamento mais completo do impacto gerado pelas medidas de desenvolvimento financeiro na distribuição condicional da variável resposta (medidas de crescimento econômico).

- Silva e Porto Junior, 2007.

Brevemente....

- ✓ A técnica de regressão quantílica permite caracterizar toda a distribuição condicional de uma variável resposta a partir de um conjunto de regressores;
- ✓ A regressão quantílica pode ser usada quando a distribuição não é gaussiana;
- ✓ A regressão quantílica é robusta a outliers;
- ✓ Por utilizar a distribuição condicional da variável resposta, podem-se estimar os intervalos de confiança dos parâmetros e do regressando diretamente dos quantis condicionais desejados;
- ✓ Como os erros não possuem uma distribuição normal, os estimadores provenientes da regressão quantílica podem ser mais eficientes que os estimadores por meio de MQO;
- ✓ A regressão quantílica pode ser representada como um modelo de programação linear, o que facilita a estimação dos parâmetros.
- ✓ Muitos pacotes econométricos já possuem comandos próprios para esta finalidade, tais como S-PLUS, Stata, SHAZAM, entre outros.
- ✓ Regressão quantílica pode ser vista como uma extensão natural dos quantis

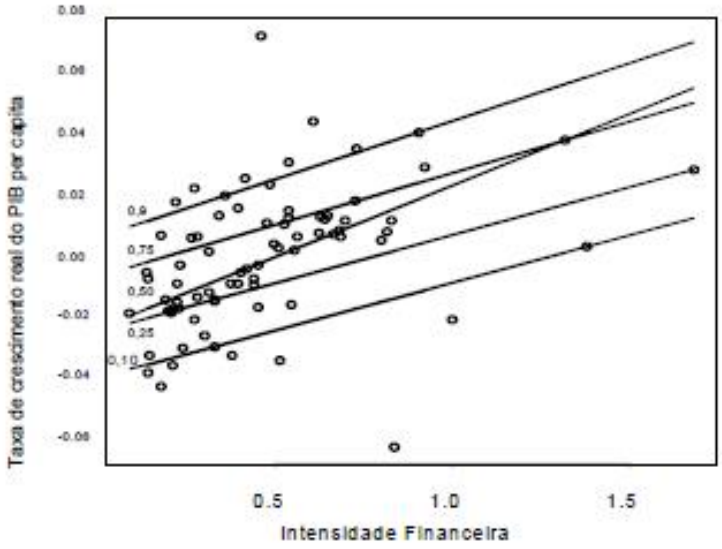
O quadro 1 mostra as estimativas obtidas pelo método de Mínimos Quadrados Ordinários (MQO) e Regressão Quantílica (RQ). A variável explicativa de interesse é a medida de *intensidade financeira* (IF) e como variáveis dependentes a taxa de crescimento real do PIB *per capita* (PIB) e a taxa de crescimento real do capital *per capita* (CAPITAL).

Quadro 1: Resultados por MQO e RQ. Variável explicativa: intensidade financeira

Variável Dependente	MQO	RQ 0,1	RQ 0,25	RQ 0,5	RQ 0,75	RQ 0,9
Taxa de crescimento real do PIB *per capita*	**.0379723*** (.0066534)	**.0330522** (.0306141)	**.0393209*** (.0118841)	**.0455272*** (.0058049)	**.0307359*** (.0075574)	**.0390612**** (.0170946)
Taxa de crescimento real do capital *per capita*	**.0350952*** (.0084814)	**.0472281***** (.0292561)	**.0383212*** (.0146699)	**.0380457*** (.0048689)	**.0279687**** (.0146657)	**.0216139** (.0226384)

Fonte: Elaboração dos autores
Nota: coeficientes em negrito e desvios-padrão em parênteses.
* significante ao nível de 1%; ** significante ao nível de 5%; *** significante ao nível de 10%.

Figura 1: Crescimento real do PIB *per capita* e Intensidade Financeira



Fonte: Elaboração dos autores

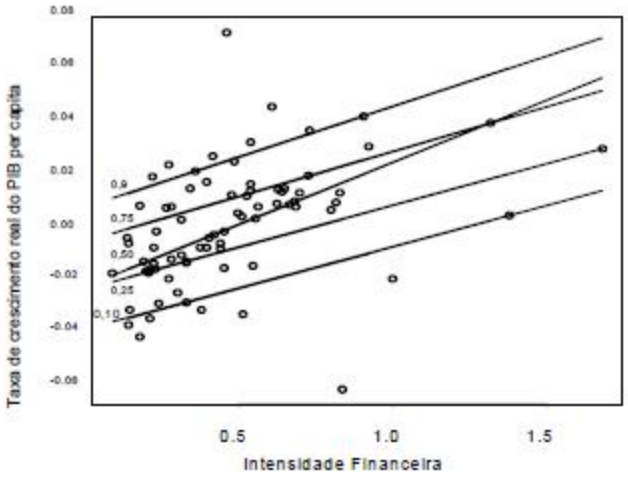
O quadro 1 mostra as estimativas obtidas pelo método de Mínimos Quadrados Ordinários (MQO) e Regressão Quantílica (RQ). A variável explicativa de interesse é a medida de *intensidade financeira* (IF) e como variáveis dependentes a taxa de crescimento real do PIB *per capita* (PIB) e a taxa de crescimento real do capital *per capita* (CAPITAL).

Quadro 1: Resultados por MQO e RQ. Variável explicativa: intensidade financeira

Variável Dependente	MQO	RQ 0,1	RQ 0,25	RQ 0,5	RQ 0,75	RQ 0,9
Taxa de crescimento real do PIB *per capita*	**.0379723*** (.0066534)	**.0330522** (.0306141)	**.0393209*** (.0118841)	**.0455272*** (.0058049)	**.0307359*** (.0075574)	**.0390612**** (.0170946)
Taxa de crescimento real do capital *per capita*	**.0350952*** (.0084814)	**.0472281***** (.0292561)	**.0383212*** (.0146699)	**.0380457*** (.0048689)	**.0279687**** (.0146657)	**.0216139** (.0226384)

Fonte: Elaboração dos autores
Nota: coeficientes em negrito e desvios-padrão em parênteses.
* significante ao nível de 1%; ** significante ao nível de 5%; *** significante ao nível de 10%.

Figura 1: Crescimento real do PIB *per capita* e Intensidade Financeira



Fonte: Elaboração dos autores

questões importantes

- ✓ estudos com técnicas de amostragem complexa
- ✓ modelos multi-nível: é feita uma partição da variância, de acordo com os níveis (cidade, hospital, médico, mãe, indivíduo)

O que fazer com os valores ignorados?

- Imputação?
- Retirar da amostra?
- Analisar como categoria?

Estudos caso-controle

	doente	não doente	TOTAL
EXPOSTO			
NÃO EXPOSTO			
TOTAL	a+c	b+d	

Estudos caso-controle

- ✓ 1950: Wynder EL; Graham EA.
  - Tobacco smoking as a possible etiologic factor in ....
- ✓ 1950: Levin ML; Goldstein H; Gerhardt PR.
  - Cancer and tobacco smoking: a preliminary report.
- ✓ 1950: Doll R; Hill B.
  - Smoking and carcinoma of lung: preliminary report.

Nathan Mantel.

William Haenszel

- ✓ 1959: Mantel N; Haenszel W.
  - Statistical aspects of the analysis of data from retrospective studies of disease.

Estudos caso-controle

- Modelos de regressão logística
  - condicional
  - não condicional

$$\text{Prob}\,(\text{Y}=1)=\frac{1}{1+\exp\left(\beta_0+\beta_1 X_1+\ldots.\beta_k X_k\right)}$$

$$OR(X_i)=\exp\left(\beta_i\right)$$

Exemplo

Exemplo

- Objetivos: analisar os fatores associados à ocorrência de baixa estatura em escolares de Campinas, utilizando o processo de modelagem hierárquico.

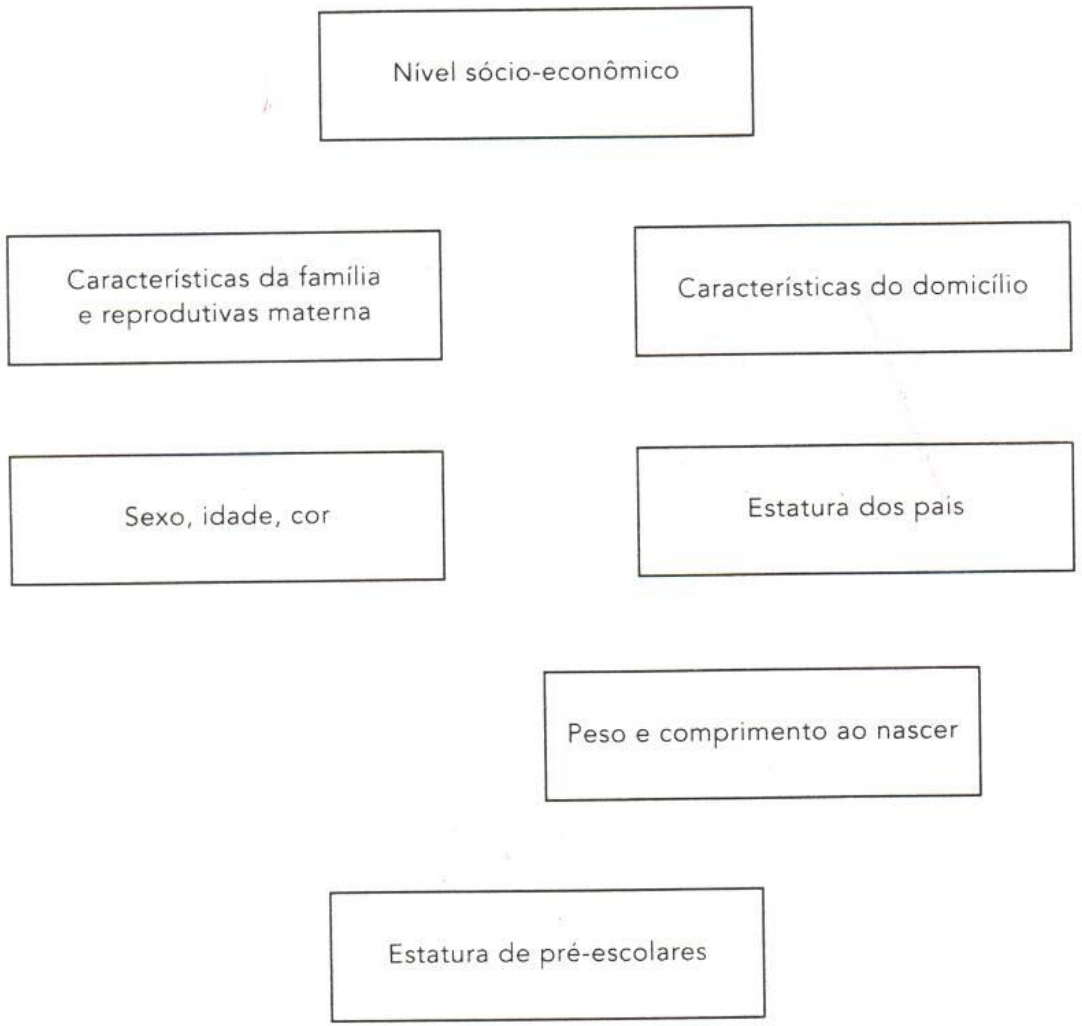
- Y: baixa estatura (sim=1 e não=0)
- X_i: características sócio-econômicas, da família, do domicílio e da criança (sexo e antropometria)

Guimarães, Latorre e Barros, 1999

Modelo hierárquico

Figura 1

Modelo de análise de estatura de pré-escolares.



- ✓ Nível 1:
  - escolaridade materna até 4ª. série (OR=2,1)
  - renda (quanto menor, maior a OR)

- ✓ Nível 2:
  - no. de pessoas que moram no domicílio (> no. > a chance)
  - no. de equipamentos (quanto menor, maior a OR)

- ✓ Nível 3:
  - sexo masculino (OR=2,1)
  - comprimento ao nascer (quanto menor, maior a chance)
  - estatura materna <156,6 (OR=5,9)
  - estatura paterna <169,8 (OR=4,2)

Estudos de coorte

	doente	não doente	TOTAL
EXPOSTO			a+b
NÃO EXPOSTO			c+d
TOTAL			

	doente	não doente	TOTAL
EXPOSTO	a	b	a+b
NÃO EXPOSTO	c	d	c+d
TOTAL	a+c	b+d	N=a+b+c+d

Medida de risco:

RR: risco relativo ➔ $$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

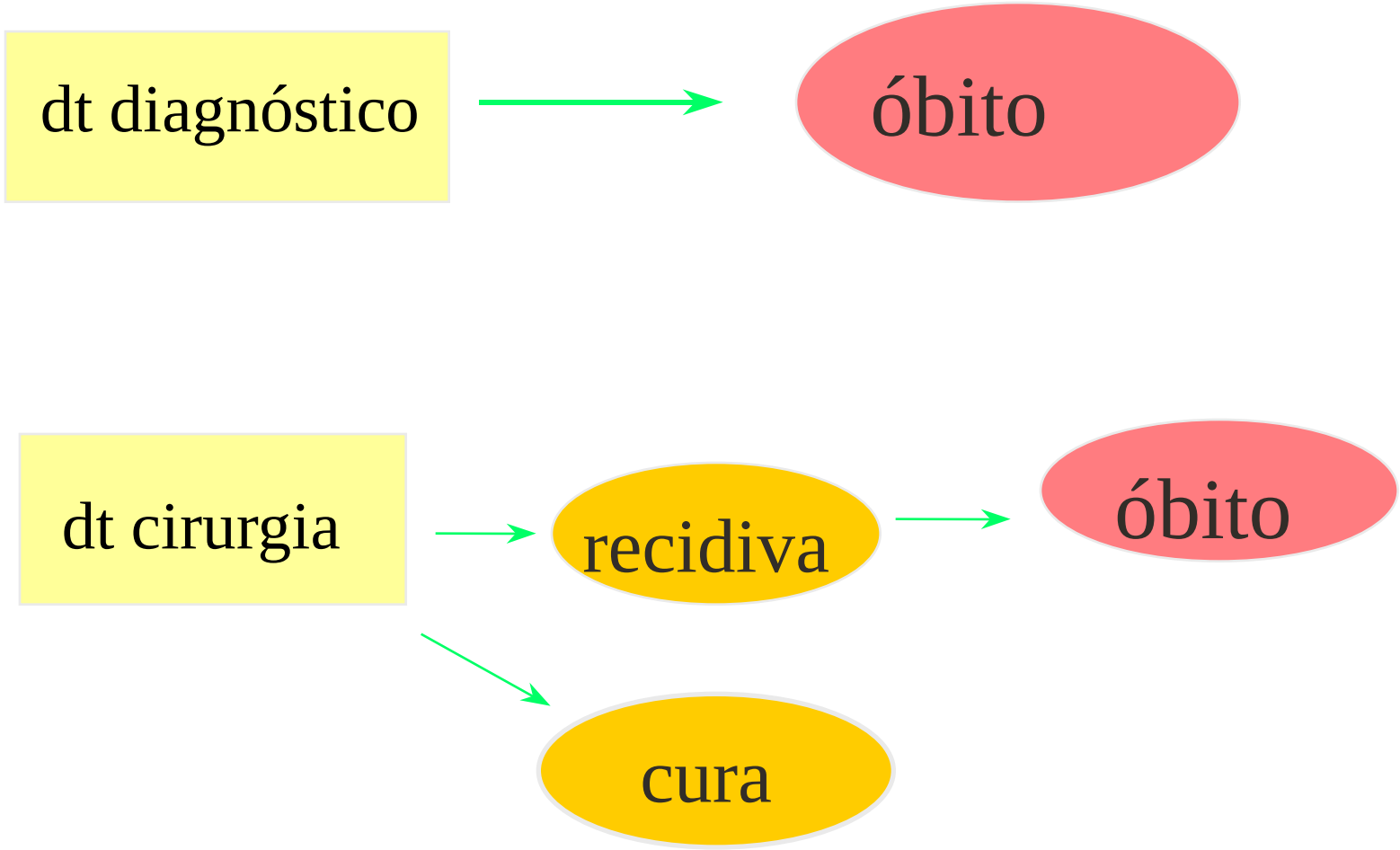
densidade de incidência, incidência acumulada.

Análise de sobrevida

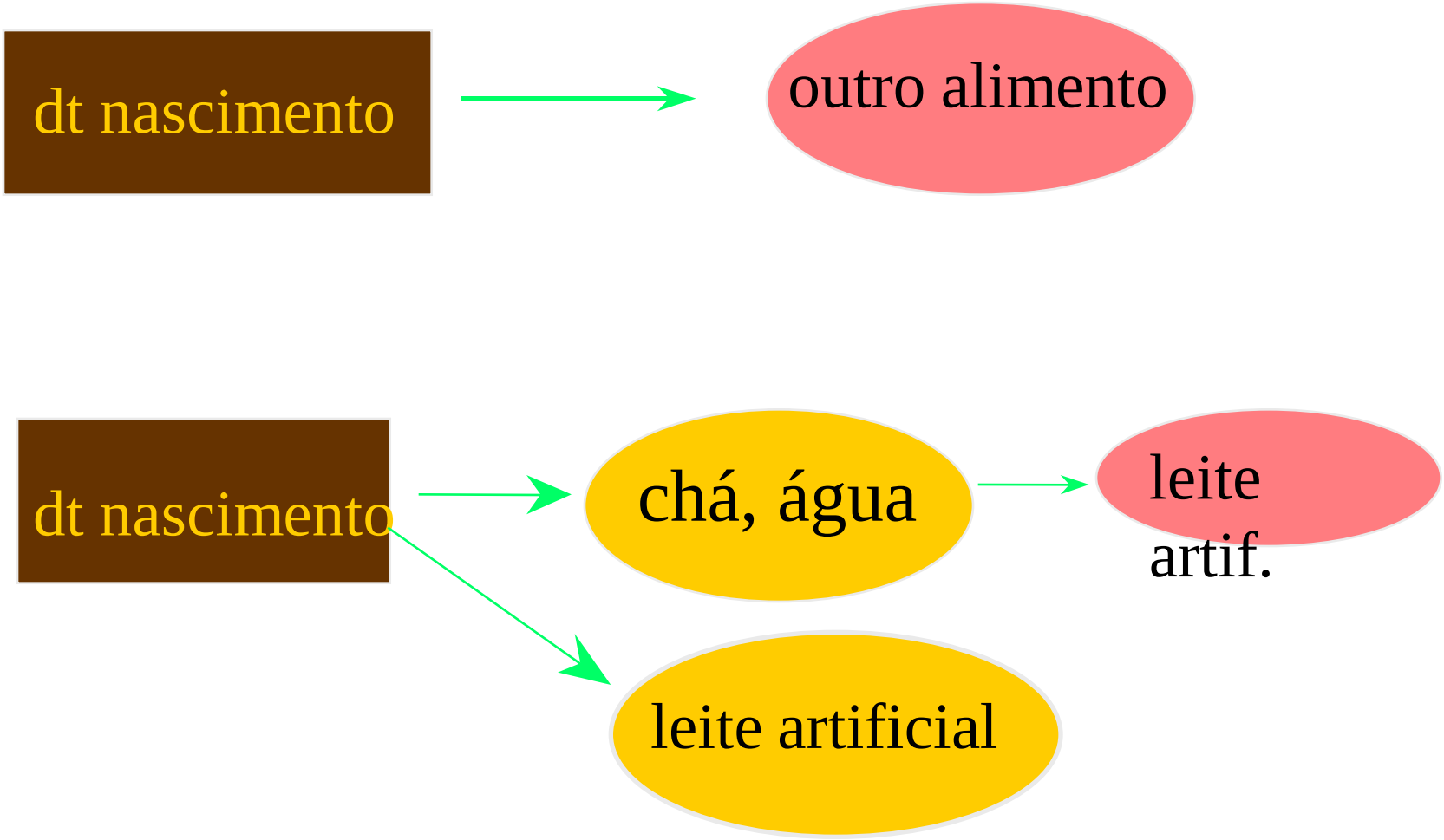
- ✓ variável dependente: o tempo até o aparecimento de um evento.
  - óbito, cura, desmame, introdução de alimento e outros.

- ✓ variáveis independentes (co-variáveis).
  - preditoras, fatores de risco, fatores prognóstico.

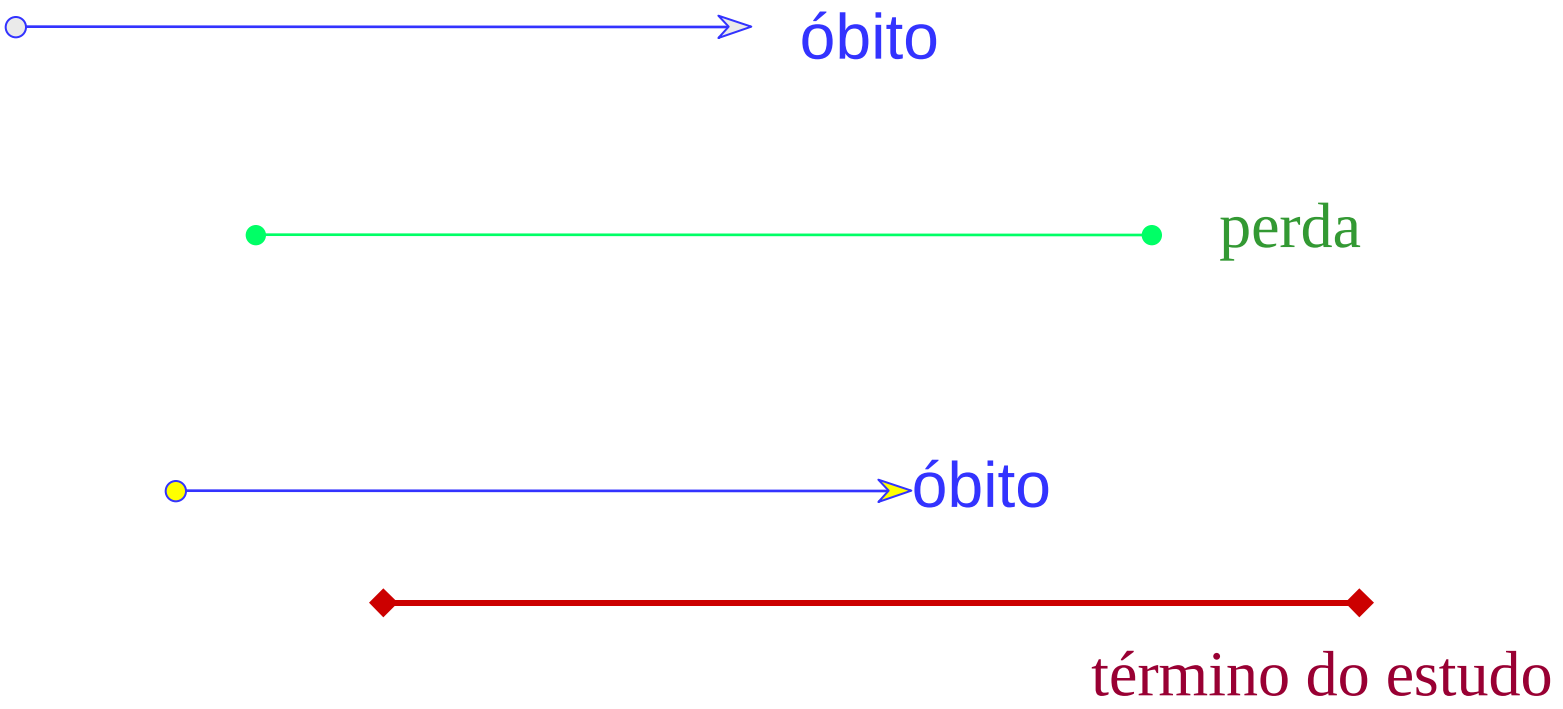
ANÁLISE DE SOBREVIDA



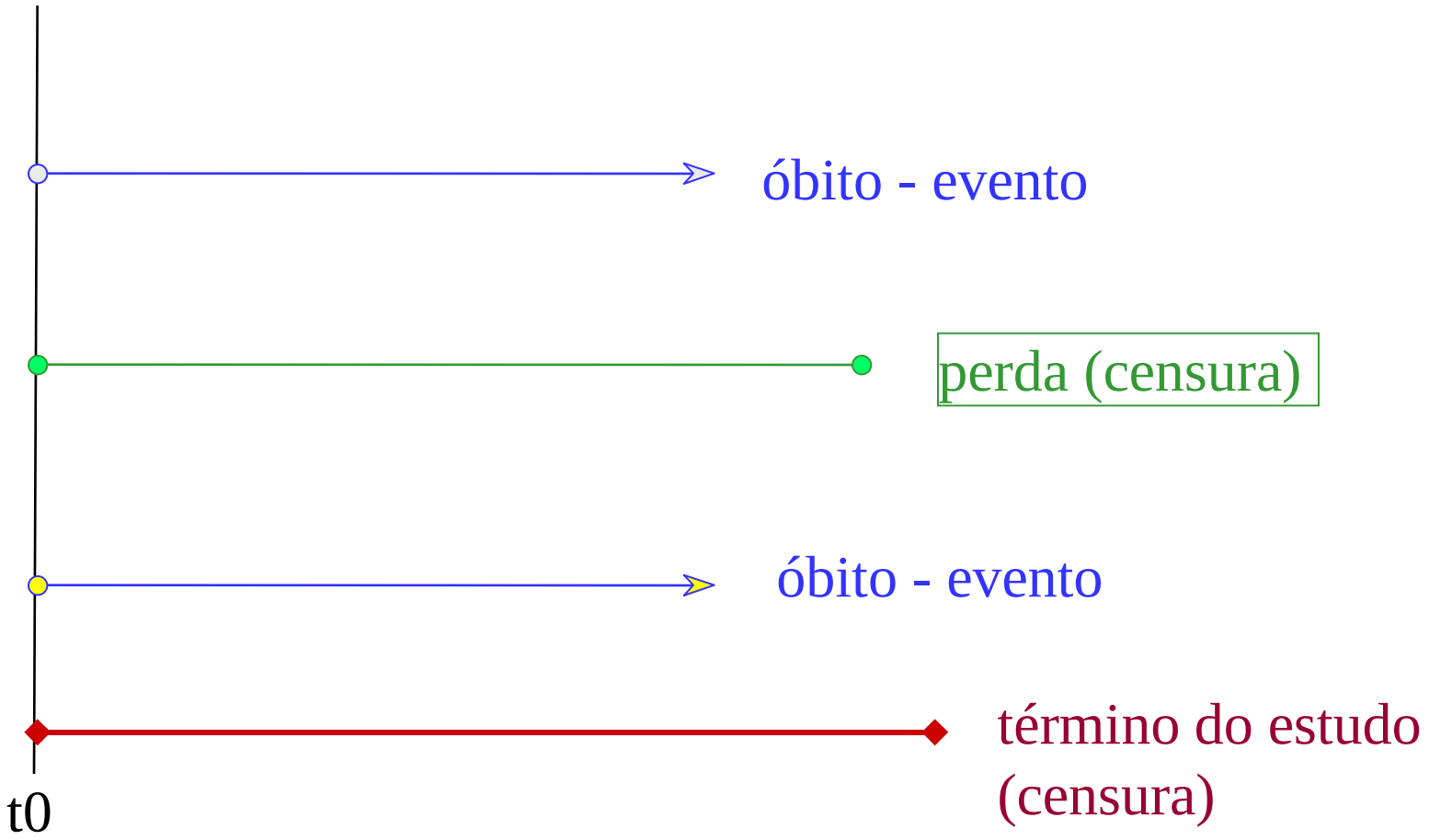
ANÁLISE DE SOBREVIDA



seguimento:
estudos de coorte, ensaios clínicos



análise


óbito - evento
perda (censura)
óbito - evento
término do estudo
(censura)
t0

Técnicas estatísticas mais utilizadas

- ✓ tábua de vida atuarial .

- ✓ modelos lineares generalizados:
  - modelos de regressão de Cox (variável tempo-dependente ou não).
  - modelos de regressão para medidas repetidas.

Estudos de coorte (longitudinais)

- ✓ não há interesse no tempo até a ocorrência do evento
  - modelos de regressão logística

- ✓ há interesse no tempo até a ocorrência do evento
  - análise de sobrevivência

Funções

- ✓ função de probabilidade de sobrevida acumulada (“taxa de sobrevida”) (S(t))

$$S(t) = P[T > t] = 1 - F(t) = \int_{t}^{\infty} f(s)\, ds$$

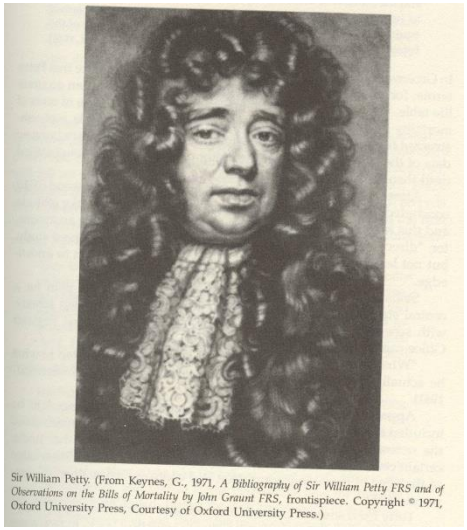
- ✓ função de riscos (*hazard function*) - [**h(t)**]

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\ P[t \le T \le t + \Delta t \,/\, T \ge t] = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

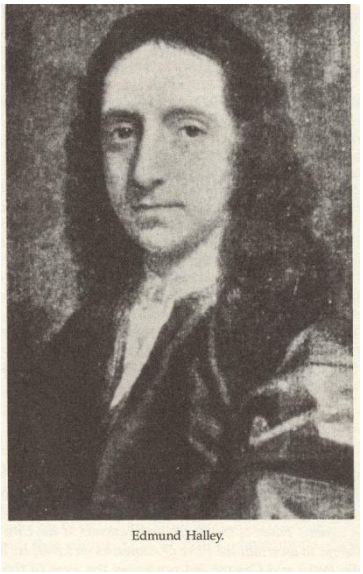
as primeiras tábuas de vida

- Sir William Petty e Sir John Graunt (1620-1674), no século XVII, foram os pioreiros da estatística vital e da estatística aplicada à medicina.

- Edmund Halley (1656-1742):1693 propôs a primeira tábua de vida de grande importância para a demografia, analisando os dados de óbitos ocorridos em Breslaw entre 1587 e 1661.



Sir William Petty. (From Keynes, G., 1971, *A Bibliography of Sir William Petty FRS and of Observations on the Bills of Mortality by John Graunt FRS*, frontispiece. Copyright © 1971, Oxford University Press, Courtesy of Oxford University Press.)



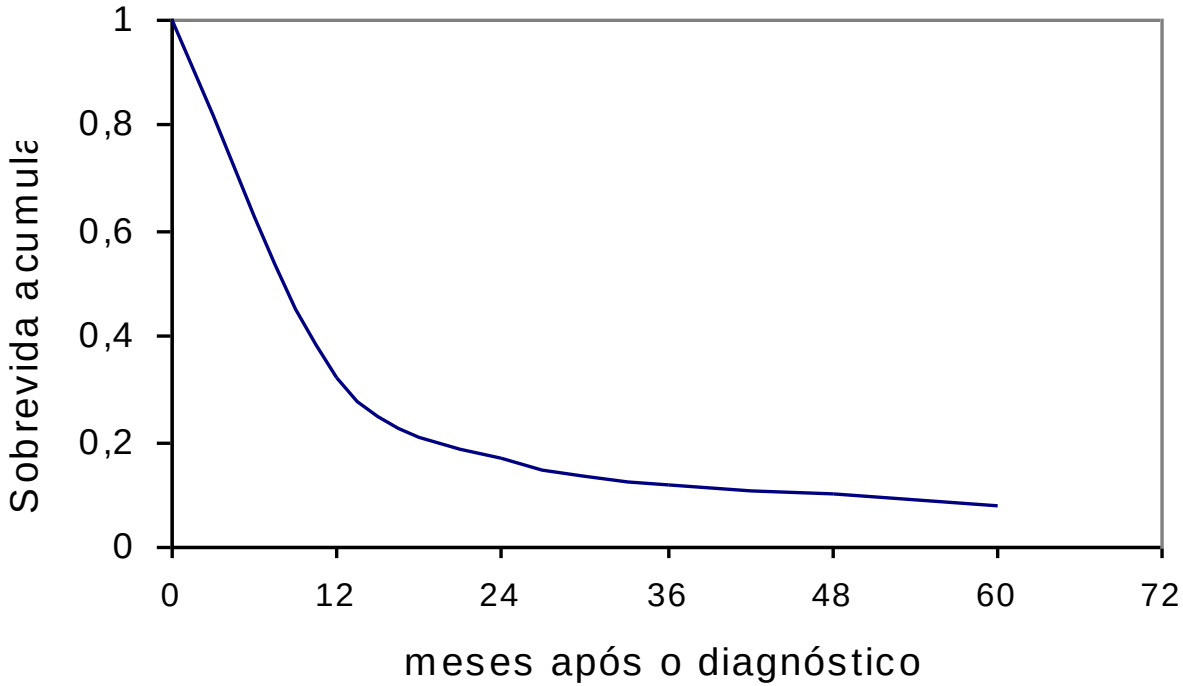
Edmund Halley.

Tábua de vida atuarial

- ✓ observações completas (s/censura)
- ✓ observações incompletas (c/ censura)
- ✓ intervalos de tempo fixos

Exemplo:

Probabilidade de sobrevida acumulada para câncer de estômago, em Campinas-SP , 1991-1994 (método atuarial com observações incompletas)



Bustamante-Teixeira, Faerstein, Latorre, Moreira-Filho, Mahayri , no prelo

Propostas de cálculo de S(t)

- ✓ 1950: Berkson J.; Gage RP.
  - Calculation of survival rates for cancer. *Proc.Staff Meet. Mayo Clin.* **25**:270-86.
- ✓ 1955: Merrel M; Shulman LE.
  - Determination of prognosis in chronic disease, illustrated by systemic lupus erythematosus. *J.Chron.Dis*.**1**:12-32.
- ✓ 1958: Kaplan EL; Meier P.
  - Nonparametric estimation from incomplete observations. *American Statistical Association Journal.***58**:457-81.

Paul Meier

Estimador produto-limite de Kaplan-Meier

(Método de Kaplan-Meier)

- caso particular da tábua de vida onde a divisão de tempo não é arbitrária, mas determinada sempre que aparece uma falha (por exemplo, o óbito). Nessa situação, o número de falhas em cada intervalo deve ser **1**.

Exemplo

- Objetivo: analisar a sobrevida dos tumores da infância em pacientes registrados no RCBP de São Paulo e de Goiânia.

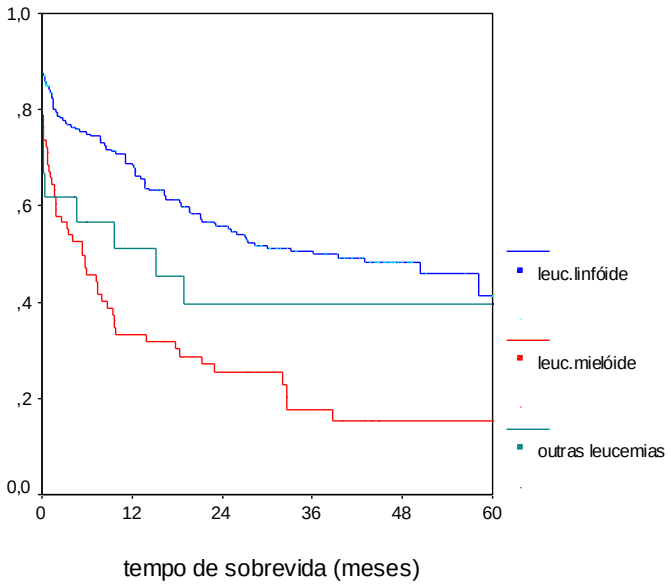
  - Y: tempo do diagnóstico até óbito
  - X_i: sexo, idade, tipo de tumor, ano do diagnóstico

Latorre, Mirra, Curado, Koifman, 2005

leucemias

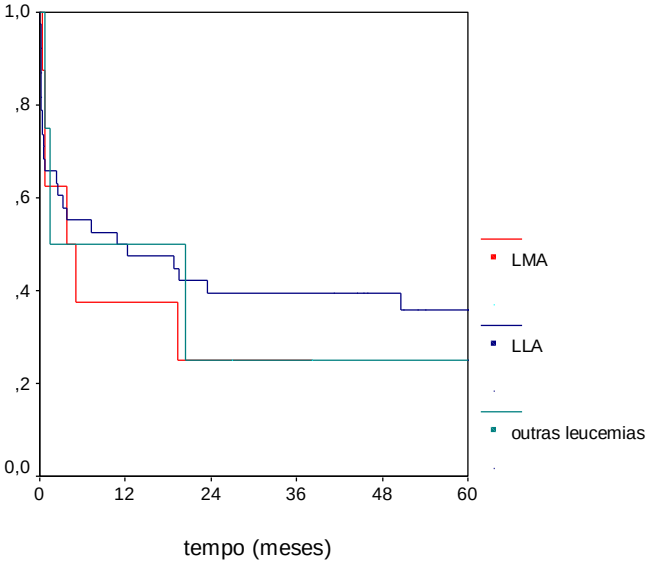
(prob. de sobrevida acumulada após 60 meses)

SP



p<0,001

Goiânia

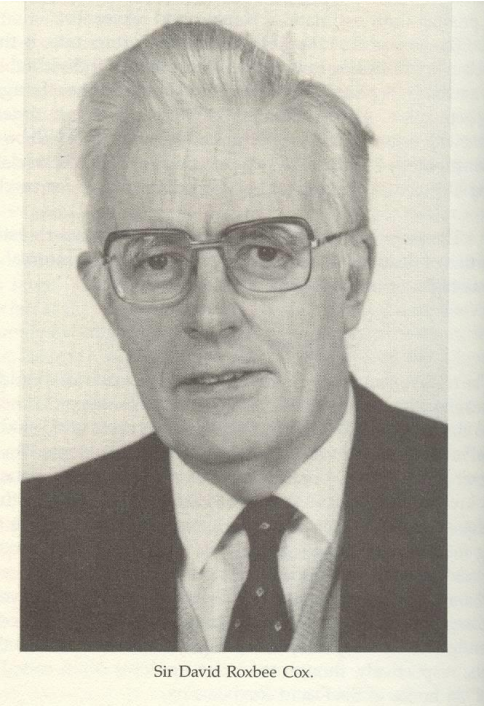


p=0,900

Mirra, Latorre, Curado, Koifman, no prelo

Modelo de riscos proporcionais de Cox

- ✓ Sir David R.Cox (1924-)
- ✓ Regression models and life-tables
  - *Journal of the Royal Statistical Society*. 34:187-202; 1975.

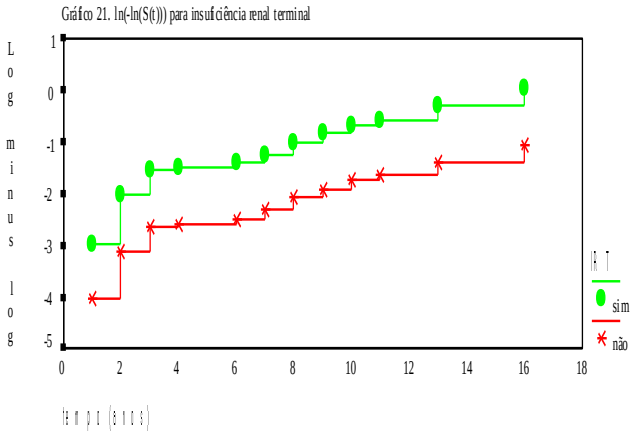


Sir David Roxbee Cox.

Modelo de riscos proporcionais de Cox

$$h(t) = h_0(t).\exp\left(\beta_1 X_1 + ... + \beta_k X_k\right)$$

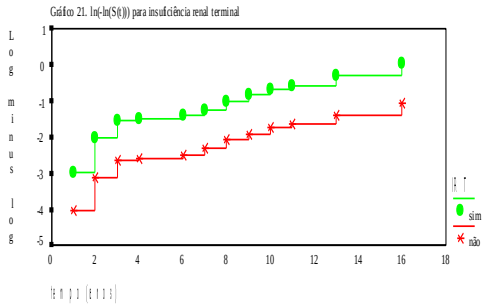
$$HR(X_i) = \exp(\beta_i)$$

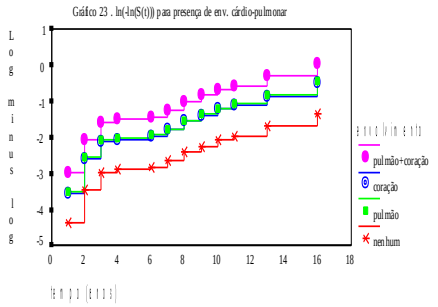


Modelo de riscos proporcionais de Cox

$$h(t) = \exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

$$\begin{cases} HR(X_1) = \exp(\beta_1) \\ HR(X_2) = \exp(\beta_2) \\ .... \\ HR(X_k) = \exp(\beta_k) \end{cases}$$





Exemplo

- Objetivo: propor um escore preditivo de recorrência em pacientes submetidas a tratamento cirúrgico radical do carcinoma do colo do útero estádios IB e IIA.

  - Y: tempo do diagnóstico até a recorrência
  - X_i: características sócio-demográficas,história reprodutiva, clínicas e do tratamento

Fregnani, Latorre, 2002.

Resultados da análise múltipla

- ✓ Idade abaixo de 35 anos (HR=3,67)
- ✓ Estar na menopausa (HR=2,19)
- ✓ Ter mais que 4 gestações anteriores (HR=2,5)
- ✓ presença de linfonodos metastáticos
  - 1ou 2: HR=2,9
  - 3 ou + : HR= 6,4
- ✓ Linfopenia relativa (<15%) – (HR=3,4)
- ✓ Eosinofilia relativa (>10%) – (HR=2,4)

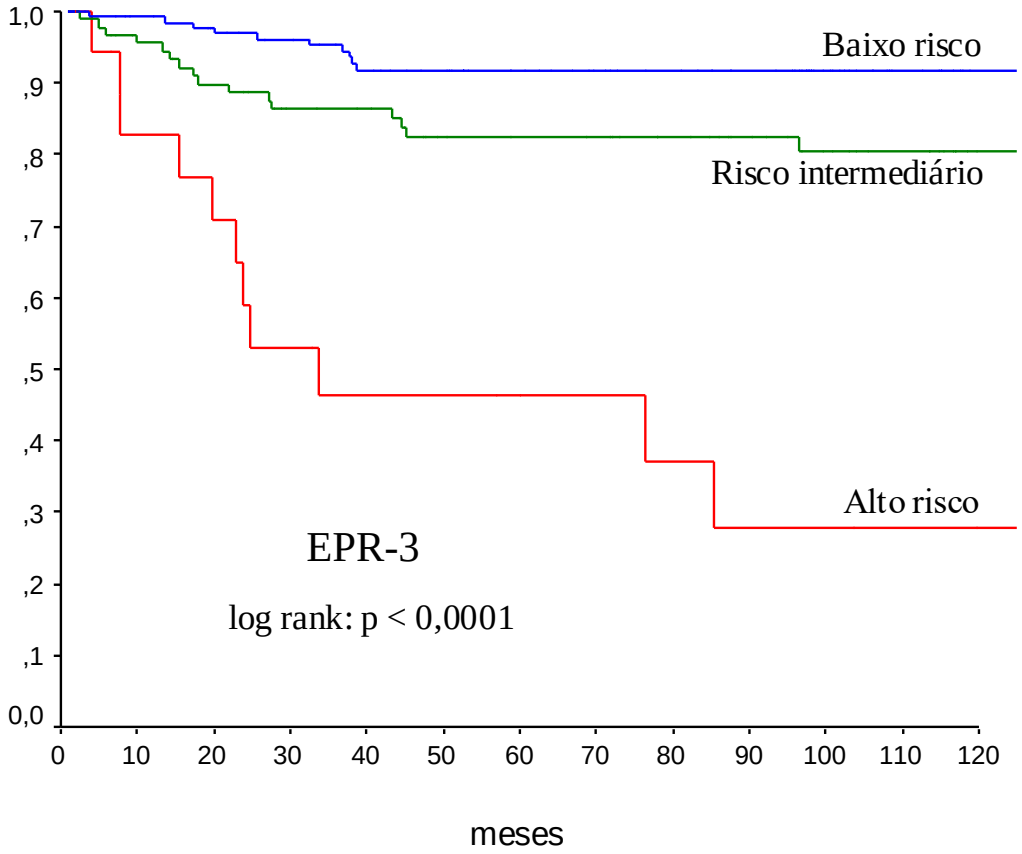
Elaboração de um escore preditivo para recidiva

- β_{min}=0,765 e β_{max}=0,1858

  Δ=1,093

  - 1o. terço: 1 ponto
  - 2o. terço: 2 pontos
  - 3o. terço: 3 pontos

- Idade< 35 anos: 2
- Menopausa: 1
- + que 4 gestações: 1
- 1 ou 2 linf.met.: 1
- 3 ou + linf. met.: 3
- Linfopenia(<15%):2
- Basofilia (>1%): 3
- Eosinofilia (>10%):1



Baixo risco: 0-1 ponto

Risco interm: 2-3 ptos

Alto risco: 4-6 pts

Variável tempo-dependente

- utilizar o modelo de Cox modificado
- utilizar modelo de Processos de Contagem
- incorporar a variável no modelo convencional.

Latorre, 1996

Problemas metodológicos que podem afetar as estimativas das probabilidades de sobrevida e que podem ser controlados na análise

- ➔ mortalidade por outras causas
- ➔ eventos ocorrem em indivíduos agrupados (residência, hospital)
- ➔ falhas no seguimento

Bustamante-Teixeira; Faerstein, Latorre,2002

Medidas repetidas

- ocorrência de vários eventos no mesmo indivíduo
- ocorrência de eventos competitivos ao longo do seguimento

- problema: a maioria dos modelos estatísticos assumem independência nos tempos de falha

- solução: modificar a estrutura de variância

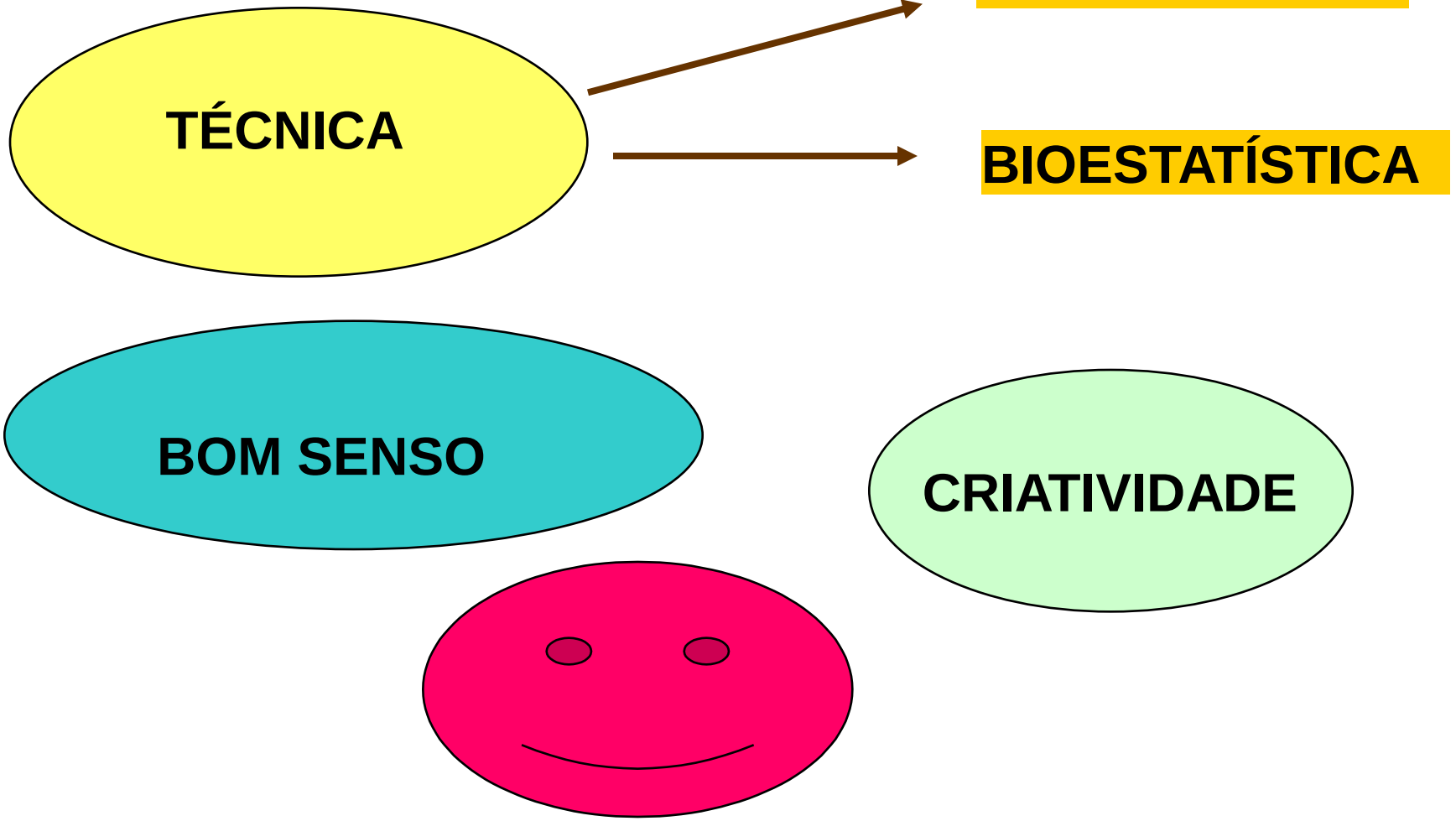
novas abordagens

- ✓ modelos MLG para dados longitudinais:
  - modelos multinível: é feita uma partição da variância, de acordo com os níveis (cidade, hospital, médico, mãe, indivíduo)
  - modelos de transição: processos de Markov

novas abordagens

- variância robusta: é feito um ajuste na variância dos estimadores tradicionais (SAS, Stata, S-Plus)

- modelos de fragilidade *(*frailty*):* assume-se que os indivíduos de um mesmo grupo têm uma *fragilidade* em comum (w), que tem um efeito multiplicativo sobre a h(t)

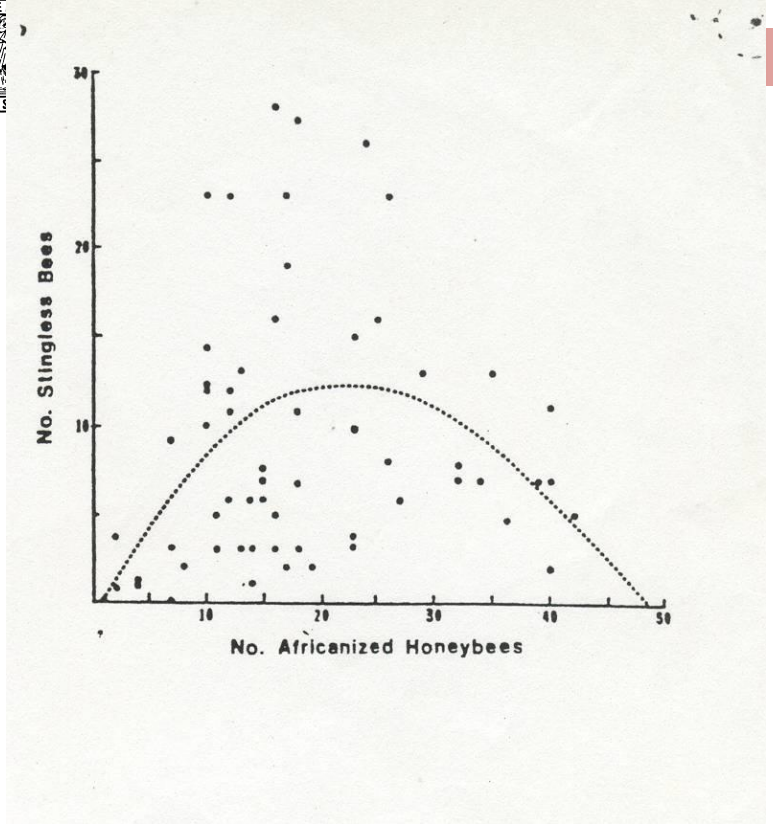

EPIDEMIOLOGIA
TÉCNICA
BIOESTATÍSTICA
BOM SENSO
CRIATIVIDADE

no planejamento da pesquisa

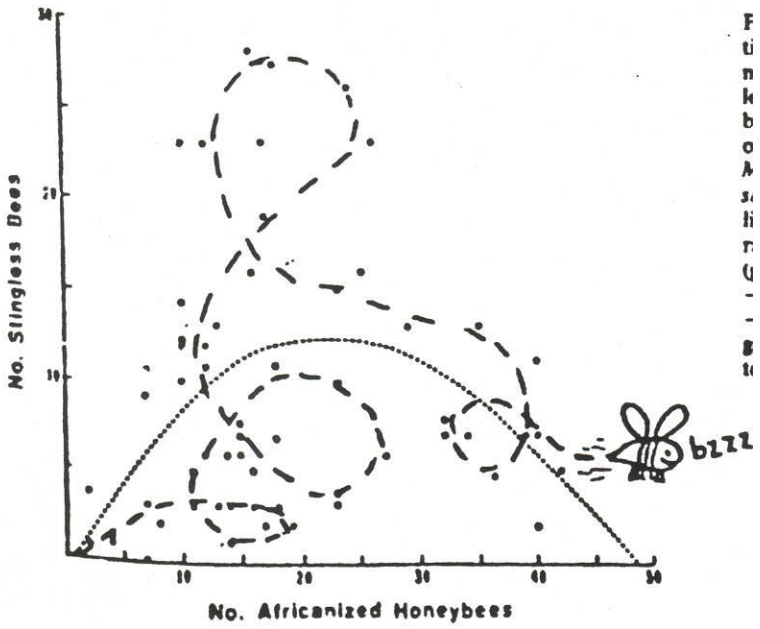
- ✓ verificar quais as variáveis de interesse
- ✓ verificar quais as possíveis variáveis de confusão que deverão estar no modelo
- ✓ verificar quais os possíveis efeitos de interação deverão ser testados
- ✓ definir/sugerir um modelo teórico
- ✓ estimar o tamanho da amostra necessário para fazer a análise múltipla

durante e após o processo de modelagem

- ✓ definir um modelo parcimonioso
- ✓ verificar os possíveis viéses do estudo
- ✓ interpretar os resultados, verificando a sua plausibilidade
- ✓ validar o modelo

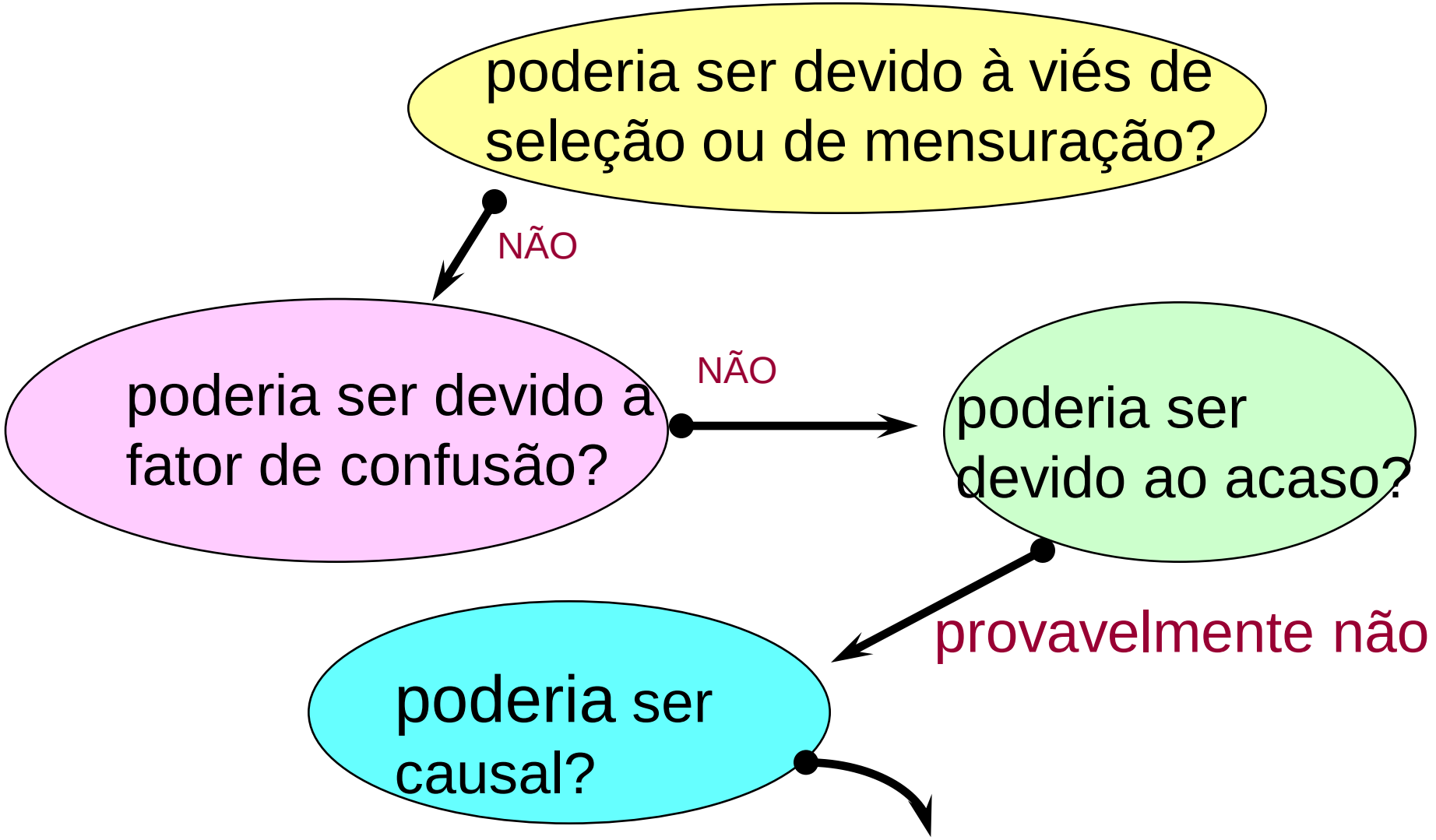


Cuidado com o modelo *abelha*!



Hazen, Science,1978

ASSOCIAÇÃO ESTATÍSTICA OBSERVADA



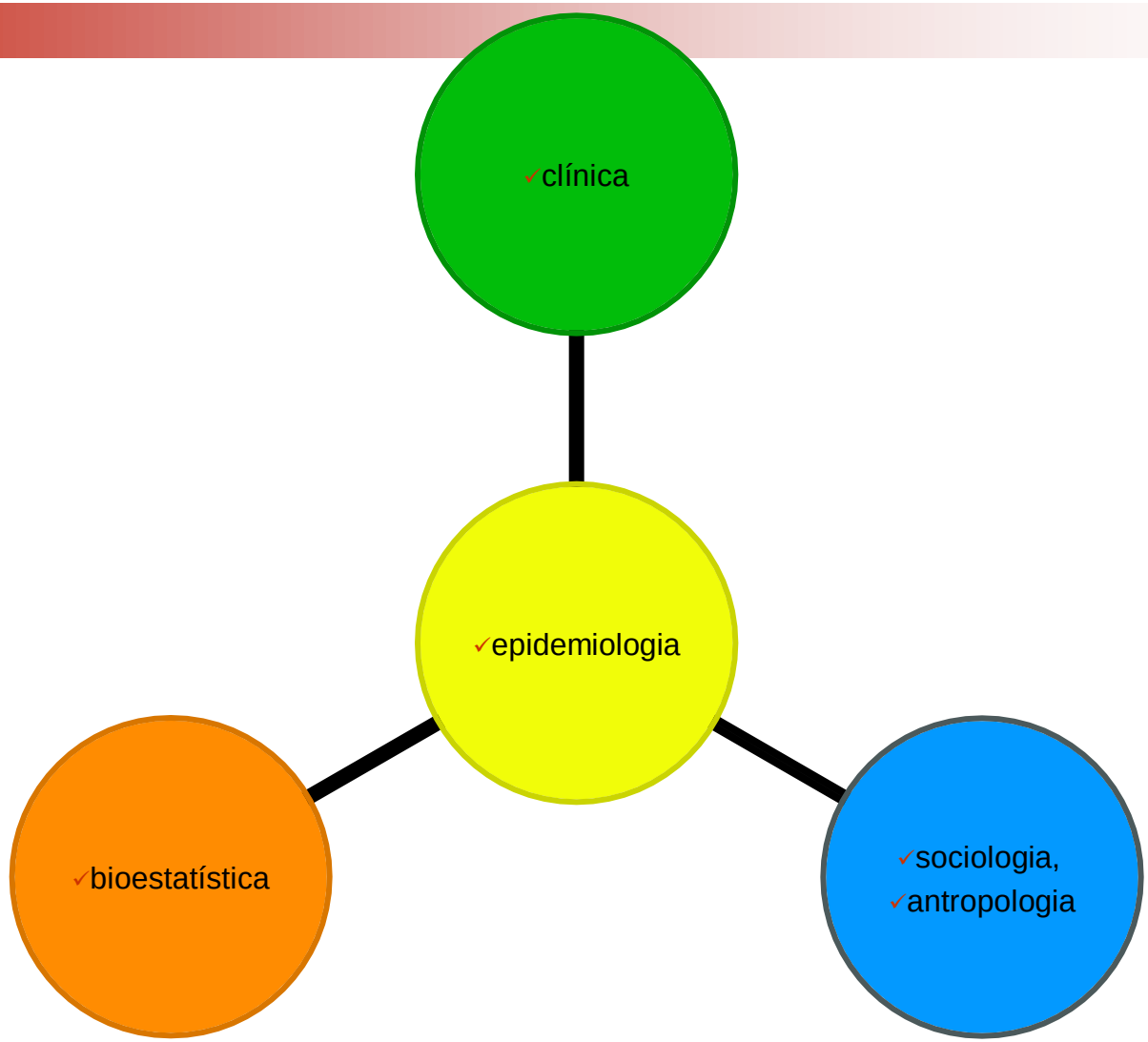
USE OS CRITÉRIOS DE HILL

BEAGLEHOLE, BONITA & KJELLSTRÖM, 1996

CRITÉRIOS PROPOSTOS POR HILL (1965)

⇨ FORÇA DA ASSOCIAÇÃO

- ⇨ CONSISTÊNCIA DOS RESULTADOS
- ⇨ ESPECIFICIDADE
- ⇨ TEMPORALIDADE
- ⇨ GRADIENTE BIOLÓGICO (EFEITO DOSE RESPOSTA)
- ⇨ PLAUSIBILIDADE BIOLÓGICA
- ⇨ COERÊNCIA DA ASSOCIAÇÃO
- ⇨ EVIDÊNCIAS EXPERIMENTAIS
- ⇨ ANALOGIA

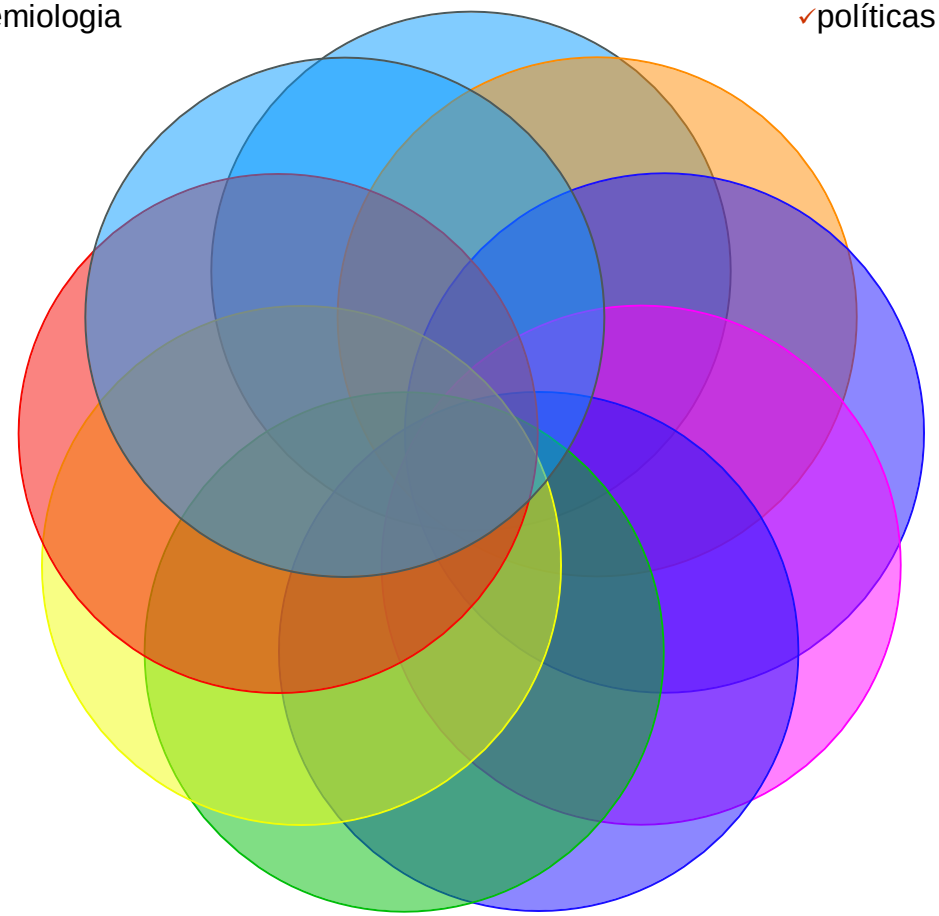

✓clínica
✓epidemiologia
✓bioestatística
✓sociologia,
✓antropologia

Susser, 1985

✓nutrição

✓epidemiologia

✓políticas públicas



✓bioestatística

✓planejamento e
✓gestão

✓sociologia,
✓antropologia

✓clínica

✓meio ambiente

✓biologia
✓ molecular
✓ e genética

Boas férias...


Obrigada
Boa sorte a todos