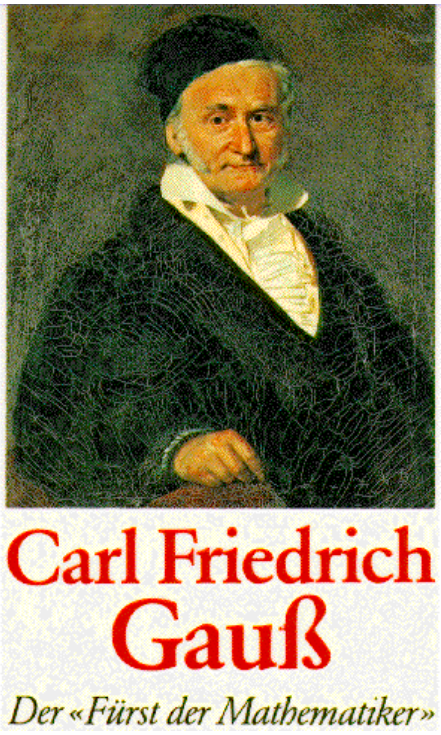
**FACULDADE DE SAÚDE PÚBLICA - USP**

**DEPARTAMENTO DE EPIDEMIOLOGIA**

# MODELOS DE REGRESSÃO APLICADOS À EPIDEMIOLOGIA REG




**Profa. Dra. MARIA DO ROSARIO DIAS DE OLIVEIRA LATORRE**

**Professora Titular do Departamento de Epidemiologia**

**Prof. Convidado: BRUNO CÉSAR SPINELI SILVA**

**Monitor: MATHEUS ABREU**

**2026**

# PROGRAMA

1. Introdução à análise de regressão;

2. Nocões de covariância e correlação;

3. Regressão linear simples e múltipla:
   - estimação dos parâmetros;
   - Tabela de análise de variância (ANOVA);
   - distribuições de probabilidades: Normal, t-Student, F-Snedecor e $\chi^2$;
   - interpretação dos coeficientes;
   - análise dos resíduos;
   - teste F-parcial;
   - correlação parcial e múltipla;
   - variáveis indicadoras*;*
   - como avaliar os efeitos de confusão e interação;
   - escolha do melhor modelo;

4.Análise de regressão logística simples e múltipla:
   - o modelo logístico;
   - estimação dos parâmetros;
   - interpretação dos coeficientes;
   - medidas de ajuste do modelo;
   - como avaliar os efeitos de confusão e interação;
   - escolha do melhor modelo;
   - análise de resíduos.

# BIBLIOGRAFIA RECOMENDADA


1. Berquó ES, Souza JMP; Gotlieb SLD. **Bioestatística**. EPU, 1ª edição revista, São Paulo, 1981.

2. Breslow NE; Day NE. **Statistical Methods in Cancer Research: vol. 1 - The Analysis of Case-Controls Studies**. IARC, Lyon, 1980.

3. DAWSON-SANDERS B; TRAPP RG. **Bioestatística Básica e Clínica**. 3a. edição, Lange - Appleton & Lange/Mc Graw-Hill, 2001.

4. Draper NR; Smith H. **Applied Regression Analysis.** John Wiley and Sons, 3rd edition. New York, 1998.

5. Hosmer DW; Lemeshow S. **Applied logistic regression**. John Wiley and Sons, 2nd edition. New York, 2000.

6. Kleinbaum DG; Kupper LL; Muller KE; Nizam A. **Applied regression analysis and other multivariable methods**. 3rd edition. Brooks/Cole Pub Co, Boston, 1997.

7. Curns AT; Mizam A. **Student solutions manual for Kleimbaum, Kupper, Muller and Nizam´s Applied regression analysis and other multivariable methods.** Brooks/Cole Pub Co, Boston, 1998.

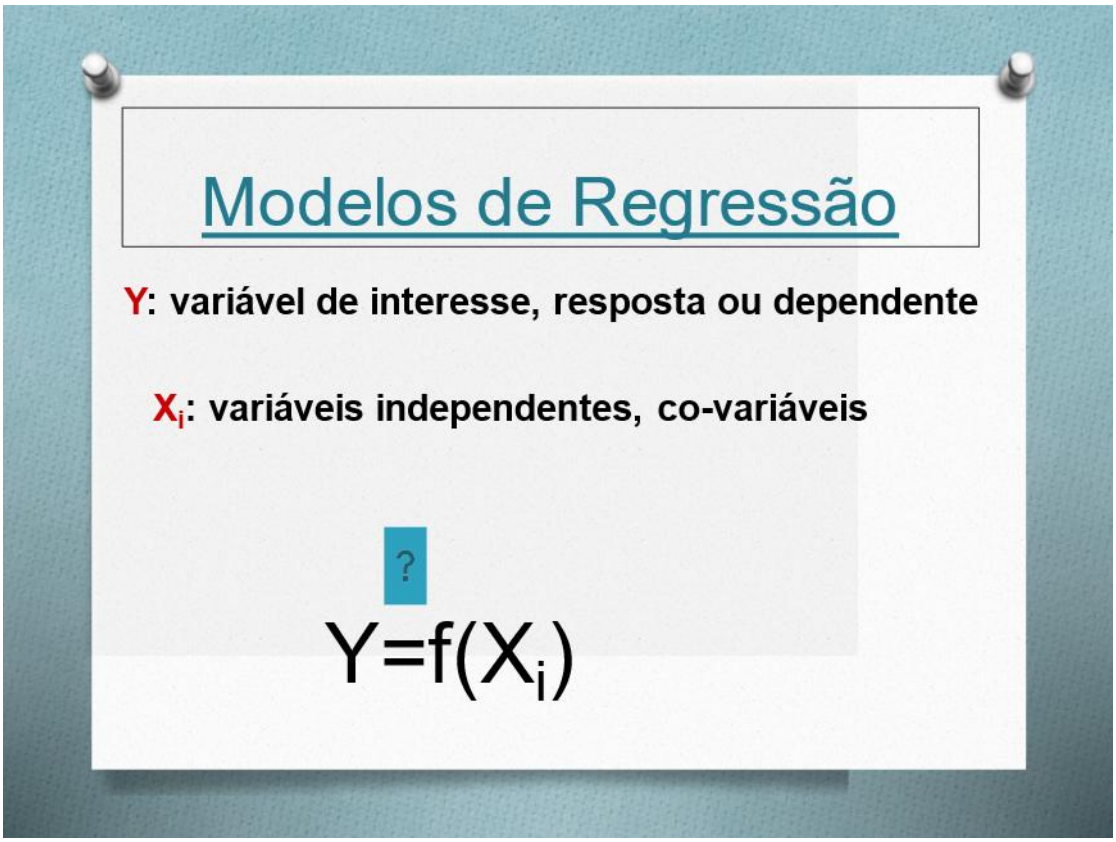
8. Kleinbaum DG; Klein M. **Logistic regression. A self-learning text**. 2nd edition. Springer-Verlag, New York, 2002.

9. Magalhães MN; Lima ACP. **Noções de Probabilidade e Estatística**. EDUSP. São Paulo, 2002.

10. Massad E; Menezes RX; Silveira PSP; Ortega NRS. **Métodos Quantitativos em Medicina.** Manole Editora Ltda. São Paulo 2004.

11. Pereira MG. **Epidemiologia Teoria e Prática**. Rio de Janeiro: Editora Guanabara Koogan, 1999.

# História natural dos alunos do curso de Regressão Logística




**E a escolha do modelo de regressão dependerá de 2 aspectos:**

- **O delineamento**
- **Qual a variável dependente.**

# INTRODUÇÃO À ANÁLISE DE REGRESSÃO

Na prática há diversas situações em que a análise de regressão é apropriada:

1. Quando se deseja caracterizar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes ($X_i$), ié, avaliar a extensão, direção e força da relação (associação).

2. Procurar uma função matemática ou equação para descrever a variável dependente (Y) como função da variáveis independentes ($X_i$), ié, predizer Y em função dos $X_i$; determinando o melhor modelo estatístico que descreva essa relação.

3. Descrever quantitativa e/ou qualitativamente a relação entre os $X_i$ e Y, controlando o efeito de outras variáveis ($C_i$).

4. Verificar o efeito interativo de 2 ou mais variáveis independentes às quais se relacionam com a variável dependente.

5. Determinar quais das muitas variáveis independentes são importantes para descrever ou predizer a variável dependente. Ordenar as variáveis independentes em sua ordem de importância em relação à variável dependente.

6. Comparar múltiplos relacionamentos derivados da análise de regressão.

É importante ser cauteloso sobre os resultados obtidos em uma análise de regressão, ou, de uma maneira mais geral, em qualquer análise utilizando técnicas estatísticas que procurem quantificar uma associação entre 2 ou mais variáveis.

A análise estatística pode estar correta, porém os dados podem estar viciados e/ou incompletos.

(vícios no delineamento, na amostragem, nas medidas, na escolha das variáveis e outros)

O achado de uma associação estatística significativa em um particular estudo não estabelece uma relação causal.

# QUESTÕES BÁSICAS

- Qual a função matemática mais apropriada a ser utilizada? (Em outras palavras: os dados se ajustam melhor a uma reta? A uma parábola? A uma função logística?)

- Como determinar o melhor modelo que se ajuste aos dados?

- Qual a validade e a precisão da(s) estimativa(s) do(s) coeficiente(s) de regressão?

- A presença, no modelo, de determinada variável independente melhora a precisão do mesmo?

- Dado um modelo específico, o que ele significa?

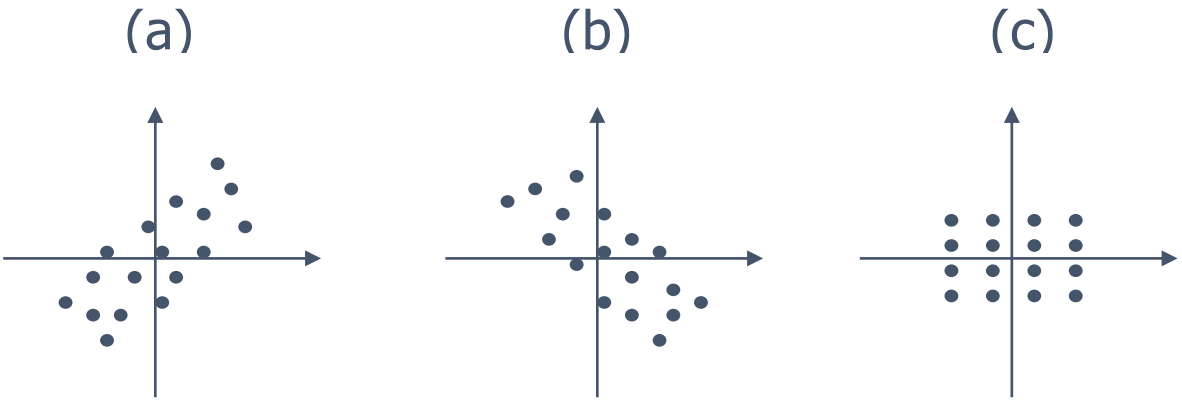
## ESTRATÉGIAS (*stepwise*):

MODELO MAIS COMPLEXO ➔ MAIS SIMPLES
*(BACKWARD SELECTION)*

MODELO MAIS SIMPLES ➔ MAIS COMPLEXO
*(FORWARD SELECTION)*

# O COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON (ρ )

É a análise da associação entre 2 variáveis quantitativas




O coeficiente de correlação linear de Pearson é definido como:

$$r = corr(X,Y) = \frac{cov(X,Y)}{S_X S_Y}$$

*onde* $cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$

*e também pode ser escrito como:*

$$r = corr(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\bar{X})}{S_X}\frac{(Y_i-\bar{Y})}{S_Y}$$

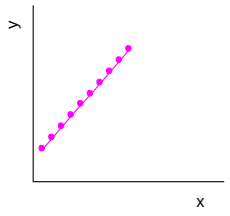
Isto é, o coeficiente de correlação é a média dos produtos dos valores padronizados das variáveis X e Y.

O coeficiente de correlação de Pearson:

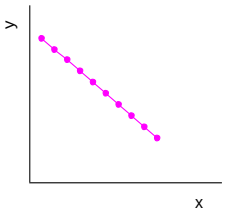
- ✓ Assume valores entre -1 e 1 ($-1 \leq corr(X,Y) \leq 1$)
- ✓ Valores próximos de 1 ou –1 indicam uma associação forte
- ✓ Valores próximos de zero quando não existe associação

O coeficiente de correlação mede:

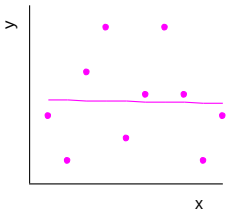
- ✓ Presença de associação linear
- ✓ Força de uma associação linear



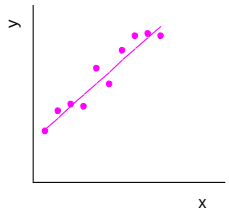

r = 1



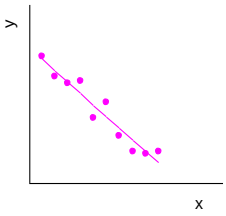

r = -1



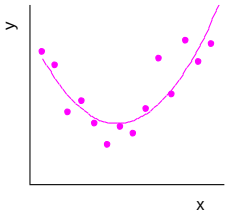

r próximo de 0




r próximo de 1




r próximo de -1




r próximo de 0

Alguns autores sugerem avaliar a presença de associação linear a partir do coeficiente de correlação do seguinte modo:

- de 0,10 a 0,39 - fraca
- de 0,40 a 0,69 - moderada
- de 0,70 até 1 - forte

Mas não há, de fato, uma norma rígida sobre isto.

Deve-se levar em conta o contexto, o tamanho da amostra e sempre avaliar a associação observando conjuntamente o coeficiente de correlação e o diagrama de dispersão.

## *TESTE DE HIPÓTESE PARA $\rho$ :*

$$\begin{cases} H_0 : \rho = 0 \\ H_a : \rho \neq 0 \end{cases}$$

$$t_o = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad ; \qquad \text{onde} \quad t_c \sim t_{n-2}$$

$$\text{obs : é equivalente ao teste } H_0 : \hat{\beta}_1 = 0,$$

$$\text{pois} \quad \beta_1 = \rho \frac{\sigma_Y}{\sigma_X} \qquad \therefore \quad \hat{\beta}_1 = r \frac{S_Y}{S_X}$$

$$\text{INTERVALO DE CONFIANÇA (IC) :}$$

$$IC\left(\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\right) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \pm \frac{z_{1-\alpha}}{\sqrt{n-3}}$$

*OBS: como H0 : $\rho$ = 0 pode ser escrito inteiramente em termos de r e de n, pode-se realizar o teste de hipótese mesmo sem o ajuste de uma linha reta.*

# O MODELO DE REGRESSÃO

A análise que utiliza um modelo de regressão é indicada quando se deseja caracterizar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (Xi), descrevendo Y como f(Xi), estimando a extensão, a direção e a força desta relação/associação.

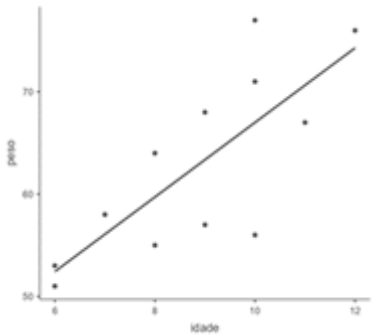
A escolha do modelo depende do **delineamento do estudo e** do **tipo de variável dependente**.

# O modelo de regressão linear simples.

O modelo de regressão linear simples refere-se à análise da relação entre a variável dependente (Y) e uma única variável independente (Xi), ambas quantitativas, assumindo que Y tem distribuição Normal.

A análise da associação entre duas variáveis quantitativas é feita através do coeficiente de correlação linear de Pearson (r). Primeiramente, deve-se construir um diagrama de dispersão para verificar qual a dispersão dos pontos.

Por exemplo, o gráfico mostra que pode-se assumir que existe uma relação linear entre as duas variáveis. Por isso, a utilização de um modelo de regressão com a função de reta é adequada.



A função que representa uma reta é o modelo de regressão linear simples:

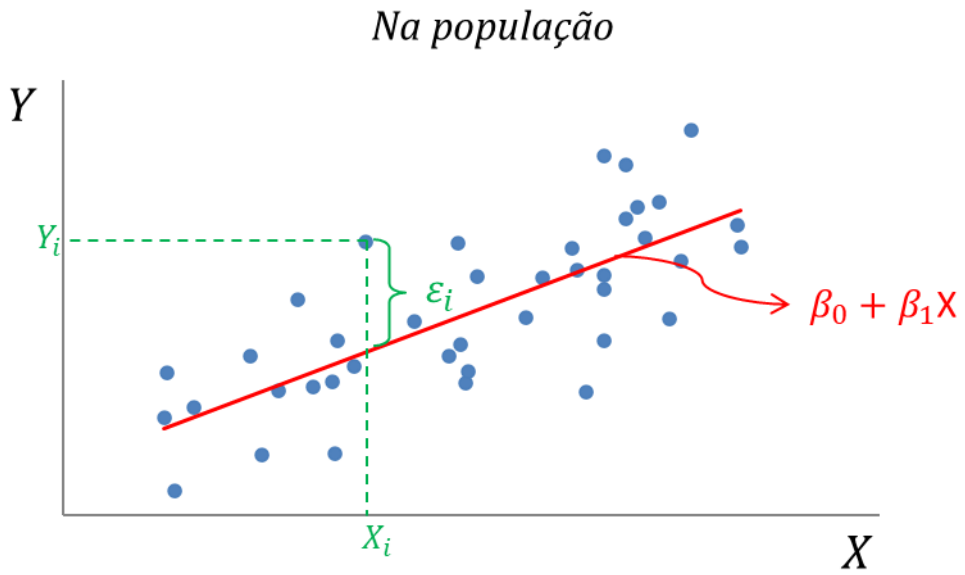
$$Y = \beta_0 + \beta_1 * X$$

Onde:

Y:variável dependente

X: variável independente

$\beta_i$: são os coeficientes de regressão a serem estimados.



# O MODELO DE REGRESSÃO LINEAR SIMPLES

## SUPOSIÇÕES

### 1. Distribuição Normal

Para um valor fixo da variável aleatória (v.a.) X (que, idealmente,deve ser contínua), Y é uma v.a. com distribuição normal, com média e variância finitas.

$$Y \approx N\left(\overline{Y}_{X_i}; S_{Y/Xi}\right)$$

### 2. Os valores de Y são independentes uns dos outros.

(às vezes esta suposição é violada quando se faz diferentes observações no mesmo indivíduo, em tempos diferentes)

### 3. Linearidade

O valor médio de Y ($\overline{Y}_{X_i}$) é uma função de linha reta sobre os $X_i$.

### 4. Homocedasticidade

A variância de Y é a mesma, qualquer que seja X.

$$S^2_{Y/X_i} = S^2_{Y/X_K}, \forall\, i \text{ e } k; \text{ie, } S^2_{Y/X_i} = S^2 \text{ para todo X.}$$

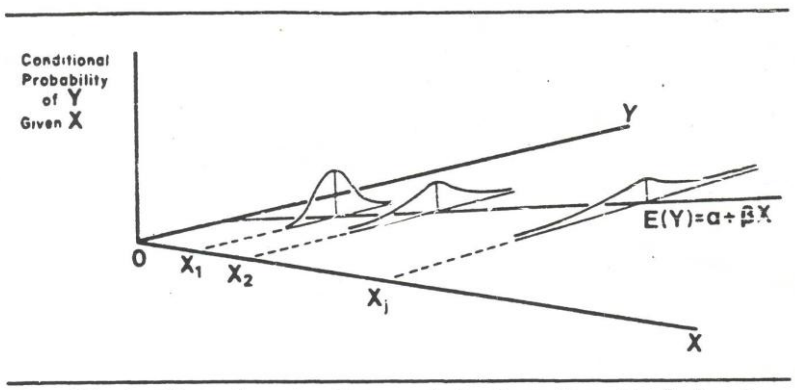



Figure 6.1: An Illustration of a Heteroscedastic Error Term Distribution for a Bivariate Regression Model: COV[VAR(ϵ), X] > 0

# MÉTODOS DE ESTIMATIVAS DE PARÂMETROS

## 1. MÉTODO DOS MÍNIMOS QUADRADOS

É o método que determina a linha reta mais apropriada, minimizando a soma dos quadrados das diferenças entre os valores estimados de Y por meio da reta de regressão ($\hat{Y}$) e os valores observados de Y.

## 2. MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Consiste em determinar uma função, denominada função de verossimilhança $\left[L(y,\theta)\right]$, que é a função de probabilidade de ocorrência daquele específico consjunto de dados e estimar os parâmetros que maximizam a mesma.

# O MODELO DE REGRESSÃO LINEAR SIMPLES

A função que determina uma reta é: $Y = \beta_0 + \beta_1 X$.

Porém, como se deseja fazer uma estimativa, a reta de regressão estimada pode ser escrita da seguinte maneira:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ , e $Y = \beta_0 + \beta_1 X + \varepsilon$ ①, onde $\varepsilon$ = erro $= Y - \hat{Y}$

$\hat{\beta}_0$ e $\hat{\beta}_1$ são estimados pelo Método dos Mínimos Quadrados da seguinte maneira:

Em uma amostra de tamanho n tem-se n pares de observações das v.a. X e Y: $(X_1, Y_1)$, ... $(X_n, Y_n)$ e n equações do tipo ①.

Somando-se todas as n equações, tem-se:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \left( \beta_0 + \beta_1 X_i + \varepsilon_i \right)$$

A soma (S) dos quadrados dos desvios ($\varepsilon$) é:

$$\sum_{i=1}^{n} \left( \varepsilon_i^2 \right) = \sum_{i=1}^{n} \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2$$

Para se encontrar os valores de $\beta_0$ e $\beta_1$ que minimizam a equação acima deve-se derivá-la em relação a $\beta_0$ e $\beta_1$, igualando as equações a zero. (Não se preocupem que não irei demonstrar isso nesse curso!!).

Dessa maneira os valores estimados para $\beta_0$ e $\beta_1$ são:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad ②$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \qquad ③$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad ④$$

Analisando melhor a equação ① ...

$Y = \beta_0 + \beta_1 X + \varepsilon$ (lembrando que $\varepsilon$ =resíduo= $Y_i - \hat{Y}_i$)

Qual o valor esperado para $\varepsilon$ ? $(\varepsilon \to 0)$

Na verdade, $\varepsilon \sim N(0, S_\varepsilon)$.

Substituindo-se o valor de $\hat{\beta}_0$ na equação ① encontra-se que:

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1 \left( X_i - \bar{X} \right).$$

Isso significa que quando $X_i \to \bar{X} \Rightarrow Y_i \to \bar{Y}$.

# PRECISÃO DA RETA ESTIMADA

Considera-se a seguinte identidade:

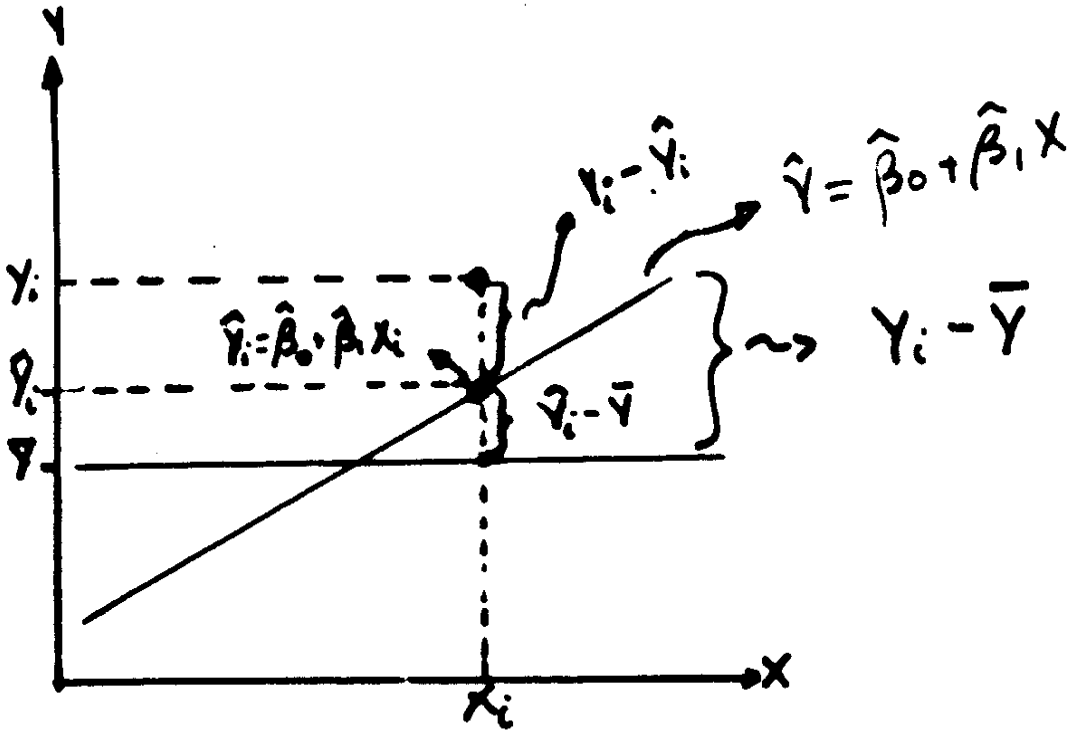
$$Y_i - \hat{Y}_i = \left( Y_i - \bar{Y} \right) - \left( \hat{Y}_i - \bar{Y} \right).$$

Elevando-se ao quadrado os 2 lados da igualdade acima e fazendo-se a soma de todas as n equações (i=1,2, ...,n), obtem-se:

$$\sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2 = \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2 + \sum_{i=1}^{n} \left( \hat{Y}_i - \bar{Y} \right)^2 + 0 \qquad ⑤$$

↓ SQT ↓ SQR ↓ SQM

- **SQT**: soma de quadrados total, ié, soma dos quadrados dos desvios do valor de Y da i-ésima observação em relação à média dos Y.
- **SQR**: soma dos quadrados devido aos resíduos, ié,a soma dos quadrados dos desvios entre o valor de Y da i-ésima observação e seu valor estimado.
- **SQM:** soma dos quadrados devido à regressão, ié, a soma dos quadrados dos desvios do valor estimado de Y para a i-ésima observação e a média dos Y.




$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

A equação ⑤ é chamada a EQUAÇÃO FUNDAMENTAL DA REGRESSÃO e pode ser reescrita como:

soma dos quadrados sobre a média (SQT) = soma de quadrados sobre a regressão (SQR) + soma de quadrados devida à regressão (SQM).

Isso significa que a variação total dos Y's sobre sua média pode ser explicada uma parte pela linha de regressão e outra pelos resíduos. Se todos os Y's caíssem sempre na linha de regressão a SQR seria zero!!

Portanto, quanto mais a SQM for próxima da SQT melhor.

Daí deriva-se uma medida quantitativa de precisão da reta estimada denominada $r^2$ **(coeficiente de determinação)**.

$$r^2 = SQM/SQT \Rightarrow 0 \leq r^2 \leq 1$$

$$\therefore$$

$$\text{quanto mais } r^2 \to 1, \text{ melhor}$$

# ANOVA

| FONTE | SQ | GL | MÉDIA QUADRÁTICA (MQ) | F |
|---|---|---|---|---|
| DEVIDO A REGRESSÃO | $\hat{\beta}_1\left[\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}\right]$ | 1 | SQM/GL | $F_c(1, n-2) =$ |
| DEVIDO AO RESÍDUO | por subtração | n-2 | $S^2 = \frac{SQR}{GL}$ | $\frac{MQM}{MQR}$ |
| TOTAL | $\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$ | n-1 | | |

SQT = SQR + SQM

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

# REGRESSÃO LINEAR SIMPLES

## 1. O MODELO

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = \bar{Y} + \hat{\beta}_1 \left( X - \bar{X} \right) \left( \text{lembrar que } \hat{Y} = \hat{Y}_i = \bar{Y}_{Y/X_i} \right)$$

$$Y_i \sim N\left( \hat{\beta}_0 + \hat{\beta}_1 X_i ; S^2 \right)$$

### 1.1. Estimativas para $S^2 \left( s^2_{Y/X} \right)$

a) $S^2_{Y/X} = \dfrac{SQR}{n-2} = \dfrac{n-1}{n-2} \left( S^2_Y - \hat{\beta}^2_1 S^2_X \right)$ , se $\beta_1 \neq 0$

b) $S^2_{Y/X} = \dfrac{SQM}{1}$ , se $\beta_1 = 0$

## 2.1.INTERVALO DE CONFIANÇA PARA $\hat{Y}_i = \bar{Y}_{Y/X_i} = \bar{Y} + \hat{\beta}_1\left(X_i - \bar{X}\right)$

$$IC = \hat{Y}_i \pm t_{n-2,1-\alpha}\, S_{Y/X}\sqrt{\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{(n-1)S_X^2}}$$

## 2.2.TESTE DE HIPÓTESES PARA $\hat{Y}_i = Y_i'$

$$\begin{cases} H_0 : \hat{Y}_i = Y_i' \\ H_a : \hat{Y}_i \neq Y_i' \end{cases}$$

$$t_o = \frac{\bar{Y} + \hat{\beta}_1\left(X_i - \bar{X}\right) - Y_i'}{S_{Y/X}\sqrt{\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{(n-1)S_X^2}}} \quad \text{onde} \quad t_c \sim t_{n-2}$$

## 2.3 INTERVALO DE PREDIÇÃO PARA $\hat{Y}_i$, onde $\hat{Y}_i$ é observação e não parâmetro.

$$IP = \bar{Y} + \hat{\beta}_1\left(X_i - \bar{X}\right) \pm t_{n-2,1-\alpha}\, S_{Y/X}\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{(n-1)S_X^2}}$$

## 3. A INCLINAÇÃO: $\beta_1$.

a) o estimador:

$$\hat{\beta}_1 = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

b) o Intervalo de Confianca (IC) :

$$IC = \hat{\beta}_1 \pm t_{n-2,1-\alpha} \frac{S_{Y/X}}{S_X\sqrt{n-1}}$$

c) o teste de hipotese :

$$\begin{cases} H_o: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$$

c.1.) teste F:

$F_o = {}^{MQM}\!/_{MQR}$ , onde $F_c \sim F_{1,n-2}$

c.2.) teste *t*:

$$t_o = \frac{\hat{\beta}_1 \; S_X\sqrt{n-1}}{S_{Y/X}} \quad , \text{ onde } t_c \sim t_{n-2}$$

d) $S^2_{\hat{\beta}_1} = \frac{S^2_{Y/X}}{S^2_X(n-1)}$

## 4. O INTERCEPTO: $\beta_0$

a) o estimador:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

b) o intervalo de confianca (IC) :

$$IC = \hat{\beta}_0 \pm t_{n-2,1-\alpha} \, S_{Y/X} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}$$

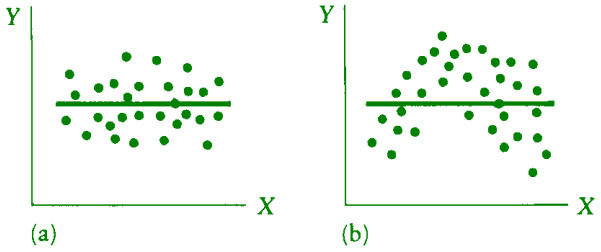
c) o teste de hipotese :

$$\begin{cases} H_0 : \hat{\beta}_0 = 0 \\ H_a : \hat{\beta}_0 \neq 0 \end{cases}$$

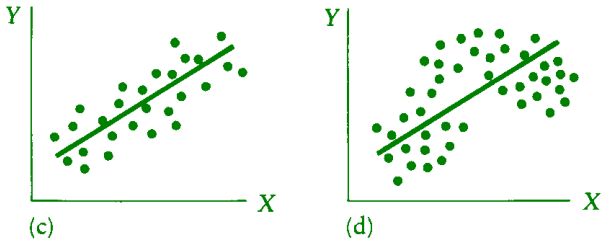
$$t_o = \frac{\hat{\beta}_0}{S_{Y/X} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}} \quad ; \text{ onde } \quad t_c \sim t_{n-2}$$

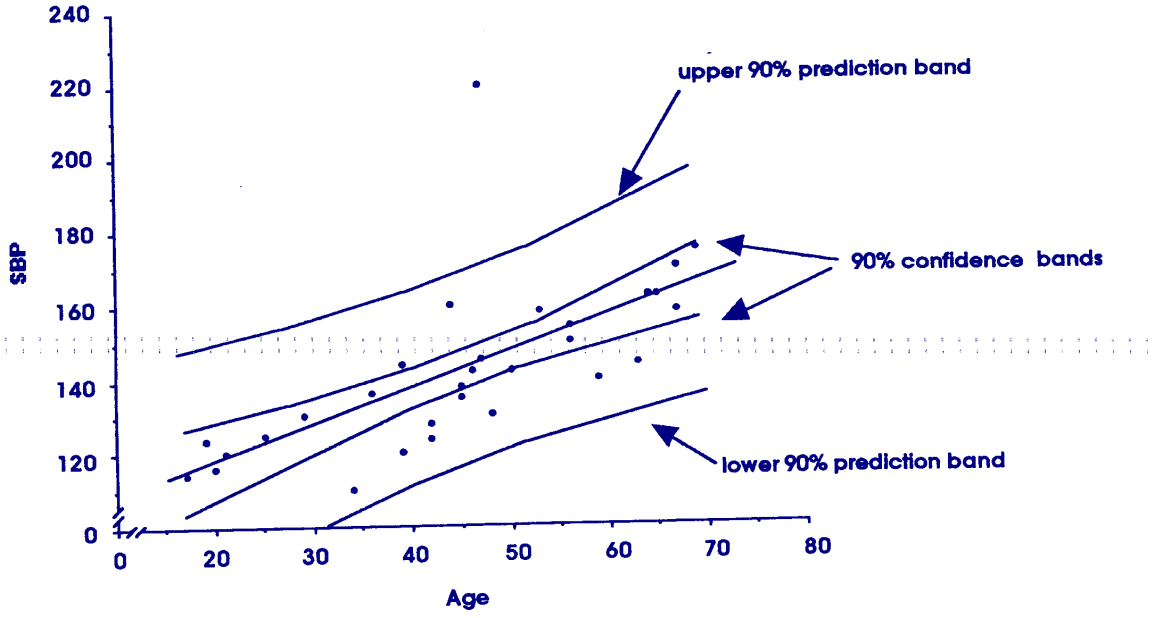
d) $$S^2_{\hat{\beta}_0} = S^2_{Y/X} \left[ \frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2} \right]$$

H0:B1=0 não é rejeitada




Examples when $H_0$: $\beta_1 = 0$ is rejected



# O COEFICIENTE DE CORRELAÇÃO (ρ ) E A ANÁLISE DE REGRESSÃO

Na análise de regressão linear, um estimador para o coeficiente de correlação é:

$$\hat{\rho} = r = \frac{S_X}{S_Y} \quad \hat{\beta}_1$$

$$\text{propriedade} \ : \ r \text{ tem o mesmo sinal de } \hat{\beta}_1$$

$$\therefore \begin{cases} \text{se } r > 0 \Rightarrow \quad \hat{\beta}_1 > 0 \\ \text{se } r = 0 \Rightarrow \quad \hat{\beta}_1 = 0 \\ \text{se } r < 0 \Rightarrow \quad \hat{\beta}_1 < 0 \end{cases}$$

Ou seja, o teste de hipótese do modelo de regressão (t ou F) é o mesmo do coeficiente de correlação.

$H_0: \rho = 0$

$H_1: \rho \neq 0$

Isto equivale a testar as hipóteses

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

## Lembram-se do $r^2$? Na verdade, $r^2 = (r)^2$.

$$R^2 = r^2 = \frac{SQM}{SQT}$$

como $-1 \le r \le +1 \Rightarrow 0 \le R^2 \le 1$

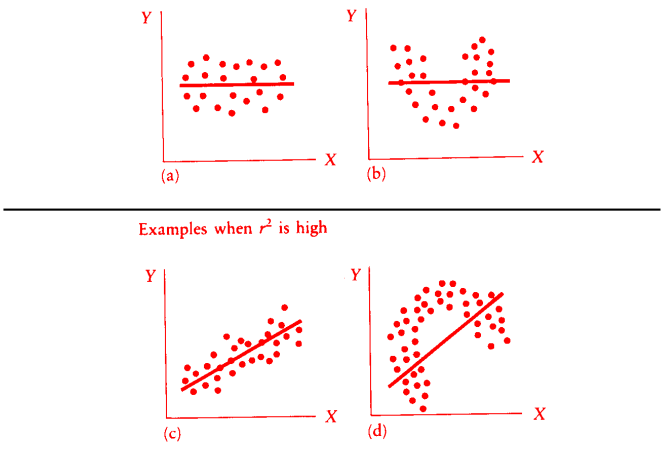
quando $R^2 = 1 \Rightarrow \hat{\beta}_1 \neq 0$ e SQR = 0 $\therefore$ o ajuste é perfeito!!!

por outro lado, quando $R^2 = 0 \Rightarrow \hat{\beta}_1 = 0$ e que SQT = SQR $\Rightarrow$ não há melhora na predição de Y, quando se utiliza X.

O que $r^2$ não mede:

1. a magnitude da inclinação de uma reta de regressão;
2. não é uma medida apropriada para avaliar a linearidade do modelo.

quando r2 é baixo



## 6. ANÁLISE DOS RESÍDUOS ($\varepsilon_i = e_i$):

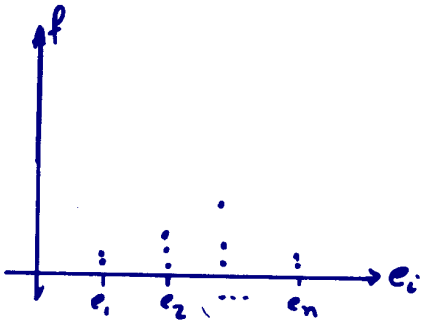
$$e_i = Y_i - \hat{Y}_i \quad , \quad i = 1,2, \ldots\ldots n$$

Suposições:

a) os $e_i$ são independentes, ié, $COV(e_i, e_k) = 0$, para $i \neq k$.

b) $e_i \sim N(0, S_e)$ , onde $S_e^2 = \text{constante}$

### 6.1. Análise Global:




o gráfico deve ter a aparência de uma curva normal

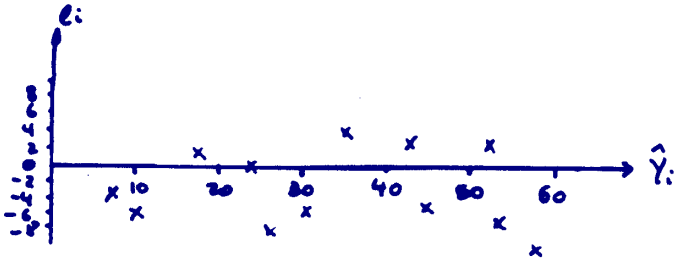
$$\text{se } e_i \sim N(0; S_e) \Rightarrow \frac{e_i - \bar{X}_e}{S_e} \sim N(0;1)$$

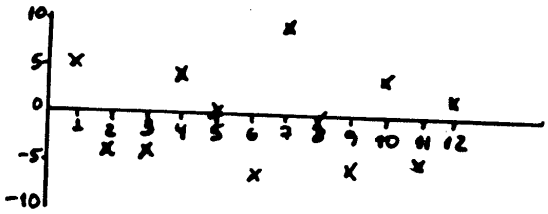
$$\text{onde } S_e^2 = \frac{\sum\left(e_i - \bar{X}_e\right)^2}{n - p} = \frac{\sum e_i^2}{n - p} \quad ; \quad p = \text{no. de variáveis indep.}$$

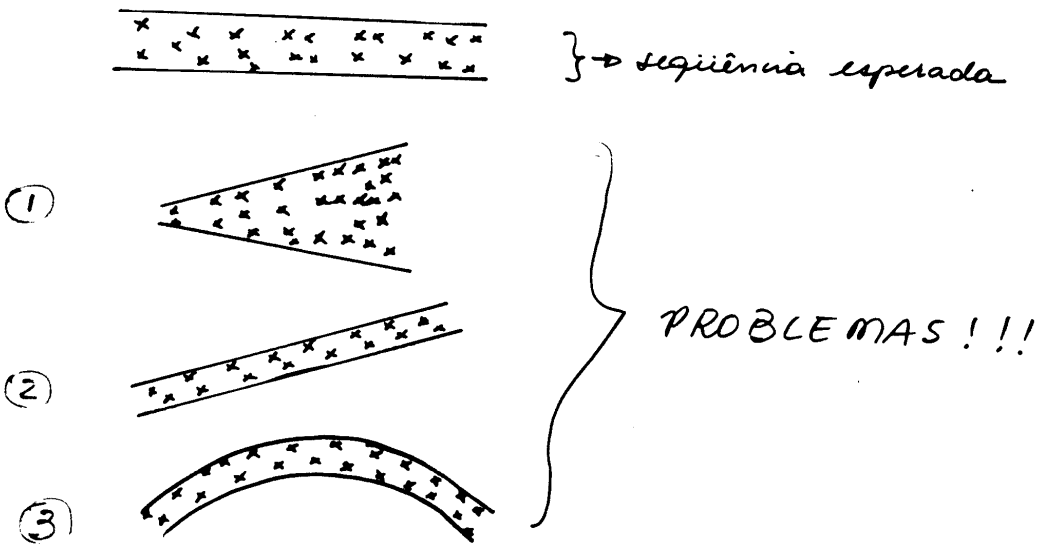
$$\therefore IC_{95\%}(e_i) = [-1.96; +1.96]$$

teste estatístico: aderência dos $e_i$ à curva Normal.

## 6.2. Gráfico $e_i \times \hat{Y}_i$



❶ a variância não é constante (conforme suposto): deve-se fazer uma transformação na variável dependente $Y_i$, antes da análise de regressão ou fazer a estimação por mínimos quadrados ponderados.

❷ erro na análise de regressão: o modelo está viciado.

❸ o modelo é inadequado. São necessários termos adicionais (ex:quadrático ou produtos cruzados) ou é necessário que se faça uma transformação na variável dependente Y antes da análise.

## 6.3. Gráfico $e_i \times X_i$ : idem ao 6.2.

## 7. VALORES ABERRANTES (*OUTLIERS*)

Um valor aberrante é um ponto peculiar do conjunto de dados e, por isso, deve ser examinado cuidadosamente para que se descubra a razão de sua particularidade.

Não é prudente descartá-lo sem antes se proceder à uma investigação. Ele pode ser descartado quando seu valor for devido à um erro de mensuração e/ou registro ou devido à outro fator externo ao estudo.

## Como iniciar a análise de um modelo de regressão linear múltipla?

Primeiramente, fazer os modelos de regressão linear simples, incluindo a análise de resíduos, selecionando:

1. Aquelas variáveis independentes com p<0,20 (ou qualquer outro ponto de corte; a sugestão é NUNCA colocar, como ponto de corte p<0,14).

2. A(s) variável (eis) independentes de interesse, que devem estar descritas nos objetivos.

3. Selecionar as possíveis variáveis independentes de controle (do ponto de vista estatístico ou do ponto de vista da epidemiologia).

4. Fazer a matriz de correlação para avaliar se existe colinearidade perfeita (r>0,95 entre as variáveis independentes) e a ordem de entrada do modelo.

# MATRIZ DE CORRELAÇÃO

É uma matriz (k+1) x (k+1), sendo k o número de variáveis independentes que serão testadas no modelo múltiplo. Nesta matriz aparecem os coeficientes de correlação (r) entre todas as variáveis de estudo, sendo que na primeira linha (ou primeira coluna) deverão estar os coeficientes de correlação entre a variável dependente e as variáveis independentes. Esta é uma matriz com a diagonal unitária.

| | Y | $X_1$ | $X_2$ | X3 | ... | $X_k$ | |
|---|---|---|---|---|---|---|---|
| Y | 1 | $r_{Y,X1}$ | $r_{Y,X2}$ | $r_{Y,X3}$ | | $r_{Y,Xk}$ | ordem de entrada das variáveis indepen |
| $X_1$ | | 1 | $r_{X1,X2}$ | $r_{X1,X3}$ | | $r_{X1,Xk}$ | |
| $X_2$ | | | 1 | $r_{X2,X3}$ | | $r_{X2,Xk}$ | colinearidade |
| ... | | | | | | | |
| ... | | | | | | | |
| $X_k$ | | | | | | 1 | |

# ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA
## (MULTIVARIADA ???? -> nunca usar esse termo)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ..... + \beta_k X_k + \varepsilon \quad ; \quad \text{k : numero de variaveis}$$

$$\mathrm{Y} = \mathrm{f}\left(\mathrm{X}_1, \mathrm{X}_2, ....., \mathrm{X}_k\right) \quad \text{, utilizando amostra de tamanho} \quad n$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ..... + \hat{\beta}_k X_k$$

Y : v.a. dependente

$X_j$ : v.a. independentes (regressores)

$\beta_j$ : coeficientes de regressão (a serem estimados)

(cada $\beta_j$ representa a mudança em $\bar{Y}_{Y/X_1,...,X_k}$ para uma unidade de cada $X_j$, quando todas as outras variáveis independentes permanecem constantes)

ex:

$$\mathrm{E}\left(Y / X_1 = X_2 = ..... = X_k = 0\right) = \beta_0$$

$$\mathrm{E}\left(Y / X_1 = 1, X_2 = ..... = X_k = 0\right) = \beta_0 + \beta_1$$

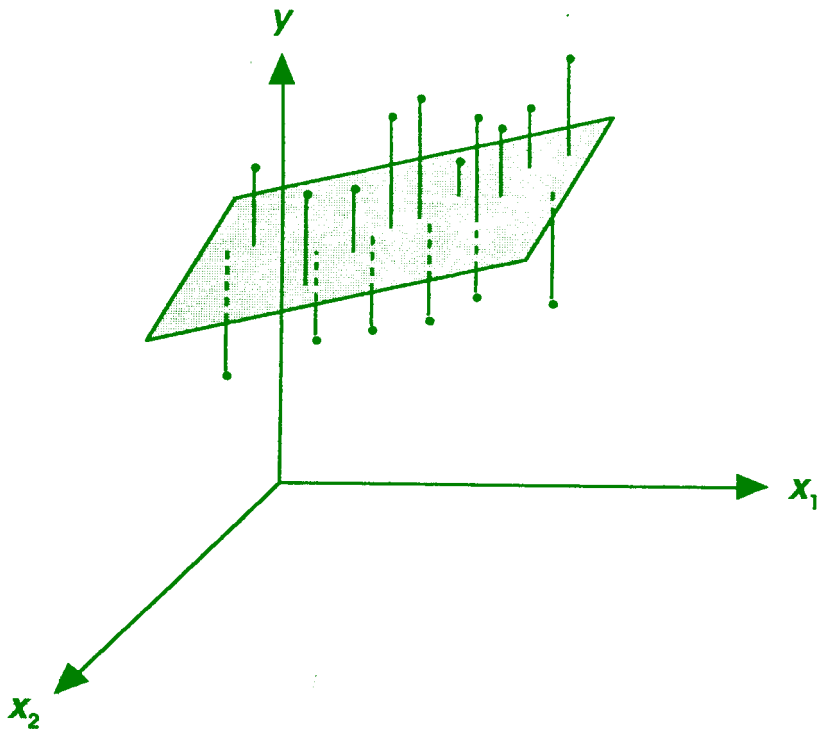
$$\mathrm{E}\left(Y / X_2 = 1, X_1 = X_3 = ..... = X_k = 0\right) = \beta_0 + \beta_2$$

$$\mathrm{E}\left(Y / X_1 = X_2 = 1, X_3 = X_4 = ..... = X_k = 0\right) = \beta_0 + \beta_1 + \beta_2$$

## ESTIMATIVA POR MÍNIMOS QUADRADOS:

$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \rightarrow$ achar os $\beta_j$ que minimizam esta expressão

ε : erro = resíduo (desvio do verdadeiro valor de Y em relação ao valor estimado pelo modelo, ié, $\left(Y_i - \hat{Y}_i\right)$



## SUPOSIÇÕES BÁSICAS

São as mesmas do modelo simples, porém com extensão para múltiplas variáveis.

1. Distribuição Normal

Para um conjunto de valores fixos das v.a. $X_j$ (que, idealmente, devem ser contínuas), Y é uma v.a. com distribuição normal, com média e variância finitas (aqui se trabalha em um espaço k-dimensional).

$$Y_i \sim N\left(\bar{Y}_{Y/X_1,X_2,\ldots,X_k};S\right)$$

2. Os valores de Y são independentes uns dos outros.

3. Linearidade

O valor médio de Y ($\bar{Y}_{Y/X_1,X_2,\ldots,X_k}$) é uma função de linear sobre os $X_j$.

4. Homocedasticidade

A variância de Y é a constante, qualquer que seja o conjunto dos $X_j$.

5. Não existe correlação entre os erros, ié, para quaisquer 2 amostras tem-se que : $COV(\varepsilon_i,\varepsilon_l)=0\,,\quad \forall\ i\neq l\,.$

6. Cada variável independente não está correlacionada com o termo de erro, ié, para cada $X_j$, $COV\left(X_j, \varepsilon_{i,j}\right) = 0$

7. Não há colinearidade perfeita entre as variáveis independentes, ié, nenhuma variável independente está relacionada linearmente, de maneira perfeita, com uma ou mais variáveis independentes.

## EQUAÇÃO GERAL DA REGRESSÃO

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + 0 \qquad ⑤$$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

**SQT** **SQR** **SQM**

SQTotal=SQ devida ao resíduo + SQ devida à regressão

# ANOVA (modelo geral)

| FONTE | SQ | GL | MQ | $F_{TOTAL}$ |
|---|---|---|---|---|
| regressão | $\sum\left(\hat{Y}_i - \bar{Y}\right)^2$ | k | $SQM/k$ | $F_o\left(k, n-k-1\right) =$ |
| resíduo | $\sum\left(Y_i - \hat{Y}_i\right)^2$ | n-k-1 | $SQR/n-k-1$ | $MQM/MQR$ |
| TOTAL | $\sum\left(Y_i - \bar{Y}\right)^2$ | n-1 | | |

$$r^2 = SQM/SQT \;\; ; F_c \sim F_{k,n-k-1}$$

# ANOVA (adição de variáveis)

| FONTE | | SQ | GL | MQ | $F_{parcial}$ |
|---|---|---|---|---|---|
| regressão | $X_1$ | * | 1 | $SQM_{X_1}/1$ | $F_o(1,n-1-1) = MQM_{X_1}/MQR$ |
| | $X_2/X_1$ | * | 1 | $SQM_{X_2}/1$ | $F_o(1,n-2-1) = MQM_{X_2}/MQR$ |
| | .... $X_k/X_1,X_2,...X_{k-1}$ | * | ..... 1 | ..... $SQM_{X_k}/1$ | ..... $F_o(1,n-k-1) = MQM_{X_k}/MQR$ |
| resíduo | | $\sum\left(Y_i - \hat{Y}_i\right)^2$ | n-k-1 | $SQR/n-k-1$ | |
| TOTAL | | $\sum\left(Y_i - \bar{Y}\right)^2$ | n-1 | | |

* fórmulas nas páginas seguintes.

# TESTES DE HIPÓTESES

## 1. Teste de significância do modelo geral

$$\begin{cases} \mathrm{H}_0: \beta_1 = \beta_2 = \ldots\ldots = \beta_k = 0 \\ \mathrm{H}_a: \text{existe pelo menos um dos } \beta_j \neq 0 \end{cases}$$

$$F_o = \frac{MQM}{MQR} \qquad \text{, onde } F_c \sim F_{k,n-k-1}$$

$$F_0 = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

## 2. teste do intercepto

$$\begin{cases} \mathrm{H}_0: \beta_0 = 0 \\ \mathrm{H}_a: \beta_0 \neq 0 \end{cases}$$

$$F_0 = \frac{\frac{SQR(\text{modelo sem } \beta_0) - SQR(\text{modelo com } \beta_0)}{1}}{\frac{SQR(\text{modelo com } \beta_0)}{\text{n - k - 1}}} \qquad , F_c \sim F_{1,n-k-1}$$

$$F_0 = \frac{\frac{n\bar{Y}^2}{1}}{\frac{\sum\left(Y_i - \bar{Y}\right)^2}{n-1}} \qquad , \qquad F_c \sim F_{1,n-1}$$

## 3. Teste do F parcial

$$\begin{cases} H_0 : \beta^* = 0 \text{ , no modelo } Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \beta^* X^* \\ H_a : \beta^* \neq 0 \\ H_a : X^* \text{ melhora significativamente a predicao de Y,} \\ \text{dado que } X_1, X_2, ..., X_p \text{ já estao no modelo} \end{cases}$$

$$SQM\left(X^* / X_1, X_2, \ldots, X_p\right) = SQM\left(X_1, X_2, \ldots, X_p, X^*\right) - SQM\left(X_1, X_2, \ldots, X_p\right)$$

$$\therefore \; F_{p_o}\left(X^* / X_1, X_2, \ldots, X_p\right) = \frac{SQM\left(X^* / X_1, X_2, \ldots, X_p\right) \Big/ 1}{MQR\left(X_1, X_2, \ldots, X_p, X^*\right)}$$

$$F_{p_c}\left(X^* / X_1, X_2, \ldots, X_p\right) \sim F_{1, n-(p+1)-1}$$

# 4. Teste múltiplo do F parcial

$$\begin{cases} H_0: \beta_1^* = \beta_2^* = \ldots = \beta_k^* = 0 \text{ no modelo} \\ \quad Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \underbrace{\beta_1^* X_1^* + \beta_2^* X_2^* + \ldots + \beta_k^* X_k^*}_{\text{bloco de variáveis}} \\ \\ H_a: \text{pelo menos um } \beta_j^* \neq 0 \\ H_a: \text{ o bloco inteiro dos } X_j^* \text{ melhora significativamente a} \\ \text{predicao de Y, dado que } X_1, X_2, \ldots, X_p \text{ já estao no modelo} \end{cases}$$

$$SQM\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) =$$
$$= SQM\left(X_1, X_2, \ldots, X_p, X_1^*, X_2^*, \ldots, X_k^*\right) - SQM\left(X_1, X_2, \ldots, X_p\right)$$

$$\therefore \; F_{mp_o}\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) = \frac{SQM\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) / k}{MQR\left(X_1, X_2, \ldots, X_p, X_1^*, X_2^*, \ldots, X_k^*\right)}$$

$$F_{mp_c}\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) \sim F_{k, n-(p+k)-1}$$

OBS:

1. como reconhecer variável de confusão?
2. como testar interação entre 2 variáveis independentes?

# CORRELAÇÃO MÚLTIPLA

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

DEF.

$$r_{Y/X_1,X_2,\ldots X_k} = r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2}}$$

$$r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n} Y_i \hat{Y}_i - n\overline{Y}^2}{\sqrt{\left(\sum_{i=1}^{n} Y_i - n\overline{Y}\right).\left(\sum_{i=1}^{n} \hat{Y}_i - n\overline{\hat{Y}}\right)}}$$

DEF: coeficiente de determinação múltipla ($r^2$)

$$r^2_{Y/X_1,X_2,\ldots X_k} = R^2_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{SQM}{SQT}$$

## Coef. de determinação múltipla ajustado ($r^2_{aj.}$)

$$r_{aj}^2 = r^2 - \frac{k}{n-k-1}\left(1-r^2\right) = \frac{(n-1)r^2 - k}{n-k-1}$$

$r_{aj}^2$ ➔ leva em conta a chance de contribuição de cada variável incluída, subtraindo-se o valor que seria esperado se nenhuma variável independente fosse associada à variável dependente.

# O COEFICIENTE DE CORRELAÇÃO PARCIAL

$r_{Y,Xi/Xj}$ ➔ é uma estimativa de $\rho_{Y,Xi/Xj}$

Vamos supor a situação em que tenho apenas duas variáveis independentes $X_1$ e $X_2$.

$$\rho^2_{Y,X_1/X_2} = \frac{\sigma^2_{Y/X_2} - \sigma^2_{Y/X_1,X_2}}{\sigma^2_{Y/X_2}}$$

Nesta situação particular, tem-se que o coeficiente de correlação parcial ao quadrado é:

$$r^2_{Y,X_1/X_2} = \frac{S^2_{Y/X_2} - S^2_{Y/X_1,X_2}}{S^2_{Y/X_2}}$$

$$r^2_{Y,X_1/X_2} = \frac{SQR(\text{do modelo so com } X_2) - SQR(\text{do modelo completo, ie, com } X_1 \text{ e } X_2)}{SQR(\text{mod elo so com } X_2)}$$

$$r^2_{Y,X_1/X_2} = \frac{\text{extra } SQ \text{ devido a adicao de } X_1, \text{ dado que } X_2 \text{ ja estava no modelo}}{SQR(\text{modelo so com } X_2)}$$

$$r_{Y,X_1/X_2} = \frac{r_{Y,X_1} - r_{Y,X_2}.r_{X_1,X_2}}{\sqrt{\left(1-r^2_{Y,X_2}\right).\left(1-r^2_{X_1,X_2}\right)}}$$

A estatística $F_{parcial}(X_p/X_1, X_2,...X_k)$ é a utilizada para testar se $r_{Y,Xp/X_1,X_2,...,X_k} = 0$.

# COLINEARIDADE

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

pode-se demonstrar que: $\beta_j = c_j \left[ \frac{1}{1 - r^2_{X_1 X_2}} \right]$ e

que $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ sao diretamente proporcionais a $1 / 1 - r^2_{X_1, X_2}$

*FIV*: fator inflacionario da variancia

$$FIV = \frac{1}{1 - R_j^2}$$

quando $FIV > 10 \Rightarrow$ ha colinearidade

$FIV > 10 \Rightarrow R_j^2 > 0.90 \Rightarrow r_j > 0.95$

Para se evitar a colinearidade pode-se "centralizar" a variável.

# VARIÁVEIS QUALITATIVAS (CATEGÓRICAS) EM MODELOS DE REGRESSÃO

Há dois métodos para se analisar variáveis categóricas em em modelos de regressão linear:

MÉTODO 1

Estimar uma equação de regressão para cada categoria da variável.

MÉTODO 2

Definir uma(algumas) variável(eis) *dummy* e incorporá-la(s) no modelo. Este método é menos poderoso.

## VARIÁVEIS INDICADORAS

Variáveis indicadoras (ou *dummy* ) são quaisquer variáveis que têm um número finito de valores que representam diferentes categorias de uma variável qualitativa.

As variáveis indicadoras são utilizadas em qualquer modelo de regressão.

Exemplo:
Y= PAS

X = idade ;

Z = sexo $\begin{cases} Z = 0 \Rightarrow \text{sexo} = \text{masculino} \\ Z = 1 \Rightarrow \text{sexo} = \text{feminino} \end{cases}$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ \qquad (1)$$

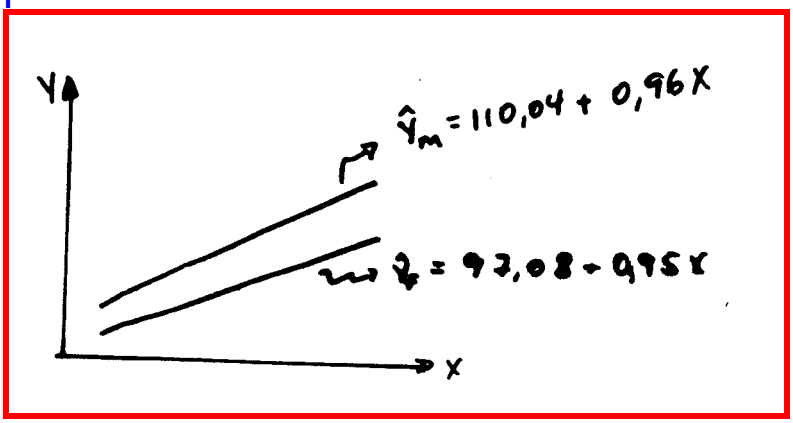
$$\text{qdo } Z = 0 \Rightarrow Y_M = \beta_0 + \beta_1 X \qquad (2)$$

$$\text{qdo } Z = 1 \Rightarrow Y_F = \beta_0 + \beta_1 X + \beta_2 + \beta_3 X \Leftrightarrow$$
$$Y_F = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X \qquad (3)$$

O modelo (1) incorpora as 2 equações de regressão separadas [(2) e (3)] em um único modelo.

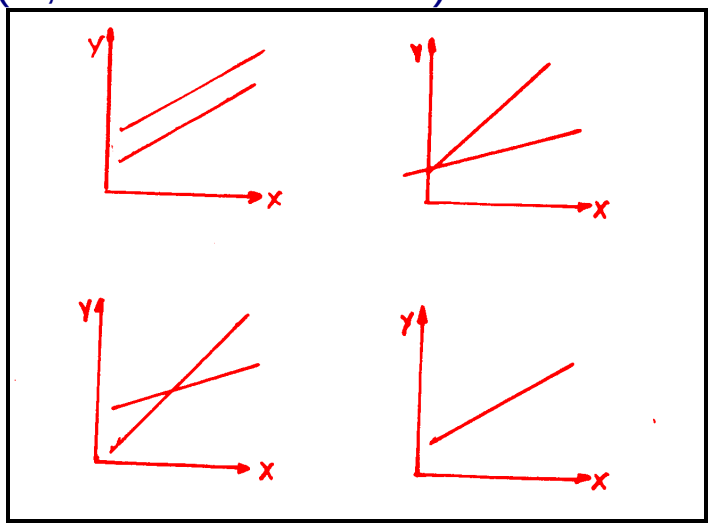
# COMPARAÇÃO DE 2 RETAS DE REGRESSÃO

Questão:será que a associação entre PAS e idade é a mesma para homens e mulheres?




Perguntas:

1. As inclinações das 2 retas são iguais?(ié, existe paralelismo?)
2. Os interceptos das 2 retas são iguais?(somente no caso das 2 retas não serem paralelas)
3. As 2 retas têm interceptos e inclinações iguais?(ié, são coincidentes?)



# "PASSOS" PARA SE FAZER MODELAGEM EM REGRESSÃO

1. Selecionar as variáveis independentes, não se esquecendo das possíveis variáveis de ajuste;

2. Codificar previamente as variáveis;

3. Fazer gráficos de dispersão (*scatter plot*) com todas as variáveis, 2 a 2;

4. Fazer a análise univariada das variáveis independentes, não se esquecendo de fazer a análise de resíduos.

5. Fazer a matriz de correlação para avaliar a colinearidade das variáveis independentes e definir a ordem de entrada das mesmas no modelo múltiplo.

6. Fazer a análise múltipla, avaliando a significância do modelo geral, de cada uma das variáveis e do incremento de cada uma delas, através do teste F e $F_{parcial}$ (ou teste t). Não se esquecer de avaliar os possíveis efeitos de confusão e a colinearidade entre as variáveis;

7.Decidir pelo melhor modelo, ié, o mais "ajustado". Fazer a estimação por ponto e por intervalo de cada um dos $\beta_j$;

8.Avaliar as interações se for o caso.

9. Fazer análise dos resíduos.

# ANÁLISE DE REGRESSÃO POLINOMIAL

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots\ldots + \beta_k X^k$$

## ANOVA (regressão polinomial)

| FONTE | | SQ | GL | MQ | $F_{parcial}$ |
|---|---|---|---|---|---|
| regressão | X | * | 1 | $SQM_X / 1$ | $F_o(1, n-1-1) = MQM_X / MQR$ |
| | $X^2/X$ | * | 1 | $SQM_{X^2} / 1$ | $F_o(1, n-2-1) = MQM_{X^2} / MQR$ |
| | .... $X^k/X, X^2, \ldots X^{k-1}$ | * | ..... 1 | ..... $SQM_{X^k} / 1$ | ..... $F_o(1, n-k-1) = MQM_{X^k} / MQR$ |
| resíduo | | $\sum \left( Y_i - \hat{Y}_i \right)^2$ | n-k-1 | $SQR / n-k-1$ | |
| TOTAL | | $\sum \left( Y_i - \bar{Y} \right)^2$ | n-1 | | |

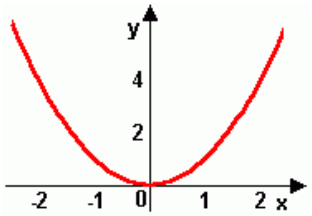
* fórmulas iguais às já citadas.

## MODELO DE REGRESSÃO LINEAR

$$Y = \beta_0 + \beta_1 X$$

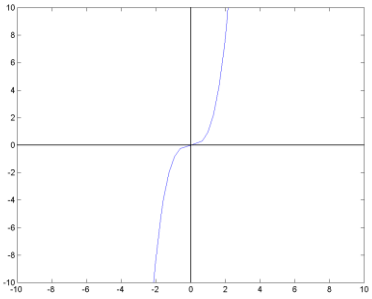
## MODELO DE REGRESSÃO DE 2a ORDEM

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$



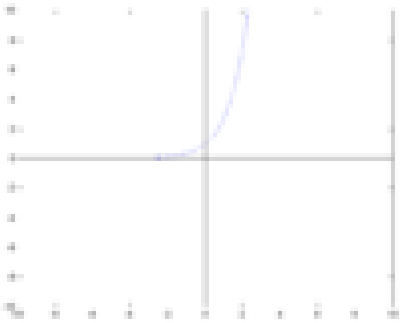

## MODELO DE REGRESSÃO DE 3a ORDEM

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$



## MODELO DE REGRESSÃO EXPONENCIAL

$$Y = \beta_0 * e^{(\beta_1 X)} \; ou \; \ln(Y) = \ln(\beta_0) + (\beta_1 X)$$



# MEDIDAS DE RISCO EM EPIDEMIOLOGIA

| | 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 |

Medidas de risco:

RP: razão de prevalências ➔ $RP = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$

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

OR: *odds ratio* ➔ $OR = \frac{a/b}{c/d} = \frac{a.d}{b.c}$

densidade de incidência, incidência acumulada.

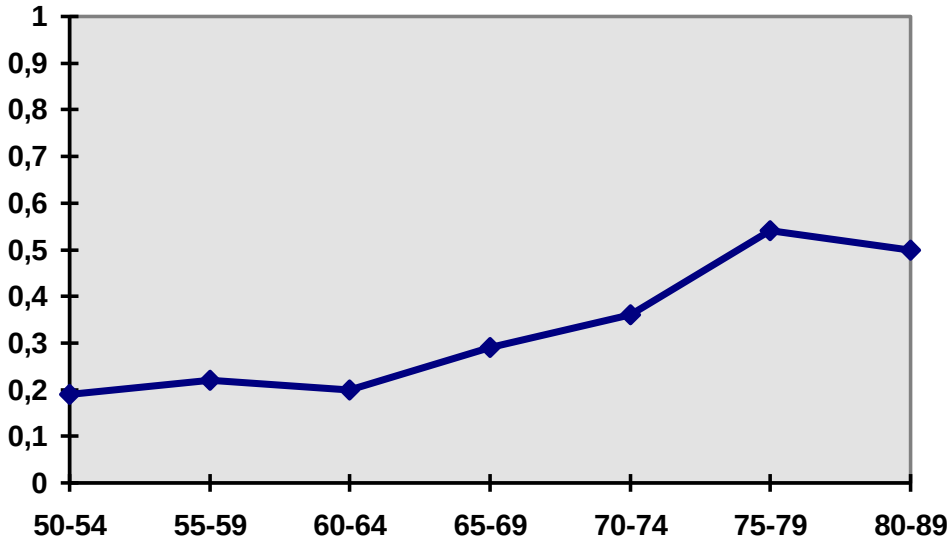
# MODELO DE REGRESSÃO LOGÍSTICA BINOMIAL

EXEMPLO

Y= doença coronariana(DC) $\begin{cases} Y = 1 \Rightarrow DC = sim \\ Y = 0 \Rightarrow DC = nao \end{cases}$

| IDADE | DC | | | |
|---|---|---|---|---|
| | SIM | NÃO | TOTAL | p=% de sim |
| 20 - 29 | 1 | 9 | 10 | 0.10 |
| 30 - 34 | 2 | 13 | 15 | 0.13 |
| 35 - 39 | 3 | 9 | 12 | 0.25 |
| 40 - 44 | 5 | 10 | 15 | 0.33 |
| 45 - 49 | 6 | 7 | 13 | 0.46 |
| 50 - 54 | 5 | 3 | 8 | 0.63 |
| 55 - 59 | 13 | 4 | 17 | 0.76 |
| 60 - 69 | 8 | 2 | 10 | 0.80 |
| Total | 43 | 57 | 100 | 0.43 |

Fonte: Kleimbaum,Klein, 2002.


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
50-54
55-59
60-64
65-69
70-74
75-79
80-89

## Y = variável dependente; variável categórica (0,1)

$$\begin{cases} Y = 1 \\ Y = 0 \end{cases} \Rightarrow Y \sim Bernoulli \Rightarrow \begin{cases} \mathrm{P}(\mathrm{Y} = 1) = \pi \\ \mathrm{P}(\mathrm{Y} = 0) = 1 - \pi \end{cases}$$

$$\mathrm{E}(\mathrm{Y}) = \sum_{i=1}^{2} y_i P(Y = y_i) = 1\, P(Y = 1) + 0\, P(Y = 0) = 1\,\pi + 0\,(1 - \pi) = \pi$$

O objetivo é escrever Y em funcao de X, porem , na regressao logistica, se escreve a probabilidade de Y como funcao de X e nao Y.

$$\pi(x) = E(Y / X = x)$$

$$\pi(x) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

Quando a f (x) e uma funcao linear, tem - se que

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Fazendo - se a transformacao para o logito de $\pi(x)$,

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 X$$

# SUPOSIÇÕES

1. Y é uma variável dicotômica (0,1). A extensão para outras variáveis categóricas não será vista neste curso.

2. Os valores de Y são independentes.

3.

$$\mathrm{E}(\mathrm{Y}) = \pi(x) \Rightarrow \hat{\mathrm{E}}(\mathrm{Y}) = \hat{\pi}(x) + \varepsilon$$

$$\varepsilon = \text{erro} = \text{resíduo}$$

$$\varepsilon \sim \textit{Binomial} \quad \text{, pois } \varepsilon = \begin{cases} 1 - \pi(x), \text{ se } \hat{\mathrm{E}}(\mathrm{Y}) = 1, \\ [\text{com prob. } \pi(x)] \\ \\ -\pi(x), \text{ se } \hat{\mathrm{E}}(\mathrm{Y}) = 0, \\ [\text{com prob. } 1 - \pi(x)] \end{cases}$$

.

$$\therefore \begin{cases} \bar{\varepsilon} = 0 \\ S_{\varepsilon}^{2} = \{\pi(x)[1 - \pi(x)]\} \rightarrow \text{variância não é constante} \end{cases}$$

4. A covariância entre dois erros quaisquer é zero.

# ESTIMATIVA DOS PARÂMETROS $\beta_i$

Na regressão logística é utilizado o Método da Máxima Verossimilhança para se estimar os parâmetros $\beta_i$ .

De uma maneira genérica, pode-se dizer que o método da máxima verossimilhança fornece os valores para os parâmetros a serem estimados, os quais maximizam a probabilidade de se obter o conjunto de dados existente.

Para se aplicar este método, em primeiro lugar precisa-se definir a função de verossimilhança. Na situação em que a variável dependente é dicotômica, tem-se:

Seja $Y = \begin{cases} 0 \\ 1 \end{cases} \Rightarrow$

$$\begin{cases} 1-\pi(x) = \dfrac{e^{\beta_0+\beta_1 X}}{1+e^{-(\beta_0+\beta_1 X)}} = \mathrm{P}(\mathrm{Y}=0/\mathrm{X}) \\ \pi(x) = \dfrac{1}{1+e^{-(\beta_0+\beta_1 X)}} = \mathrm{P}(\mathrm{Y}=1/\mathrm{X}) \end{cases}$$

para um arbitrario valor de $\underset{\sim}{\beta} = (\beta_0, \beta_1) \quad \Rightarrow$

A funcao de probabilidades de Y é

$$f(\mathrm{Y_i}) = \pi_i^{Y_i}\,(1-\pi_i)^{1-Y_i}, \quad \text{onde} \begin{cases} \mathrm{Y} = 0,1 \\ \mathrm{i} = 1,2,\ldots \mathrm{n} \end{cases}$$

Assim, para aqueles pares $(\mathrm{x_i}, 1)$, a contribuicao para a funcao de verossimilhanca é $\pi(x)$ e naqueles onde $\mathrm{Y_i} = 0$, a contribuicao e $1 - \pi(x)$.

A funcao de verossimilhanca e definida pelo produto dos termos dados acima, ié ,

$$L\left(\underset{\sim}{\beta}\right) = \prod_{i=1}^{n} f(Y_i)$$

No entanto, e mais facil maximizar o $\ln\left[L\left(\underset{\sim}{\beta}\right)\right]$.

$$\ln\left[L\left(\underset{\sim}{\beta}\right)\right] = \sum_{i=1}^{n}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]$$

Para encontrar os valores dos βi que maximizam a função acima deve-se derivar $\left[\ln L\left(\underset{\sim}{\beta}\right)\right]$ em relação a cada um dos βi e igualar a zero. Como estas equações não são lineares, são necessários métodos interativos e sua solução não é fácil! Porém os *softwares* fazem isso por nós !!!!

As equações são:

$$\begin{cases} \sum_{i=1}^{n}\left[y_i - \pi\left(x_i\right)\right] = 0 \\ \text{e} \\ \sum_{i=1}^{n} x_i \left[y_i - \pi\left(x_i\right)\right] = 0 \end{cases}$$

Estas são as chamadas equações de verossimilhança.

Normalmente as saídas de computador fornecem não só os valores dos $\beta_i$, mas, também, os respectivos erros padrão ($SE_{\beta i}$). Os valores dos $SE_{\beta i}$ serão utilizados para os testes de significância dos coeficientes e para o cálculos dos respectivos intervalos de confiança.

No caso do pior modelo (modelo só com $\beta_0$), o logaritmo da função de verossimilhança pode ser calculado por:

$$\ln\left[\mathrm{L}\left(\beta_{\mathrm{o}}\right)\right] = n_1.\ln\left(n_1\right) + n_0\,\ln\left(n_0\right) - n\ln(n)$$

onde: $n_1$: número de casos de Y=1

$n_0$: número de casos de Y=0

$n=n_1+n_0$= total da amostra

# TESTES DE HIPÓTESES

Na regressão logística a comparação entre o valor observado e o valor predito pela regressão não é feita através da ANOVA, mas é baseada no logarítmo da função de verossimilhança já definida $\left[\ln L\left(\underset{\sim}{\beta}\right)\right]$.

## 1. Teste da razão de verossimilhança

É feita a comparação entre a função de verossimilhança dos valores observados na amostra e a função de verossimilhança do modelo saturado. O modelo saturado é aquele que contém tantos parâmetros quanto o número de pontos da amostra (ex: ajustar uma linha reta com 2 pontos).

$$D = deviance$$

$$D = -2\left\{\ln\left[L(\text{modelo reduzido})\right] - \ln\left[L(\text{modelo saturado})\right]\right\} \Rightarrow$$

$$D = -2\ln\left[\underbrace{\frac{L(\text{modelo reduzido})}{L(\text{modelo saturado})}}_{\text{razão de verossimilhança}}\right]$$

Para verificar a significância de uma variável independente, compara-se o valor de D com e sem a variável independente na equação. A mudança de D devido à inclusão da variável independente é:

$$G = D(\text{para o modelo sem a variavel}) - D(\text{para o modelo com a variavel})$$

$$G = \left\{ -2\ln\left[\frac{L(\text{mod.sem variavel})}{L(\text{mod elo saturado})}\right] - 2\ln\left[\frac{L(\text{mod. com variavel})}{L(\text{mod elo saturado})}\right]\right\}$$

$$G = -2\ln\left[\frac{L(\text{modelo sem variável})}{L(\text{mod elo com variável})}\right]$$

$$G \sim \chi^2_1 \rightarrow \text{para o teste de significância de 1 variável com 2 categorias}$$

## 1.1. no caso do modelo univariado:

$$H_0 : \beta_1 = 0$$

## 1.2. no caso múltiplo

Utilizar o teste da razão de verossimilhança para verificar a adequação do modelo como um todo, ié:

$$\begin{cases} \text{H}_o : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ \text{H}_a : \text{o modelo é adequado, ie existe pelo um } \beta \neq 0 \end{cases}$$

$$G \sim \chi^2_k \text{ , onde k : numero de } \beta\text{'s do modelo}$$

## 2. Teste de Wald

Para testar a significância de cada coeficiente, utilizar o teste Wald:

$$\begin{cases} \mathrm{H}_0 : \hat{\beta}_i = 0 \\ \mathrm{H}_a : \hat{\beta}_i \neq 0 \end{cases}$$

$$W_i = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}, \text{ onde } W_{ic} \sim N(0,1)$$

# ESTIMATIVAS DAS MEDIDAS DE RISCO

## 1.Estimativa da *odds ratio* (OR) a partir do modelo de regressão logística

chance: $$\frac{\Pr ob(Y=1)}{\Pr ob(Y=0)} = \frac{p}{1-p}$$

$$OR(X_1) = \frac{\dfrac{p_{X_1=1}}{1-p_{X_1=1}}}{\dfrac{p_{X_1=0}}{1-p_{X_1=0}}} = \frac{e^{(\beta_0+\beta_1(X_1=1)+\beta_2X_2+.....+\beta_kX_k)}}{e^{(\beta_0+\beta_1(X_1=0)+\beta_2X_2+.....+\beta_kX_k)}} =$$

$$e^{(\beta_0+\beta_1+\beta_2X_2+.....+\beta_kX_k)-(\beta_0+\beta_2X_2+.....+\beta_kX_k)} = e^{\beta_1}$$

## 2. Intervalo de Confiança

$$IC_{(1-\alpha)\%}(\beta_i) = \hat{\beta}_i \pm z_{1-\alpha} \; x \; SE_{\hat{\beta}_i}$$

*no caso da OR,* "exponenciar" *o* $IC(\beta)$

# 3. Cálculo do RR

Vamos supor o caso mais simples em que a variável dependente X é dicotômica. Então,

$$RR = \frac{\Pr ob(Y = 1/\ X = 1)}{\Pr ob(Y = 1/\ X = 0)} = \frac{\dfrac{1}{1+\exp^{-(\beta\ +\beta_1 x1)}}}{\dfrac{1}{1+\exp^{-(\beta_0+\beta_1 x0)}}} = \frac{1+\exp^{-(\beta_0)}}{1+\exp^{-(\beta_0+\beta_1)}}$$

logo,

$$H_0 : \beta_1 = 0 \Leftrightarrow H_0 : OR(X_1) = 1 \Leftrightarrow H_0 : RR(X_i) = 1$$

**Logo, a hipótese avaliada no teste de Wald é:**

$$\begin{cases} \mathrm{H}_0 : \hat{\beta}_i = 0 \ \Leftrightarrow \ H_0 : OR(X_i) = 1 \Leftrightarrow H_0 : RR(X_i) = 1 \\ \mathrm{H}_a : \hat{\beta}_i \neq 0 \ \Leftrightarrow \ H_0 : OR(X_i) \neq 1 \Leftrightarrow H_0 : RR(X_i) \neq 1 \end{cases}$$

$$W_i = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}, \text{ onde } W_{ic} \sim N(0,1)$$

# Análise dos efeitos de confusão ou interação na regressão logística

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

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

Outra maneira de testar interação: criar uma 3a. variável (Z), que é a combinação de $X_1$ e $X_2$.

| $X_1$ | $X_2$ | Z | $Z_1$ | $Z_2$ | $Z_3$ |
|---|---|---|---|---|---|
| 1 | 1 | 3 | 0 | 0 | 1 |
| 1 | 0 | 2 | 0 | 1 | 0 |
| 0 | 1 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

# MODELOS DE REGRESSÃO LOGÍSTICA

- Não condicional: estudos transversais, coorte e caso-controle não pareado

- Condicional: estudos caso-controle e outros onde haja pareamento. Nestes casos, no banco de dados deverá existir a variável "**par**".

---

## ANÁLISE DOS RESÍDUOS

### 1. Estatística do $\chi^2$ de Pearson

### 2. Teste de Hosmer-Lemeshow

---------- Hosmer and Lemeshow Goodness-of-Fit Test----------

| | LOW = 0 | | LOW = 1 | | |
|---|---|---|---|---|---|
| Group | Observed | Expected | Observed | Expected | Total |
| 1 | 35.000 | 34.180 | 3.000 | 3.820 | 38.000 |
| 2 | 25.000 | 26.537 | 9.000 | 7.463 | 34.000 |
| 3 | 29.000 | 29.743 | 10.000 | 9.257 | 39.000 |
| 4 | 16.000 | 14.736 | 6.000 | 7.264 | 22.000 |
| 5 | 10.000 | 9.460 | 7.000 | 7.540 | 17.000 |
| 6 | 8.000 | 9.877 | 12.000 | 10.123 | 20.000 |
| 7 | 7.000 | 5.466 | 12.000 | 13.534 | 19.000 |

| | Chi-Square | df | Significance |
|---|---|---|---|
| Goodness-of-fit test | 2.3862 | 5 | .7935 |

---