INTRODUÇÃO À MODELOS DE REGRESSÃO LINEAR

MAT236 – MÉTODOS ESTATÍSTICOS

Introdução

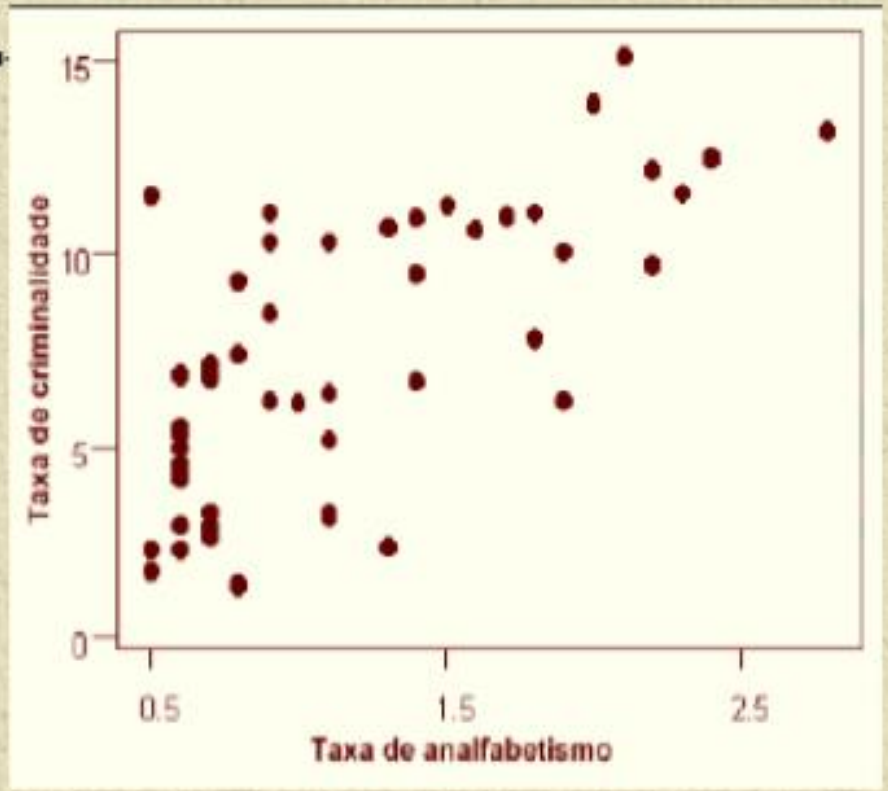
- No estudo de análise de regressão, deseja-se avaliar o efeito de uma variável independente sobre uma variável dependente.
- Variável dependente ou resposta (Y).
- Variáveis independente, explanatórias ou preditoras (X): variáveis cujo efeito sobre Y se quer conhecer.

Análise de Regressão

- Ferramenta estatística para avaliação das relações entre uma ou mais variáveis independentes ($X_1, X_2, \ldots, X_k$) e uma variável dependente (Y).
- Utilizada para predição de Y a partir de $X_1, X_2, \ldots, X_k$.
- Exemplos:
  - A população de bactérias pode ser predita pelo tempo de cultivo.
  - Peso de recém-nascido e hábito de fumar da mãe.

Exemplo: Relação entre Criminalidade e Analfabetismo

- QC: "Maiores taxas de analfabetismo aumentam os índices de criminalidade?"
- Hipótese: "Regiões com maiores taxas de analfabetismo enfrentam maiores taxas de criminalidade"
- População alvo: População residente em todos os estados americanos



O que a análise desse gráfico sugere?

Modelos de Regressão Linear

- A relação linear entre a variável dependente e uma variável independente é definida pelo modelo linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,2,\ldots n$$

onde ε_i é o termo de erro.

- Pressupõe-se para este modelo que:

(i) A relação entre Y e X é linear (Linearidade)

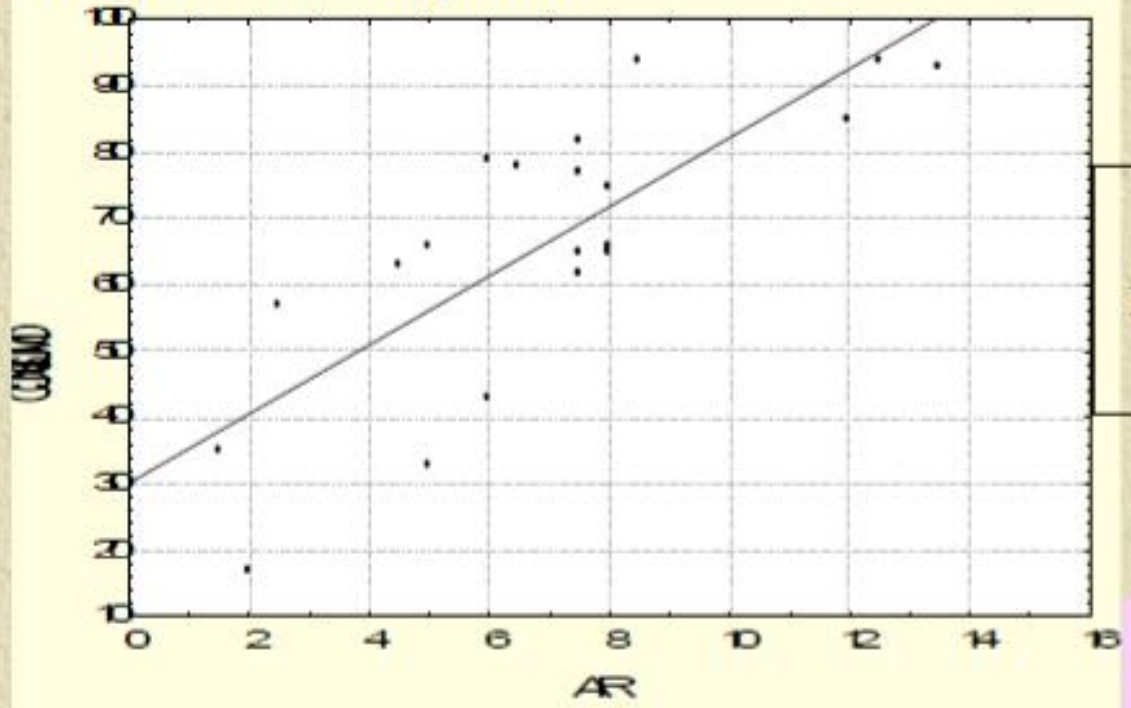
(ii) Os valores de X são fixos (ou controlados)

(iii) As observações são independentes.

(iv) Os erros têm distribuição normal, com média zero e variância constante → $\varepsilon_i \sim N(0, \sigma^2)$

Modelo de Regressão Linear Simples

Os pares de valores (X_1, Y_1), (X_2, Y_2), …. ,(X_n, Y_n) serão utilizados para avaliação da relação existente entre X e Y: construção de diagrama de dispersão.

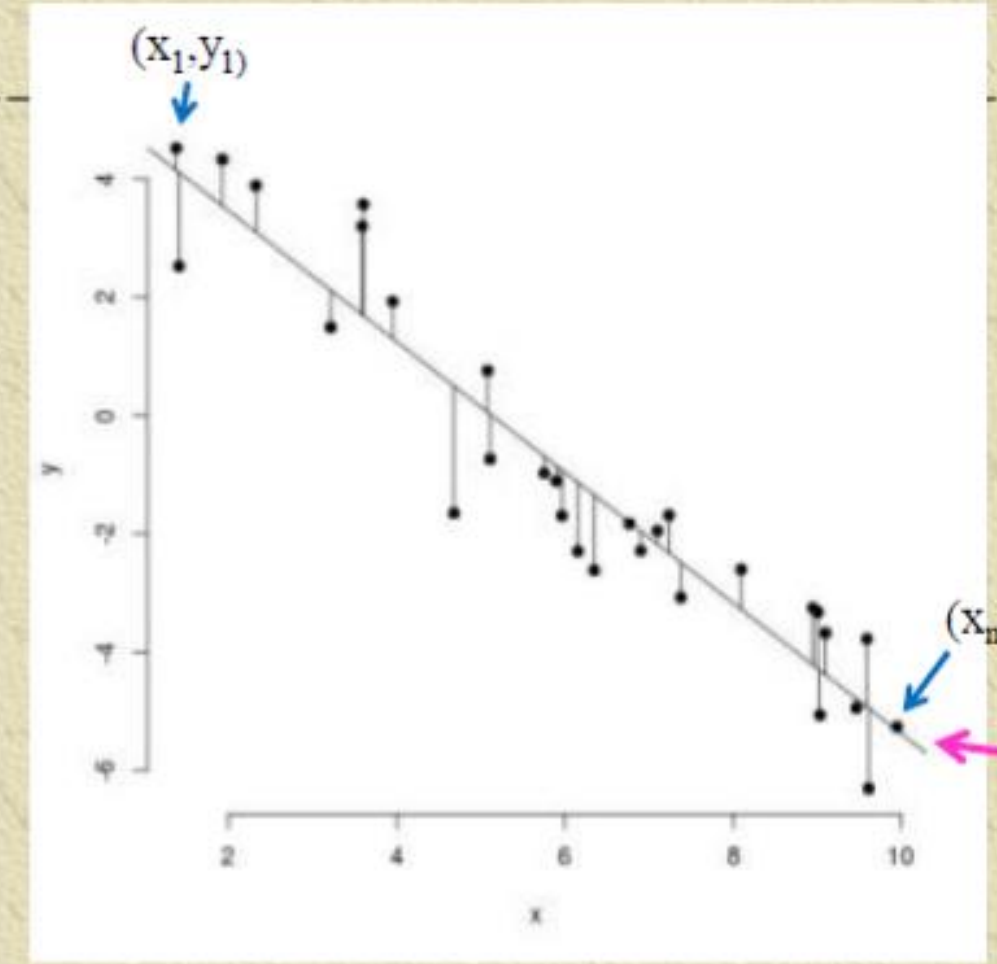


Uso de modelo de regressão para mensurar efeito de X em Y.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Resposta = Componente sistemático + parte aleatória

Qual a idéia básica do método?



Método dos Mínimos Quadrados

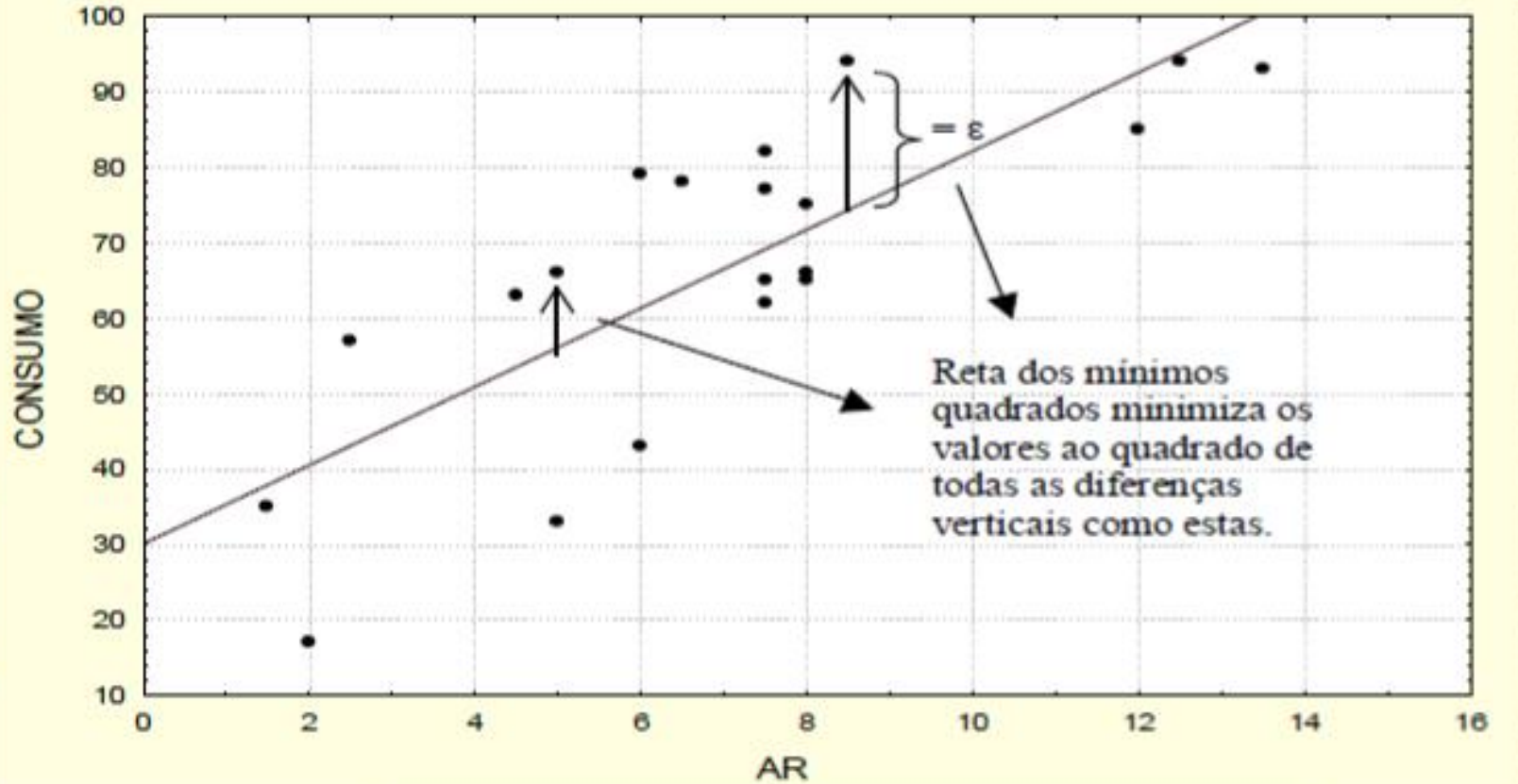


Figura 5 – Processo de Mínimos Quadrados

Método dos Mínimos Quadrados: ajuste a uma reta

- Objetivo é ajustar pontos a uma curva definida por f(x)=y=a+bx, onde a e b são os parâmetros a serem determinados.

Meta: minimizar a distância entre os pontos e a reta

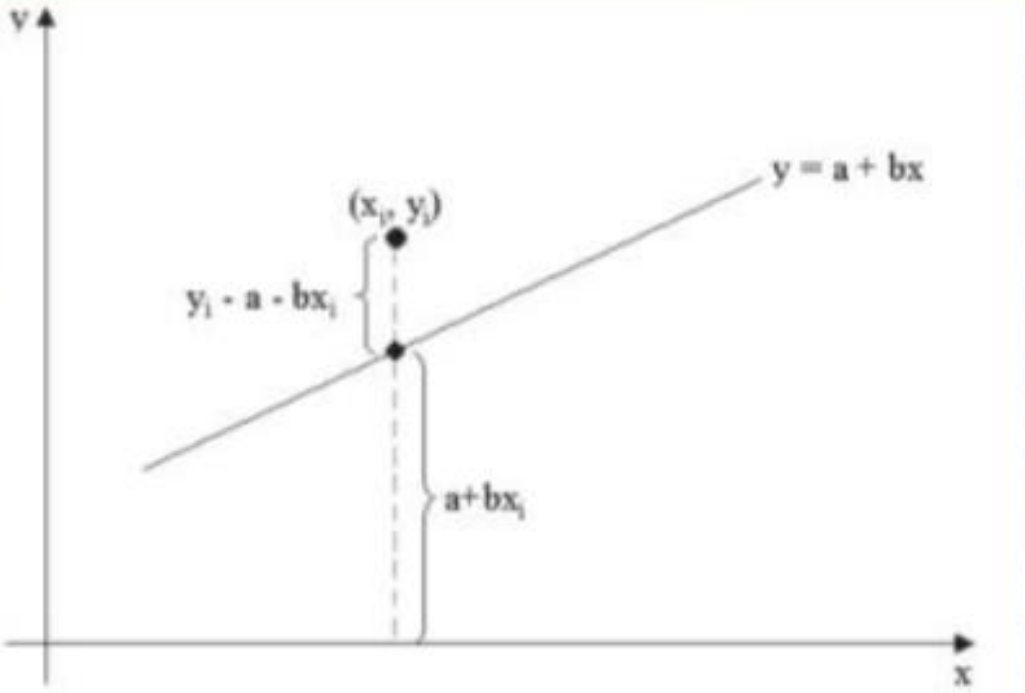
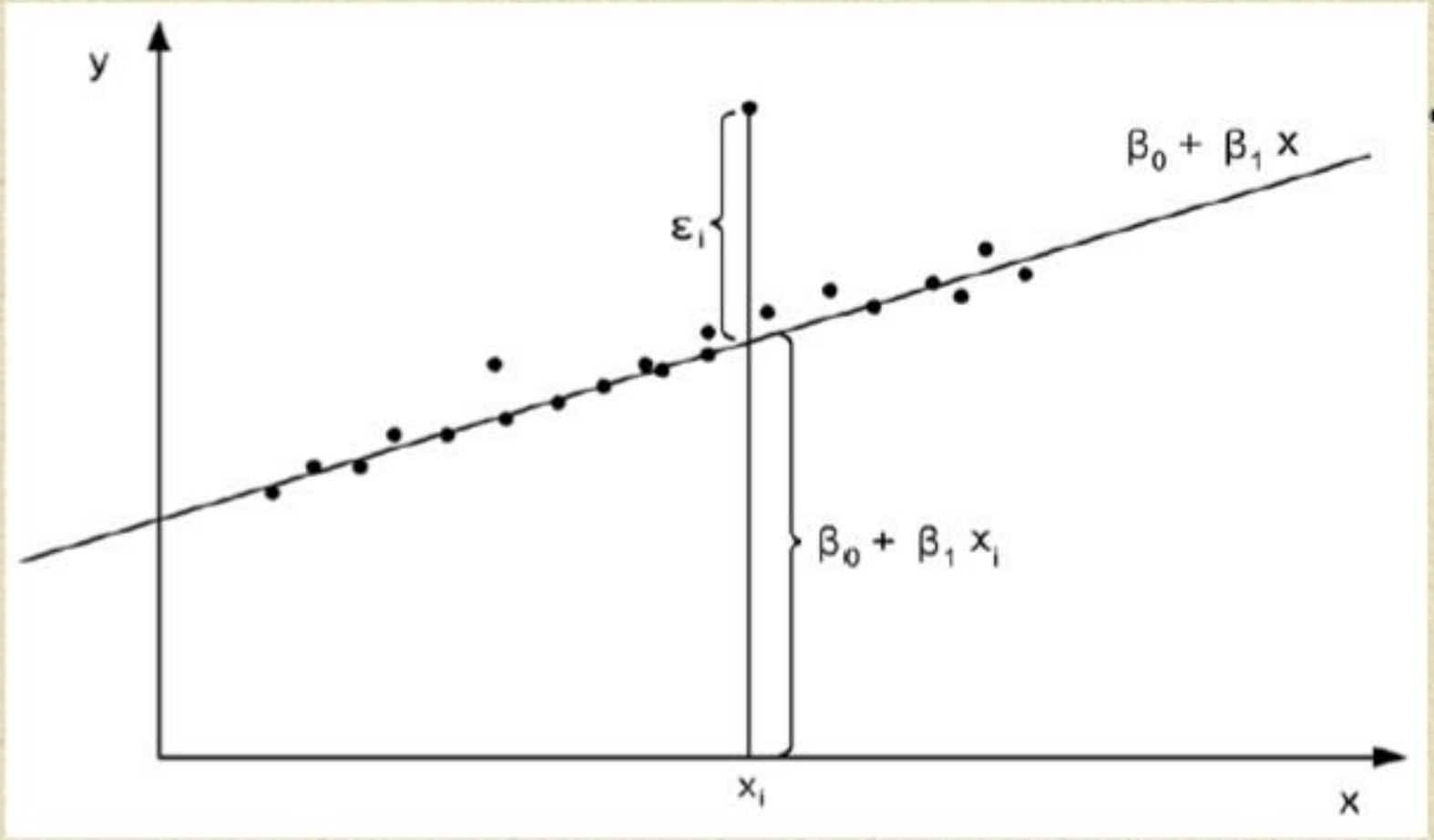


Figura 1: Distância de um ponto (x_i, y_i) à reta $y = a + bx$

Mesma lógica usando a notação de modelos de regressão:



$$D = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} [y_k - (\beta_o + \beta_1 x_k)]^2$$

Estimação no Modelo Linear

- Usando o método dos mínimos quadrados, pode-se estimar os parâmetros do modelo através das seguintes equações:

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Dos dados amostrais para o modelo populacional: Como ajustar o modelo hipotético para os dados amostrais?

Modelo populational

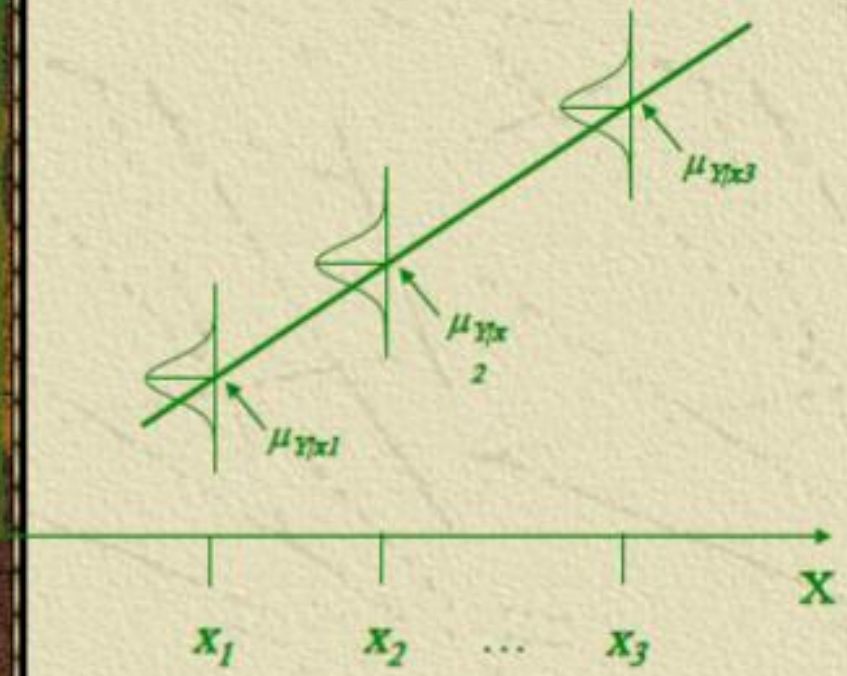
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

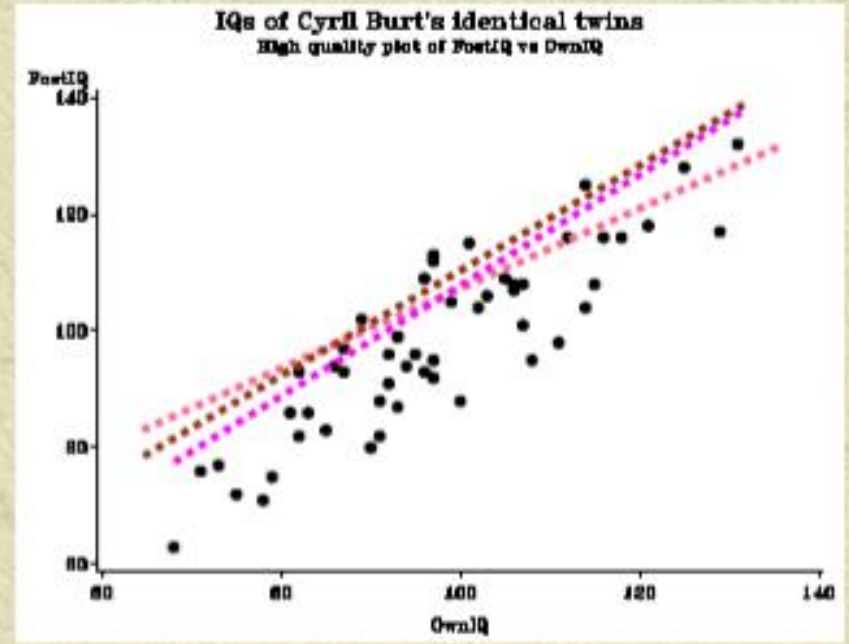
Modelo ajustado

"Chapéus" denotam estimativas dos parâmetros

Sem termo de erro

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$





Exemplo: Relação entre Comprimento e Peso de Ursos

Comprimento(x)	Peso(y)	x.y	x^2	y^2
53,0	80	4240	2809,00	6400
67,5	344	23220	4556,25	118336
72,0	416	29952	5184,00	173056
72,0	348	25056	5184,00	121104
73,5	262	19257	5402,25	68644
68.5	360	24660	4692,25	129600
73,0	332	24236	5329,00	110224
37,0	34	1258	1369,00	1156
516,5	2176	151879	34525,75	728520

Totais

Calculando correlação linear

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{8(151879) - (516{,}5)(2176)}{\sqrt{8(34525{,}75) - (516{,}5)^2}\sqrt{8(728520) - (2176)^2}}$$

$$r = 0{,}897$$

Calculando estimativas

$$\hat{\beta}_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$\hat{\beta}_1 = \frac{8(151879) - (516{,}5)(2176)}{8(34525{,}75) - (516{,}5)^2} = 9{,}66$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_o = \frac{2176}{8} - 9{,}66\frac{516{,}5}{8} = -352$$

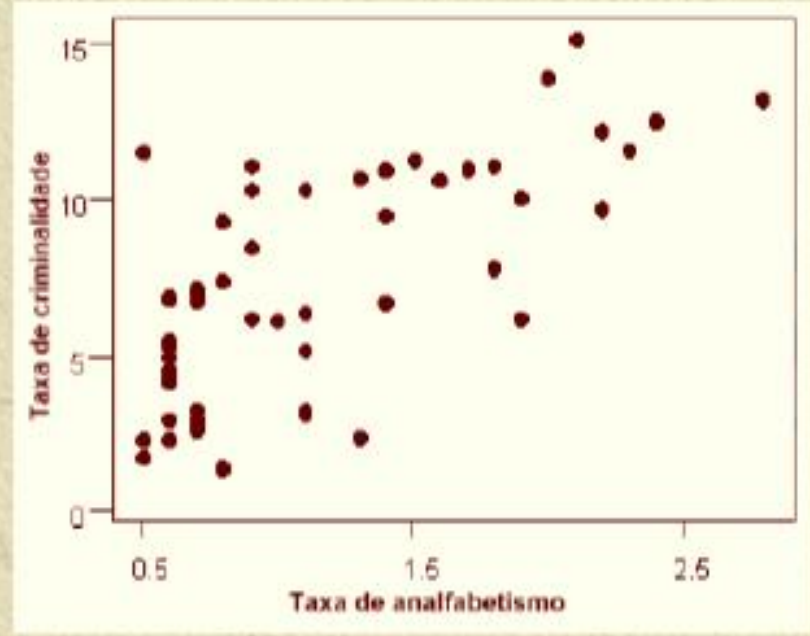
Exemplo: Relação entre Criminalidade e Analfabetismo

Considere as duas variáveis observadas nos 50 estados norte-americanos:

Y : taxa de criminalidade

X : taxa de analfabetismo

r=0.702



Exemplo: Relação entre Criminalidade e Analfabetismo

Com o uso de modelo de regressão linear simples para esse exemplo, obtém-se o seguinte resultado:

$$\hat{Y} = 2{,}397 + 4{,}257\,X$$

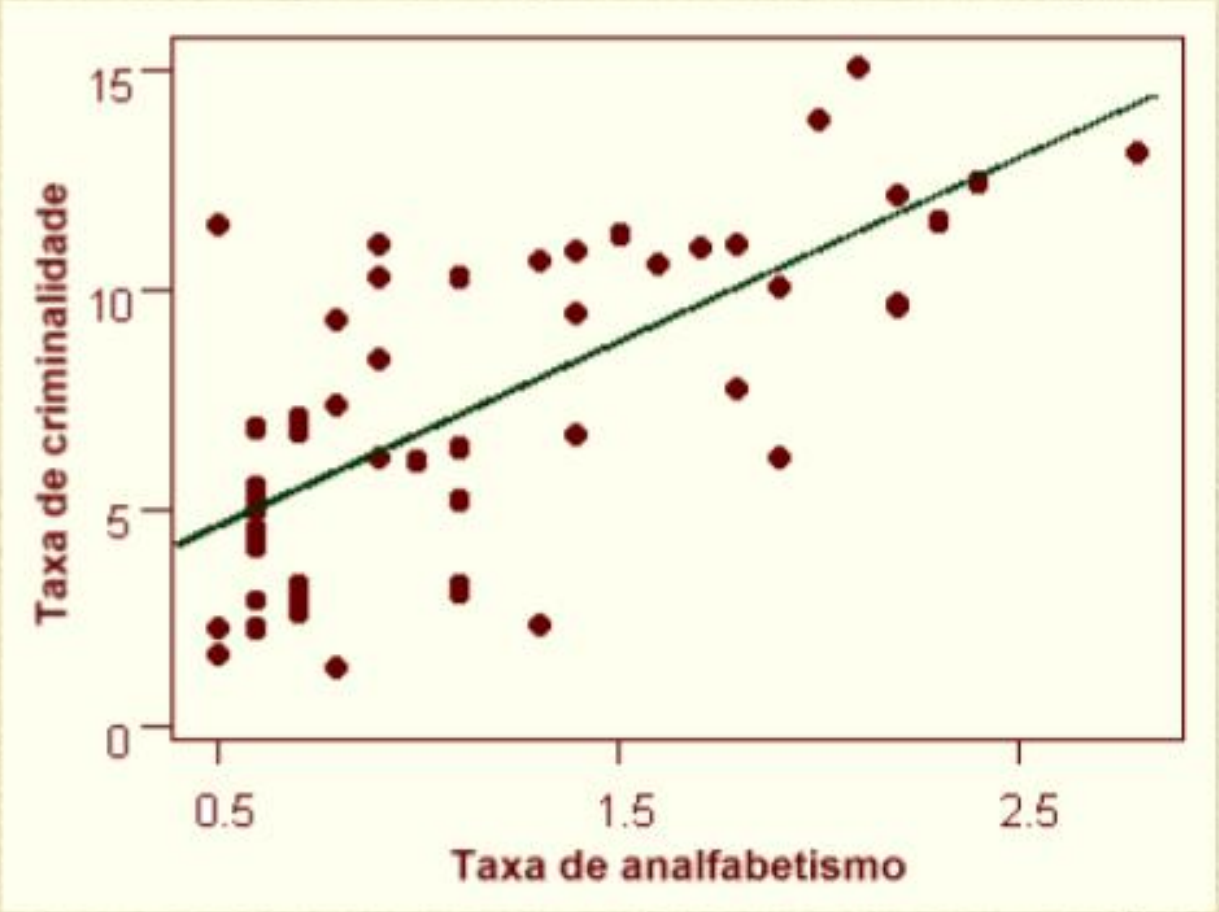
Interpretação:

- Para o aumento de uma unidade na taxa de analfabetismo, a taxa de criminalidade aumenta, em média, 4,257 unidades.

- Em estados com taxa de analfabetismo igual a zero, prevê-se que a taxa de criminalidade seja igual a 2,397.

Exemplo: Relação entre Criminalidade e Analfabetismo

Reta ajustada pelo modelo de regressão

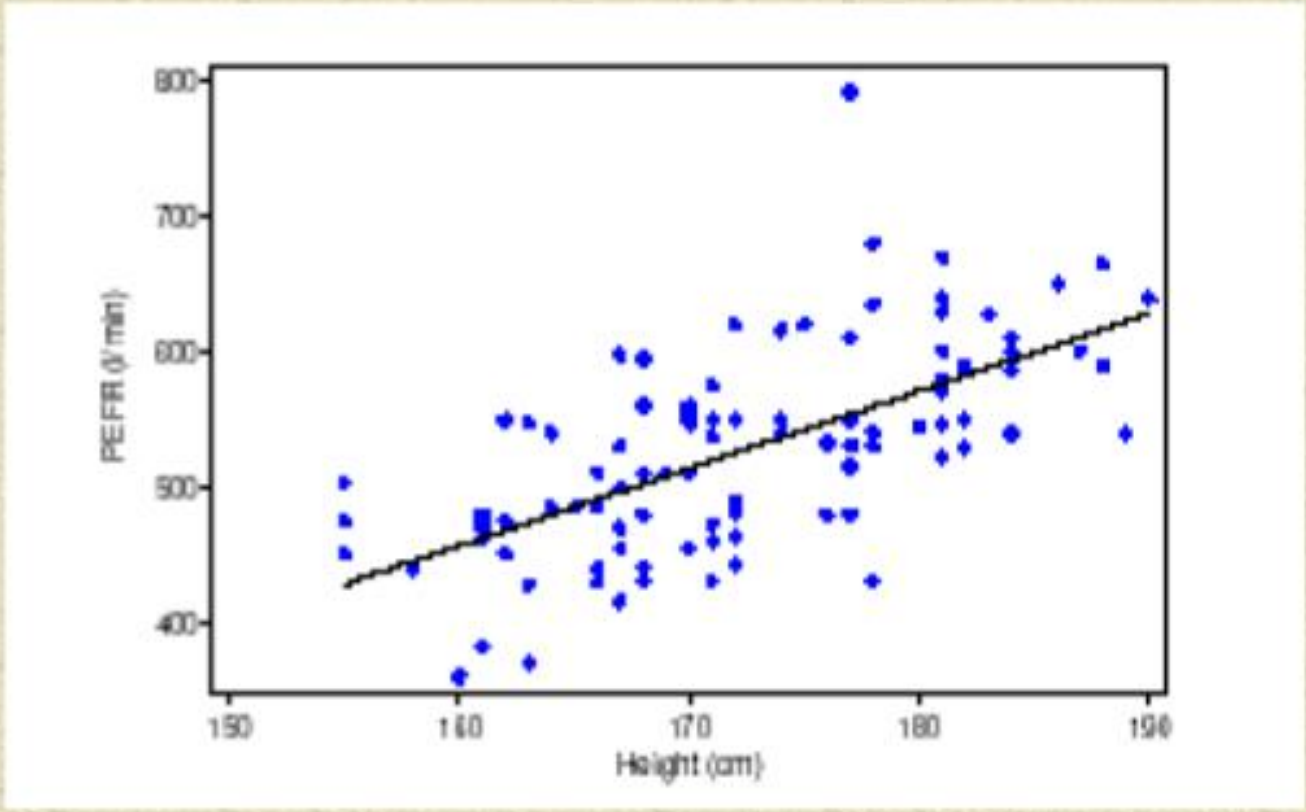


Predição?

Procedimento padrão

- **Construção do diagrama de dispersão** para visualizar a relação entre duas variáveis quantitativas;
- **Calculo do coeficiente de correlação** para quantificar a força da relação linear entre duas variáveis quantitativas;
- **Definição de modelo de regressão linear**: formulação matemática da relação linear entre as variáveis;
- **Ajuste do modelo**: através do método dos mínimos quadrados estima-se os parâmetros do modelo com os dados da amostra;
- **Interpretação dos parâmetros**: avaliar influência da variável independente sobre a dependente.

Exemplo: Relação entre Capacidade Pulmonar (Espirometria) e Altura (cm)



$r = 0.64$

$$PEFR_i = \beta_0 + \beta_1 \cdot height_i + E_i \quad E_i \sim N\left(0, \sigma^2\right)$$

Resultados da análise

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-456,921	117,957		-3,874	,000	-690,972	-222,9
	Height (cm)	5,712	,683	,643	8,361	,000	4,356	7,067

a. Dependent Variable: PEFR (l/min)

$\hat{\beta}_0$

$\hat{\beta}_1$

Interpretando os resultados da análise

- Interpretação das estimativas dos parâmetros do modelo:
  - ✓ $\hat{\beta}_1$ =5.71, ou seja, a cada aumento de 1 cm de altura há um aumento de 5.71 na capacidade média pulmonar (esperada).
  - ✓ $\hat{\beta}_0$ = - 456.92, ou seja, é a estimativa da capacidade média pulmonar quando os indivíduos tëm altura igual a zero (??).

- É comum que a estimativa do intercepto (β_0) não tenha uma interpretação plausível como nesse exemplo.

Predição?