Exemplo

O departamento de RH de uma empresa deseja avaliar a eficácia dos testes aplicados para a seleção de funcionários.

Para tanto, foi sorteada uma amostra aleatória de 50 funcionários que fazem parte da empresa e que passaram pelo processo de seleção que utilizou os tais testes.

Para cada um dos funcionários foi registrada a nota média nos testes de criatividade, raciocínio mecânico, raciocínio abstrato e habilidade matemática (notas de 0 a 26). Ainda, após 6 meses da contratação, foi calculado um escore que indica o seu desempenho profissional (0 a 120).

Pergunta: existe alguma relação entre o escore de desempenho dos funcionários e a nota média nos testes?

Associação
entre duas variáveis quantitativas

- **Diagrama de dispersão**: recurso gráfico que nos permite visualizar o comportamento conjunto das duas variáveis.

- Coeficiente de **correlação linear**: mede a intensidade da associação linear existente entre as variáveis.

Coeficiente de Correlação Linear

Definição: Medida de associação linear entre duas variáveis quantitativas (varia entre –1 e +1).

- Valores próximos a +1: indicam forte relação linear positiva;
- Valores próximos a -1: indicam forte relação linear negativa;
- Valores próximos a zero: indicam ausência de relação linear.

Um breve parênteses...

Diferença entre correlação e causalidade

- **A correlação não implica necessariamente uma relação de causalidade. Ou seja, um dos eventos não necessariamente causa a ocorrência do outro. Todavia, a correlação pode ser uma pista...**

- **Não é porque (A) acontece juntamente com (B) que podemos afirmar que (A) causa (B).**

- **Por outro lado, se (A) e (B) apresentam relação de causalidade, então eles apresentarão correlação.**

Diferença entre correlação e causalidade

- **Assim, determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer as seguintes situações:**
  - **(A) causa realmente (B);**
  - **(B) pode ser a causa de (A);**
  - **Um terceiro fator (C) pode ser causa tanto de (A) quanto de (B);**
  - **A correlação pode ser apenas uma coincidência, ou seja, os dois eventos não têm qualquer relação para além do fato de ocorrerem ao mesmo tempo.**

Diferença entre correlação e causalidade

- **Bolsa Família é paraquedas eleitoral de Dilma no Norte/Nordeste (Estadão, 11/10/2010)**

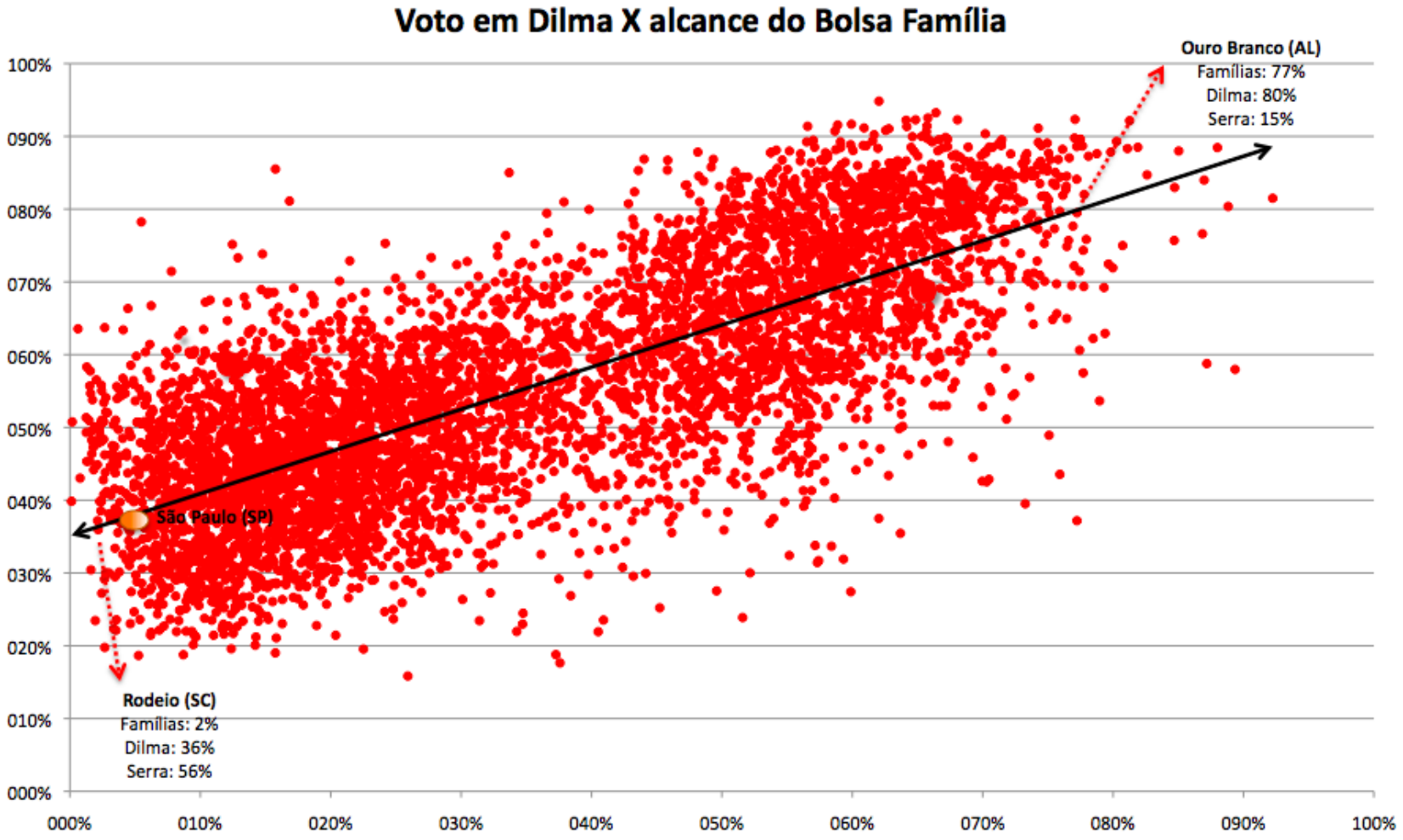
  – **“**Quanto maior o peso do Bolsa Família no município, maior a votação de Dilma Rousseff (PT).**”**

  – **“**A petista tem uma espécie de paraquedas eleitoral que lhe garante um patamar mínimo de votos, especialmente nas regiões onde o programa é mais importante para a economia local.**”**

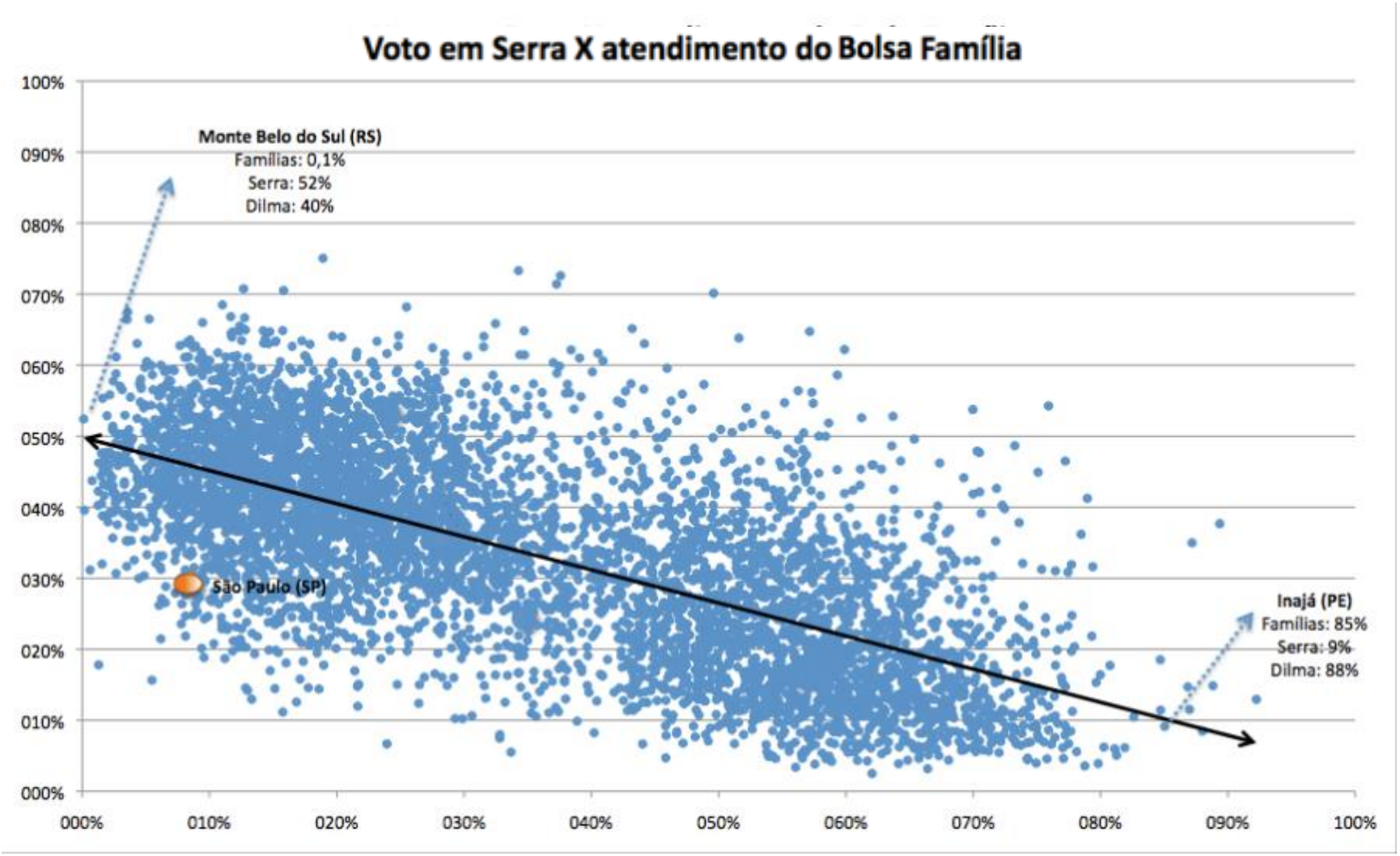
  – **“**Em cerca de metade dos municípios brasileiros, o Bolsa Família atende pelo menos um terço das famílias (...)**”**

http://blogs.estadao.com.br/vox-publica/2010/10/11/bolsa-familia-e-paraquedas-eleitoral-de-dilma-no-nortenordeste/

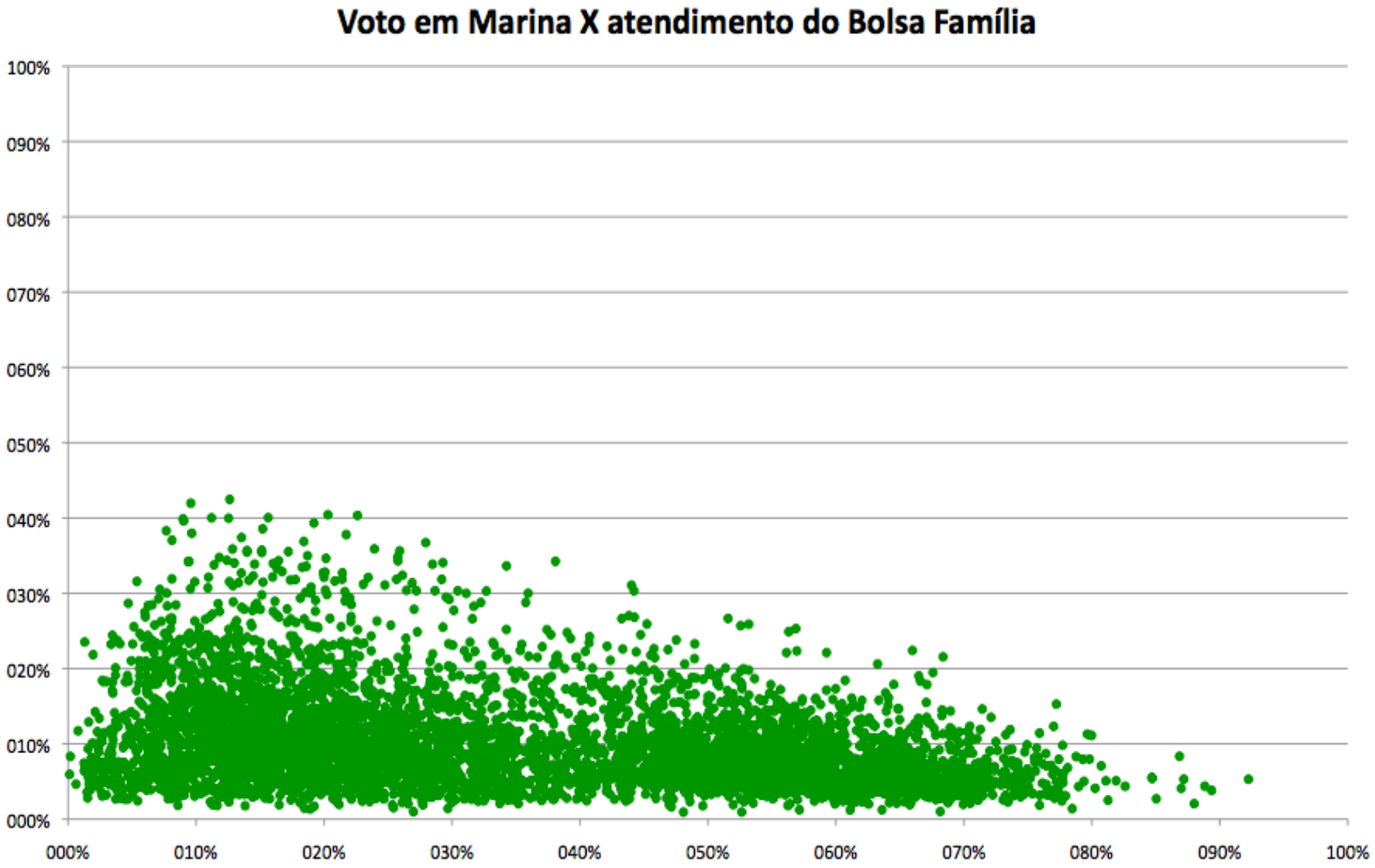
Diferença entre correlação e causalidade



Diferença entre correlação e causalidade



Diferença entre correlação e causalidade



Diferença entre correlação e causalidade

- **Então, para você, o governo usou o Bolsa-Família como moeda de troca eleitoral?**

- **Uma vez que o Bolsa-Família existe, ele gerou votos adicionais para a presidente?**

- **Os gráficos anteriores são suficientes para responder a estas perguntas?**

Voltando ao Exemplo

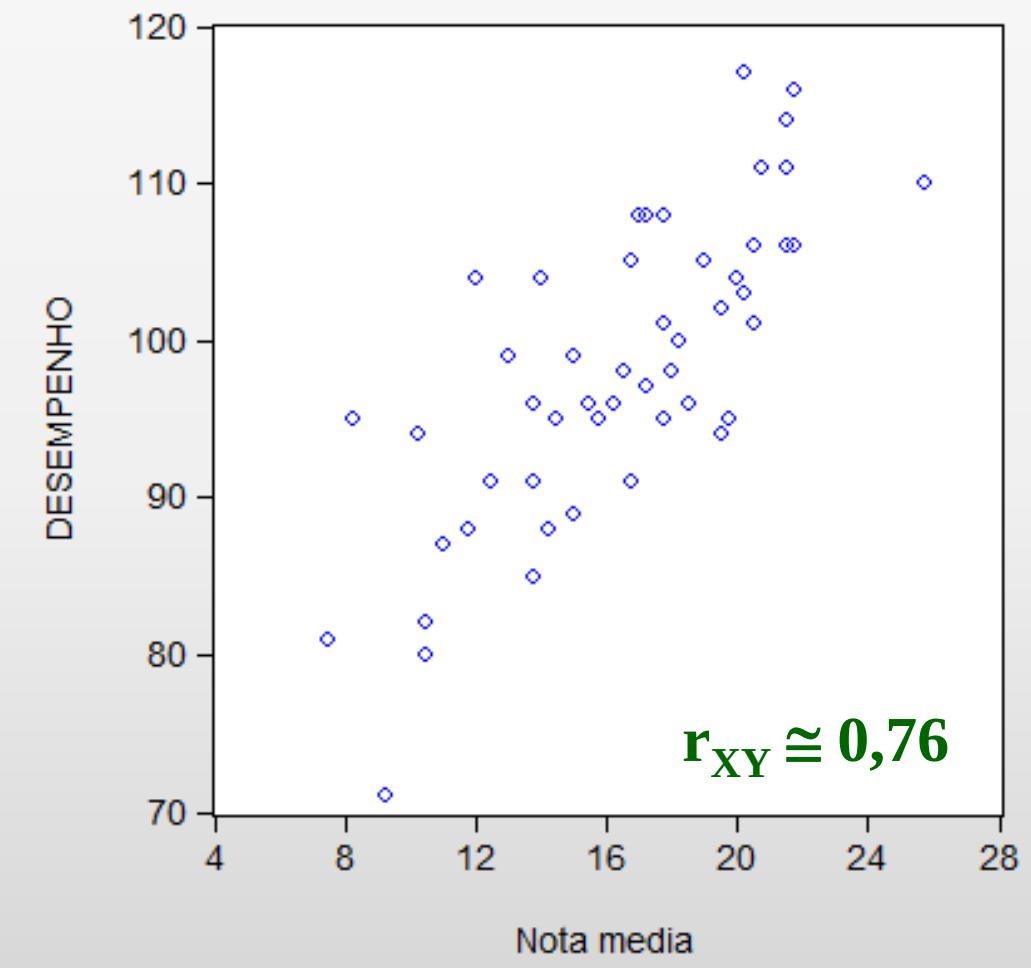
O departamento de RH de uma empresa deseja avaliar a eficácia dos testes aplicados para a seleção de funcionários.

Para tanto, foi sorteada uma amostra aleatória de 50 funcionários que fazem parte da empresa e que passaram pelo processo de seleção que utilizou os tais testes.

Para cada um dos funcionários foi registrada a nota média nos testes de criatividade, raciocínio mecânico, raciocínio abstrato e habilidade matemática (notas de 0 a 26). Ainda, após 6 meses da contratação, foi calculado um escore que indica o seu desempenho profissional (0 a 120).

Pergunta: existe alguma relação entre o escore de desempenho dos funcionários e a nota média nos testes?

Voltando ao Exemplo



Desempenho vs Nota Média

Voltando ao Exemplo

Perguntas:

a) Qual modelo estatístico você proporia para estudar a relação entre o escore de desempenho dos funcionários e a nota média nos testes?

b) Qual método de estimação você utilizaria para encontrar os estimadores dos parâmetros do modelo proposto? Esses estimadores apresentam boas propriedades?

c) A variável nota média nos testes é relevante para explicar o escore médio de desempenho dos funcionários?

d) Qual a estimativa para o escore de desempenho de funcionários que obtiveram nota média igual a 13 nos testes?

Insper | Instituto de Ensino e Pesquisa

Análise de Regressão Linear Simples I

Aula 01

Gujarati e Porter – Capítulos 2 e 3

Wooldridge – Seções 2.2 e 2.3

Análise de Regressão

Regressão – Técnica Estatística utilizada para investigar e modelar a relação entre variáveis.

Objetivo – Na situação em que muitas variáveis estão envolvidas, estudar o efeito que algumas variáveis exercem nas outras. Este estudo consistiria na construção e análise de uma relação matemática entre as variáveis (no geral, uma variável em função das outras).

Análise de Regressão

Na terminologia de regressão, a variável que está sendo estudada é chamada de variável dependente ou resposta, comumente denotada por Y.

Já as variáveis (ou a variável) que estão sendo usadas para explicar a variável dependente são chamadas de variáveis independentes, explicativas ou regressores, comumente denotadas por X_1, X_2, ..., X_k.

A análise de regressão consiste em estudar como alterações nas variáveis explicativas influenciam a variável resposta.

Análise de Regressão

O tipo mais simples de análise de regressão, envolvendo uma variável explicativa (ou independente) e uma variável resposta (ou dependente), é chamado de regressão linear simples.

A análise de regressão envolvendo duas ou mais variáveis explicativas é chamada de análise de regressão linear múltipla.

Regressão Linear Simples

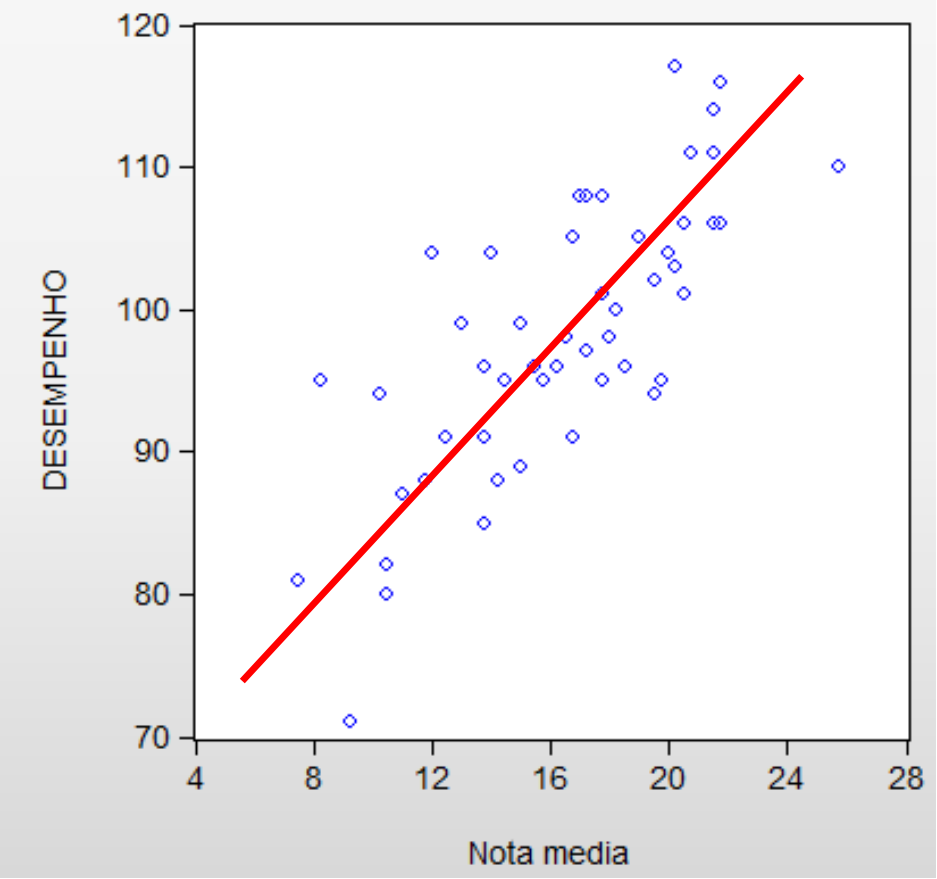
Definição – A função E(Y|X) é chamada regressão de Y em X.

Aqui, será abordado um importante modelo de regressão, **o modelo de regressão linear**, no qual **E(Y|X)** é uma **função linear nos parâmetros**.

Vale observar que a **relação matemática existente entre Y (variável resposta) e X (variável explicativa) pode ser qualquer**.

Voltando ao Exemplo

Qual forma funcional você proporia para estudar a relação entre o escore de desempenho dos funcionários e a nota média nos testes?



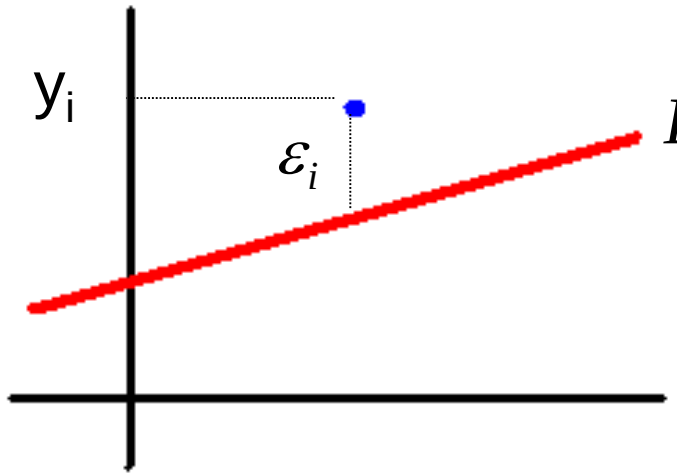
$$E(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

Observações

Duas amostras obtidas do mesmo teste de aptidão (X) não teriam obrigatoriamente que apresentar o mesmo resultado no que diz respeito ao desempenho (Y), mas valores em torno de um valor $\beta_0 + \beta_1 x$ (reta).

Não esperamos uma relação perfeita entre as variáveis nota média nos testes e escore de desempenho dos funcionários, uma vez que outros fatores não controlados como, por exemplo, tempo de experiência na função também podem influenciar na explicação da variável escore de desempenho.

Regressão Linear Simples



$$E(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

Modelo

$$y_i = E(Y|X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Característica comum

Característica específica

Regressão Linear Simples

Observação 1

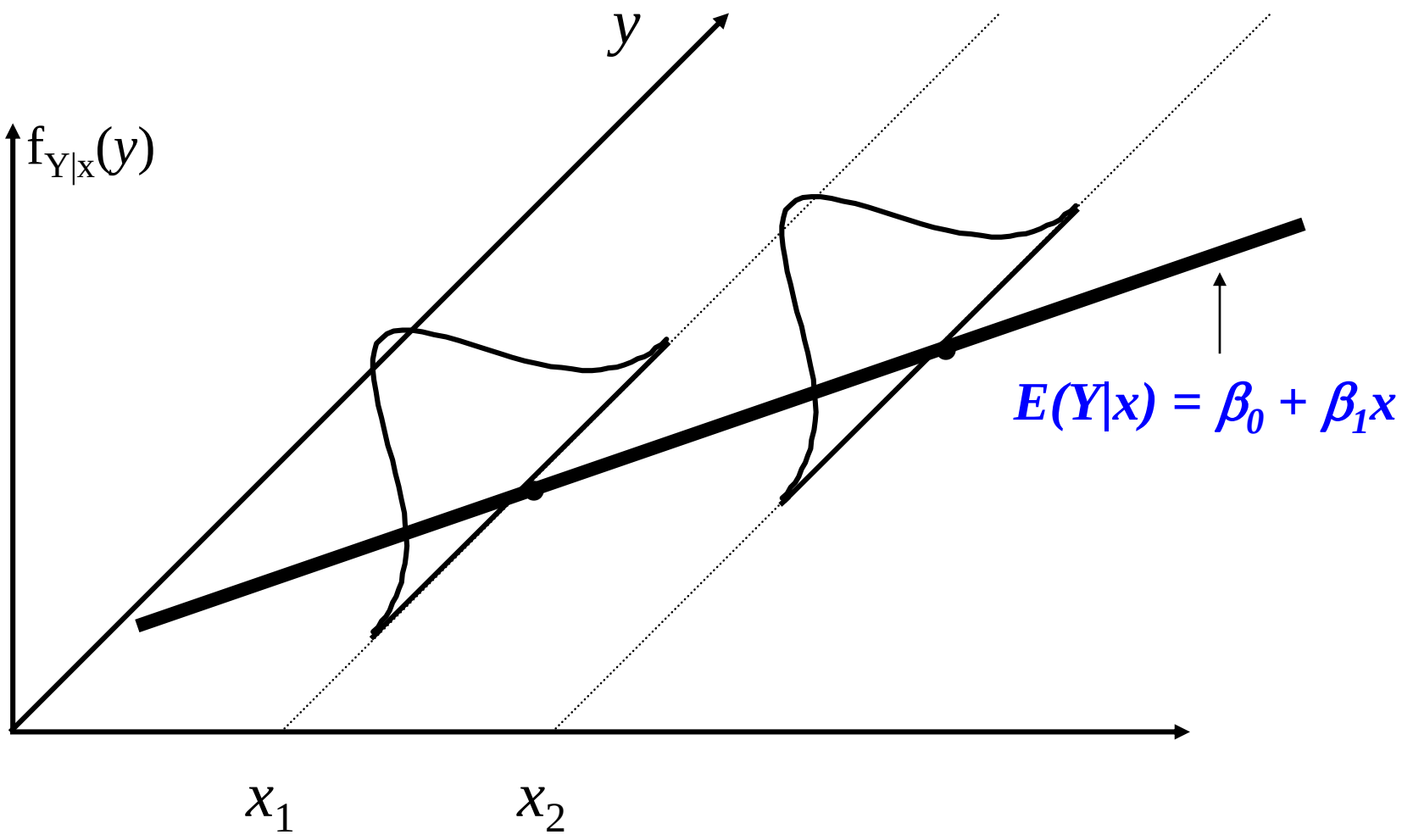
É comum supor que os

$$\varepsilon_i,\ i = 1, 2, ..., n,$$

sejam variáveis aleatórias identicamente distribuídas, qualquer que seja o valor do regressor, que em muitos casos é considerado aleatório.

E(*Y*|*x*) como uma função linear de *x*,

onde para todo *x* a distribuição de *Y* é centrada sobre E(*Y*|*x*)



Regressão Linear Simples

Observação 2

Vale salientar que o termo regressão linear significa regressão linear nos parâmetros, ou seja, modelos da forma

$$y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$$

ou da forma

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$$

também serão considerados regressões lineares.

Regressão Linear Simples

O parâmetro

$$E(Y|X=x) = \beta_0 + \beta_1 x,$$

que representa a média da v.a. Y, condicional a X = x, será estimado por

Abuso de notação

$$\underbrace{E(Y|X=x)}_{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 x = b_0 + b_1 x$$

em que

$$\hat{\beta}_0 = b_0 \;\; e \;\; \hat{\beta}_1 = b_1 \;\; são\ estimativas\ para\ \beta_0 \;\; e \;\; \beta_1.$$

Regressão Linear Simples

Ainda, a quantidade

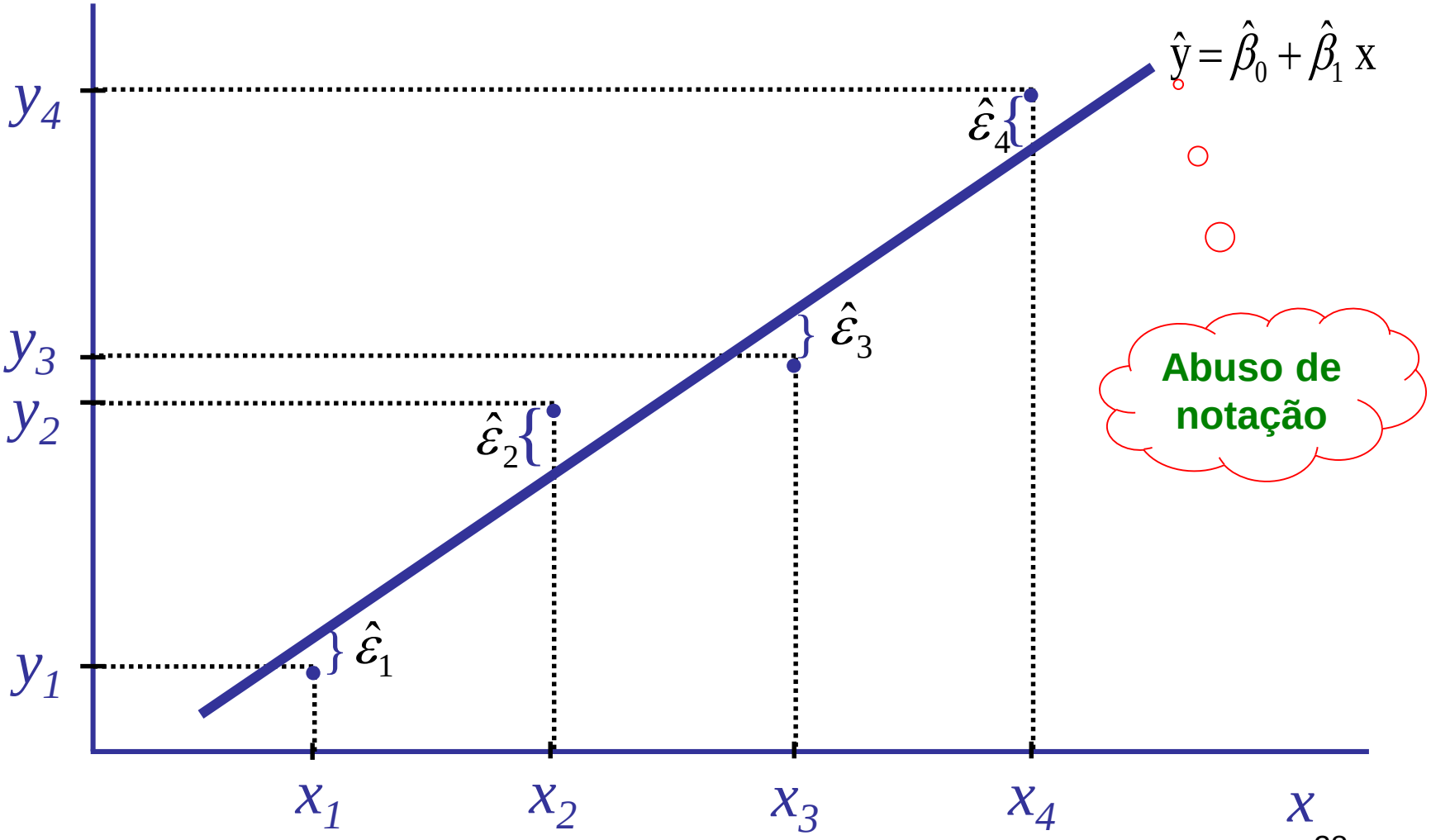
Abuso de notação

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i),\ i = 1,\ 2,\ ...,\ n.$$

será chamada de resíduo.

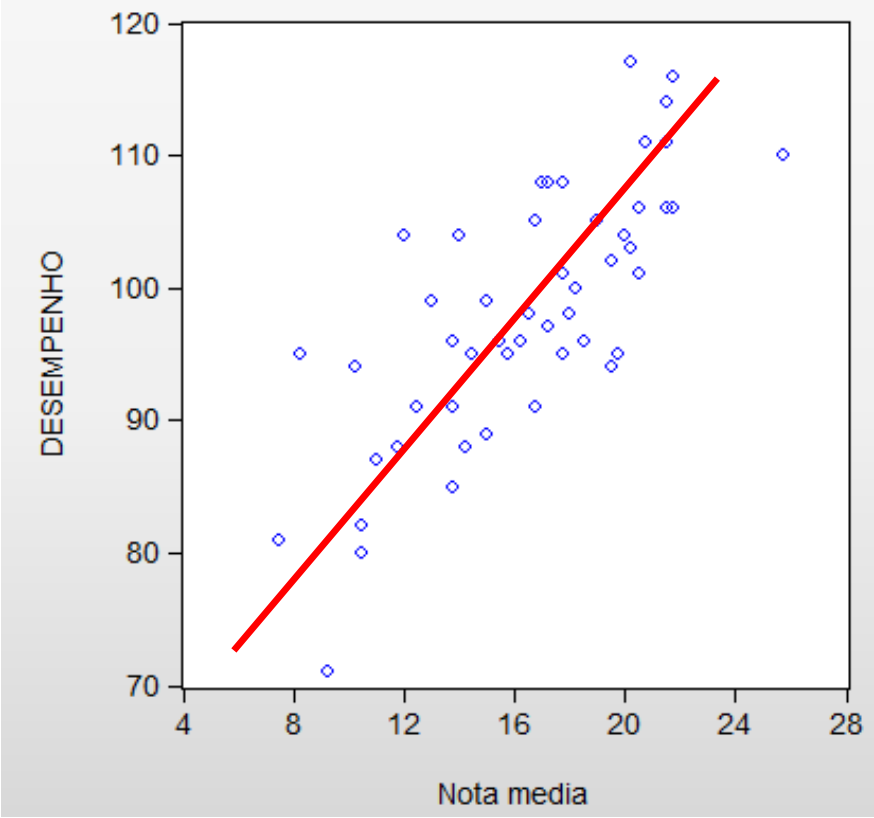
Assim, o valor $\hat{\varepsilon}_i$ pode ser encarado como o erro cometido por prever y_i (i = 1, 2, ..., n) a partir de $\hat{y}_i$.

Regressão Linear Simples



Voltando ao Exemplo

Qual método de estimação você utilizaria para, com base numa dada amostra, encontrar as estimativas dos parâmetros do modelo de regressão linear simples anteriormente proposto?



$$E(Y|x_i) = \beta_0 + \beta_1 x_i$$

Estimação

Qual método de estimação utilizar?

Um procedimento bastante utilizado em Econometria para obter estimadores é aquele que se baseia no princípio dos mínimos quadrados ordinários (MQO), introduzido por Gauss em 1794.

Mínimos Quadrados Ordinários

Ideia!

Quanto menor for o erro quadrático total ($\Sigma\varepsilon_i^2$), melhor será a estimativa. Isso nos sugere procurar a estimativa que torne mínima essa soma de quadrados. Matematicamente, o problema passa a ser o de encontrar os valores de β_0 e β_1 que minimizem a função

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Mínimos Quadrados Ordinários

O mínimo da função é obtido derivando-a em relação a β_0 e β_1, e igualando o resultado a zero, o que resulta

$$\frac{\partial}{\partial \beta_0} S(\beta_0; \beta_1) = 0$$

e

$$\frac{\partial}{\partial \beta_1} S(\beta_0; \beta_1) = 0$$

Mínimos Quadrados Ordinários

Voltando à função de interesse:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Derivando...

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} \{[y_i - (\beta_0 + \beta_1 x_i)] x_i\}$$

Mínimos Quadrados Ordinários

Igualando a zero a derivada em relação ao parâmetro β_0,

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = 0$$

vem que:

$$-2\sum_{i=1}^{n}\left[y_i - \underbrace{\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}_{\hat{y}_i} \right] = 0$$

Nota: via condição de primeira ordem, notamos que a soma dos resíduos, no modelo de regressão linear com intercepto, é sempre igual a zero.

Mínimos Quadrados Ordinários

Igualando a zero a derivada em relação ao parâmetro β_1,

$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = 0$$

vem que:

$$-2\sum_{i=1}^{n} \left\{ \underbrace{\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]}_{\hat{\varepsilon}_i} x_i \right\} = 0$$

Nota: via condição de primeira ordem, notamos que a covariância entre os resíduos e o regressor é sempre igual a zero.

Mínimos Quadrados Ordinários

Abrindo o somatório da igualdade

$$-2\sum_{i=1}^{n}\left[y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right]=0$$

vem que:

$$\sum_{i=1}^{n} y_i=\sum_{i=1}^{n}\hat{\beta}_0+\sum_{i=1}^{n}\hat{\beta}_1 x_i \Rightarrow \sum_{i=1}^{n} y_i=n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n} x_i$$

Assim,

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

Mínimos Quadrados Ordinários

Ainda, abrindo o somatório da igualdade

$$-2\sum_{i=1}^{n}\left\{\left[y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right]x_i\right\}=0$$

vem que:

$$\sum_{i=1}^{n}x_i y_i-\hat{\beta}_0\sum_{i=1}^{n}x_i-\hat{\beta}_1\sum_{i=1}^{n}x_i^2=0$$

Substituindo $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$ **na igualdade anterior, não é difícil obter:**

Mínimos Quadrados Ordinários

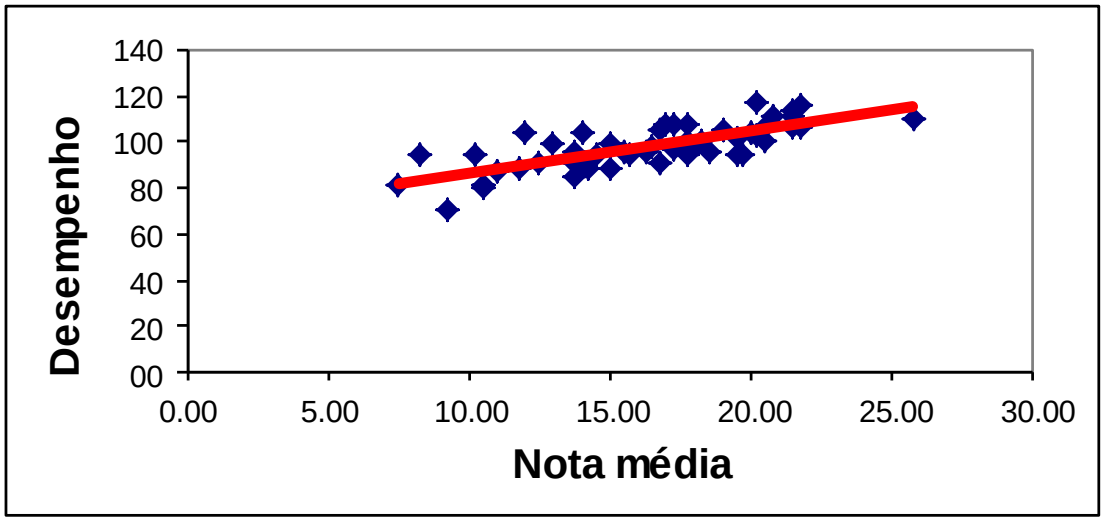
$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{XY}}{S_X^2} = r_{XY}\frac{S_Y}{S_X}$$

Dessa forma, a equação estimada por mínimos quadrados fica dado por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

Voltando ao Exemplo

Quais as estimativas dos parâmetros do modelo de regressão linear simples de interesse?



Abuso de notação

$$\hat{y} = 68{,}51 + 1{,}81x$$

Como tais estimativas devem ser interpretadas?

Regressão Linear Simples

$$y_i = E(Y|X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Parâmetros

β_0 – é o intercepto;

β_1 – coeficiente angular da reta de regressão

$$\frac{\partial E(Y|X = x)}{\partial x} = \beta_1$$

Observação

Na prática, nem sempre β_0 (intercepto) apresenta interpretação.

Voltando ao Exemplo

Abuso de notação

$$\hat{y} = 68{,}51 + 1{,}81x$$

68,51: valor médio do desempenho dos funcionários que tiraram média igual a zero nos testes de admissão.

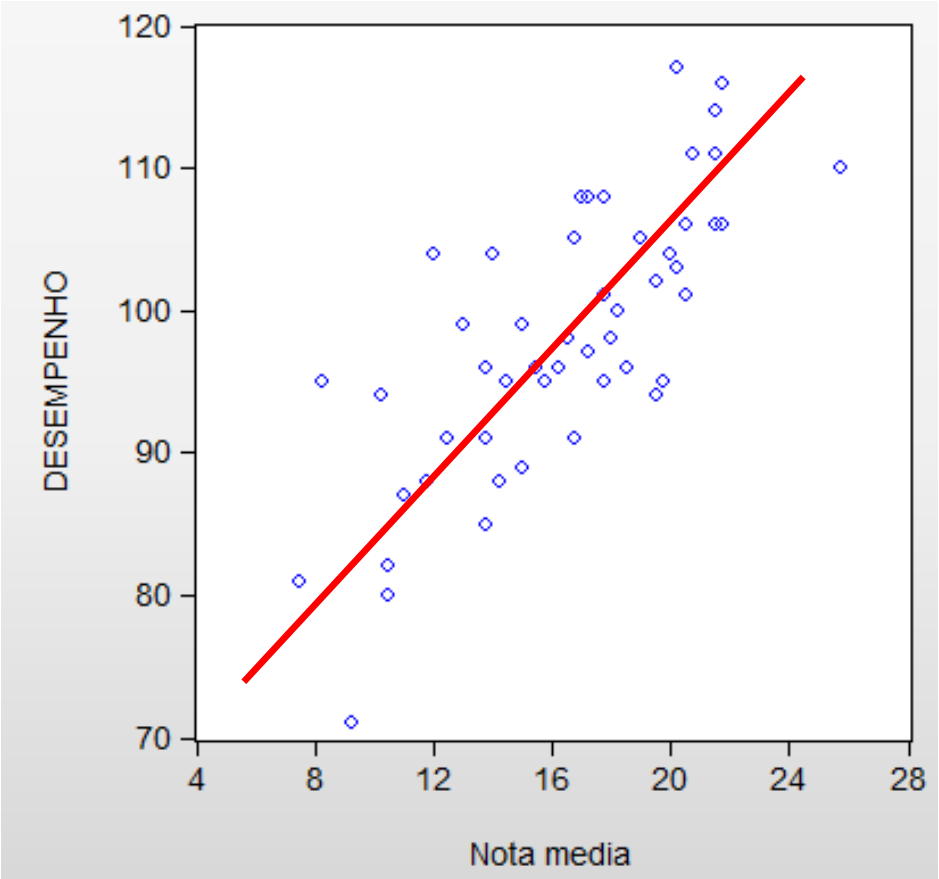
1,81: variação média no desempenho dos funcionários, quando aumenta-se a nota média obtida nos testes de admissão em 1 unidade.

Mínimos Quadrados Ordinários

Exercício

Encontre a matriz hessiana e verifique sob quais condições a mesma é definida como positiva. Ainda, discuta se os estimadores encontrados geram o mínimo da função de interesse.

Voltando ao Exemplo



Abuso de notação

$$\hat{y} = 68{,}51 + 1{,}81x$$

O modelo de regressão proposto está bem ajustado?

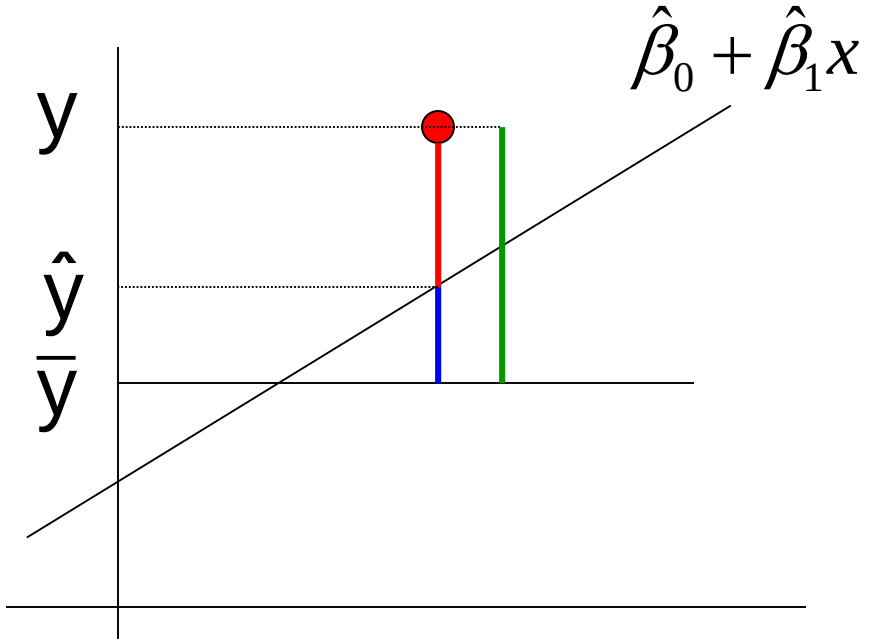
Como medir a qualidade de ajuste do modelo?

Objetivo

Construir uma medida que indique, mesmo que de modo imperfeito, a qualidade do ajuste do modelo de regressão.

Coeficiente de determinação (ou de explicação) – R^2

Somas de Quadrados

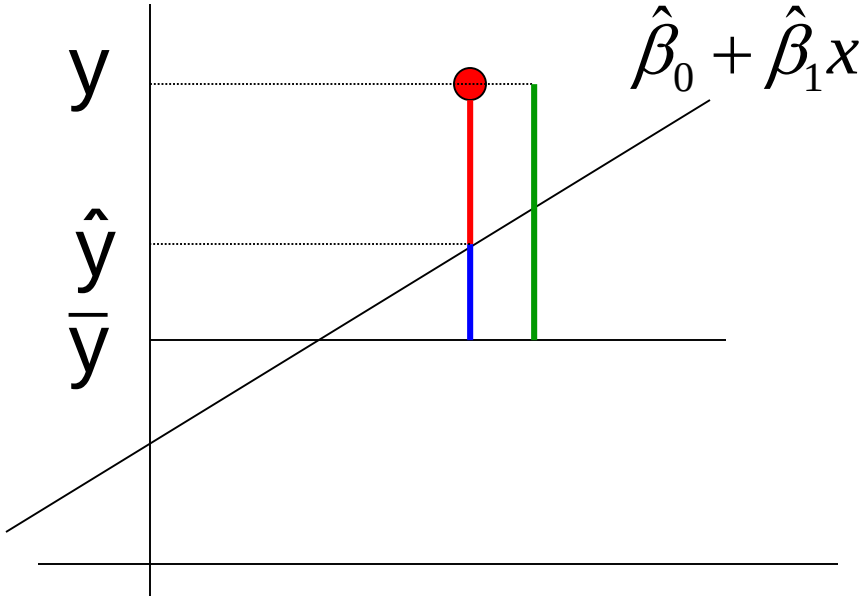


$y - \bar{y}$: erro ao se prever y pela média geral

$y - \hat{y}$: erro ao se prever y pelo valor estimado para E(Y|X)

$\hat{y} - \bar{y}$: "ganho" ao se prever y pelo valor estimado para E(Y|X) em comparação ao se prever y pela média geral

Somas de Quadrados



$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

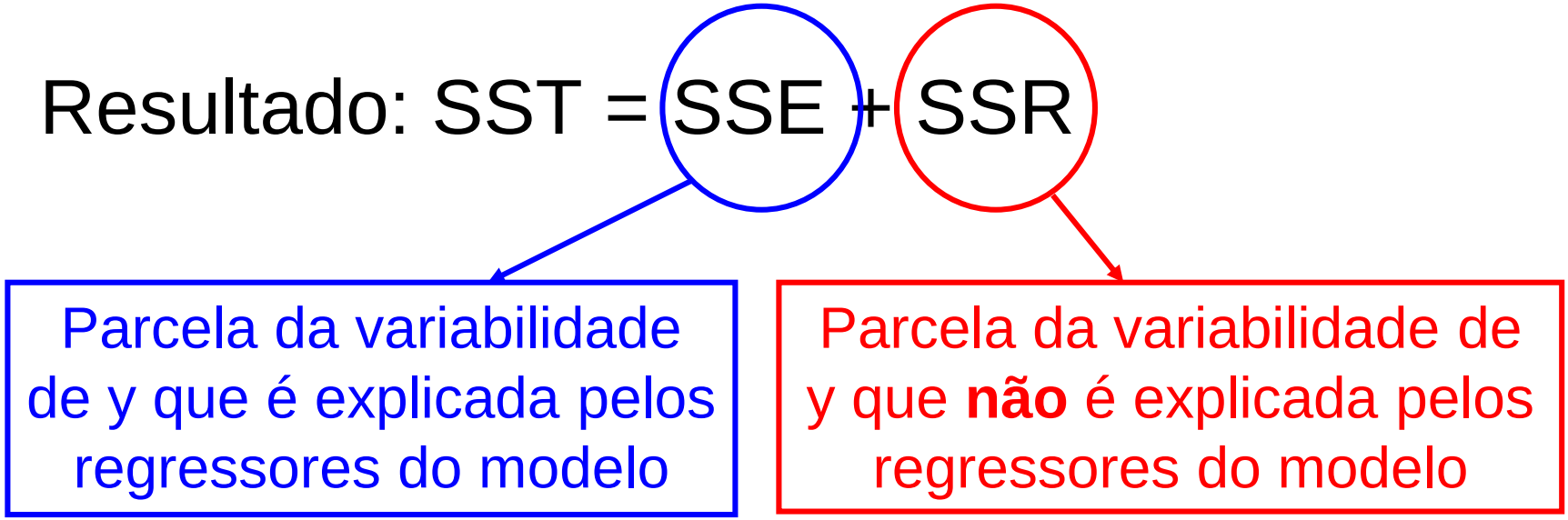
$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SST: soma de quadrados total

SSR: soma de quadrados devido aos resíduos

SSE: soma de quadrados devido à explicação (modelo de regressão)

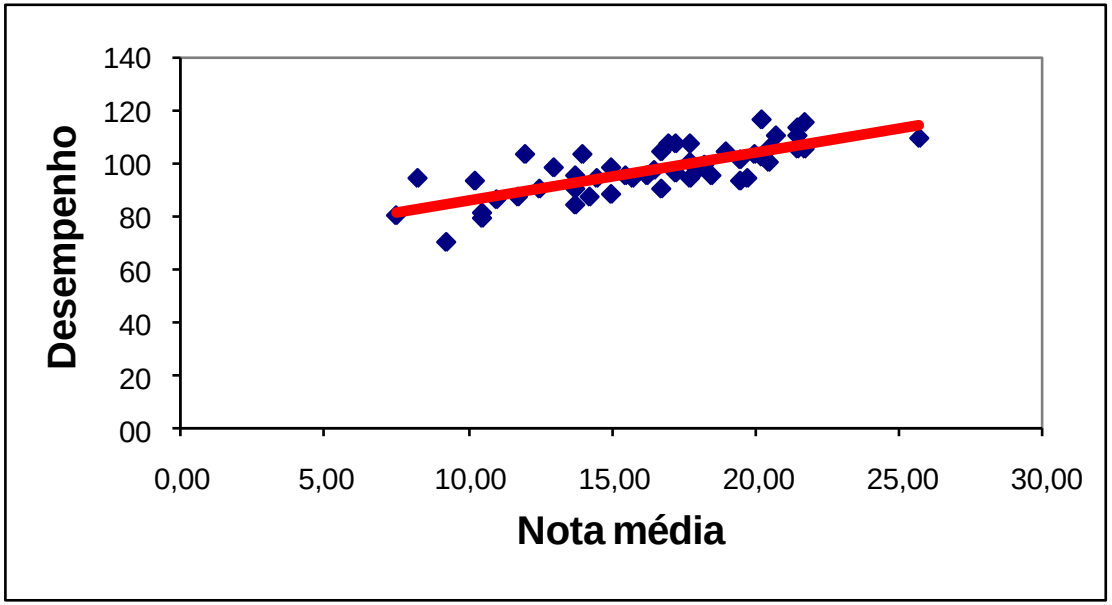
Coeficiente de Determinação (R^2)



$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Proporção da variabilidade total de y que é explicada pelos regressores do modelo adotado.

Voltando ao Exemplo



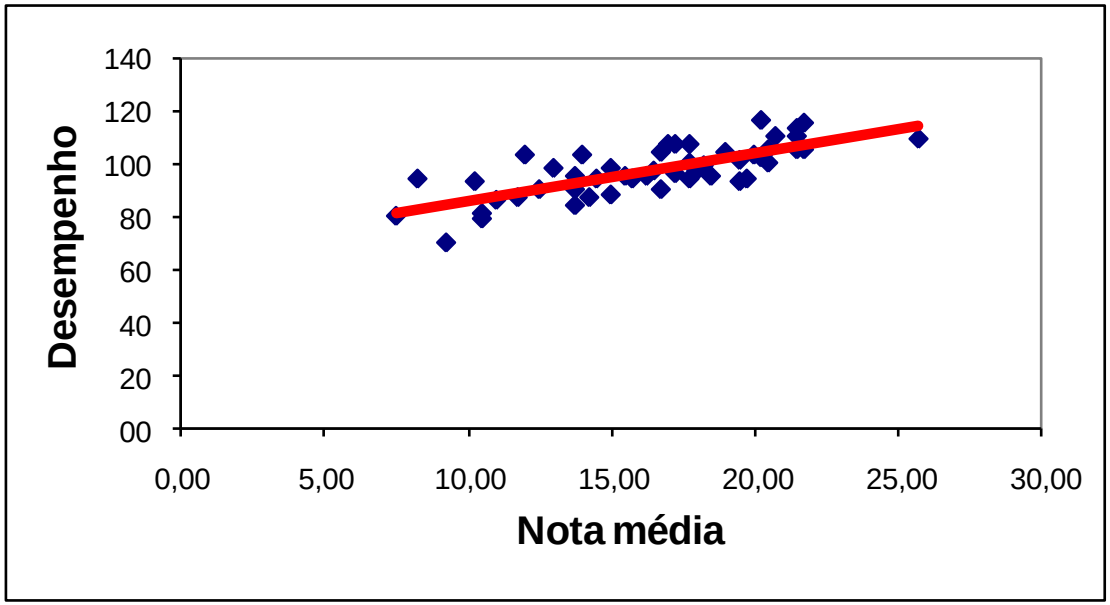
$$R^2 = \frac{SSE}{SST} = 0{,}5808$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 4.593{,}1$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 1925{,}3$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2.667{,}7 = SST - SSE$$

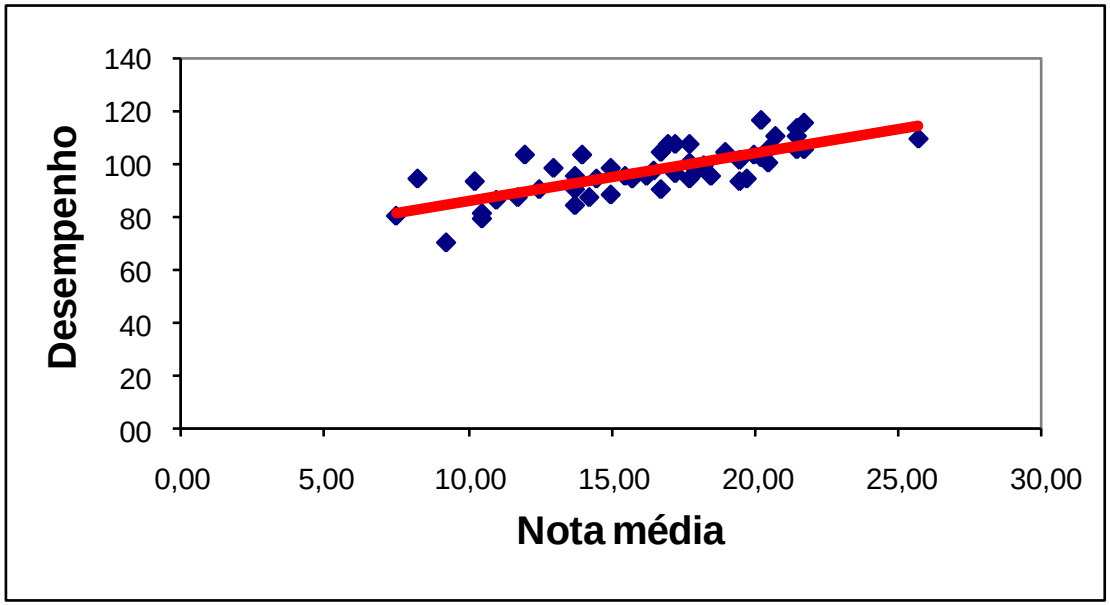
Voltando ao Exemplo



$$R^2 = \frac{SSE}{SST} = 0{,}5808$$

Interpretação: **58% das variações no desempenho dos funcionários após 3 meses de trabalho são explicadas pela nota média obtida nos testes de admissão.**

Voltando ao Exemplo



$$R^2 = \frac{SSE}{SST} = 0{,}5808$$

Conclusão: **Parece que a nota média obtida é relevante para a explicação do desempenho dos funcionários, uma vez que tal regressor explica mais da metade das variações da variável resposta.**

Coeficiente de Determinação (R^2)

Exercício

Prove que, no caso do modelo de regressão linear simples com intercepto, o coeficiente de correlação linear de Pearson elevado ao quadrado é igual ao coeficiente de explicação (ou determinação) – R^2. Ou seja,

$$\mathrm{R}^2 = \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{S_{XY}^2}{S_{xx}S_{YY}} = \hat{\beta}_2\left(\frac{S_{XY}}{S_{YY}}\right)$$