

# **AULAS 02 E 03**

# **Modelo de Regressão**

# **Simple**

**Ernesto F. L. Amaral**

**04 e 09 de março de 2010**

**Métodos Quantitativos de Avaliação de Políticas Públicas (DCP 030D)**

**Fonte:**

**Wooldridge, Jeffrey M. "Introdução à econometria: uma abordagem moderna". São Paulo: Cengage Learning, 2008. pp.20-63.**

## ESTRUTURA DO LIVRO

- **Parte 1:** trata de análise de regressão com dados de corte transversal (capítulos 2 ao 9).
- **Parte 2:** análise de regressão com dados de séries temporais (capítulos 10 ao 12).
- **Parte 3:** tópicos avançados (capítulos 13 ao 19). Iremos até o 15.

# MODELO DE REGRESSÃO SIMPLES

- O modelo de regressão linear simples explica uma variável ( $y$ ) com base em modificações em outra variável ( $x$ ).
- Ou seja, é usado para avaliar a relação entre duas variáveis.
- Esse tipo de regressão não é muito utilizada em econometria aplicada, devido à sua simplicidade.
- No entanto, serve como ponto de partida, já que sua álgebra e interpretações são fáceis de entender.
- O entendimento do modelo de regressão simples é importante para estudar a regressão múltipla.

## PREMISSA E EXEMPLOS

- Premissa da análise econométrica:
  - $y$  e  $x$  são duas variáveis que representam uma população.
  - Estamos interessados em explicar  $y$  em termos de  $x$ .
  - Ou seja, queremos estudar como  $y$  varia com variações em  $x$ .
- Exemplos:
  - $y$  é o rendimento do trabalhador, e  $x$  são os anos de escolaridade.
  - $y$  é a fecundidade da mulher, e  $x$  é a idade da mulher.
  - $y$  é a taxa de criminalidade em uma comunidade, e  $x$  é o número de policiais.

## PERGUNTAS IMPORTANTES

- Como nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam  $y$ ?
- Qual é a relação funcional entre  $y$  e  $x$ ?
- Como podemos estar certos de que estamos capturando uma relação *ceteris paribus* (outros fatores constantes) entre  $y$  e  $x$ ?

# MODELO DE REGRESSÃO LINEAR SIMPLES

- Também chamado de modelo de regressão linear de duas variáveis ou modelo de regressão linear bivariada.

$$y = \beta_0 + \beta_1 x + u$$

- Terminologia:

<b>y</b>	<b>x</b>	<b>Uso</b>
Variável Dependente	Variável Independente	Econometria
Variável Explicada	Variável Explicativa	
Variável de Resposta	Variável de Controle	Ciências Experimentais
Variável Prevista	Variável Previsora	
Regressando	Regressor	
	Covariável	

## VOLTANDO ÀS PERGUNTAS IMPORTANTES

- **Como nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam  $y$ ?**
  - Variável  $u$  é o termo erro ou perturbação da relação.
  - Na análise de regressão simples, todos fatores (além de  $x$ ) que afetam  $y$  são tratados como não-observados.
- **Qual é a relação funcional entre  $y$  e  $x$ ?**
  - Se os outros fatores em  $u$  são mantidos fixos, de modo que a variação em  $u$  é zero ( $\Delta u=0$ ), então  $x$  tem um efeito linear sobre  $y$ , tal como:  $\Delta y = \beta_1 \Delta x$ ; se  $\Delta u=0$ .
  - A linearidade do modelo de regressão linear simples implica que uma variação de uma unidade em  $x$  tem o mesmo efeito sobre  $y$ , independentemente do valor inicial de  $x$ .
  - Isso não é realista. Por exemplo, o próximo ano de escolaridade teria um efeito maior sobre os salários, em relação ao anterior. Esse problema será tratado adiante.

## E O PROBLEMA DO *CETERIS PARIBUS*?

- Estamos capturando uma relação *ceteris paribus* (outros fatores constantes) entre  $y$  e  $x$ ?
  - A variação em  $y$  é  $\beta_1$  multiplicado pela variação em  $x$ .
  - $\beta_1$ : **parâmetro de inclinação** da relação entre  $y$  e  $x$ , mantendo fixos os outros fatores em  $u$ .
  - $\beta_0$ : **parâmetro de intercepto** é raramente analisado.
  - $\beta_1$  mede o efeito de  $x$  sobre  $y$ , mantendo todos os outros fatores (em  $u$ ) fixos.
  - No entanto, estamos ignorando todos os outros fatores.
  - Os estimadores de  $\beta_0$  e  $\beta_1$  serão confiáveis em uma amostra aleatória, se o termo não-observável ( $u$ ) estiver relacionado à variável explicativa ( $x$ ) de modo que o erro seja zero.
  - É preciso assumir que o valor médio de  $u$  na população é zero:  $E(u)=0$  (exemplos pág. 23).



## HIPÓTESE SOBRE A RELAÇÃO ENTRE $x$ E $u$

- Se  $u$  e  $x$  não estão correlacionados, então (como variáveis aleatórias) não são linearmente relacionados.
- No entanto, a correlação mede somente a dependência linear entre  $u$  e  $x$ .
- Na correlação, é possível que  $u$  seja não-correlacionado com  $x$  e seja correlacionado com funções de  $x$ , tal como  $x^2$ .
- Melhor seria pensar na distribuição condicional de  $u$ , dado qualquer valor de  $x$ .
- Para um valor de  $x$ , podemos obter o valor esperado (ou médio) de  $u$  para um grupo da população.
- A hipótese é que o valor médio de  $u$  não depende de  $x$ :

$$E(u|x) = E(u) = 0$$

- Ou seja, para qualquer valor de  $x$ , a média dos fatores não-observáveis é a mesma e, portanto, é igual ao valor médio de  $u$  na população (**hipótese de média condicional zero**).

# FUNÇÃO DE REGRESSÃO POPULACIONAL

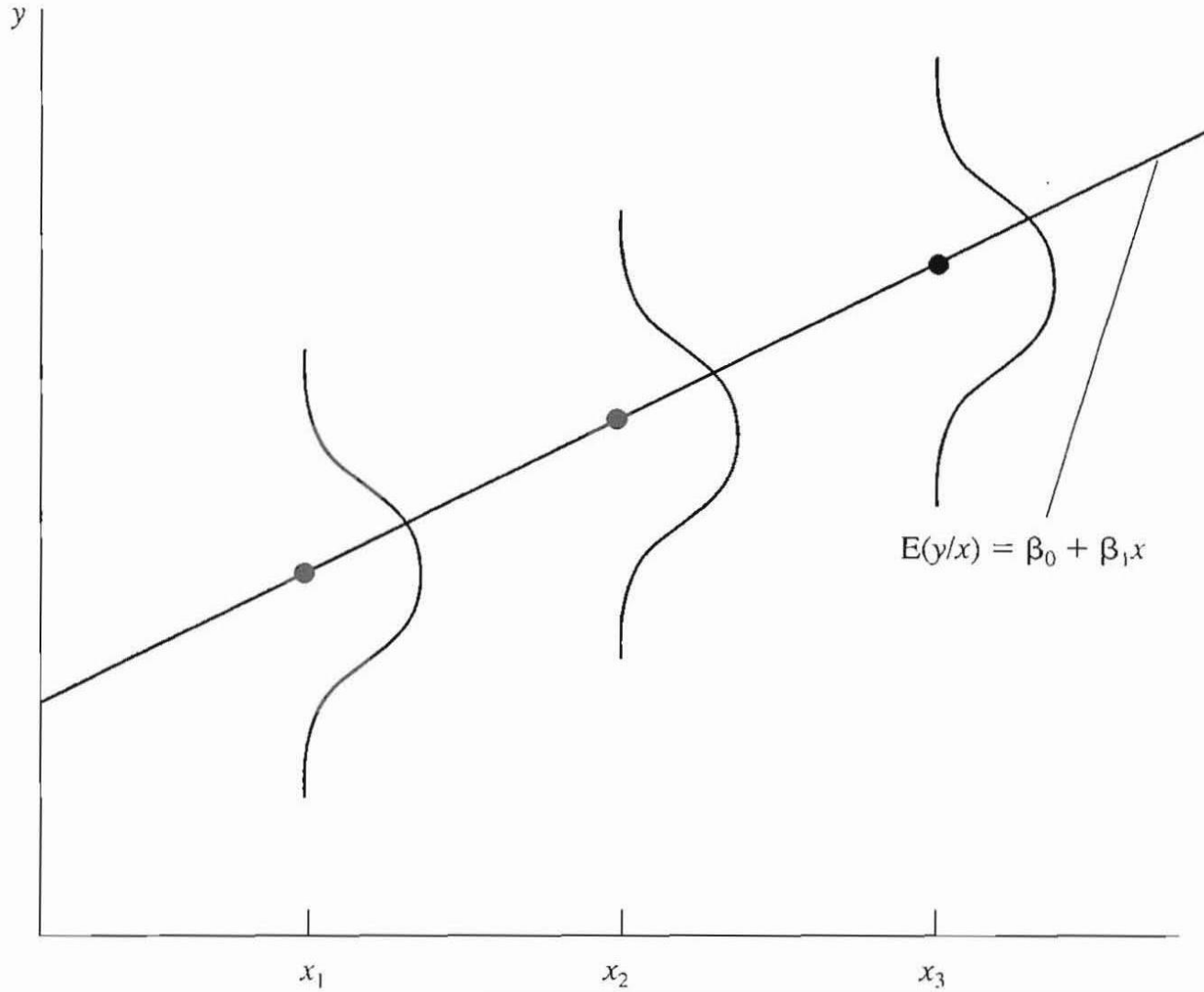
- Quando  $E(u|x)=E(u)=0$  é verdadeiro, é útil dividir  $y$  em:
  - Parte sistemática (parte de  $y$  explicada por  $x$ ):  $\beta_0 + \beta_1 x$
  - Parte não-sistemática (parte de  $y$  não explicada por  $x$ ):  $u$
- Considerando o valor esperado de  $y=\beta_0+\beta_1 x+u$  condicionado a  $x$ , e usando  $E(u|x)=0$ , temos a **função de regressão populacional** (FRP), que é uma função linear de  $x$ :

$$E(y|x) = \beta_0 + \beta_1 x$$

- **Linearidade**: o aumento de uma unidade em  $x$  faz com que o valor esperado de  $y$  varie segundo a magnitude de  $\beta_1$ .
- Para qualquer valor de  $x$ , a distribuição de  $y$  está centrada ao redor de  $E(y|x)$ .

**Figura 2.1**

$E(y|x)$  como função linear de  $x$ .



# ESTIMATIVA DE MÍNIMOS QUADRADOS ORDINÁRIOS

- Para a estimação dos parâmetros  $\beta_0$  e  $\beta_1$ , é preciso considerar uma amostra da população:

$$\{(x_i, y_i): i=1, \dots, n\}$$

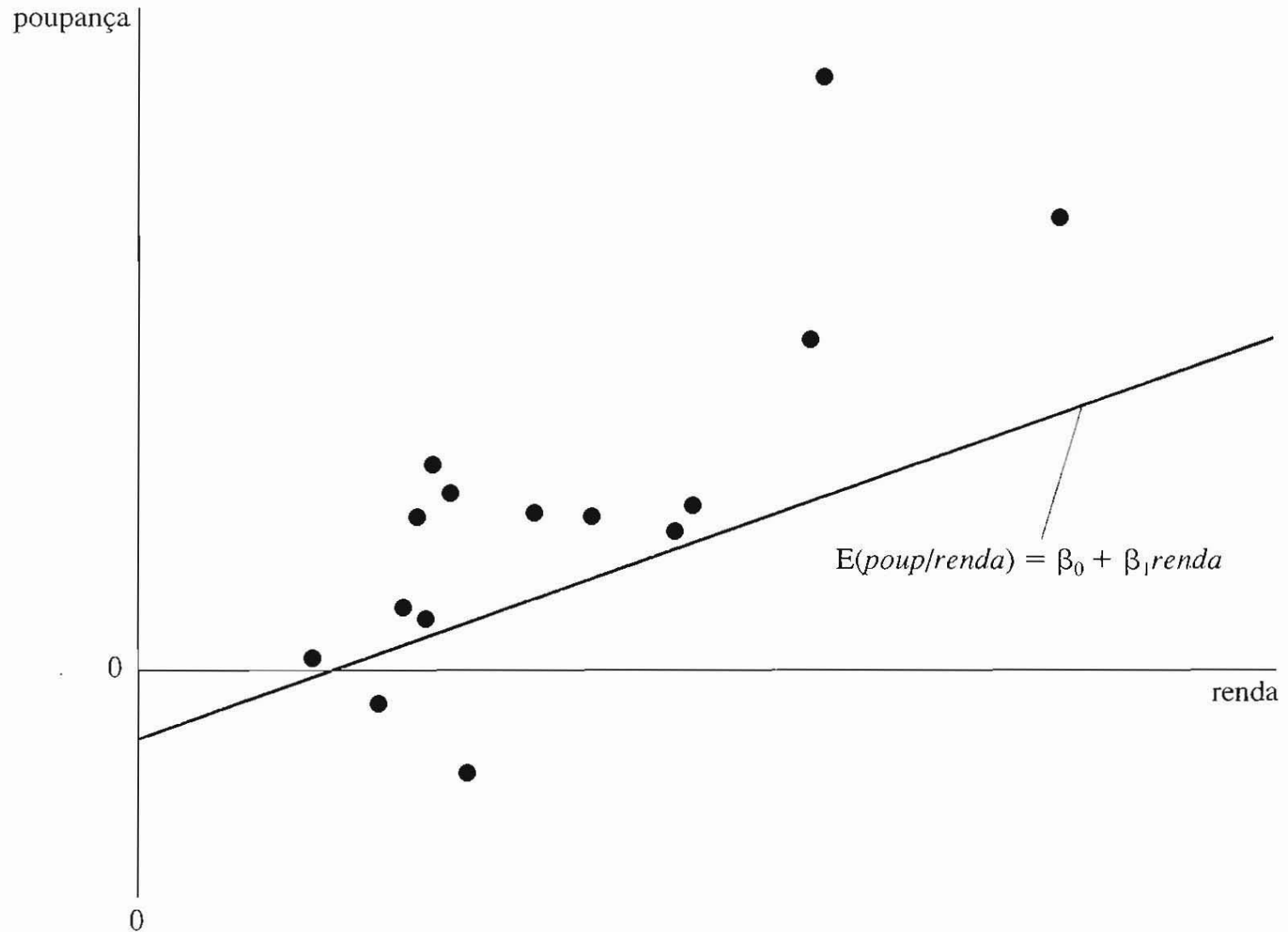
- A equação do modelo de regressão simples é escrito como:

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

- $u_i$  é o termo erro para a observação  $i$ , já que contém todos os fatores, além de  $x_i$ , que afetam  $y_i$ .
- Um exemplo é a poupança anual para a família  $i$  ( $y_i$ ), dependendo da renda anual desta família ( $x_i$ ), em um determinado ano.

**Figura 2.2**

Gráfico da dispersão de poupança e renda de 15 famílias e a regressão populacional  $E(\text{poup}|\text{renda}) = \beta_0 + \beta_1 \text{renda}$ .



# ESTIMATIVA DE MÍNIMOS QUADRADOS ORDINÁRIOS

- Como obter estimativas do intercepto ( $\beta_0$ ) e da inclinação ( $\beta_1$ ) na regressão populacional da poupança sobre a renda?
- Na população,  $u$  tem média zero. O valor esperado de  $u$  é zero:  $E(u)=0$
- Além disso,  $u$  é não-correlacionado com  $x$ . A covariância entre  $x$  e  $u$  é zero:  $Cov(x,u)=E(xu)=0$
- $E(u)=0$  pode ser escrita como:  $E(y-\beta_0-\beta_1x)=0$
- $Cov(x,u)=E(xu)=0$  pode ser escrita como:  $E[x(y-\beta_0-\beta_1x)]=0$
- Como há dois parâmetros desconhecidos para estimar ( $\beta_0$  e  $\beta_1$ ), é possível utilizar uma amostra de dados para calcular as estimativas:

$$\hat{\beta}_0 \quad \text{e} \quad \hat{\beta}_1$$

# EQUAÇÕES DA POPULAÇÃO E AMOSTRA

– Média de  $u$  na população:

$$E(y - \beta_0 - \beta_1 x) = 0$$

– Média de  $u$  na amostra:

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

– Covariância entre  $x$  e  $u$  na população:

$$E[x(y - \beta_0 - \beta_1 x)] = 0$$

– Covariância entre  $x$  e  $u$  na amostra:

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

ESTIMATIVAS DE  $\hat{\beta}_0$  E  $\hat{\beta}_1$ 

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

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

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



# ESTIMATIVAS DE MQO DE $\hat{\beta}_0$ E $\hat{\beta}_1$

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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



$$\hat{\beta}_1 = \frac{\text{Covariância amostral entre x e y}}{\text{Variância amostral de x}}$$

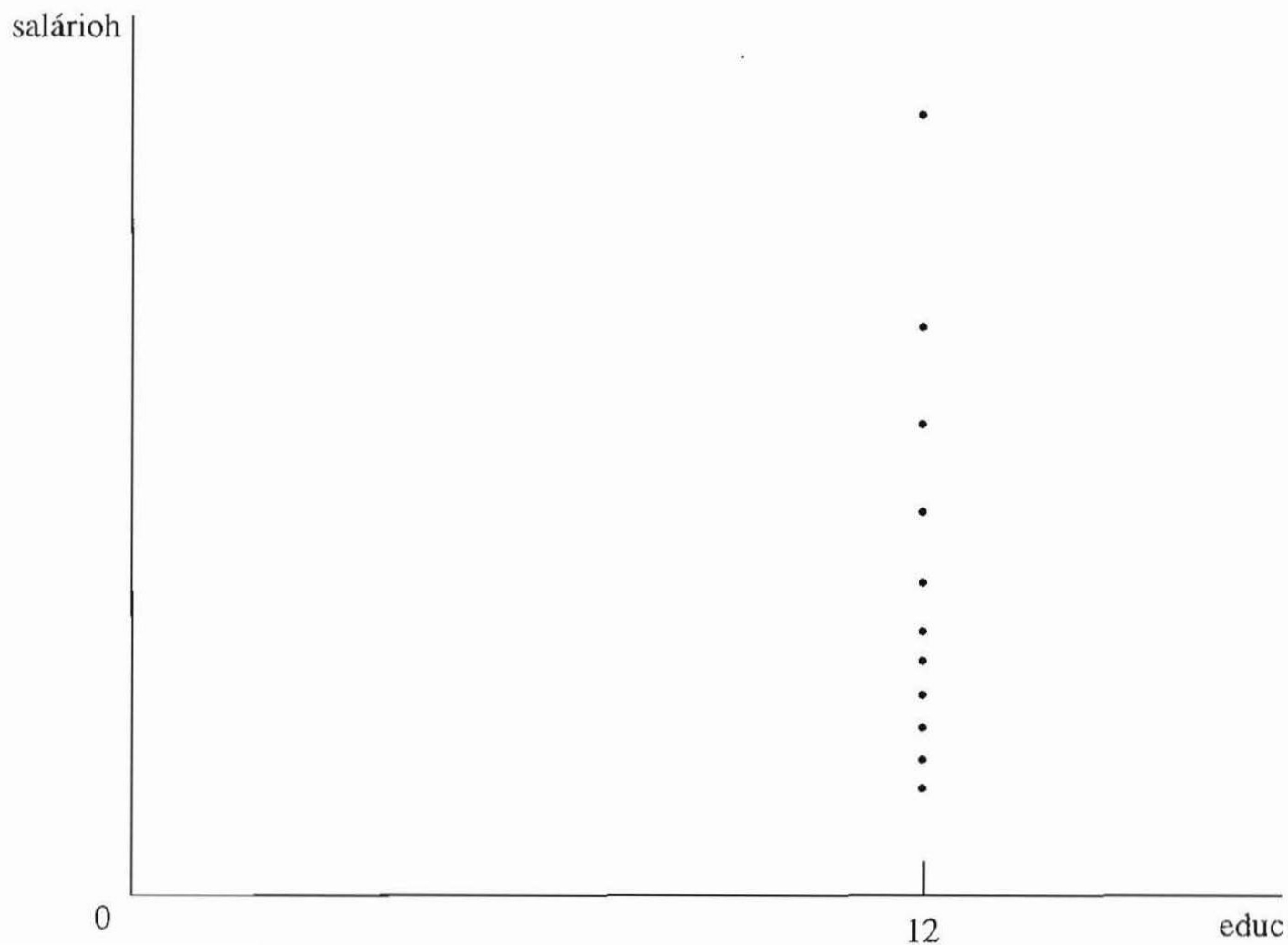
- Se x e y são positivamente correlacionados na amostra,  $\hat{\beta}_1$  é positivo e vice-versa.

## VARIÂNCIA DE $x$ DEVE SER MAIOR QUE ZERO

- A hipótese necessária para calcular estimativas de mínimos quadrados ordinários (MQO) é que a variância amostral de  $x$  seja maior que zero.
- Ou seja, os valores de  $x_i$  na amostra não devem ser todos iguais a um mesmo valor.

**Figura 2.3**

Gráfico da dispersão de salários e educação, quando  $educ_i = 12$  para todo  $i$ .



## VALORES ESTIMADOS E RESÍDUOS

- Encontrados o intercepto e a inclinação, teremos um valor estimado para  $y$  para cada observação ( $x$ ) na amostra:

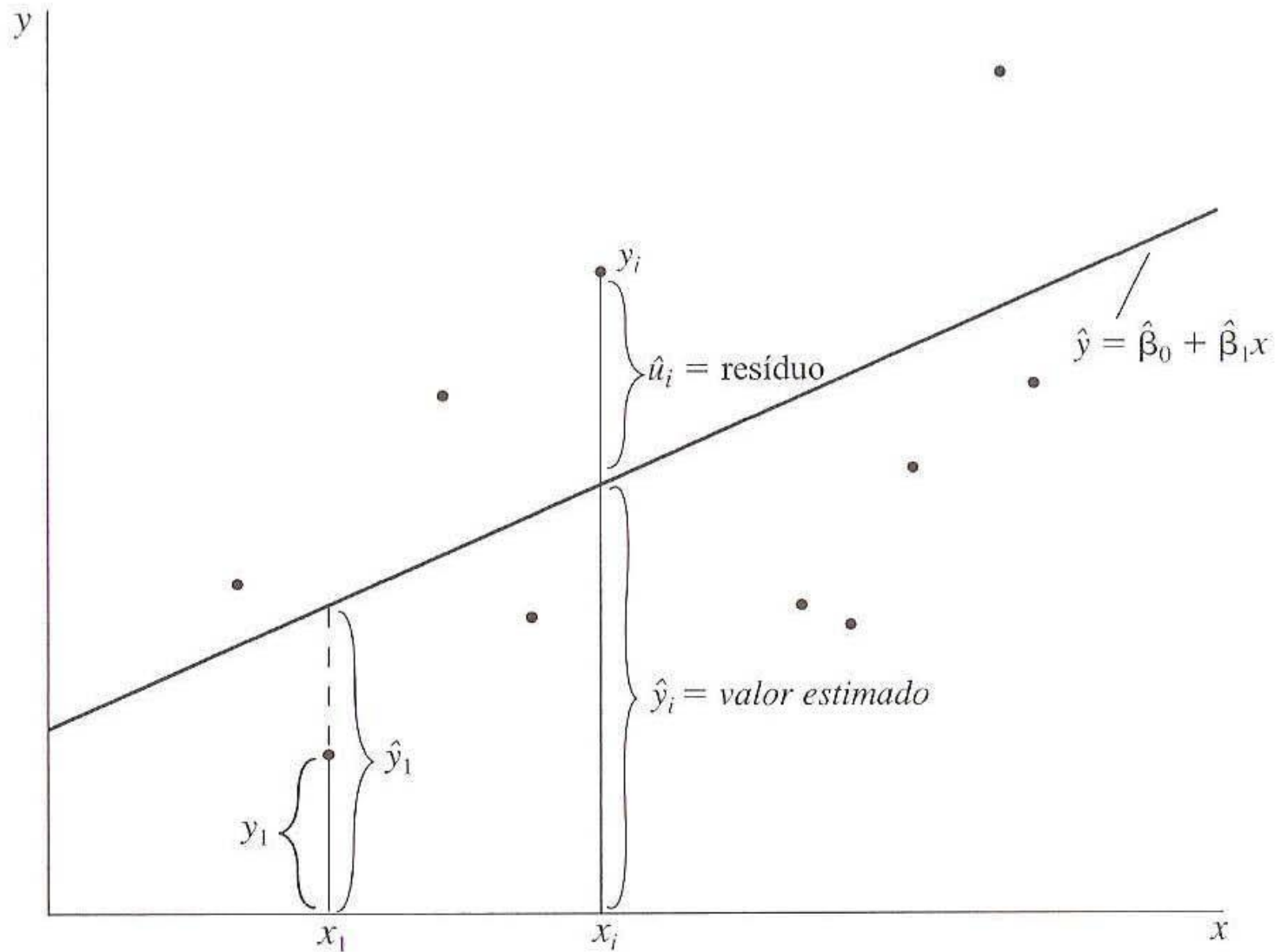
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- O resíduo é a diferença entre o valor verdadeiro de  $y_i$  e seu valor estimado:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

**Figura 2.4**

Valores estimados e resíduos.



## MINIMIZANDO A SOMA DOS RESÍDUOS QUADRADOS

- Suponha que escolhemos o intercepto e a inclinação estimados com o propósito de tornar a soma dos resíduos quadrados:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- O nome “mínimos quadrados ordinários” é utilizado porque as estimativas do intercepto e da inclinação minimizam a soma dos resíduos quadrados.
- Não é utilizada a minimização dos valores absolutos dos resíduos, porque a teoria estatística para isto seria muito complicada

## MINIMIZANDO A SOMA DOS RESÍDUOS QUADRADOS

- Reta de regressão de MQO ou função de regressão amostral (FRA) é a versão estimada da função de regressão populacional (FRP):

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

- O coeficiente de inclinação indica o quanto o valor estimado (previsto) de  $y$  varia quando  $x$  aumenta em uma unidade:

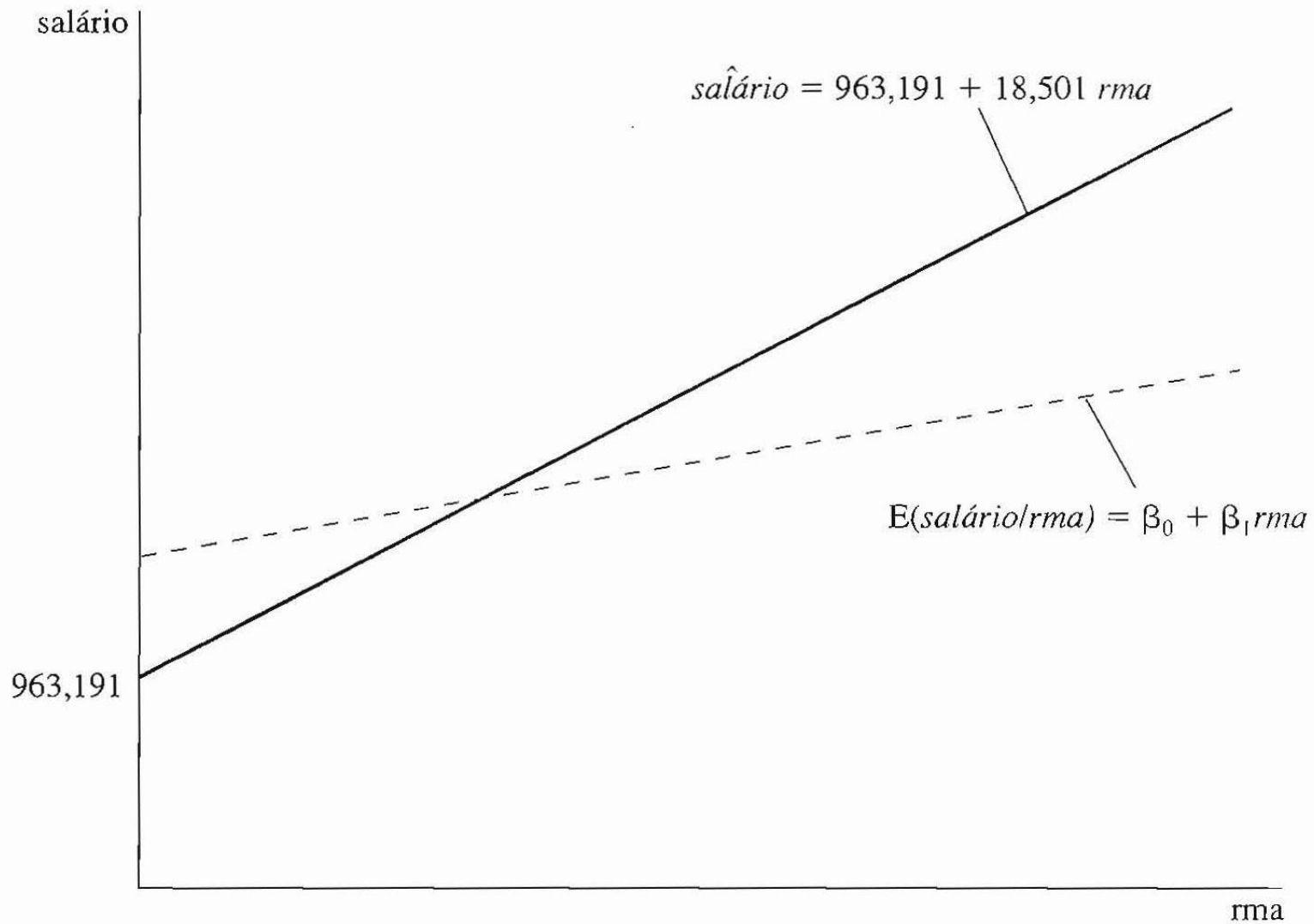
$$\hat{\beta}_1 = \Delta \hat{y} / \Delta x$$

- Da mesma forma, dada qualquer variação em  $x$ , podemos calcular a variação prevista em  $y$ :

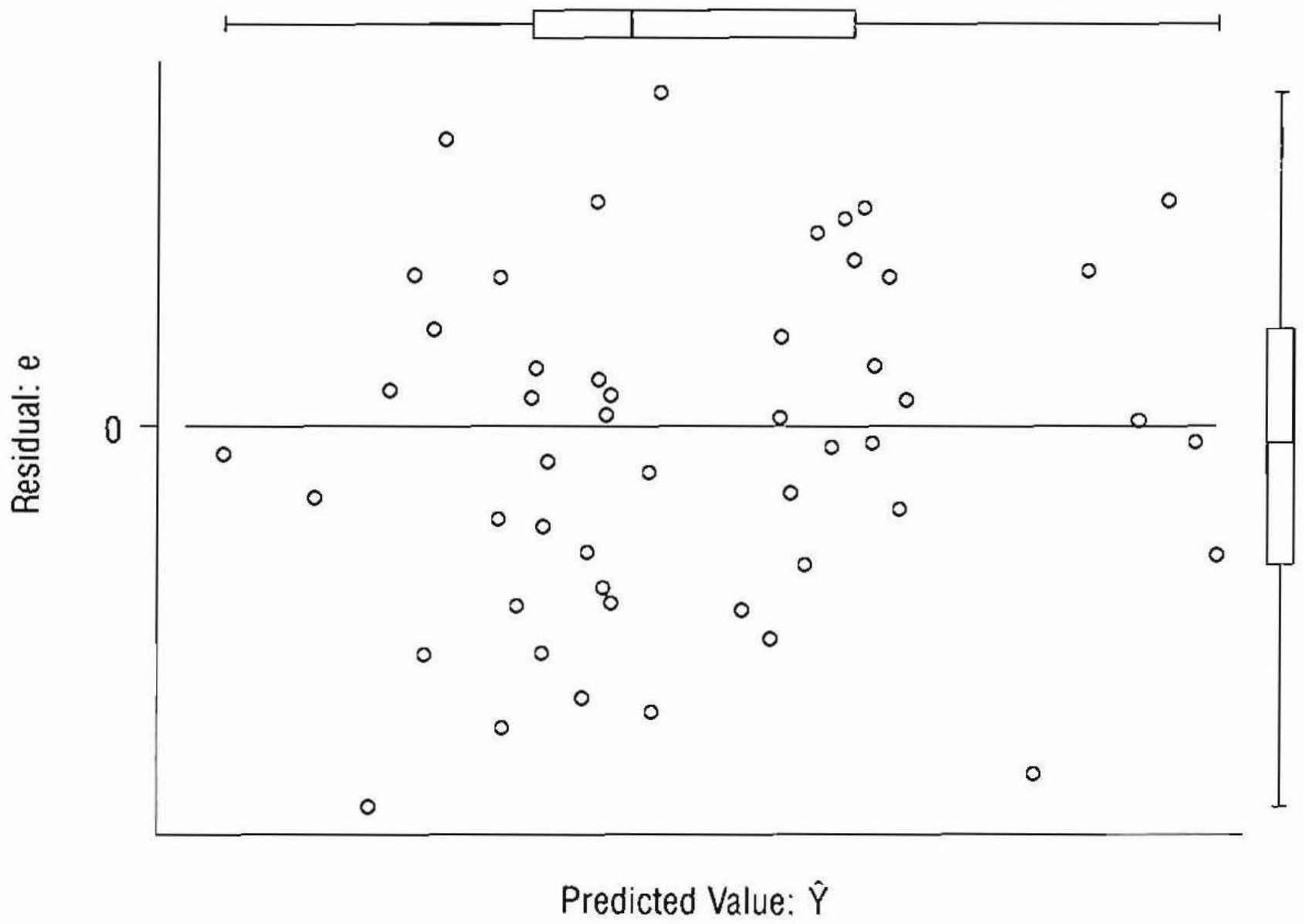
$$\Delta \hat{y} = \hat{\beta}_1 \Delta x$$

**Figura 2.5**

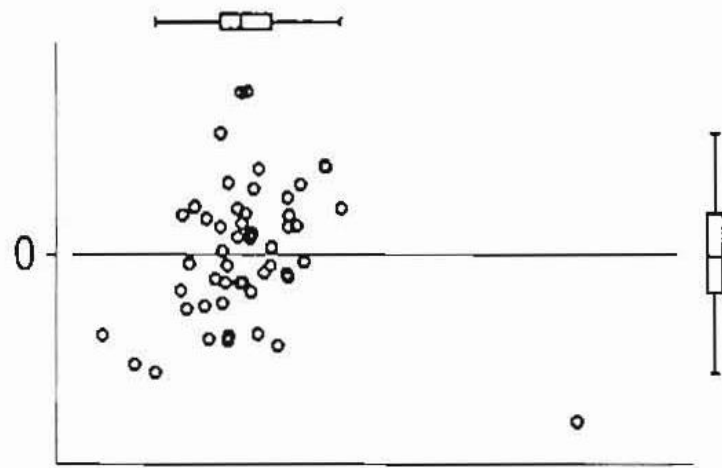
A reta de regressão de MQO  $\hat{\text{salário}} = 963,191 + 18,501 rma$  e a função de regressão populacional (desconhecida).



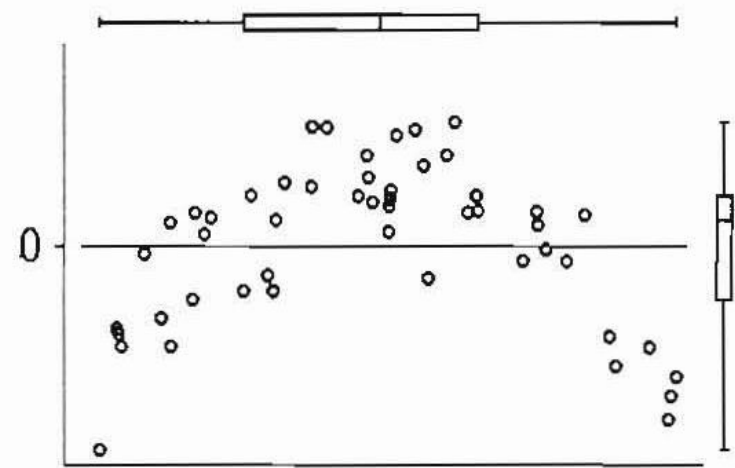




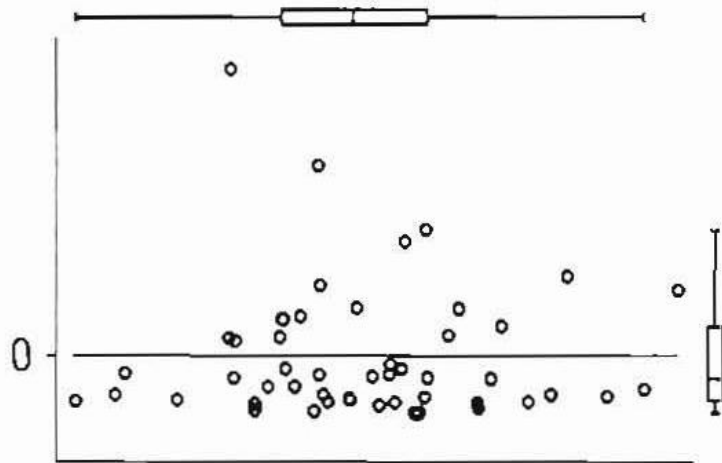
**Figure 2.10** “All clear”  $e$ -versus- $\hat{Y}$  plot (artificial data).



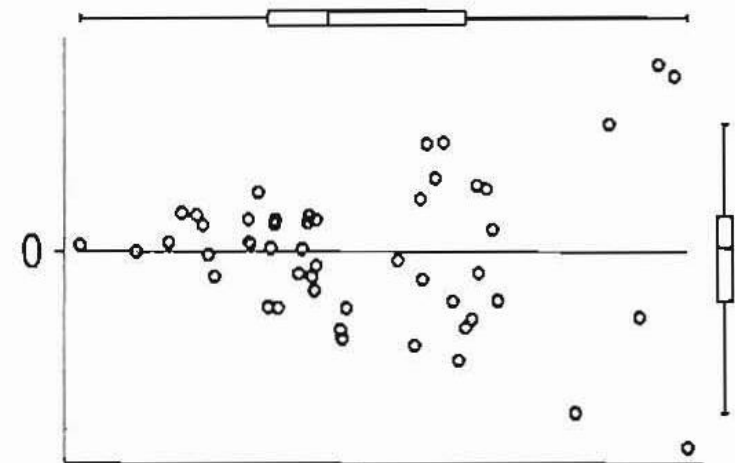
Influential Case



Curvilinear Relation



Nonnormal Residual Distribution



Heteroscedasticity

**Figure 2.11** Examples of trouble seen in  $e$ -versus- $\hat{Y}$  plots (artificial data).

# PROPRIEDADES ALGÉBRICAS DAS ESTATÍSTICAS

- A soma dos resíduos de MQO é zero, já que as estimativas de MQO de  $\hat{\beta}_0$  e  $\hat{\beta}_1$  são escolhidas para fazer com que a soma dos resíduos seja zero:

$$\sum_{i=1}^n \hat{u}_i = 0$$

- A covariância amostral entre os regressores e os resíduos de MQO é zero:

$$\frac{\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n} = \sum_{i=1}^n x_i \hat{u}_i = 0$$

- Se inserirmos a média de  $x$  no lugar de  $x_i$ , o valor estimado é a média de  $y$  (este ponto está sempre sobre a reta):

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

## SOMAS DOS QUADRADOS

- Soma dos quadrados total (SQT) é uma medida da variação amostral total em  $y_i$  (mede a dispersão dos  $y_i$  na amostra):

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

- Soma dos quadrados explicada (SQE) mede a variação amostral em :

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

- Soma dos quadrados dos resíduos (SQR) mede a variação amostral em :

$$SQR = \sum_{i=1}^n \hat{u}_i^2$$

- Variação total em  $y$  é a soma da variação explicada e da variação não-explicada:

$$SQT = SQE + SQR$$

## GRAU DE AJUSTE

- Visa mensurar o quanto bem a variável independente (x) explica a variável dependente (y).
- É um número que resume o quão bem a reta de regressão de MQO se ajusta aos dados.
- $R^2$ : razão entre a variação explicada (SQE) e a variação total (SQT).
- $R^2$ : fração da variação amostral em y que é explicada por x.

$$SQT = SQE + SQR$$

$$SQT/SQT = (SQE + SQR)/SQT$$

$$1 = SQE/SQT + SQR/SQT$$

$$SQE/SQT = 1 - SQR/SQT$$

- Usar o  $R^2$  como principal padrão de medida de sucesso de uma análise econométrica pode levar a confusões.

## MUDANÇAS DAS UNIDADES DE MEDIDA

- Ao mudar unidades de medida das variáveis dependente e/ou independente, estimativas de MQO são afetadas.
- Se a variável dependente é multiplicada pela constante  $c$  (cada valor na amostra é multiplicado por  $c$ ), então as estimativas de MQO de intercepto e de inclinação também são multiplicadas por  $c$ .
- Se a variável independente é dividida (ou multiplicada) por alguma constante diferente de zero ( $c$ ) então o coeficiente de inclinação de MQO é multiplicado (ou dividido) por  $c$ , respectivamente.
- Mudar as unidades de medida da variável independente não afeta o intercepto.
- O grau de ajuste do modelo ( $R^2$ ) não depende das unidades de medida das variáveis.

# NÃO-LINEARIDADE NA REGRESSÃO SIMPLES

- Formas funcionais populares usadas em economia podem ser incorporadas à análise de regressão.
- Até agora foram analisadas relações lineares entre as variáveis dependente e independente.
- No entanto, relações lineares não são suficientes para todas as aplicações econômicas e sociais.
- É fácil incorporar não-linearidade na análise de regressão simples.

## EXEMPLO DE NÃO-LINEARIDADE

- Para cada ano adicional de educação, há um aumento fixo no salário. Esse é o aumento tanto para o primeiro ano de educação quanto para anos mais avançados:

$$\textit{salário} = \beta_0 + \beta_1 \textit{educ} + u$$

- Suponha que o aumento percentual no salário é o mesmo, dado um ano a mais de educação formal. Um modelo que gera um efeito percentual constante é dado por:

$$\log(\textit{salário}) = \beta_0 + \beta_1 \textit{educ} + u$$

- Se  $\Delta u = 0$ , então:

$$\% \Delta \textit{salário} = (100 * \beta_1) \Delta \textit{educ}$$

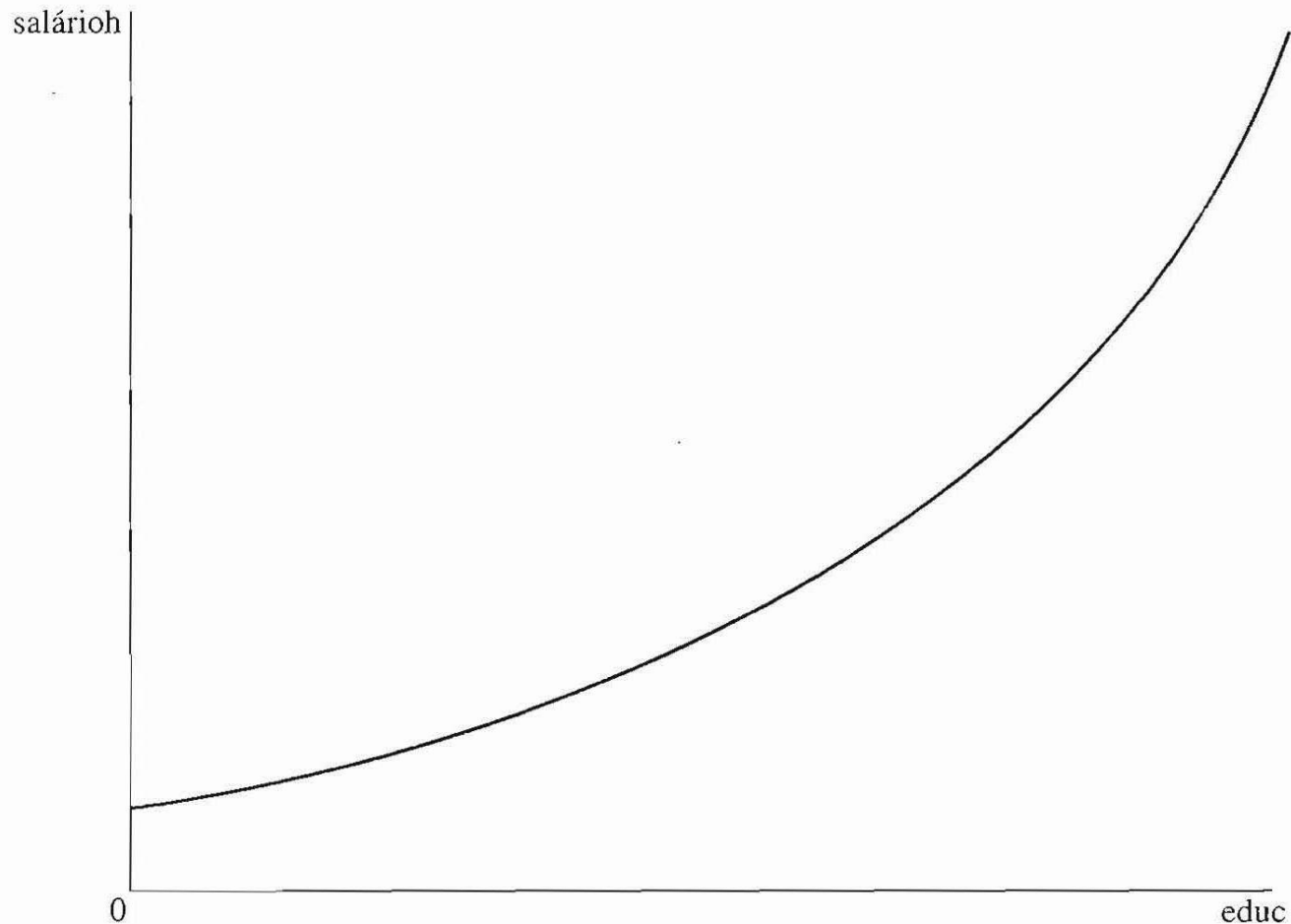
- Para cada ano adicional de educação, há um aumento de ?% sobre o salário.



- Como a variação percentual no salário é a mesma para cada ano adicional de educação, a variação no salário aumenta quando a educação formal aumenta.

**Figura 2.6**

$$\text{saláριο}_h = \exp(\beta_0 + \beta_1 \text{educ}), \text{ com } \beta_1 > 0.$$



## INTERPRETAÇÃO DOS COEFICIENTES

- O aumento de uma unidade em  $x$  aumenta  $y$  em  $\beta_1$  vezes:

$$y = \beta_0 + \beta_1 x + u$$

- O aumento de uma unidade em  $x$  aumenta  $y$  em  $(100 * \beta_1)\%$ :

$$\log(y) = \beta_0 + \beta_1 x + u$$

- O aumento de 1% em  $x$  aumenta  $y$  em  $\beta_1\%$ :

$$\log(y) = \beta_0 + \beta_1 \log(x) + u$$

- Este último é o modelo de elasticidade constante.
- Elasticidade é a razão entre o percentual de mudança em uma variável e o percentual de mudança em outra variável.

# FORMAS FUNCIONAIS ENVOLVENDO LOGARITMOS

Modelo	Variável Dependente	Variável Independente	Interpretação de $\beta_1$
nível-nível	y	x	$\Delta y = \beta_1 \Delta x$
nível-log	y	$\log(x)$	$\Delta y = (\beta_1 / 100) \% \Delta x$
log-nível	$\log(y)$	x	$\% \Delta y = (100 \beta_1) \Delta x$
log-log	$\log(y)$	$\log(x)$	$\% \Delta y = \beta_1 \% \Delta x$

# SIGNIFICADO DE REGRESSÃO LINEAR

- O modelo de regressão linear permite relações não-lineares.
- Esse modelo é linear nos parâmetros:  $\beta_0$  e  $\beta_1$ .
- Não há restrições de como  $y$  e  $x$  se relacionam com as variáveis dependente e independente originais, já que podemos utilizar: logaritmo natural, quadrado, raiz quadrada...
- A interpretação dos coeficientes depende das definições de como  $x$  e  $y$  são construídos.
- “É muito mais importante tornar-se proficiente em interpretar coeficientes do que eficiente no cálculo de fórmulas.”  
(Wooldridge, 2008: 45)

# DOCUMENTAÇÃO DO LIVRO

– UCLA Academic Technology Services:

<http://www.ats.ucla.edu>

– Introductory Econometrics: A Modern Approach  
by Jeffrey M. Wooldridge:

<http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge.html>

## DOCUMENTAÇÃO PARA EXERCÍCIO

- Vamos utilizar os dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007 para Minas Gerais.
- Os bancos de dados, questionário, livro de códigos e demais arquivos estão disponíveis no site do Consórcio de Informações Sociais (CIS), organizado pelo Núcleo de Apoio à Pesquisa sobre Democratização e Desenvolvimento da Universidade de São Paulo (NADD-USP) e pela Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais (ANPOCS):

<http://www.nadd.prp.usp.br/cis/index.aspx>