

# O Modelo de Regressão Simples

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

# Terminologia

- No modelo de regressão linear simples, onde  $y = \beta_0 + \beta_1 x + u$ , tipicamente nos referimos a  $y$  como:
  - Variável dependente, ou
  - Variável explicada, ou
  - Regressando

# Terminologia

- No modelo de regressão linear simples de  $y$  sobre  $x$ , nos referimos a  $x$  como:
  - Variável independente
  - Variável explicativa
  - Regressor
  - Variável de controle

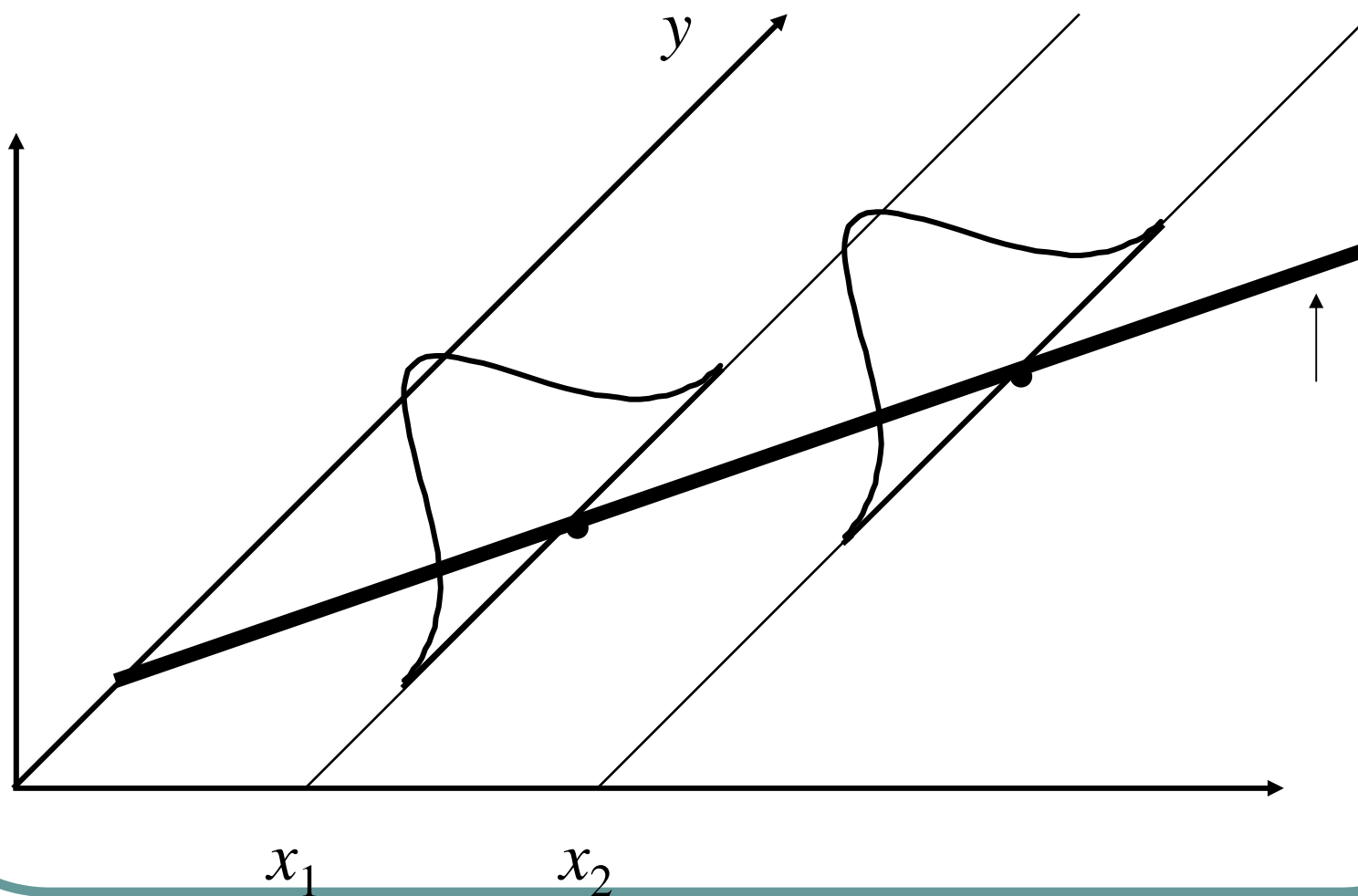
# Um Hipótese Simples

- A média de  $u$ , o termo de erro, é zero para a população. Isto é,
- $E(u) = 0$
- Esta não é uma hipótese restritiva, uma vez que podemos sempre usar  $\beta_0$  para normalizar  $E(u)$  à 0.

# Média Condicional Zero

- Precisamos fazer uma hipótese crucial sobre como  $u$  e  $x$  são relacionadas
- Desejamos que  $x$  e  $u$  não sejam relacionados de nenhuma forma. Isto é:
- $E(u|x) = E(u) = 0$ , que implica em
- $E(y|x) = \beta_0 + \beta_1 x$

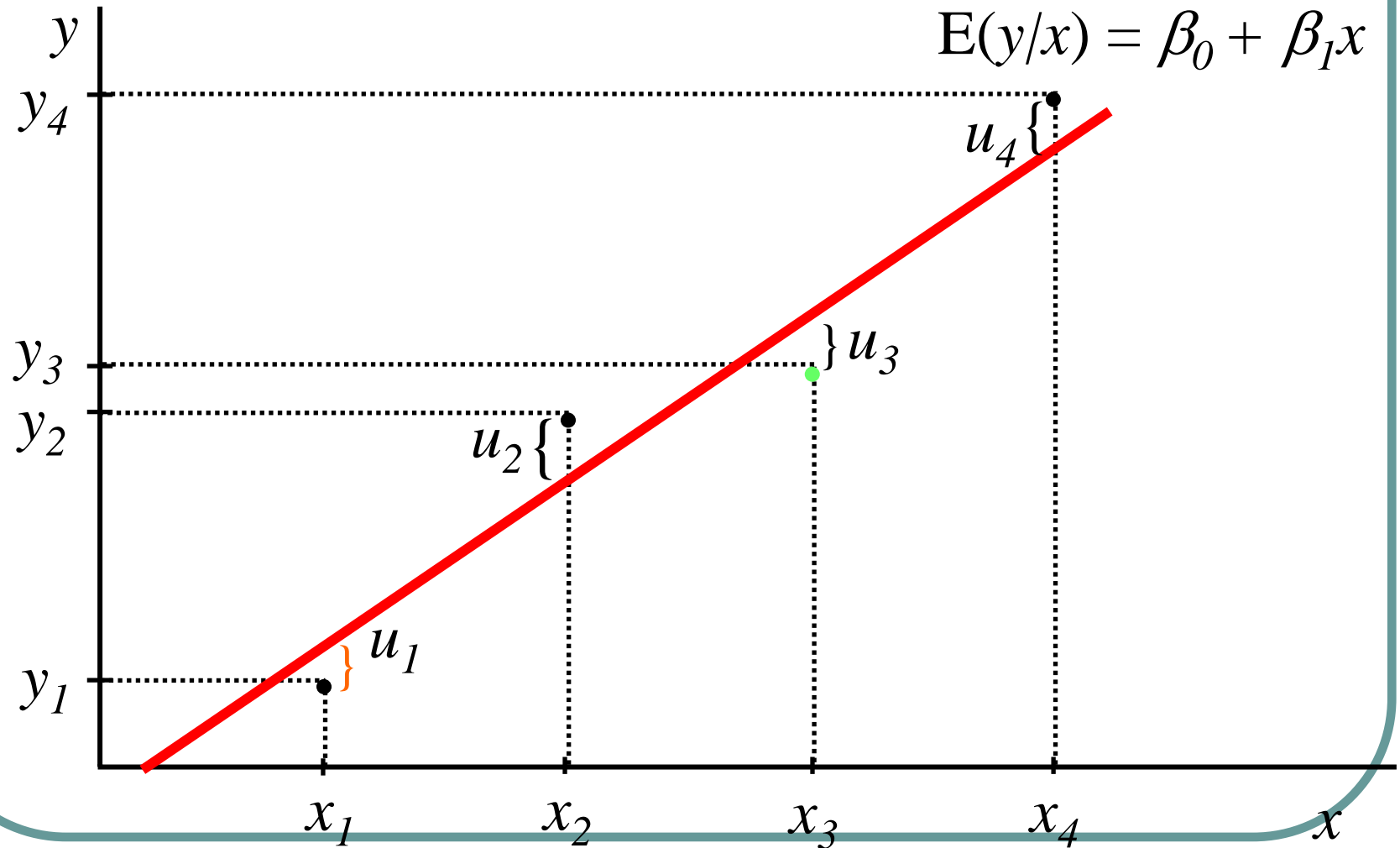
$E(y/x)$  como uma função linear de  $x$ , onde para cada  $x$  a distribuição de  $y$  seja centrada em torno de  $E(y/x)$



# Mínimos Quadrados Ordinários (OLS)

- A idéia básica da regressão é estimar os parâmetros da população a partir de uma amostra
- Faça  $\{(x_i, y_i): i = 1, \dots, n\}$  representar uma amostra aleatória de tamanho  $n$  da população
- Para cada observação na amostra, este será o caso que
- $y_i = \beta_0 + \beta_1 x_i + u_i$

# Linha de regressão da população, pontos amostrais e o termo de erro associado



# Derivando o estimador OLS

- Para derivar a estimativa OLS precisamos lembrar que a nossa hipótese principal de  $E(u|x) = E(u) = 0$  também implica que
- $Cov(x,u) = E(xu) = 0$
- Por quê? Lembre de probabilidade que  $Cov(X,Y) = E(XY) - E(X)E(Y)$

# Derivando OLS

- Podemos escrever nossas 2 restrições apenas em termos de  $x$ ,  $y$ ,  $\beta_0$  e  $\beta_1$ , dado que  $u = y - \beta_0 - \beta_1 x$
- $E(y - \beta_0 - \beta_1 x) = 0$
- $E[x(y - \beta_0 - \beta_1 x)] = 0$
- Estas são chamadas de restrições dos momentos

# Derivando OLS

- O procedimento de estimação conhecido como métodos dos momentos implica em impor as restrições dos momentos populacionais sobre os momentos amostrais
- Qual o significado? Relembre que para  $E(X)$ , a média da distribuição de uma população, um estimador amostral de  $E(X)$  é simplesmente a média aritmética da amostra

# Derivando OLS

- Desejamos escolher os valores dos parâmetros que irão garantir versões amostrais das nossas restrições de momentos sejam verdadeiras
- As versões amostrais são as seguintes:

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

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

# Derivando OLS

- Dadas as definições de uma média amostral, e as propriedades de **soma**, podemos reescrever as condições de primeira ordem como

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

ou

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

# Derivando OLS

$$\sum_{i=1}^n x_i \left( y_i - \left( \bar{y} - \hat{\beta}_1 \bar{x} \right) - \hat{\beta}_1 x_i \right) = 0$$

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

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

# Inclinação estimada do OLS é

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

dado que  $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$

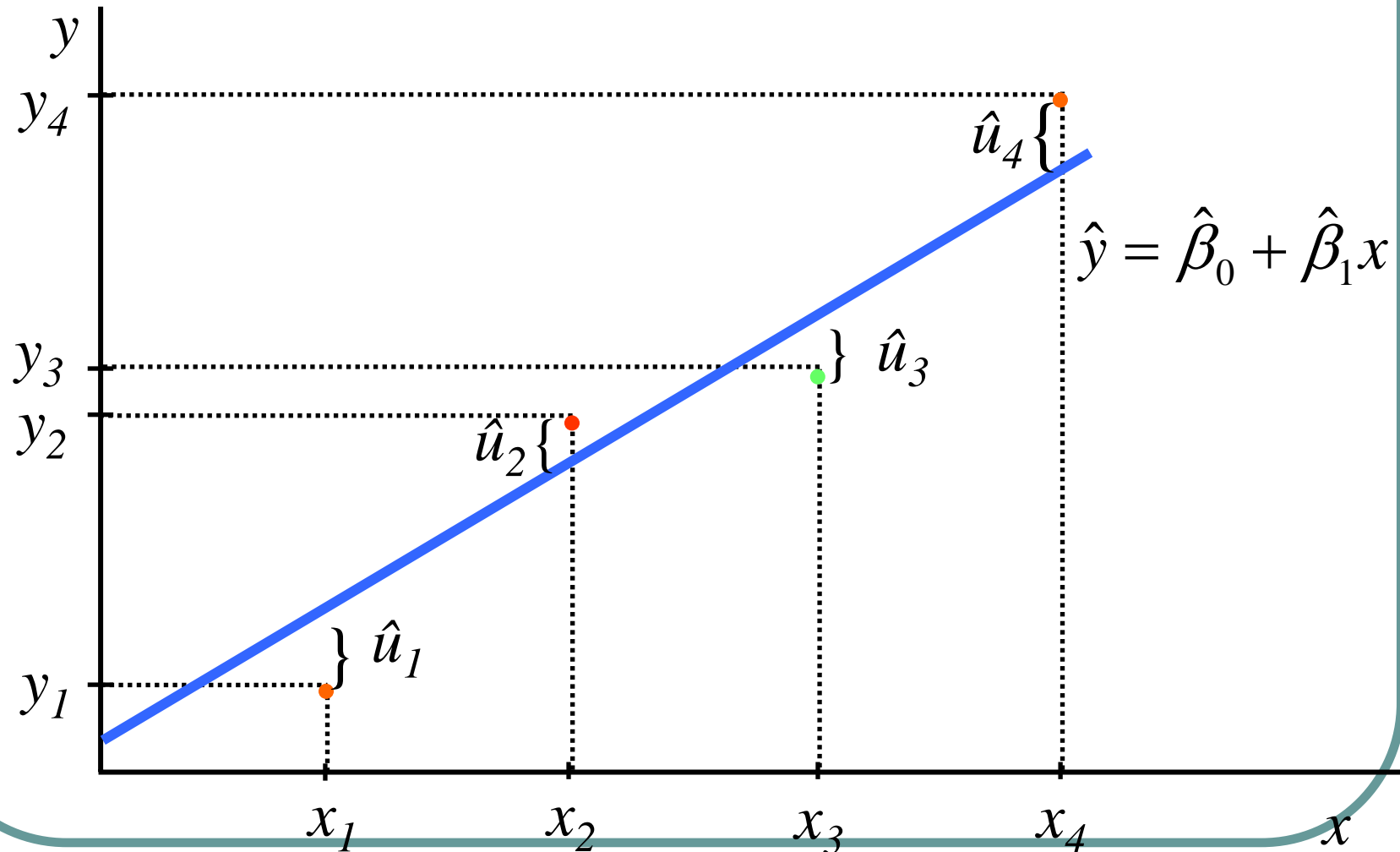
# Sumário da estimativa OLS da inclinação

- A estimativa da inclinação é a covariância amostral entre  $x$  e  $y$  dividido pela variância amostral de  $x$
- Se  $x$  e  $y$  são positivamente correlacionados então a inclinação será positiva
- Se  $x$  e  $y$  são negativamente correlacionadas, a inclinação será negativa

# Mais sobre OLS

- Intuitivamente, OLS está ajustando uma linha pelo meio dos pontos amostrais tal que a soma do quadrado do resíduo é a menor possível – assim o termo mínimos quadrados
- O resíduo,  $\hat{u}$ , é uma estimativa do termo de erro,  $u$ , e é a diferença entre a linha ajustada (função regressão amostral) e os pontos amostrais

# Reta de regressão amostral, pontos amostrais e os termos associados ao erro estimado



# Método Alternativo

- Dada a idéia intuitiva de aproximar uma reta, podemos montar um problema formal de minimização
- Isto é, desejamos escolher nossos parâmetros tal que minimizemos o seguinte:

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

# Método Alternativo

- Se desejar utilizar cálculo para resolver o problema de minimização para os dois parâmetros, obteremos as mesmas equações que obtemos com o método dos momentos antes de serem multiplicadas por  $n$ :

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

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

# Propriedades Algébricas do OLS

- A soma dos resíduos do OLS é zero
- Portanto, a média amostral do resíduo OLS também é zero
- A covariância entre os regressores e o resíduo OLS é zero
- A reta de regressão OLS sempre passa pela média da amostra

# Propriedades Algébricas

$$\sum_{i=1}^n \hat{u}_i = 0 \text{ e portanto, } \frac{\sum_{i=1}^n \hat{u}_i}{n} = 0$$

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

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

# Mais terminologia

Podemos pensar que cada observação como sendo composta de uma parte explicada e uma não-explicada

$y_i = \hat{y}_i + \hat{u}_i$  Assim definimos:

$\sum (y_i - \bar{y})^2$  como a soma total dos quadrados (SST)

$\sum (\hat{y}_i - \bar{y})^2$  soma dos quadrados explicada (SSE)

$\sum \hat{u}_i^2$  soma do quadrado dos resíduos (SSR)

Então  $SST = SSE + SSR$

# Prova que $SST = SSE + SSR$

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum \hat{u}_i^2 + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \\ &= SSR + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + SSE \\ \text{e sabemos que } \sum \hat{u}_i (\hat{y}_i - \bar{y}) &= 0\end{aligned}$$

# Grau de Ajuste

- Como pensar sobre como nossa reta de regressão amostral se ajusta aos dados amostrais?
- Podemos computar a fração da soma dos quadrados **SST** que é explicada pelo modelo, chamamos isso de R-quadrado da regressão
- $R^2 = SSE/SST = 1 - SSR/SST$

# Usando Stata para regressão OLS

- Agora que devemos a fórmula para a estimativa OLS, podemos estimá-lo
- Exemplo de regressão OLS para um modelo simples, onde  $y$  é a variável dependente e  $x$  o regressor

**regress y x**

ou

**reg y x**

# Não-viesidade do OLS

- Assuma que o modelo da população é linear nos parâmetros como

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

- Assuma que podemos usar uma amostra aleatória de tamanho  $n$ ,  $\{(x_i, y_i) : i=1, 2, \dots, n\}$ , do modelo populacional. Portanto, podemos escrever o modelo amostral como  $y_i = \beta_0 + \beta_1 x_i + u_i$
- Assuma  $E(u|x) = 0$  e portanto  $E(u_i|x_i) = 0$
- Assuma que existe variação em  $x_i$

# Não-viesidade do OLS

- Para pensar sobre não-viesidade, temos que re-escrever nosso estimador em termos do parâmetro da população
- Vamos re-escrever a fórmula do estimador OLS

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{s_x^2}, \text{ onde}$$

$$s_x^2 \equiv \sum (x_i - \bar{x})^2$$

# Não-viesidade do OLS

$$\begin{aligned}\sum (x_i - \bar{x})y_i &= \sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) = \\ &\sum (x_i - \bar{x})\beta_0 + \sum (x_i - \bar{x})\beta_1 x_i \\ &+ \sum (x_i - \bar{x})u_i = \\ &\beta_0 \sum (x_i - \bar{x}) + \beta_1 \sum (x_i - \bar{x})x_i \\ &+ \sum (x_i - \bar{x})u_i\end{aligned}$$

# Não-viesidade do OLS

$$\sum (x_i - \bar{x}) = 0,$$

$$\sum (x_i - \bar{x}) x_i = \sum (x_i - \bar{x})^2$$

então, o numerador pode ser re-escrito como

$$\beta_1 s_x^2 + \sum (x_i - \bar{x}) u_i, \text{ e portanto}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \bar{x}) u_i}{s_x^2}$$

# Não-viesidade do OLS

faça  $d_i = (x_i - \bar{x})$ , tal que

$$\hat{\beta}_1 = \beta_1 + \left( \frac{1}{s_x^2} \right) \sum d_i u_i, \text{ então}$$

$$E(\hat{\beta}_1) = \beta_1 + \left( \frac{1}{s_x^2} \right) \sum d_i E(u_i) = \beta_1$$

# Resumo

- O estimador OLS de  $\beta_1$  e  $\beta_0$  não são viesados
- Prova de viés depende de nossas 4 hipóteses – se qualquer hipótese falhar, então a estimativa OLS não é necessariamente não-viesada
- Lembre-se de não-viesidade é uma descrição do estimador – em uma dada amostra podemos estar perto ou longe do parâmetro verdadeiro

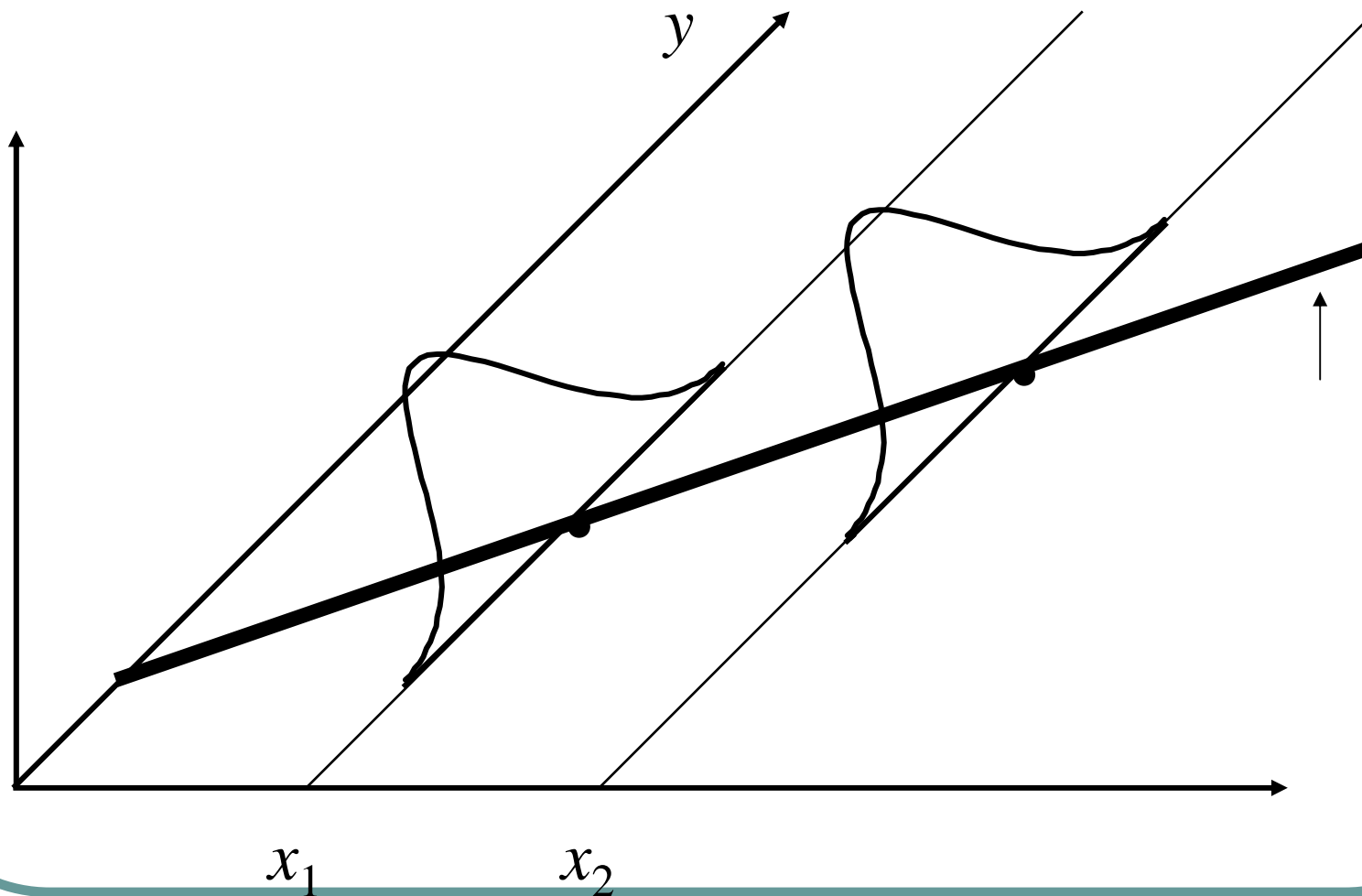
# Variância dos Estimadores OLS

- Agora sabemos que a distribuição amostral do nosso estimador é centrada em torno do parâmetro verdadeiro
- Desejamos pensar sobre o quão dispersa é esta distribuição
- É muito fácil pensar sobre esta variância sobre uma hipótese adicional
- Assuma que  $\text{Var}(u/x) = \sigma^2$   
(Hipótese de Homocedasticidade)

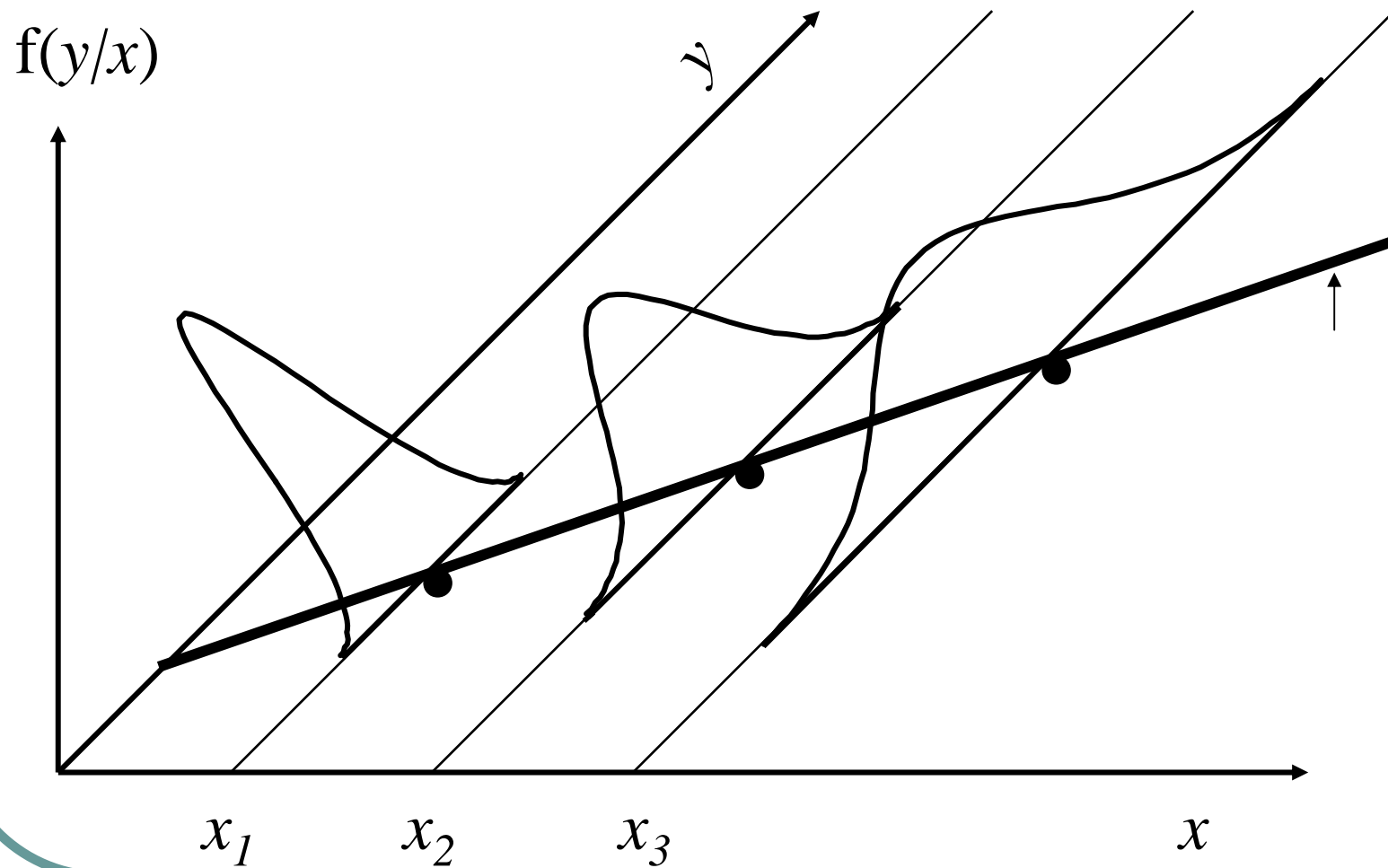
## Variância do OLS (cont)

- $\text{Var}(u|x) = E(u^2|x) - [E(u|x)]^2$
- $E(u|x) = 0$ , então  $\sigma^2 = E(u^2|x) = E(u^2) = \text{Var}(u)$
- $\sigma^2$  é também a variância incondicional, chamada de variância do erro
- $\sigma$ , é chamado de desvio-padrão do erro
- Podemos dizer:  $E(y|x) = \beta_0 + \beta_1 x$  e  $\text{Var}(y|x) = \sigma^2$

# Caso Homocedástico



# Caso Heterocedástico



# Variância do Estimador OLS

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\beta_1 + \left(\frac{1}{s_x^2}\right) \sum d_i u_i\right) = \\ &\left(\frac{1}{s_x^2}\right)^2 \text{Var}\left(\sum d_i u_i\right) = \left(\frac{1}{s_x^2}\right)^2 \sum d_i^2 \text{Var}(u_i) \\ &= \left(\frac{1}{s_x^2}\right)^2 \sum d_i^2 \sigma^2 = \sigma^2 \left(\frac{1}{s_x^2}\right)^2 \sum d_i^2 = \\ &\sigma^2 \left(\frac{1}{s_x^2}\right)^2 s_x^2 = \sigma^2 / s_x^2 = \text{Var}(\hat{\beta}_1) \end{aligned}$$

# Variância do OLS: Resumo

- Quanto maior for a variância do erro,  $\sigma^2$ , maior será a variância da inclinação estimada
- Quanto maior a variabilidade em  $x_i$ , menor será a variância da inclinação estimada
- Como um resultado, uma amostra de tamanho maior deve reduzir a variância da inclinação
- O problema é que a variância do erro é desconhecida

# Estimando a Variância do Erro

- Não sabemos a variância do erro,  $\sigma^2$ , pois não observamos os erros,  $u_i$ ,
- O que observamos são os resíduos,  $\hat{u}_i$
- Nós podemos usar os resíduos para formar uma estimativa da variância do erro

# Estimando a Variância do Erro

$$\begin{aligned}\hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\end{aligned}$$

Então, um estimador não viesado de  $\sigma^2$  é

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum \hat{u}_i^2 = SSR / (n-2)$$

# Estimando a Variância do Erro

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$  = Desvio-padrão da regressão

relembre que  $SD(\hat{\beta}) = \frac{\sigma}{s_x}$

se nós substituirmos  $\hat{\sigma}$  por  $\sigma$  então temos

o erro padrão de  $\hat{\beta}_1$ ,

$$SE(\hat{\beta}_1) = \hat{\sigma} / \left( \sum (x_i - \bar{x})^2 \right)^{1/2}$$