Regressão Linear Simples

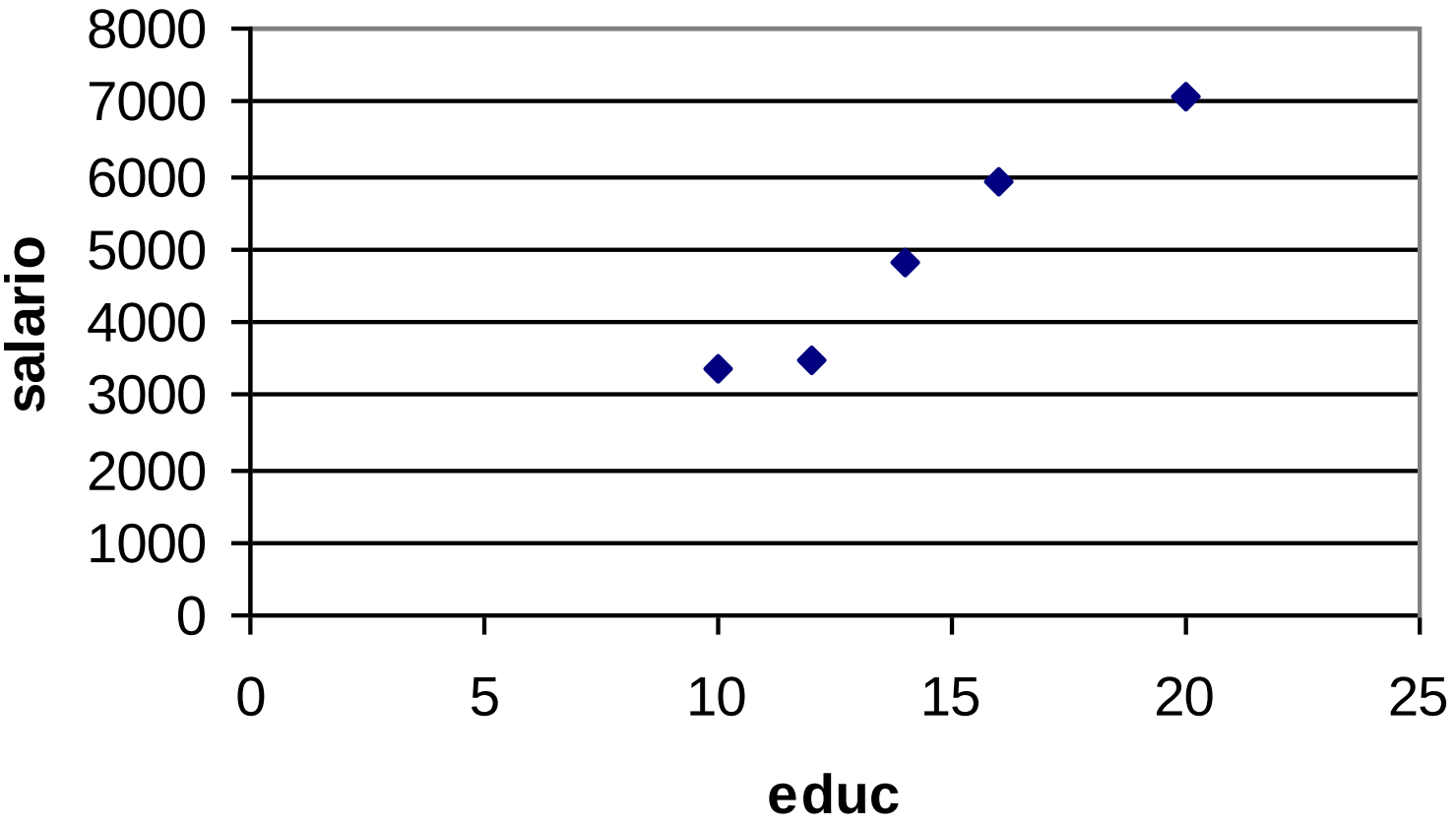
$$y = \beta_0 + \beta_1 x + u$$

Baseado em Woodridge

Exemplos

- Gasto em viagem em função do número de dias;
- Peso e altura
- Altura do filho e do pai (Galton)
- Número de anúncios e número de páginas
- Idade e Altura de árvore

Exemplo


8000
7000
6000
5000
4000
3000
2000
1000
0
salario
0
5
10
15
20
25
educ

Terminologia

- Em regressão linear simples,

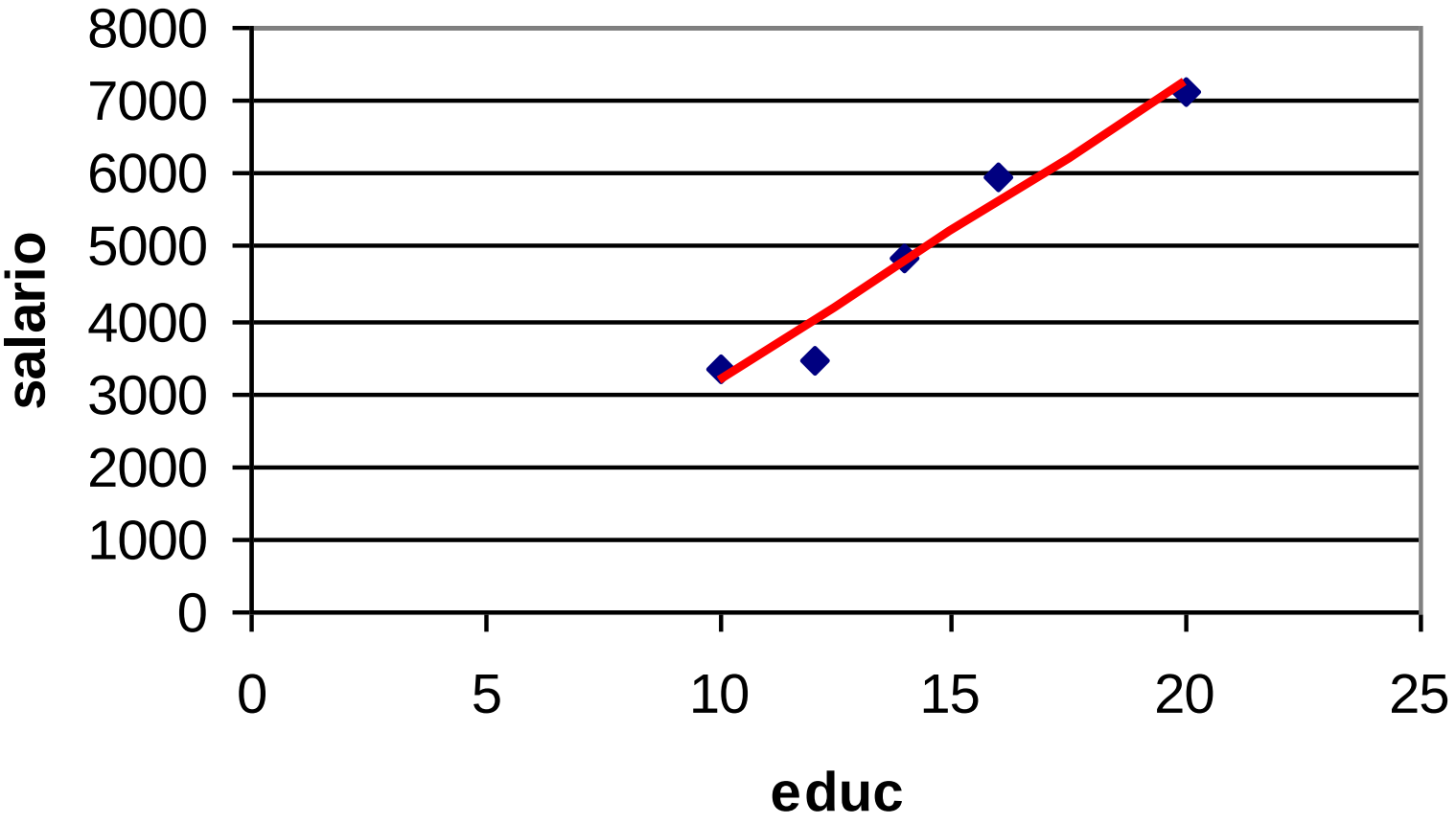
  $y = \beta_0 + \beta_1 x + u$, chamamos

  - **y** de Variável Resposta, Dependente, Regressando ou Endógena, Desfecho (salário)
  - **x** de Variável Explicativa, Independente, Regressora ou Exógena, ou ainda, Covariável e Variável de Controle, Exposição (educ= escolaridade)

Suposição

- A média de *u*, o erro, na população é 0.
- E(*u*) = 0
- Interpretação: Em média, o erro é zero.

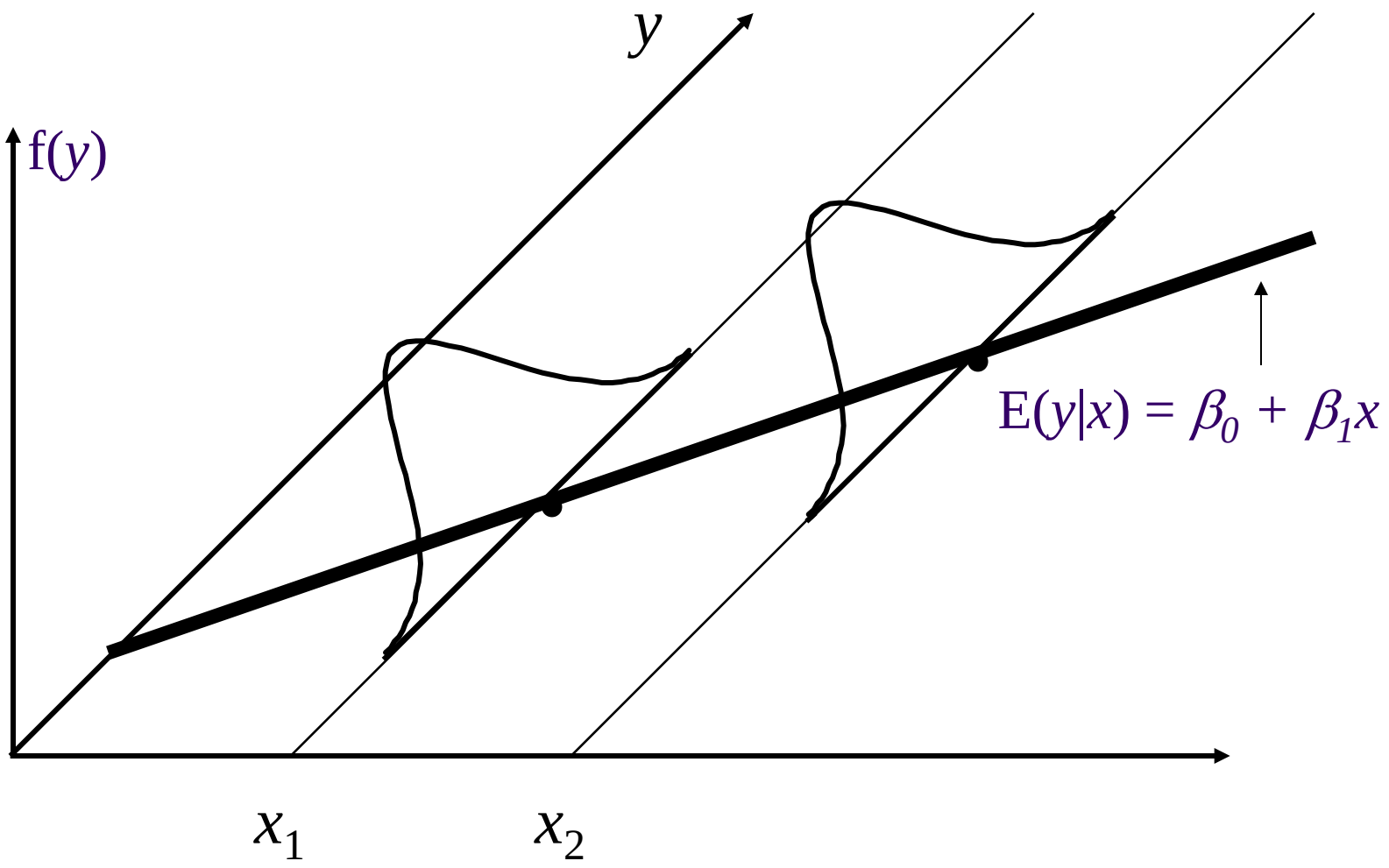
Exemplo


8000
7000
6000
5000
4000
3000
2000
1000
0
salario
0
5
10
15
20
25
educ

Média Condicional Zero

- Suposição sobre a relação entre u e x: Conhecer x não nos dá informação sobre u.
- E(u|x) = E(u) = 0, implicando
- E(y|x) = $\beta_0 + \beta_1 x$

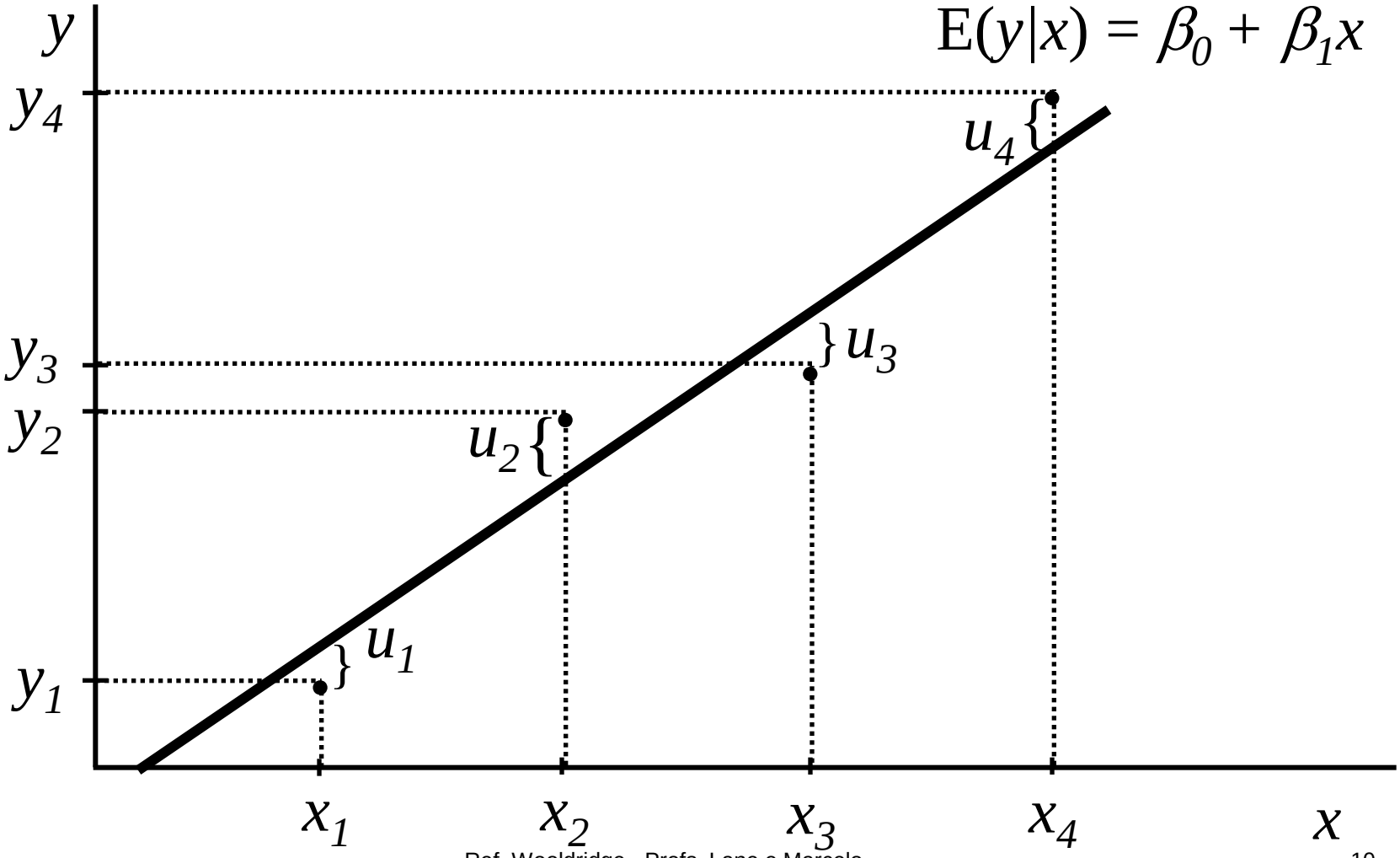
E($y|x$) é função linear de x, onde para cada x a distribuição de y é centrada em E($y|x$)



Mínimos Quadrados Ordinários

- Idéia: Estimar os parâmetros desconhecidos, β_0 *e* β_1 , a partir dos dados de uma amostra (Gauss 1794)
- Seja $\{(x_i,y_i): i=1, \ldots,n\}$ uma amostra aleatória de tamanho n retirada da população
- Para cada observação:
- $y_i = \beta_0 + \beta_1 x_i + u_i$

Reta de regressão, valores observados e erros



Mínimos Quadrados Ordinários

- A idéia intuitiva é a de ajustar uma reta que passe entre os pontos, o que cai em um problema de minimização
- Queremos os valores para os parâmetros que minimizem:

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

MQO

- Derivando com relação a β_0 e β_1, e igualando a zero obtemos:

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

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

MQO

- Usando a primeira equação, temos que:

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

ou

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

MQO – 2a equação

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

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

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

Inclinação MQO

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

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

Propriedades do MQO

- A soma e a média dos resíduos é zero
- A covariância entre os regressores, x, e os resíduos é zero
- A reta de regressão MQO sempre passa pela média da amostra

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

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

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

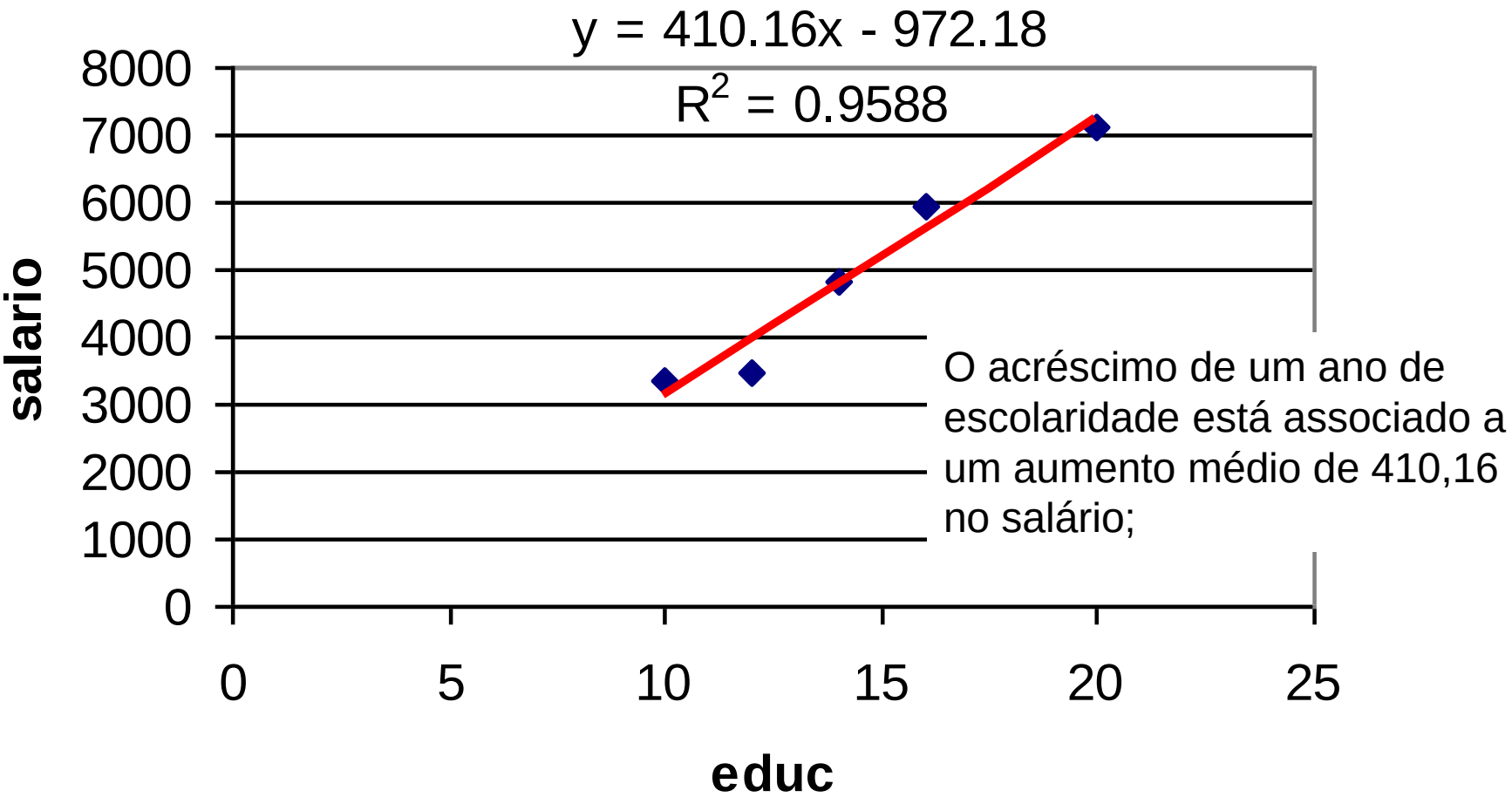
Inclinação

- A estimativa da inclinação é a covariância amostral entre x e y dividido pela variância amostral de x.
- Se x e y estão positivamente correlacionados, a inclinação será positiva
- Se x e y estão negativamente correlacionados, a inclinação será negativa
- Precisamos de variação em x na amostra

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

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

Vamos interpretar!



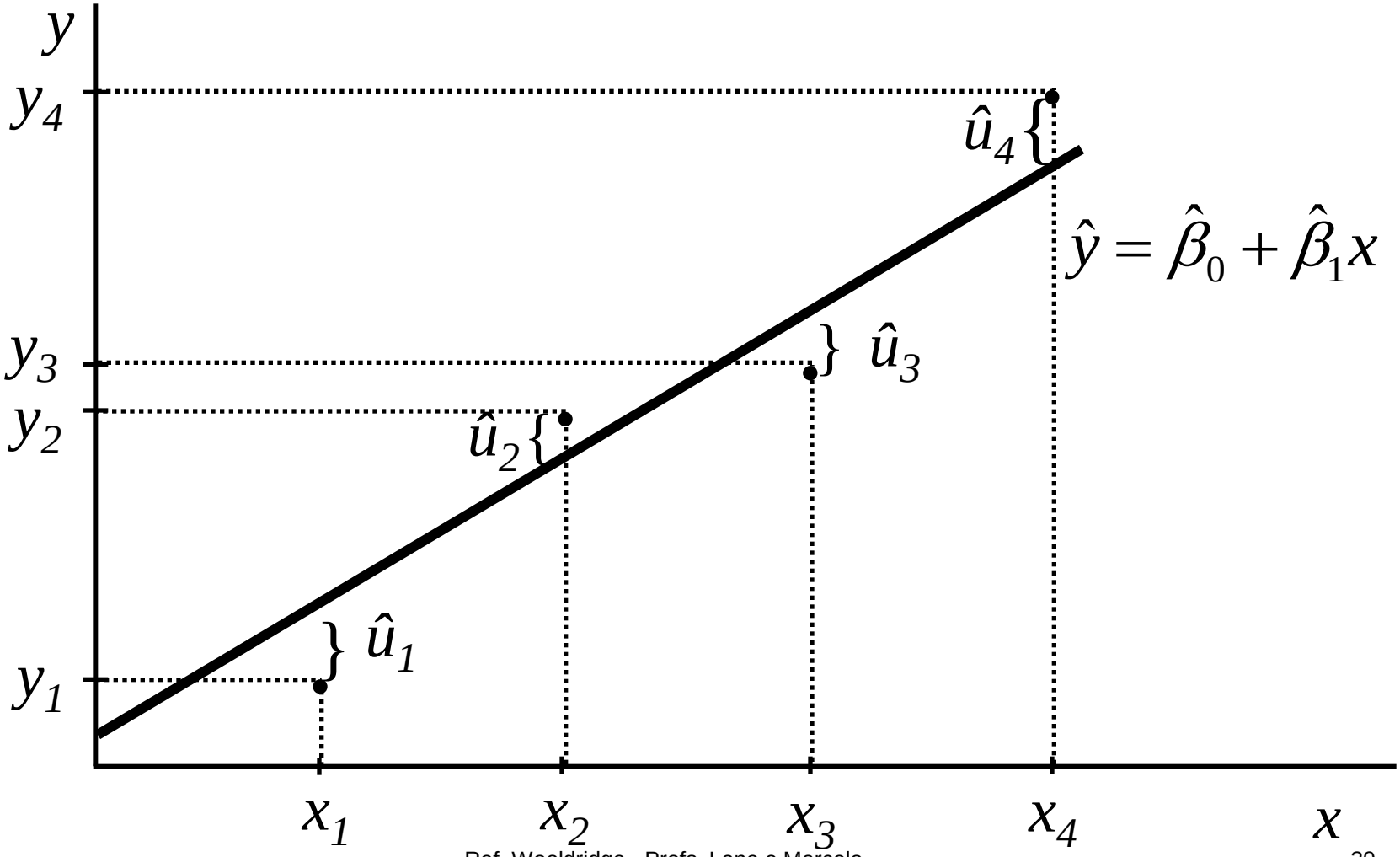
O acréscimo de um ano de escolaridade está associado a um aumento médio de 410,16 no salário;

Método dos Momentos

- 2 restrições de momentos:
- $E(y - \beta_0 - \beta_1 x) = 0$
- $E[x(y - \beta_0 - \beta_1 x)] = 0$

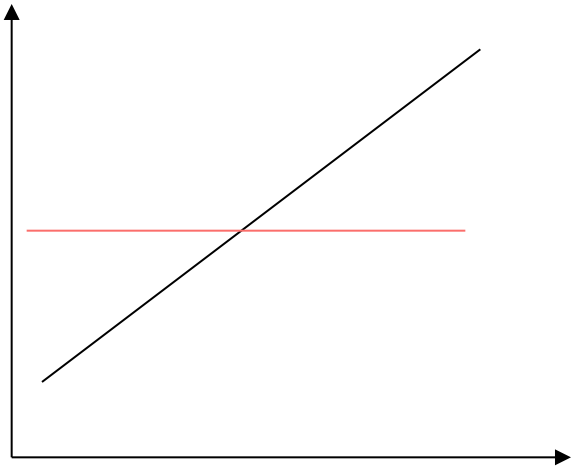
- Obteria os estimadores iguais aos de MQO

Valores observados e ajustados e resíduos

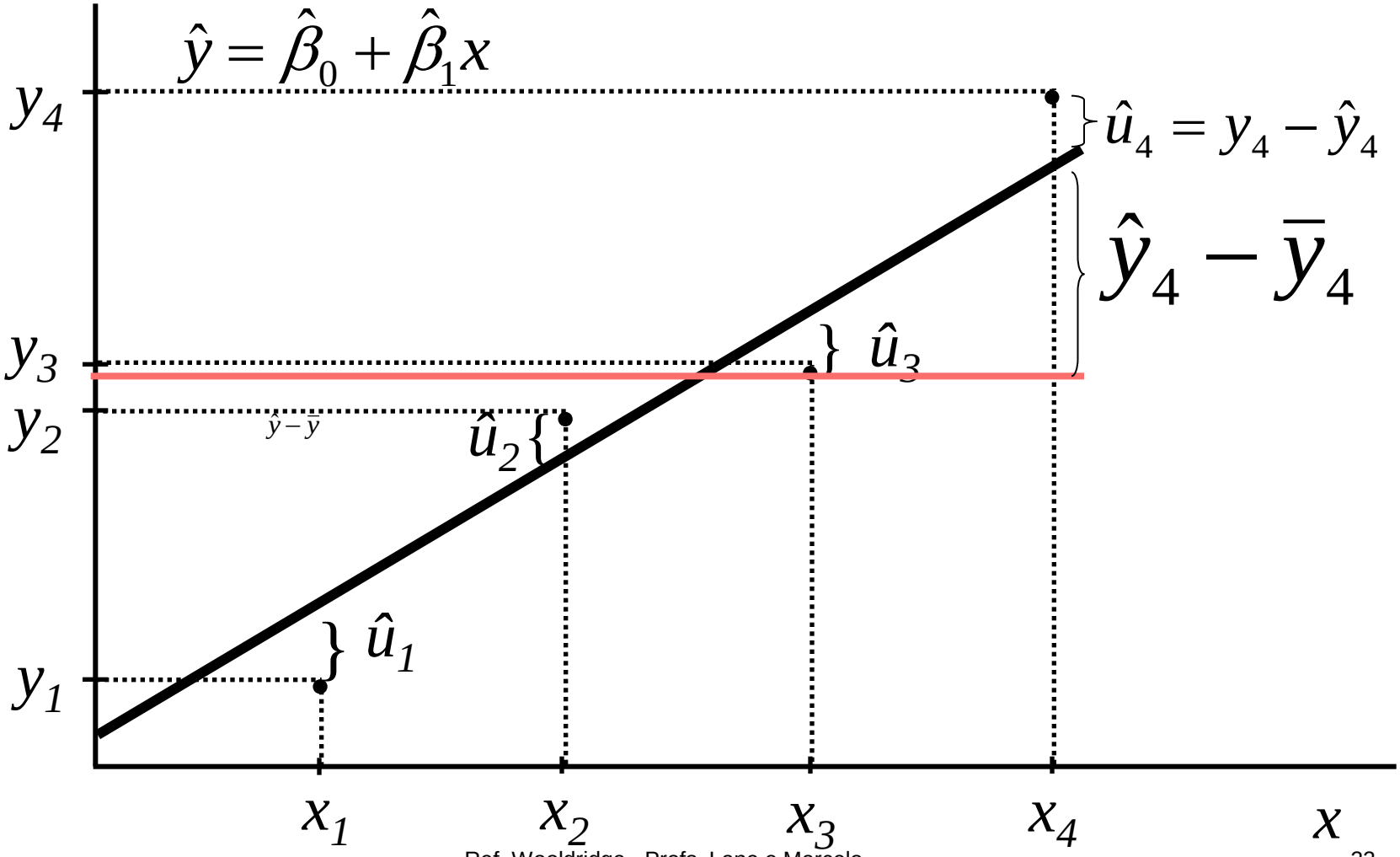


Observação

- Se a reta é inclinada, grande parte da variabilidade de y é explicada por x. Se a inclinação for nula, y é bem explicado por $\bar{y}$



Variabilidades



Variabilidades

$$\sum (y_i - \bar{y})^2 = \text{SST} = \text{Variabilid ade Total}$$

$$\sum (y_i - \hat{y}_i)^2 = \text{SSR} = \text{Variabilid ade do Resíduo}$$

$$\sum (\hat{y}_i - \bar{y})^2 = \text{SSE} = \text{Variabilid ade Explicada}$$

SST = SSE + SSR

$$\sum (y_i - \bar{y})^2 = \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2$$

$$= \sum \hat{u}_i^2 + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2$$

$$= \text{SSR} + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \text{SSE}$$

$$= \text{SSR} + \text{SSE},$$

$$\sum \hat{u}_i (\hat{y}_i - \bar{y}) = 0$$

Qualidade do ajuste

- Como medir a qualidade do ajuste do modelo aos dados amostrais?

- Podemos calcular a fração da soma de quadrados total (SST) que é explicada pelo modelo, e chamamos isso de R-quadrado da regressão

- $R^2 = SSE/SST = 1 – SSR/SST$

Regressão usando Excel e R

- No Excel, deve-se usar o Ferramentas > Análise de Dados.
- No R o comando lm

Suposições do modelo

1. Assumindo que o modelo populacional é linear nos parâmetros, $y = \beta_0 + \beta_1 x + u$ e que usamos amostra de tamanho n, $\{(x_i, y_i): i=1, 2, \ldots, n\}$ temos $y_i = \beta_0 + \beta_1 x_i + u_i$
2. $E(u|x) = 0$ e portanto $E(u_i|x_i) = 0$
3. Há variação em x_i

Estimadores não viesados

- Escrevendo os parâmetros em função da variável aleatória $y_i = \beta_0 + \beta_1 x_i + u_i$

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

Estimador Não Viesado

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

$$\sum (x_i - \bar{x}) \beta_0 + \sum (x_i - \bar{x}) \beta_1 x_i$$

$$+ \sum (x_i - \bar{x}) u_i =$$

$$\beta_0 \sum (x_i - \bar{x}) + \beta_1 \sum (x_i - \bar{x}) x_i$$

$$+ \sum (x_i - \bar{x}) u_i$$

Estimador Não Viesado

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

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

então, o numerador pode ser escrito como

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

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

Estimador Não Viesado

Seja $d_i = (x_i - \bar{x})$, então

$$\hat{\beta}_i = \beta_1 + \left(1 / s_x^2 \right) \sum d_i u_i, \text{ e}$$

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

Estimador Não Viesado

- Os estimadores MQO de β_1 e β_0 são não viesados
- A prova depende das 3 suposições – se alguma falha, então os estimadores não são necessariamente não viesados
- Lembre que não viesado é descrição do estimador – em uma dada amostra podemos (a estimativa) estar longe ou perto do valor verdadeiro do parâmetro

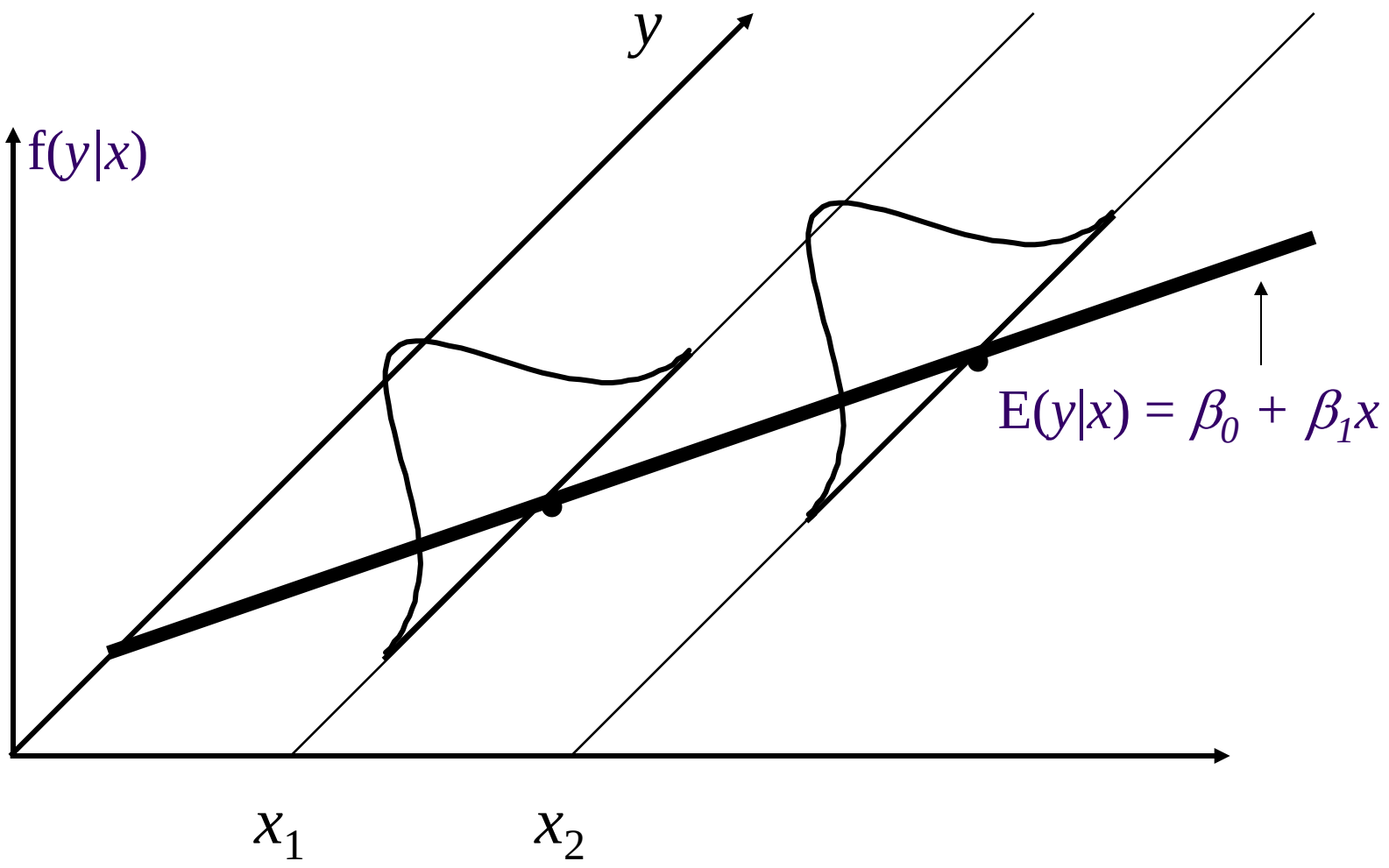
Variância dos Estimadores MQO

- Agora sabemos que a distribuição do nosso estimador está centrada em torno do verdadeiro parâmetro
- Queremos saber quão dispersa essa distribuição é
- Mais fácil pensar sobre essa variância, supondo também que
- Suposição: Var($u|x$) = σ^2 (Homocedasticidade)

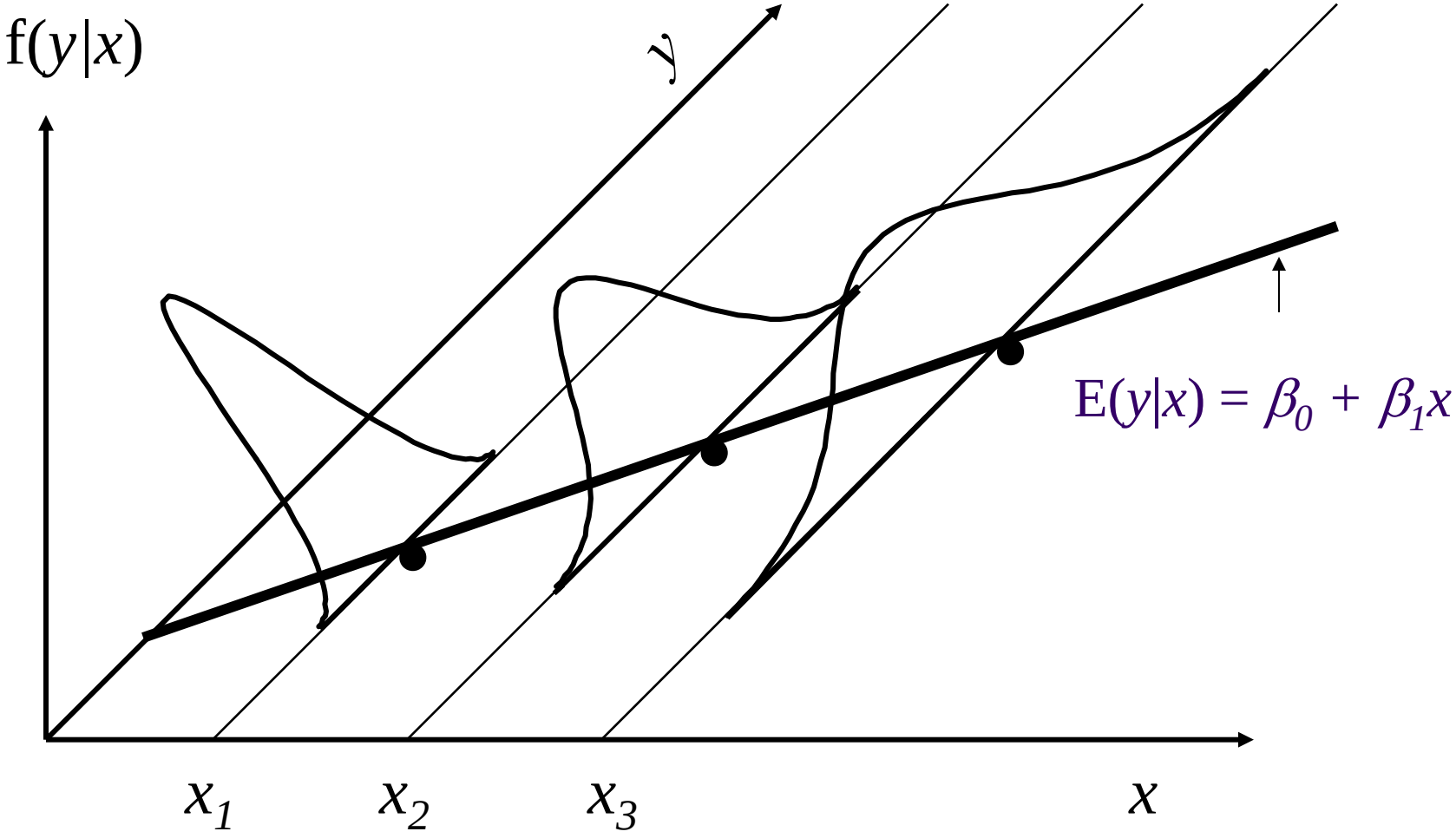
Variância

- Var($u|x$) = E($u^2|x$)-[E($u|x$)]^2
- E($u|x$) = 0, então σ^2 = E($u^2|x$) = E(u^2) = Var(u)
- Então σ^2 é também a variância incondicional, chamada de variância do erro
- σ, a raiz quadrada da variância do erro é chamada de desvio padrão do erro
- Logo: E($y|x$)=$\beta_0 + \beta_1 x$ e Var($y|x$) = σ^2

Caso Homocedástico



Caso Heterocedástico



Variância (cont)

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

$$\left(1/s_x^2\right)^2 Var\left(\sum d_i u_i\right) = \left(1/s_x^2\right)^2 \sum d_i^2 Var\left(u_i\right)$$

$$= \left(1/s_x^2\right)^2 \sum d_i^2 \sigma^2 = \sigma^2 \left(1/s_x^2\right)^2 \sum d_i^2 =$$

$$\sigma^2 \left(1/s_x^2\right)^2 s_x^2 = \sigma^2 / s_x^2 = Var\left(\hat{\beta}_1\right)$$

Resumo Variância

- Quanto maior a variância do erro, σ^2, maior a variância do estimador da inclinação
- Quanto maior a variabilidade do x_i, menor a variância do estimador da inclinação
- Como um resultado, uma amostra maior deveria diminuir a variância do estimador da inclinação
- Problema que a variância do erro é desconhecida

Estimador da variância do erro

- Não conhecemos quanto é a variância do erro, σ^2, porque não observamos os erros, u_i

- O que observamos são os resíduos, $\hat{u}_i$

- Podemos usar os resíduos para construir estimador da variância do erro

Estimador da variância do erro (cont)

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$
$$= \left(\beta_0 + \beta_1 x_i + u_i\right) - \hat{\beta}_0 - \hat{\beta}_1 x_i$$
$$= u_i - \left(\hat{\beta}_0 - \beta_0\right) - \left(\hat{\beta}_1 - \beta_1\right)$$

Um estimador não viesado de σ^2 é

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

Estimador da variância do erro (cont)

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \text{Erro padrão do erro da regressão}$$

$$\text{Lembrando que } \text{DP}\left(\hat{\beta}\right) = \sigma / s_x$$

$$\text{se substituir mos } \hat{\sigma} \text{ por } \sigma \text{ temos}$$

$$\text{o erro padrão de } \hat{\beta}_1 \text{ ,}$$

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

Teorema Gauss Markov

1. Modelo populacional é linear nos parâmetros, $y = \beta_0 + \beta_1 x + u$

   Temos amostra de tamanho n, $\{(x_i, y_i): i=1, 2, \ldots, n\}$, do modelo populacional. Portanto, escrevemos o modelo para a amostra $y_i = \beta_0 + \beta_1 x_i + u_i$
2. $E(u|x) = 0$ and portanto $E(u_i|x_i) = 0$
3. Há variação em x_i
4. $Var(u|x) = \sigma^2$ (Homocedasticidade)

Teorema Gauss Markov

- Dadas as 4 Suposições Gauss-Markov, os estimadores MQO são “BLUE”
- Best
- Linear
- Unbiased
- Estimator
- Melhor Estimador Não Viesado Linear
- Por que melhor? Menor variância

Predição

- A predição para a variável resposta é

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

Um estimador não viesado de σ^2 é

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

Estimador da média pop de y

O estimador regressão para μ_y é

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

Sua variância é

$$\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(\mu_x - \bar{x})^2}{\sum_{i \in s} (\mu_x - \bar{x})^2} \right] (1 - f)$$

- Lohrs

Referências

- Wooldridge. Introduction to econometrics.

- Lohrs. Design Sampling.