Regressão Linear Simples – Aula 06b

Statistics for Business and Economics 7 edi..o, by Paul Newbold , William Carlson , Betty Thorne (cap. Simple Regression)

Bussab e Morettin (Cap 16 Regress.o Linear Simples)

Statistics for Economics, Accounting and Business Studies, cap.tulo 7, Barrow (Correlation and Simple Regression)

Marislei Nishijima

Visão geral dos modelos lineares

- Uma equação pode ser adequada para mostrar a melhor relação linear entre duas variáveis:

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

Sendo Y a variável dependente

X a variável independente

β_0 é o intercepto em Y

β_1 a inclinação

Regressão de Mínimos quadrados

- As estimativas para os coeficientes β_0 e β_1 são encontradas usando uma técnica de regressão de mínimos quadrados
- A linha de regressão de mínimos quadrados, com base em dados de amostra, é

$$\hat{y} = b_0 + b_1 x$$

- Sendo b_1 a inclinação e b_0 o intercepto em y:

$$b_1 = \frac{Cov(x,y)}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Introdução à Análise de Regressão

- A análise de regressão é usada para:
- Prever o valor de uma variável dependente com base no valor de pelo menos uma variável independente
- Explicar o impacto das mudanças em uma variável independente na variável dependente

- Variável dependente: a variável que desejamos explicar
- (também chamada de variável endógena)
- Variável independente: a variável usada para explicar a variável dependente
- (também chamada de variável exógena)

Linear Regression Model

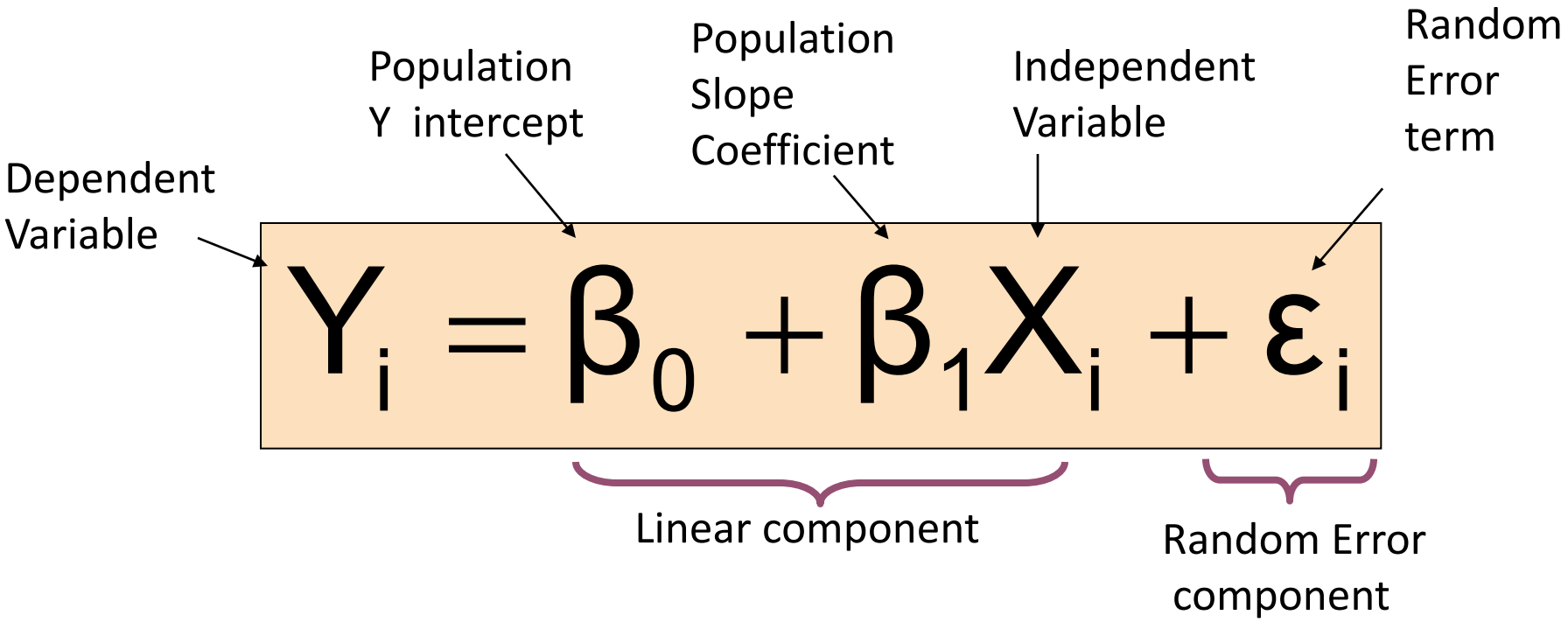
- A relação entre X e Y é descrita por uma função linear
- As mudanças em Y são consideradas causadas por mudanças em X
- Modelo de equação de regressão linear populacional

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

- Onde β_0 and β_1 são os coeficientes do modelo populacional e ε é um termo de erro aleatório.

Modelo de regressão linear simples

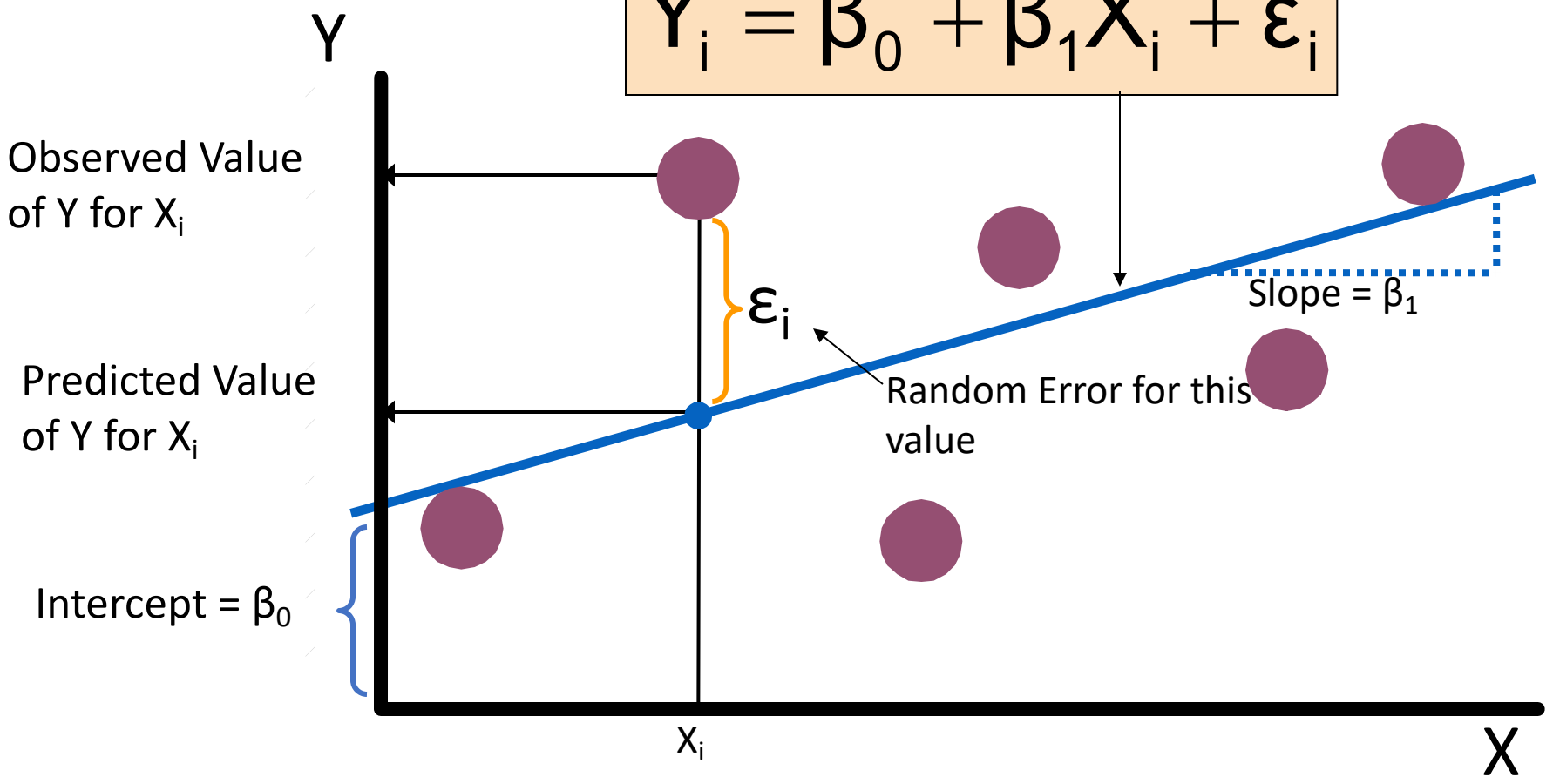
O modelo de regressão populacional:



Modelo de regressão linear simples

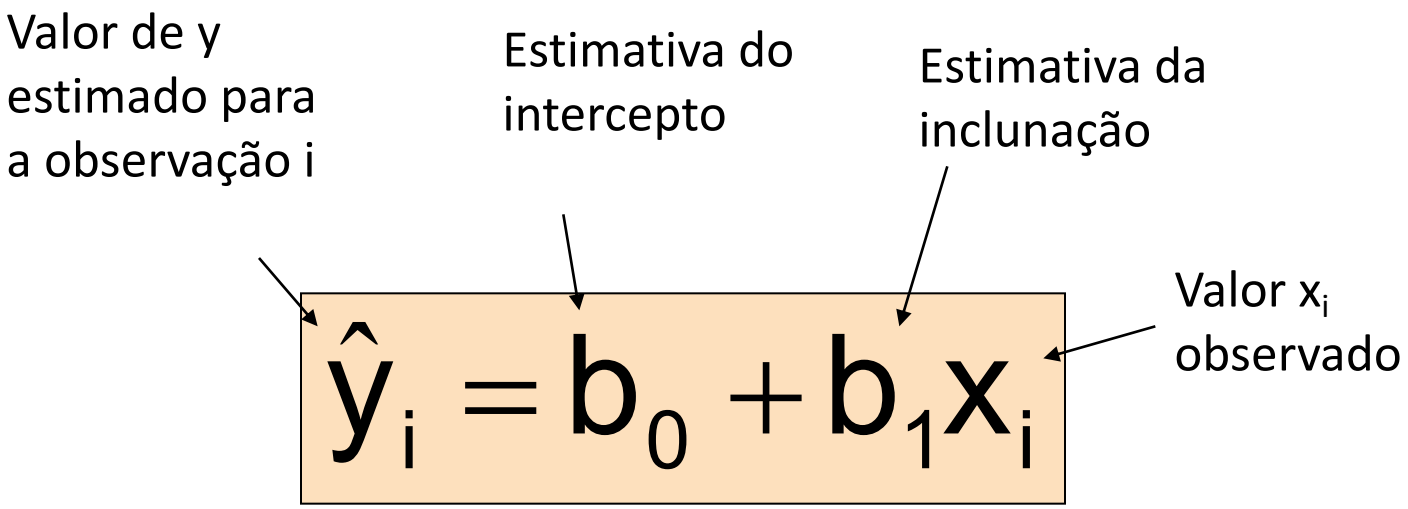
(continued)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Equação do Modelo de regressão linear simples

A equação de regressão linear simples fornece uma estimativa da linha de regressão da população



Os termos de erro aleatório individuais e_i tem media zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Estimadores de MQ

- b_0 e b_1 são obtidos encontrando oas valores de b_0 e b_1 que minimizam a soma dos quadrados das diferenças entre y e $\hat{y}$:

- $\min SQE = \min \sum e_i^2$
- $= \min \sum (y_i - \hat{y}_i)^2$
- $= \min \sum (y_i - b_0 + b_1 x_i)^2$

Cálculo Diferencial é usado para obter os estimadores dos coeficiente b_0 e b_1 que minimizam o erro ao quadrado

Estimadores de MQ

(cont.)

- O estimador do coeficiente de inclinação é:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2} = r_{xy}\frac{s_y}{s_x}$$

- E do intercepto é:

$$b_0 = \bar{y} - b_1\bar{x}$$

- A reta de regressão sempre passa por $\bar{x}$, $\bar{y}$

Obtendo as estimativas da equação de MQ

- Os coeficiente b_0 e b_1 , e outros resultados da regressão serão calculados por excel ou stata

Pressupostos do modelo de regressão linear

- A verdadeira forma da relação é linear (Y é uma função linear de X, mais o erro aleatório)
- Os termos de erro, εi são independentes dos valores x
- Os termos de erro são variáveis aleatórias com média 0 e variância constante, σ2. (a propriedade de variância constante é chamada de homocedasticidade)

$$E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{for } (i = 1, \ldots, n)$$

- Os termos de erro aleatório, εi, não estão correlacionados uns com os outros, de modo que

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{for all } i \neq j$$

Interpretação da inclinação e do intercepto

$$\hat{y}_i = b_0 + b_1 x_i$$

- b_0 é o valor médio estimado de y quando o valor de x for zero (se x = 0 estiver na faixa dos valores x observados)

- b_1 é a mudança estimada no valor médio de y como resultado de uma mudança de uma unidade em x

Exemplo de regressão linear simples

- Um corretor de imóveis deseja examinar a relação entre o preço de venda de uma casa e seu tamanho (medido em metros quadrados)

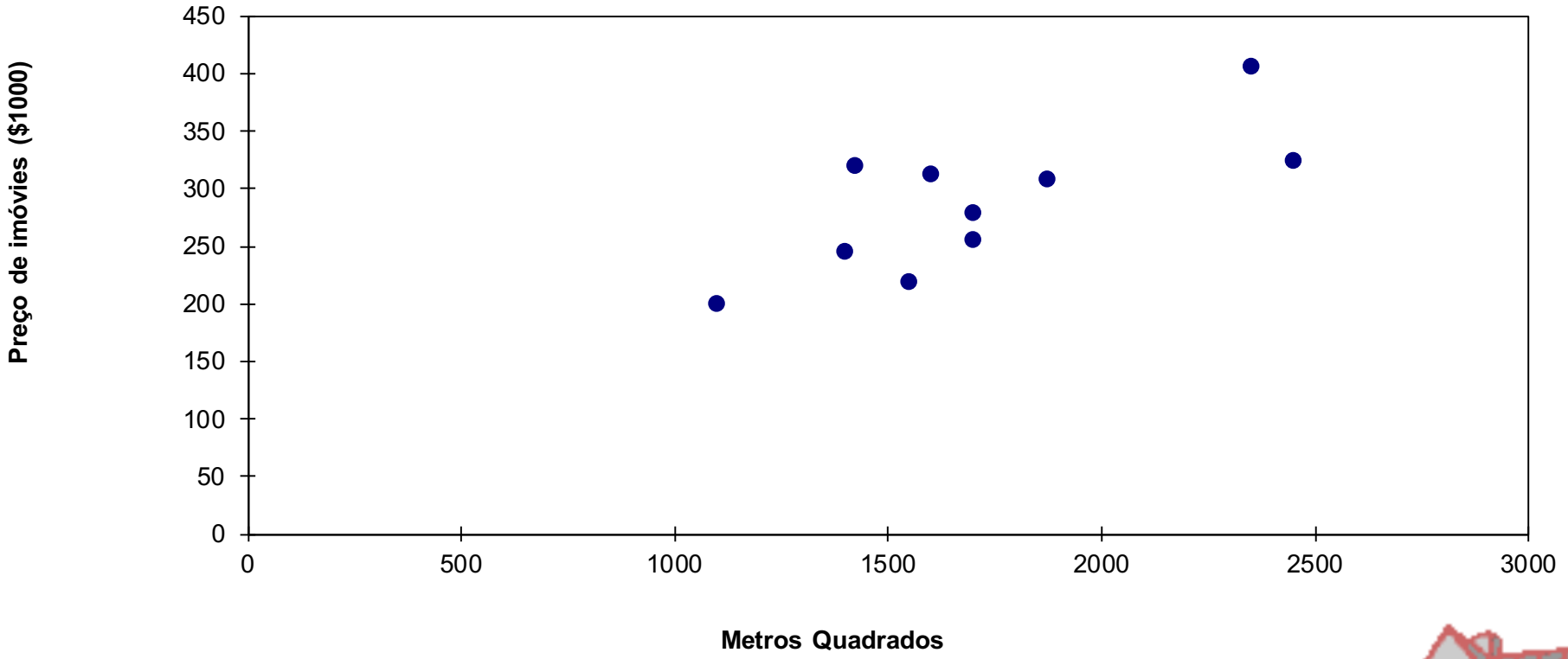
- Uma amostra aleatória de 10 casas é selecionada
- Variável dependente (Y) = preço da casa em $ 1000s
- Variável independente (X) = metros quadrados

Amostra para o medelo de preços de imóveis

Preços de imóvies em $1000 (Y)	Metros quadrados (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

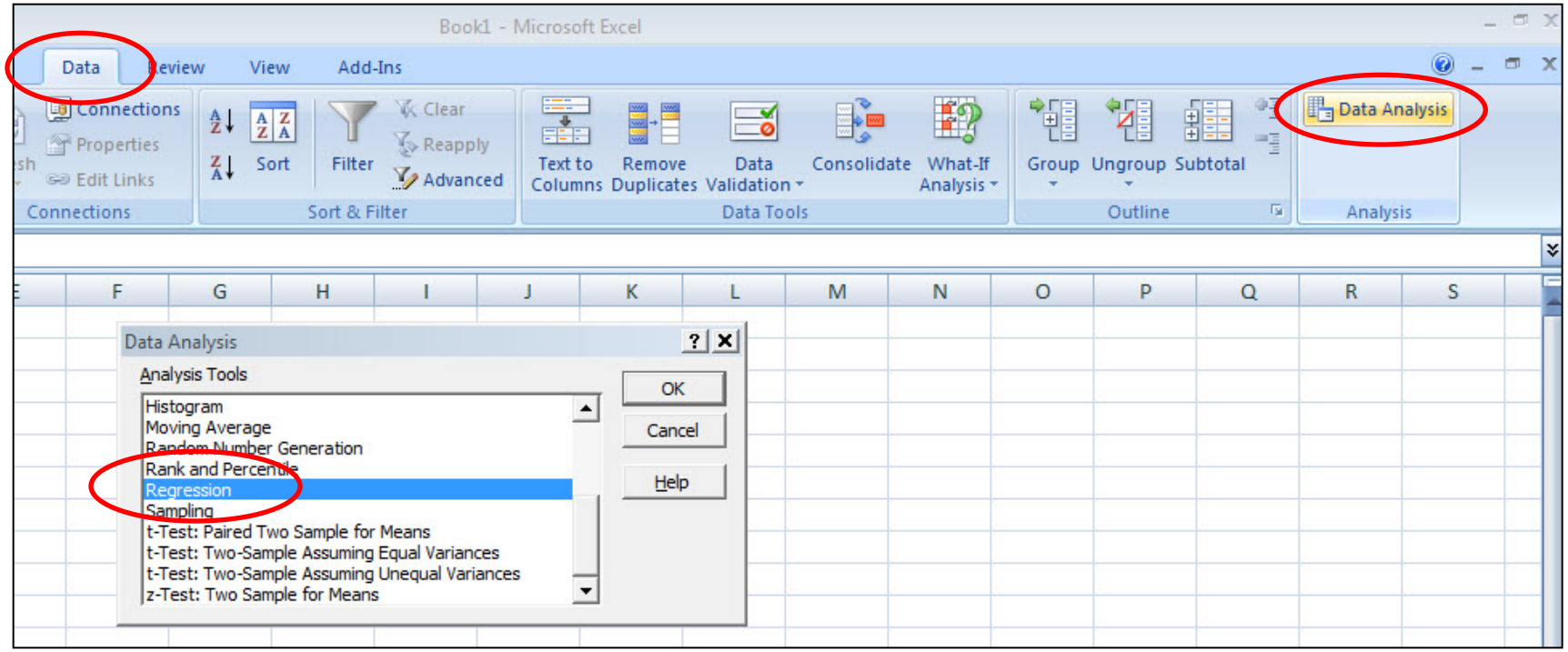
Representação Gráfica

- Modelo de preços de imóveis: scatter plot



Regressão usando Excel

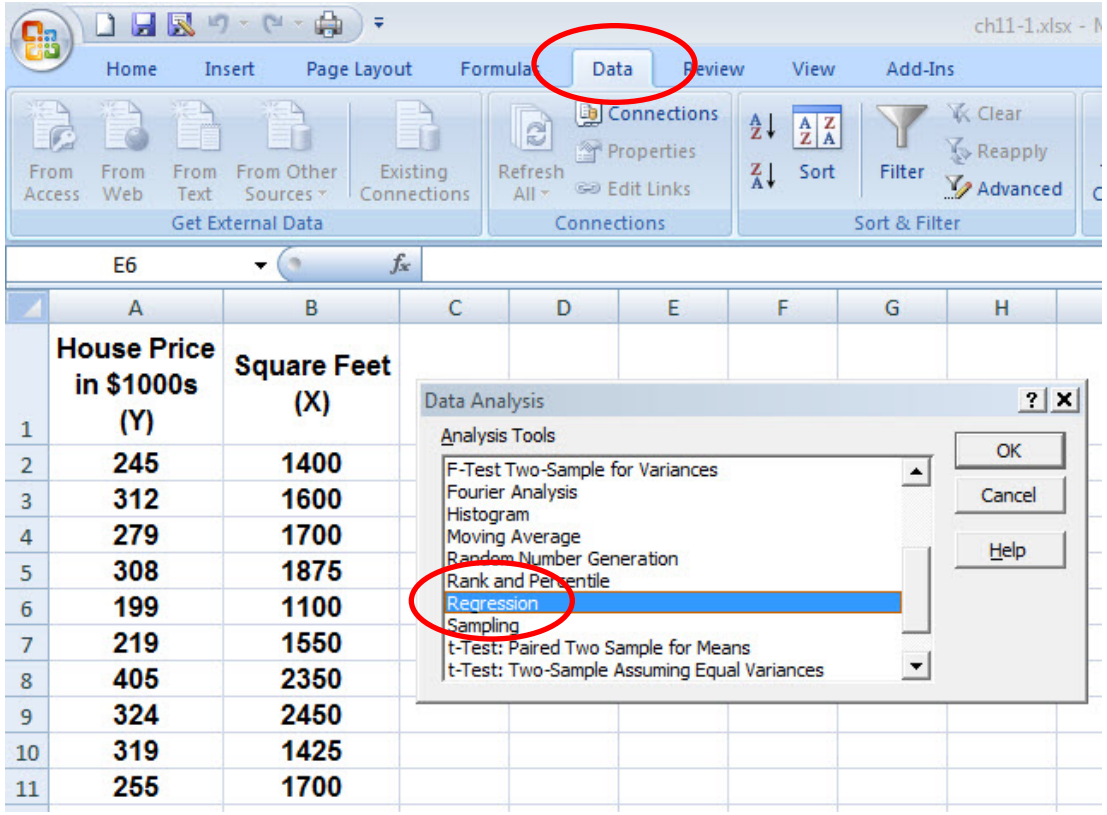
- Excel will be used to generate the coefficients and measures of goodness of fit for regression
  - Data / Data Analysis / Regression



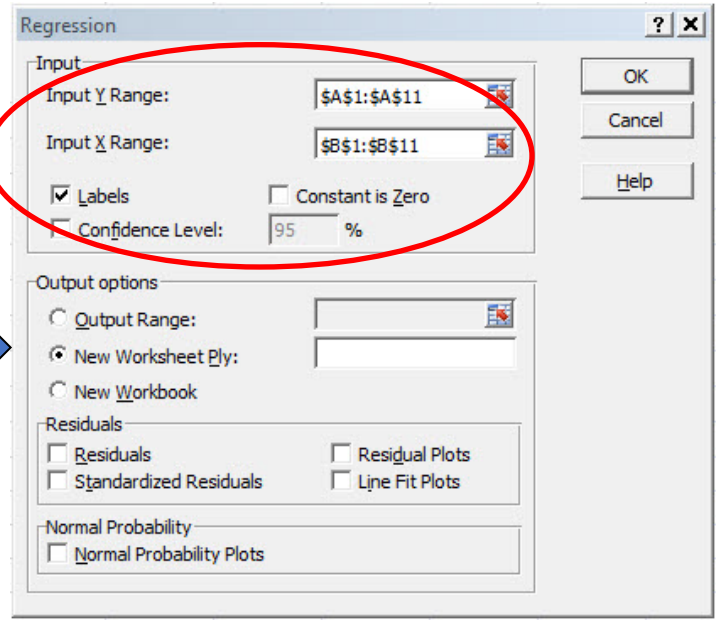
Regressão usando Excel

(cont.)

- Data / Data Analysis / Regression



Provide desired input:



Saída do Excel

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.762113713					
5	R Square	0.580817312					
6	Adjusted R Square	0.528419476					
7	Standard Error	41.33032365					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	18934.9348	18934.9348	11.0848	0.01039	
13	Residual	8	13665.5652	1708.1957			
14	Total	9	32600.5				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	98.24833	58.03348	1.69296	0.12892	-35.57711	232.07377
18	Square Feet (X)	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Saída do Excel

(cont.)

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

A regressão de regressão é:

$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Meters	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Saída Stata

$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$

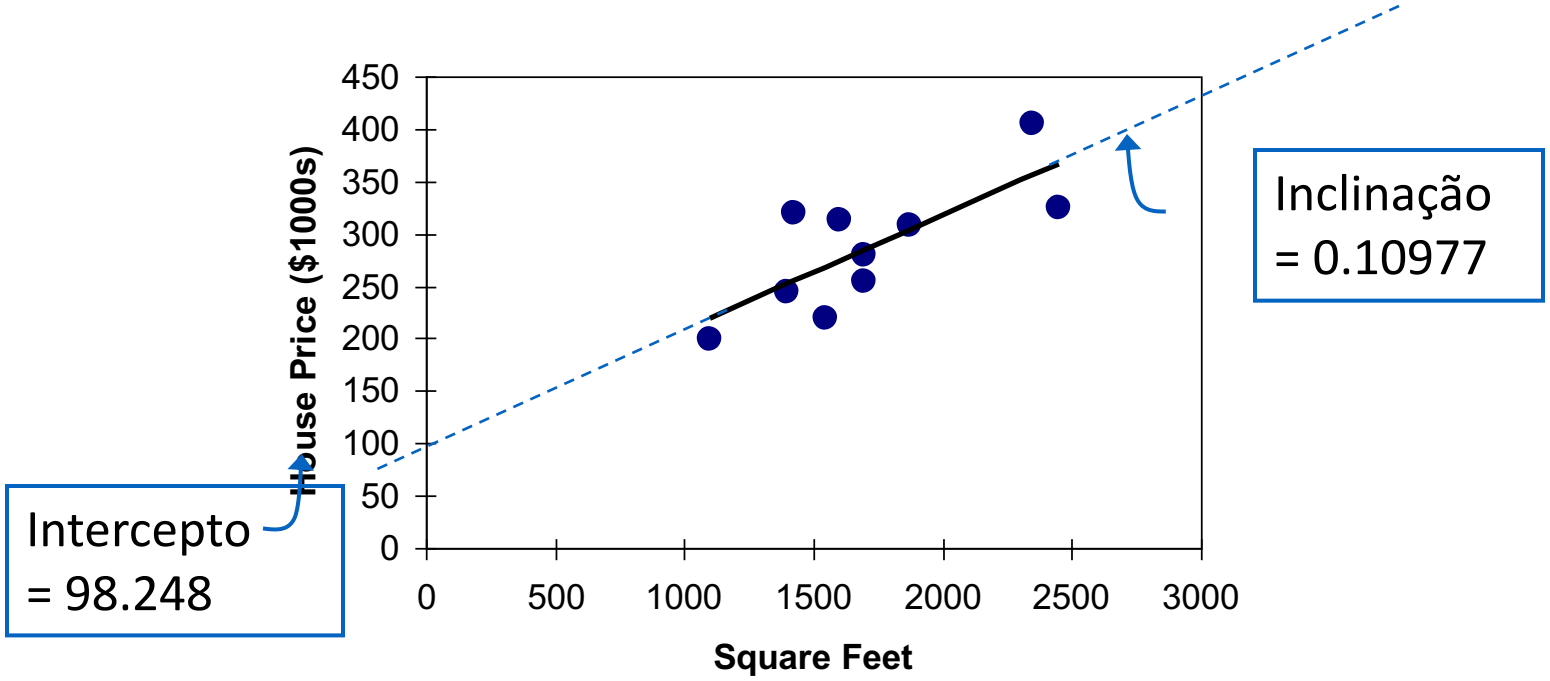
reg y x

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(1, 8)       =   11.08
       Model |  18934.9348     1  18934.9348            Prob > F      =  0.0104
    Residual |  13665.5652     8  1708.19565            R-squared     =  0.5808
-------------+------------------------------           Adj R-squared =  0.5284
       Total |     32600.5     9  3622.27778            Root MSE      =  41.33

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1097677   .0329694     3.33   0.010     .0337401    .1857954
       _cons |   98.24833   58.03348     1.69   0.129    -35.57711    232.0738
------------------------------------------------------------------------------
```

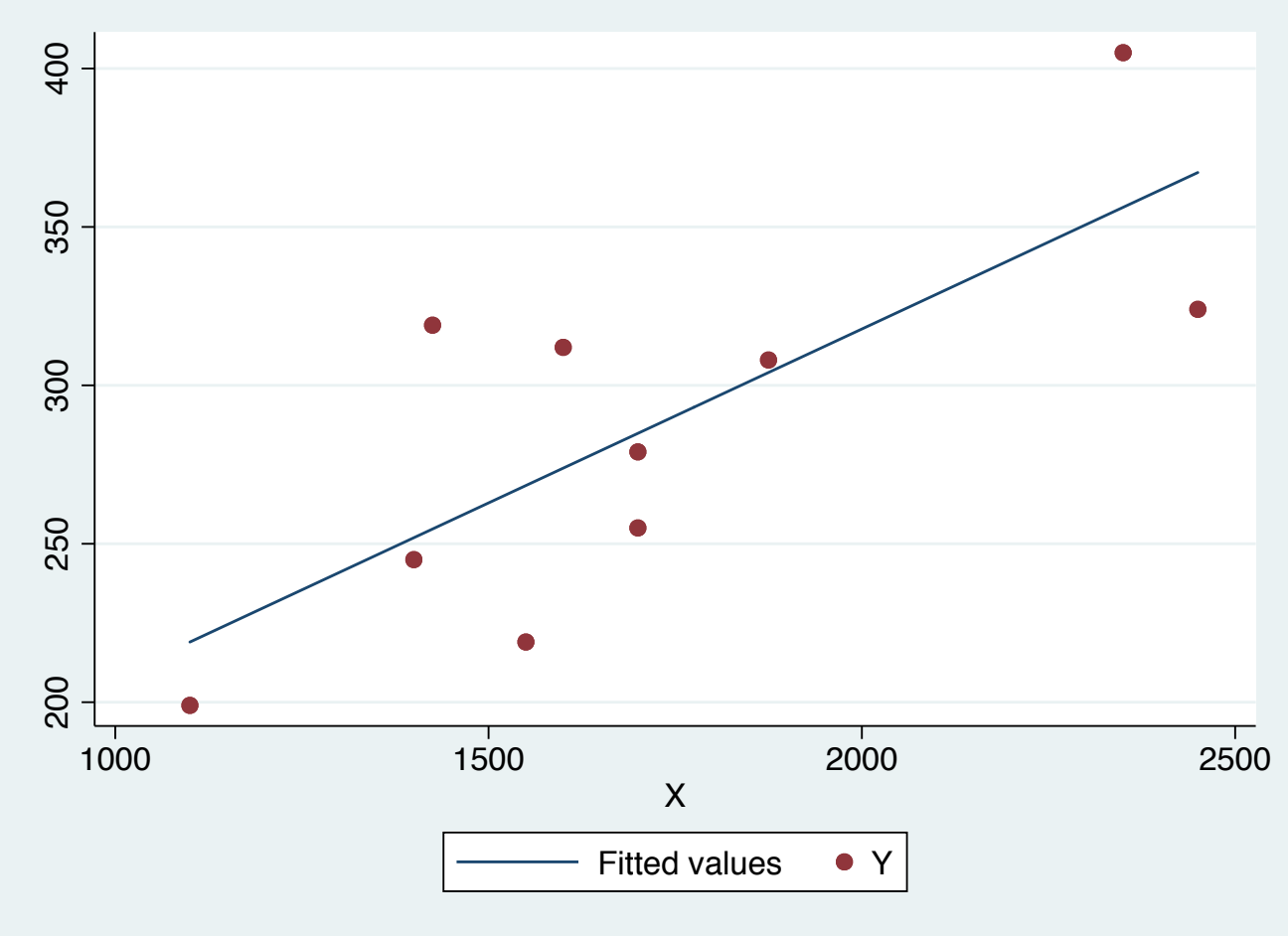
Representação Gráfica

- Modelo de preço de imóvel: gráfico de dispersão e linha de regressão



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

graph twoway (lfit y x) (scatter y x)



Interpretação do Intercepto, b_0

$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$

- b_0 é o valor médio estimado de Y quando o valor de X é zero (se X = 0 estiver no intervalo de valores X observados)
  - Aqui, nenhuma casa tinha 0 m2, então b0 = 98,24833 apenas indica que, para imóveis dentro da faixa de tamanhos observada, $ 98.248,33 é a parte do preço da casa não explicada pela metragem quadrados

Interpretação do coeficiente de inclinação, b_1

$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$

- b_1 mede a mudança estimada no valor médio de Y como resultado de uma mudança de uma unidade em X
- Aqui, b1 = 0.10977 nos diz que o valor médio de uma casa aumenta em .10977 ($ 1000) = $ 109,77, em média, para cada metro quadrado adicional de tamanho

Medidas de Variação

- A variação total é composta por duas partes:

$$SQT = SQR + SQE$$

Soma de Quadrados Total	Soma de quadrados da Regressão	Soma de quadrados dos Erros
$SQT = \sum (y_i - \bar{y})^2$	$SQR = \sum (\hat{y}_i - \bar{y})^2$	$SQE = \sum (y_i - \hat{y}_i\ )^2$

where:

$\bar{y}$ = Valor médio da variável dependente

y_i = Valores observados da variável dependente

$\hat{y}_i$ = Valor previsto de y para um dado valor de x_i

Medidas de Variação

(cont.)

SQT = Soma de Quadrados Total

- Mede a variação dos valores de y_i em torno de sua média, $\overline{y}$

SQR = Soma de Quadrados da Regressão

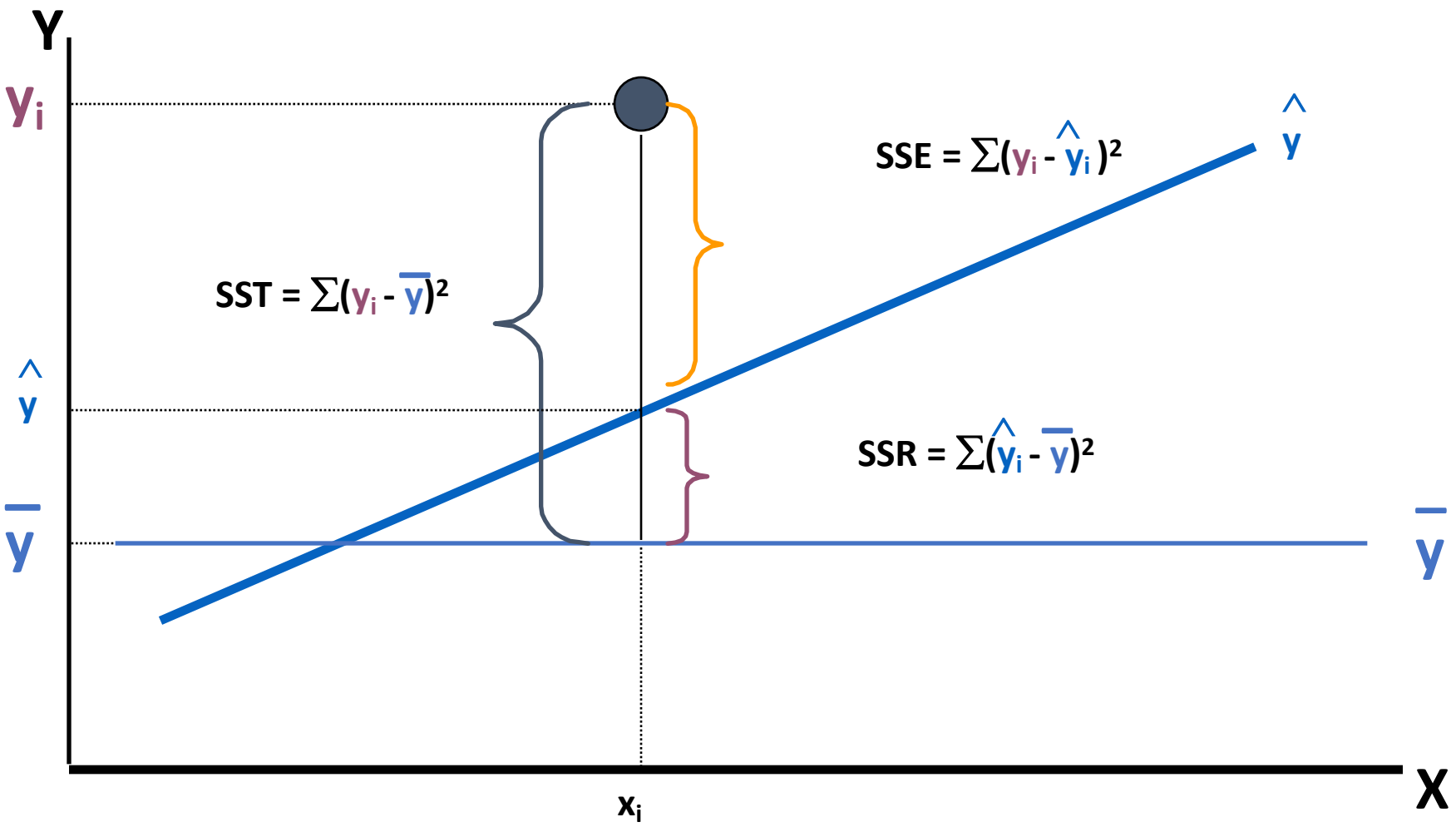
- Variação explicada atribuível à relação linear entre x e y

SQE = Soma de Quadrados dos Erros

- Variação atribuível a outros fatores que não a relação linear entre x e y

Medidas de Variação

(cont.)



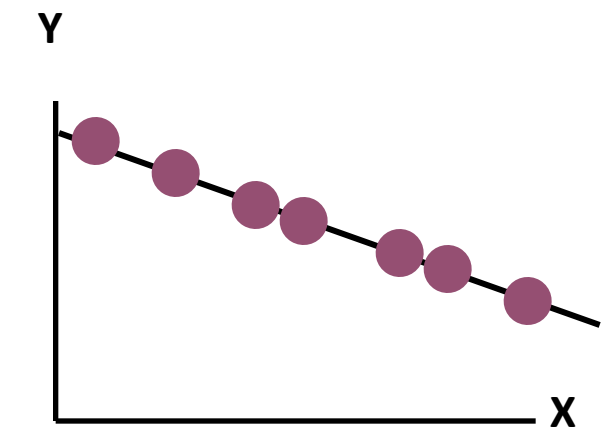
Coeficiente de Determinação, R^2

- O Coeficiente de Determinação é a porção da variação total na variável dependente que é explicada pela variação na variável independente
- O coeficiente de determinação também é chamado de R ao quadrado e é denotado como R^2

$$R^2 = \frac{SQR}{SQT} = \frac{Soma\ dos\ quadrados\ da\ Regressão}{Soma\ dos\ quadrados\ total}$$

note: $0 \leq R^2 \leq 1$

Exemplos de valores de r^2

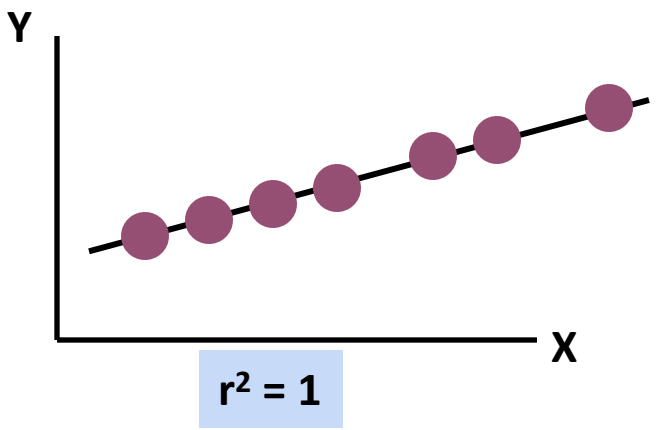


$r^2 = 1$

$r^2 = 1$

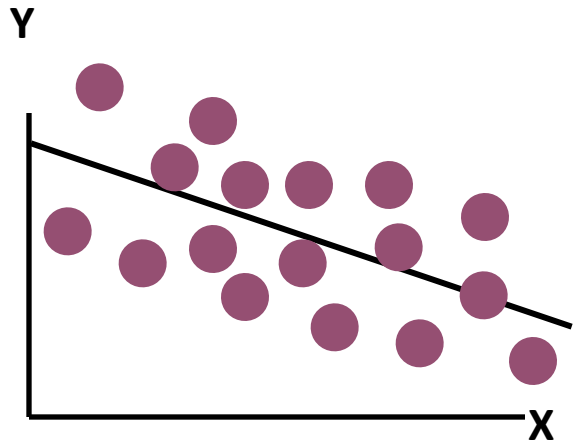
Relação linear perfeita entre X e Y:

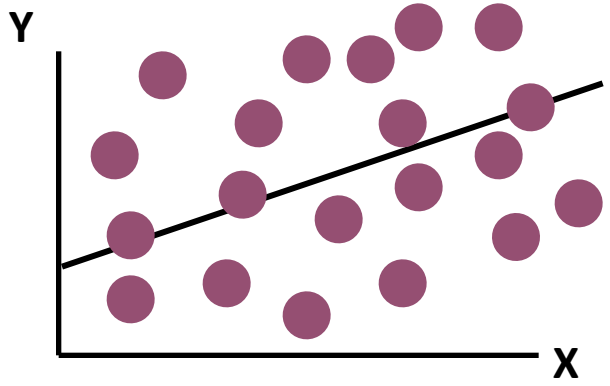
100% da variação em Y é explicada pela variação em X



$r^2 = 1$

Exemplos de valores de r^2



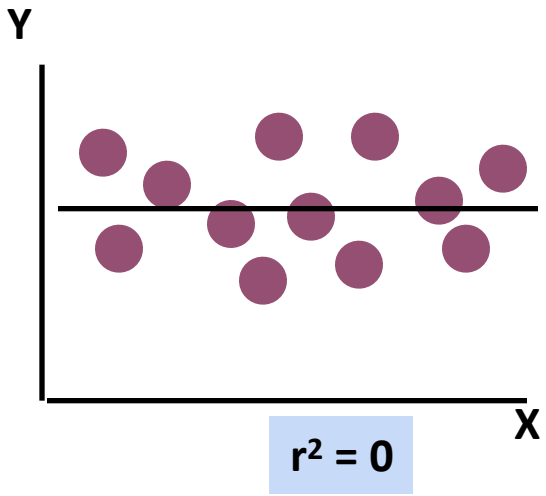


$0 < r^2 < 1$

Relações lineares mais fracas entre X e Y:

Algumas, mas não todas as variações em Y são explicadas pela variação em X

Exemplos de valores de r^2



$r^2 = 0$

Sem relação linear entre X e Y:

O valor de Y não depende de X.
(Nenhuma das variações em Y é explicada pela variação em X)

Medidas de Variação

- A variação total é composta por duas partes:

$$SQT = SQR + SQE$$

Soma de Quadrados Total	Soma de quadrados da Regressão	Soma de quadrados dos Erros
$SQT = \sum (y_i - \bar{y})^2$	$SQR = \sum (\hat{y}_i - \bar{y})^2$	$SQE = \sum (y_i - \hat{y}_i\ )^2$

where:

$\bar{y}$ = Valor médio da variável dependente

y_i = Valores observados da variável dependente

$\hat{y}_i$ = Valor previsto de y para um dado valor de x_i

Saída do Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SQR}{SQT} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% é a variação do preço dos imóveis explicada pela metragem quadrada

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Meters	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Saída Stata

$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$

reg y x

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(1, 8)       =   11.08
       Model |  18934.9348     1  18934.9348            Prob > F      =  0.0104
    Residual |  13665.5652     8  1708.19565            R-squared     =  0.5808
-------------+------------------------------           Adj R-squared =  0.5284
       Total |     32600.5     9  3622.27778            Root MSE      =  41.33

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1097677   .0329694     3.33   0.010     .0337401    .1857954
       _cons |   98.24833   58.03348     1.69   0.129    -35.57711    232.0738
------------------------------------------------------------------------------
```

Correlação e R_2

- O coeficiente de determinação, R^2, para o modelo de regressão simpels é igual ao coeficiente de correlação simples ao quadrado.

$$R^2 = r_{xy}^2$$

Estimando a Variância do Erro do modelo

Um estimador para a variância do erro do modelo populacional é

$$\hat{\sigma}^2 = S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SQE}{n-2}$$

- A divisão por n – 2 ao invés de n – 1 ocorre porque o modelo de regressão simples estima dois parâmetros, b_0 e b_1, ao invés de 1

$s_e = \sqrt{s_e^2}$ é chamado de erro padrão estimado

Saída do Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

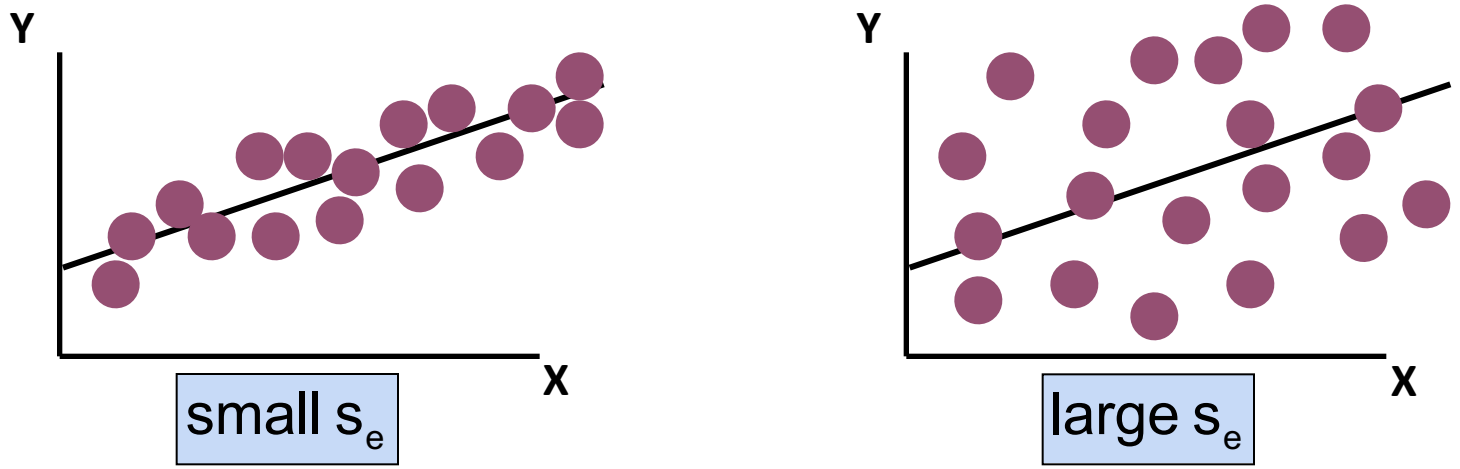
$$s_e = 41.33032$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparando erros padrão

s_e é uma medida da variação dos valores de y observados a partir da linha de regressão



O tamanho de s_e deve sempre ser julgado em relação ao tamanho dos valores y nos dados da amostra

Ou seja: s_e = \$41.33K é moderadamente pequeno em relação ao intervalo de preços das casas \$200 - \$300K

Inferência sob o modelo de regressão

- A variância da inclinação do modelo de regressão (b_1) é estimada por

$$s^2_{b_1} = \frac{s^2_e}{\sum (x_i - \bar{x})^2} = \frac{s^2_e}{(n-1)s^2_x}$$

sendo:

s_{b_1} = Estimativa do erro padrão da inclinação dos MQO

$S_e = \sqrt{\frac{SQE}{n-2}}$ = Erro padrão da estimativa

Saída do Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

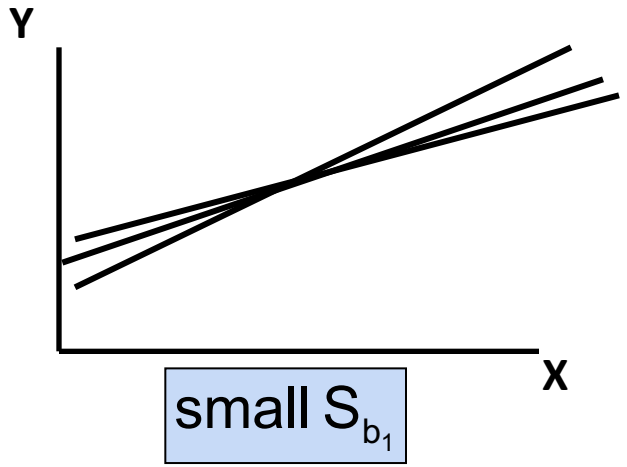
$$s_{b_1} = 0.03297$$

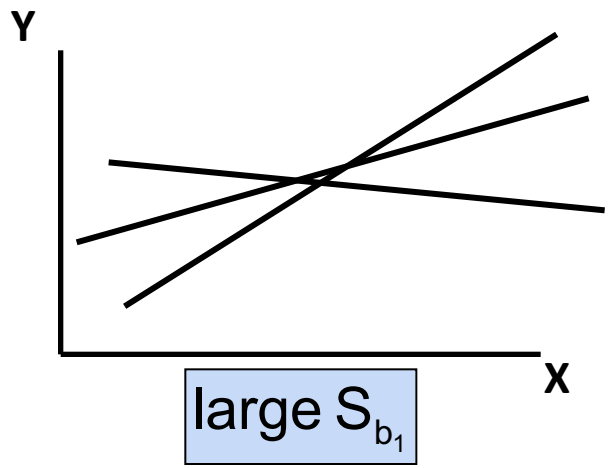
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparando o Erro Padrão da Inclinação

S_{b_1} é a medida de variação da inclinação da reta de regressão a partir de diferentes amostras.





Inferência sobre a inclinação: Teste t

- Teste t para a inclinação populacional
  - Existe uma relação linear entre X e Y?
- Hipóteses nula e alternativa

  H_0: $\beta_1 = 0$ (não há relação linear)

  H_1: $\beta_1 \neq 0$ (há relação linear)
- Estatística de Teste

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$d.f. = n - 2$$

sendo:

b_1 = coeficiente de inclinação da regressão

β_1 = inclinação hipotética

s_{b1} = erro padrão da inclinação

Inferência sobre a inclinação: Teste t

(cont.)

Preço do imóvel em $1000s (y)	m2 (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Equação de Regressão Estimada:

$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$

A inclinação deste modelo é 0,1098

A metragem quadrada da casa afeta seu preço de venda?

Inferência sobre a inclinação: Teste t Exemplo

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Das saídas dos programas:

b_1 S_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferência sobre a inclinação: Teste t Exemplo

(cont.)

Estatística de Teste: **t = 3.329**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

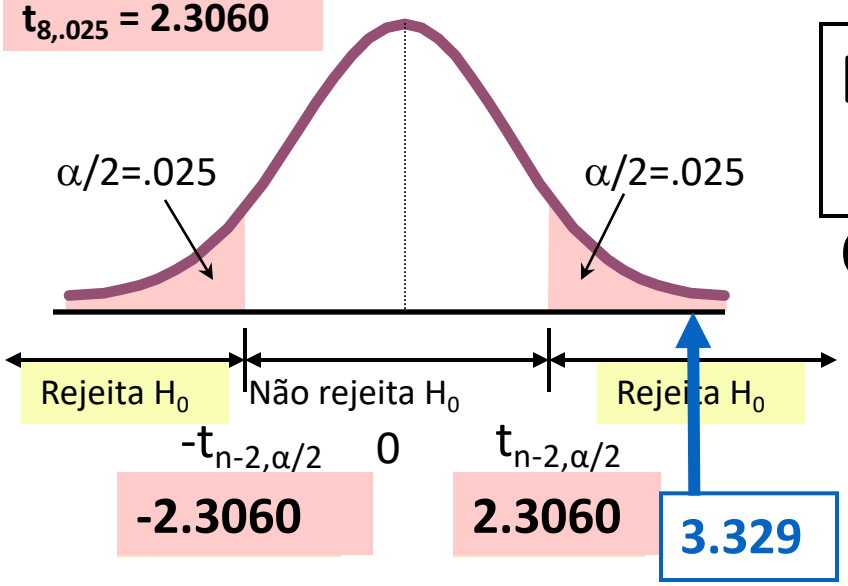
From Excel output:

b_1 S_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

g.l. = 10-2 = 8

$t_{8,.025}$ = 2.3060



Decisão:

Reject H_0

Conclusão:

Há evidências suficientes de que a metragem quadrada afeta o preço dos imóveis

Inferência sobre a inclinação: Teste t Exemplo

(cont.)

P-valor = **0.01039**

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Da saída do Excel:

P-valor

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

Este é um teste bicaudal, então o p- valor

$P(t > 3.329)+P(t < -3.329) = 0.01039$

(para 8 g.l.)

Decisão: P-valpr < α então

Reject H_0

Conclusão:

Há evidências suficientes de que a metragem quadrada afeta o preço dos imóveis

Estimativa do intervalo de confiança para a inclinação

Intervalo de Confiança para a inclinação:

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

g.l. = n - 2

Saída do Excel para preços internos

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Ao nível de confiança de 95%, o intervalo de confiança para a inclinação é (0,0337, 0,1858)

Estimativa do intervalo de confiança para a inclinação

(continued)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Como as unidades da variável de preço da casa são $ 1000s, estamos 95% confiantes de que o impacto médio no preço de venda está entre $ 33,70 e $ 185,80 por metro quadrado do tamanho da casa

Este intervalo de confiança de 95% não inclui 0.
Conclusão: há uma relação significativa entre o preço da casa e os pés quadrados no nível de significância de 0,05

Teste F para Significância Conjunta

- Estatística de Teste F:

$$F = \frac{MSR}{MSE}$$

sendo

$$MSR = \frac{SQR}{k}$$

$$MSE = \frac{SQE}{n - k - 1}$$

onde F segue uma distribuição F com k numerador e (n - k - 1) denominador graus de liberdade

(k = o número de variáveis independentes no modelo de regressão)

Saída do Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

Com 1 e 8 graus de liberdade

P-valor para p Teste F

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Teste F para Significância Conjunta

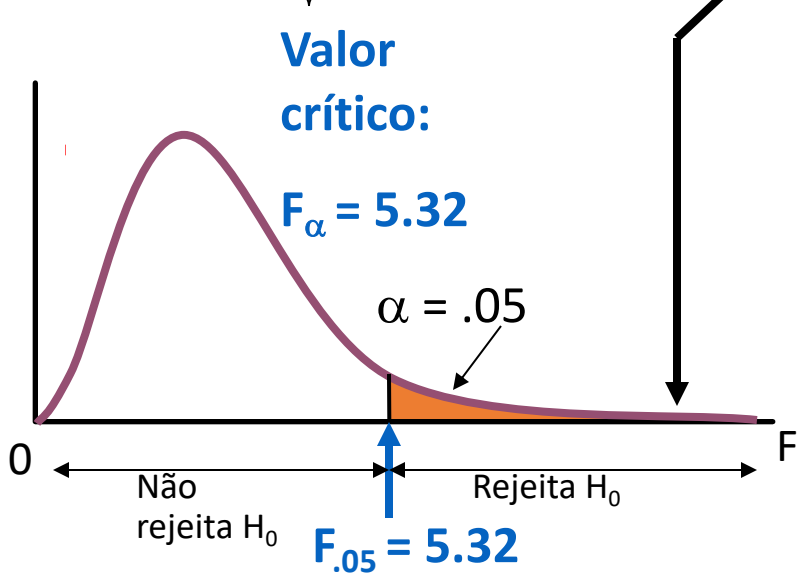
(cont.)

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \quad df_2 = 8$



Estatística de Teste:

$$F = \frac{MSR}{MSE} = 11.08$$

Decisão:

Rejeita H_0 a $\alpha = 0.05$

Conclusão:

Há evidências suficientes de que o tamanho da casa afeta o preço de venda

Previsão

- A equação de regressão pode ser usada para prever um valor para y, dado um x particular

- Para um valor específico, x_{n+1} , o valor previsto é

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Previsõesc usando análise de regressão

Previsão do preço de uma casa com 2.000 metros quadrados:

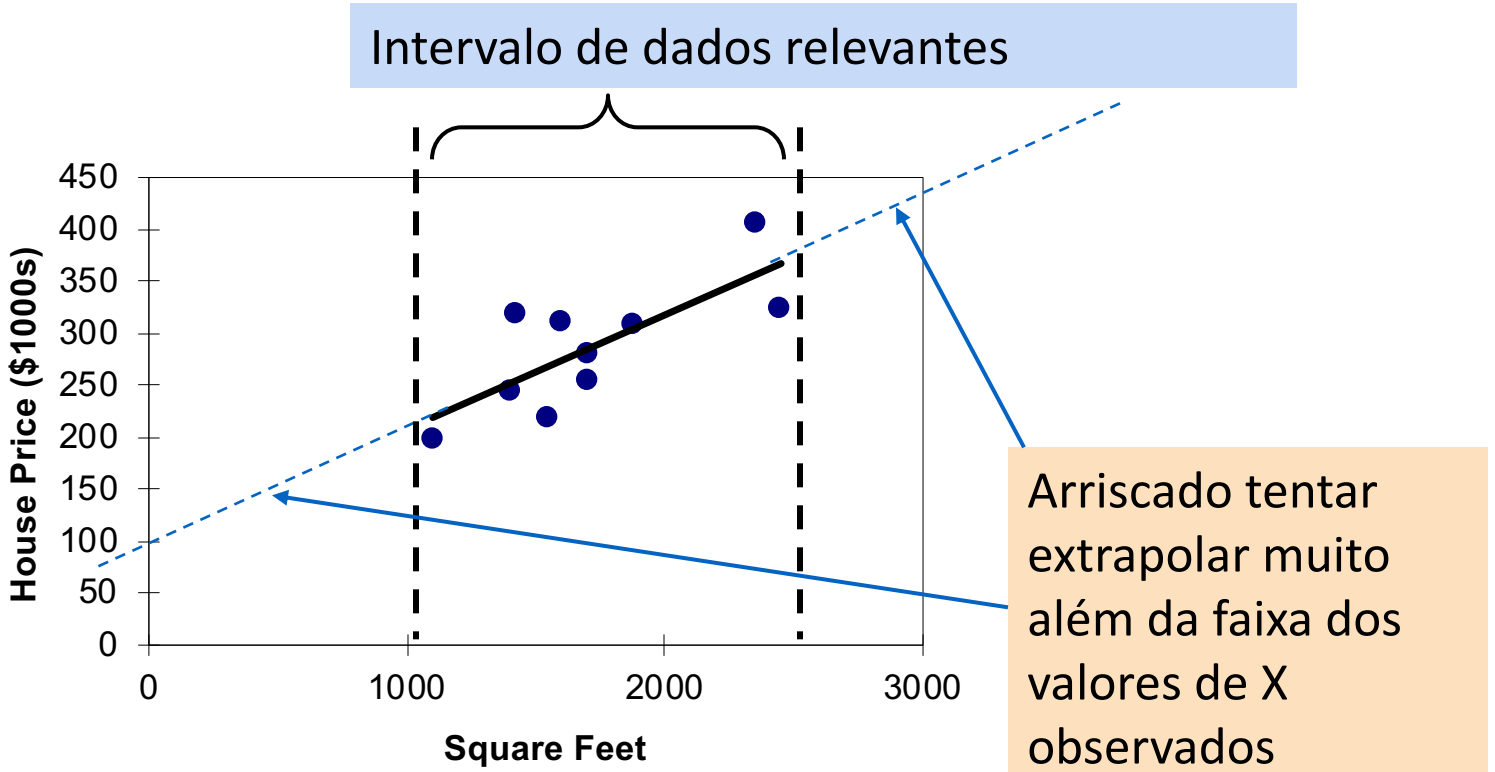
$$\widehat{preço\ imóvel} = 98.24833 + 0.10977 metros2$$
$$= 98.24833 + 0.10977(2000)$$
$$= 317.85$$

O preço previsto para uma casa com 2.000 pés quadrados é

317.85($1,000s) = $317,850

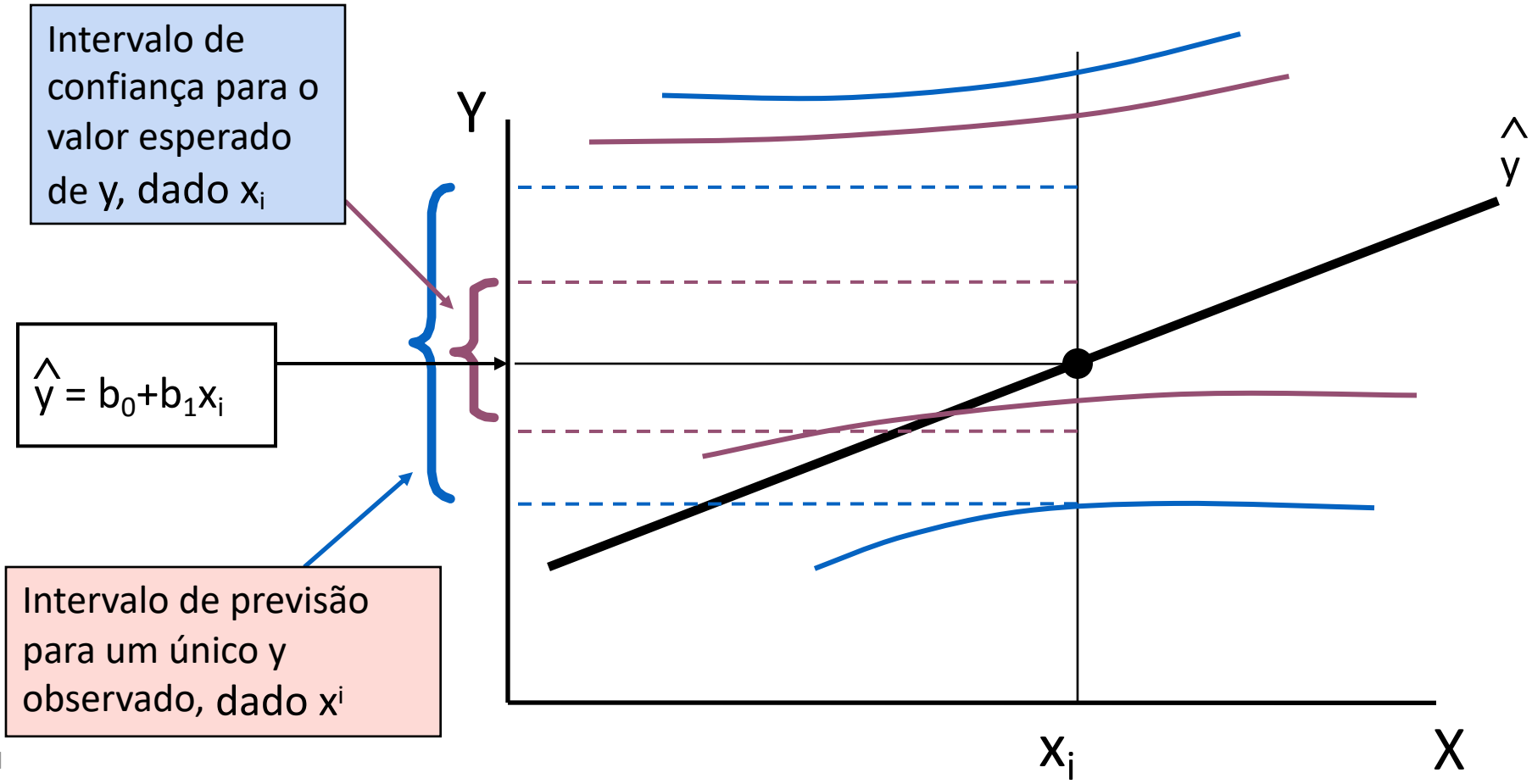
Intervalo de dados relevantes

- Ao usar um modelo de regressão para previsão, apenas preveja valores dentro do intervalo relevante de dados



Estimativa de valores médios e previsão de valores individuais

Objetivo: formar intervalos em torno de y para expressar a incerteza sobre o valor de y para um determinado x_i



Intervalo de confiança para a media de y, dado x_i

Intervalo de confiança estimado para o valor esperado de y, dado um particular valor de x_i

Confidence interval for $E(Y_{n+1} \mid X_{n+1})$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

Noe que a fórmula envolve o termo $(x_{n+1} - \bar{x})^2$ então o tamanho do interval varia de acordo com a distância de x_{n+1} de sua média, $\bar{x}$

Intervalo de previsão para um valor indivídual Y dado X

IC estimado para um **verdadeiro valor observado de y** dado um particular valor de x_i

Confidence interval for $\hat{y}_{n+1}$:

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

Este termo extra adiciona à largura do intervalo para refletir a incerteza adicionada para um caso individual

Estimativa de valores médios: exemplo

Estimativa de IC para $E(Y_{n+1} | X_{n+1})$

Encontre o IC de 95% para a média de precos de imóveis com 2,000 metros quadrados

Preço previsto y_i = $\hat{317.85}$ ($1,000s)

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 37.12$$

Os IC são [280.66; 354.90] ou [$280,660; $354,900]

Intervalo de previsão para um valor indivídual Y dado X: Exemplo

IC estimado para $\hat{y}_{n+1}$

Encontre o IC com 95% um imóvel individual com 2,000 metros quadrados

Preço Previsto y_i = 317.85 ($1,000s)

$$\hat{y}_{n+1} \pm t_{n-1,\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 102.28$$

Os IC são [215.50; 420.07], ou [$215,500; $420,070]

Análise de Correlação

- A análise de correlação é usada para medir a força da associação (relação linear) entre duas variáveis
- A correlação está apenas preocupada com a força do relacionamento
- Nenhum efeito causal está implícito na correlação

Análise de Correlação

O coeficiênte de correlação populacional é denominado ρ (rho)

- O coeficiente de correlação amostral é

$$r = \frac{s_{xy}}{s_x s_y}$$

sendo

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Teste de Hipótese para a Correlação

- Para testar a hipótese nula de nenhuma associação linear, $H_0 : \rho = 0$

a estatística de teste segue a distribuição t de Student com (n - 2) graus de liberdade:

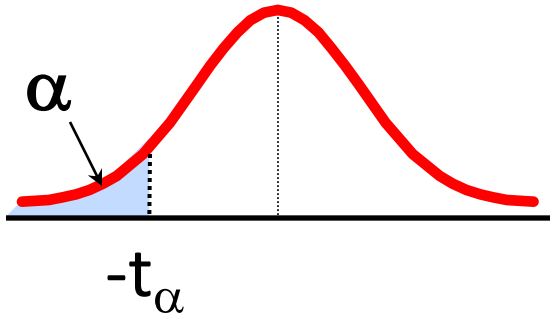
$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

Regras de Decisão

ste de Hipótese para a Correlação

Teste de cauda inferior:

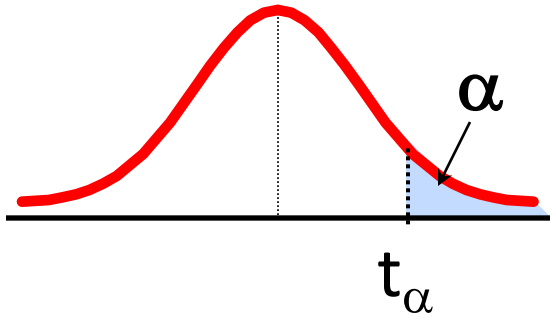
H_0: $\rho \geq 0$
H_1: $\rho < 0$



Rejeita H_0 if $t < -t_{n-2,\,\alpha}$

Teste de cauda superior:

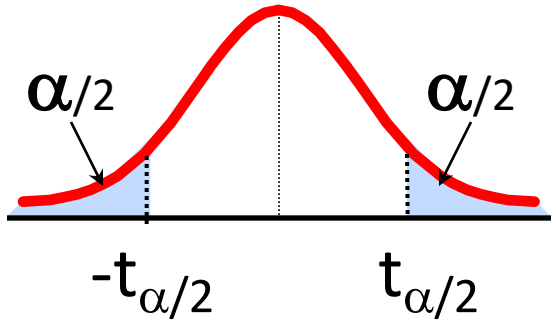
H_0: $\rho \leq 0$
H_1: $\rho > 0$



Rejeita H_0 if $t > t_{n-2,\,\alpha}$

Teste de duas caudas:

H_0: $\rho = 0$
H_1: $\rho \neq 0$



Rejeita H_0 if $t < -t_{n-2,\,\alpha/2}$ or $t > t_{n-2,\,\alpha/2}$

Sendo que $t = \dfrac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$ tem n - 2 g.l.

Graphical Analysis

- O modelo de regressão linear é baseado na minimização da soma dos erros quadrados
- Se existirem discrepâncias, seus erros quadráticos potencialmente grandes podem ter uma forte influência na linha de regressão ajustada
- Certifique-se de examinar seus dados graficamente em busca de outliers e pontos extremos
- Decida, com base em seu modelo e lógica, se os pontos extremos devem permanecer ou ser removidos