

Regressão Linear Múltipla – Aula 07

Statistics for Business and Economics 11 ed., by Paul Newbold , William Carlson , Betty Thorne (cap. Simple Regression)

Statistics for Economics, Accounting and Business Studies, capítulo 8, Barrow (Multiple Regression)
Marislei Nishijima

O modelo de regressão múltipla

Ideia: Examinar a relação linear entre uma variável dependente (Y) e 2 ou mais variáveis independentes (X_i)

modelo de regressão múltipla com k variáveis independentes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Diagram illustrating the components of the multiple regression model equation:

- β_0 : Y-intercept
- $\beta_1, \beta_2, \dots, \beta_k$: Inclinações populacionais (Population Slopes)
- ε : Erro aleatório (Random Error)

O modelo de regressão múltipla

Os coeficientes do modelo de regressão múltipla são estimados usando dados de amostra

Equação de regressão múltipla com k variáveis independentes:

Valor Estimado
(ou previsto) de y

Intercepto
estimado

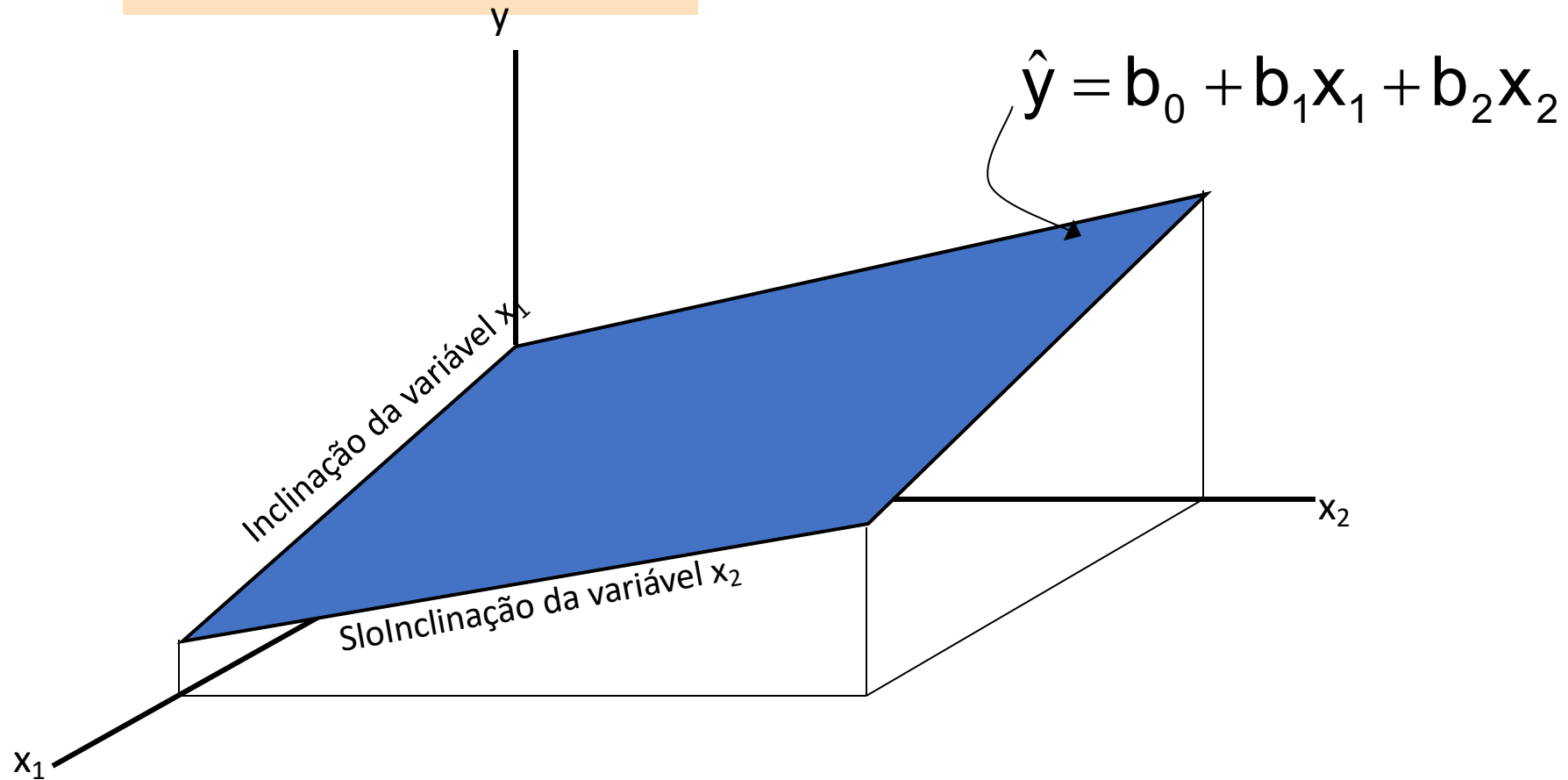
Coeficientes de inclinação
estimados

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki}$$

A equação de regressão múltipla

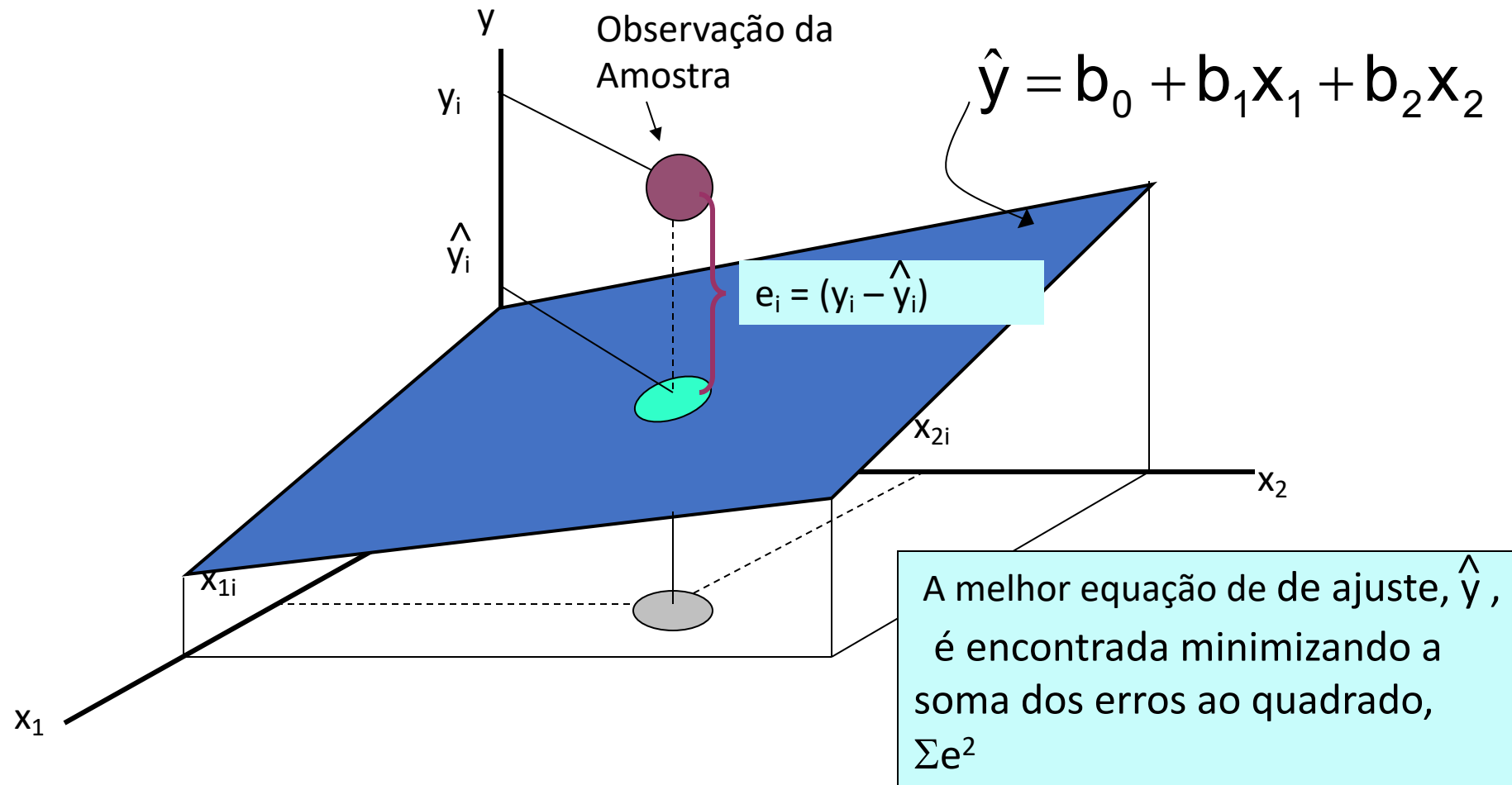
(cont.)

Duas variáveis de interesse



O modelo de regressão múltipla

Modelo com duas variáveis



Pressupostos do modelo de regressão múltipla

- Os valores de x_i e os termos do erro ε_i são independentes
- Os termos de erro são variáveis aleatórias com média 0 e uma variância constante, σ^2 .

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

(A propriedade de variância constante é chamada homocedasticidade)

Pressupostos do modelo de regressão múltipla

(cont.)

- Os termos aleatórios, ε_i não são correlacionados entre si, então

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

- Não é possível encontrar um conjunto de números, c_0, c_1, \dots, c_k , tal que

$$c_0 + c_1 X_{1i} + c_2 X_{2i} + \dots + c_k X_{ki} = 0$$

(Esta é a propriedade de não relação linear para os X_j 's)

Exemplo: 2 Variáveis Independentes

- Um distribuidor de tortas congeladas quer avaliar fatores que influenciam a sua demanda
- Variável dependente: vendas de tortas (unidades por semana)
 - Variáveis independentes: { Preço (em \$)
 - { Propaganda (gastos em \$100's)
- Dados são coletados para 15 semanas



Exemplo de vendas de Tortas

semana	Qtortas	Preço (\$)	Propag (\$100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

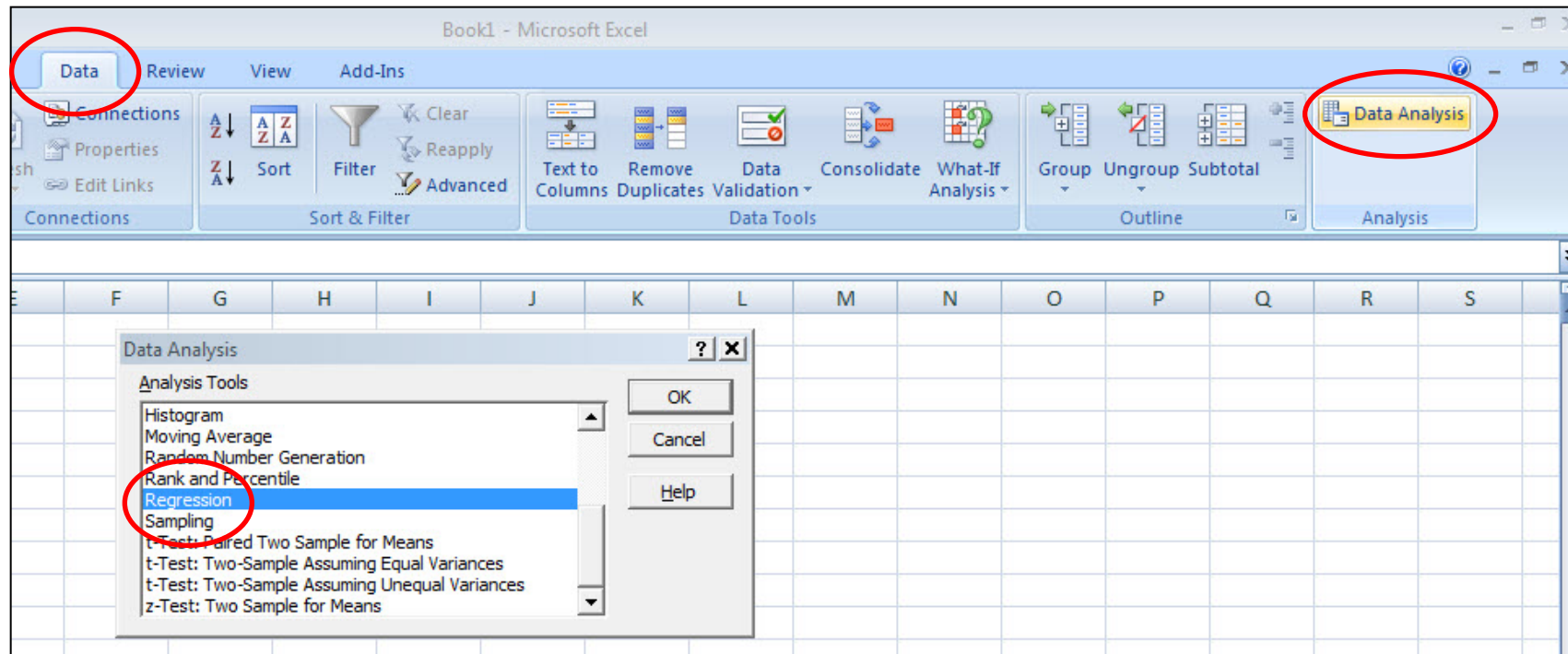
Equação da Regressão Múltipla:

$$\widehat{\text{vendas}} = b_0 + b_1 (\text{Preço}) + b_2 (\text{Propaganda})$$



Estimando a Regressão Linear Múltipla

- Excel e Stata serão usados para gerar os coeficientes e medidas de qualidade de ajuste para regressão múltipla
 - Data / Data Analysis / Regression



Saída da Regressão Múltipla



<i>Regression Statistics</i>	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{Qtortas} = 306.526 - 24.975 * Preco + 74.131 * Propaganda$$

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Teste F para Significância Conjunta

- Estatística de Teste F:

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

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

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

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)

Equação da Regressão Múltipla

$$\widehat{Qtortas} = 306.526 - 24.975 * Preço + 74.131 * Propaganda$$

Sendo
vendas é o número de tortas por semana
Preço em \$
Propaganda em \$100's.

$b_1 = -24.975$: as vendas diminuirão, em média, em 24.975 tortas por semana para cada aumento de \$ 1 no preço de venda, líquido dos efeitos das mudanças devido à publicidade

$b_2 = 74.131$: as vendas aumentarão, em média, em 74.131 tortas por semana para cada \$ 100 de aumento em publicidade, líquido dos efeitos das mudanças devido ao preço



Coeficiente de Determinação, R^2

- Relata a proporção da variação total em y explicada por todas as variáveis x tomadas em conjunto


$$R^2 = \frac{SQR}{SQT} = \frac{\textit{Soma dos Quadrados da Regressão}}{\textit{Soma dos Quadrados Totais}}$$

- Esta é a razão entre a variabilidade explicada e a variabilidade total da amostra.

Coeficiente de Determinação, R²

(cont.)

<i>Regression Statistics</i>	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SQR}{SQT} = \frac{29460.0}{56493.3} = 0.52148$$


52.1% da variação nas vendas de tortas é explicada pela variação de preço e publicidade

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Coeficiente de Determinação, R^2

reg q_torta preco proga

Source	SS	df	MS	Number of obs	=	15
Model	29460.0261	2	14730.0131	F(2, 12)	=	6.54
Residual	27033.3072	12	2252.7756	Prob > F	=	0.0120
Total	56493.3333	14	4035.2381	R-squared	=	0.5215
				Adj R-squared	=	0.4417
				Root MSE	=	47.463

q_torta	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
preco	-24.97509	10.83213	-2.31	0.040	-48.57626	-1.373916
proga	74.13096	25.96732	2.85	0.014	17.55303	130.7089
_cons	306.5262	114.2539	2.68	0.020	57.58834	555.4641

Estimativa da Variância do Erro

- Considere o modelo de regressão populacional

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

- A estimativa não viesada da variância dos erros é

$$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-K-1} = \frac{\text{SSE}}{n-K-1}$$

Sendo

$$e_i = y_i - \hat{y}_i$$


- A raiz quadrada da variância, s_e , é chamado de **erro padrão da estimativa**

Erro Padrão, s_e

<i>Regression Statistics</i>	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$s_e = 47.463$

A magnitude deste valor pode ser comparada ao valor médio de y



ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Coeficiente de Determinação Ajustado, \bar{R}^2

- R^2 nunca diminui quando uma nova variável X é adicionada ao modelo, mesmo se a nova variável não for uma variável preditora importante
- Isso pode ser uma desvantagem ao comparar modelos
- Qual é o efeito líquido de adicionar uma nova variável?
- Perdemos um certo grau de liberdade quando uma nova variável X é adicionada
- A nova variável X adicionou poder explicativo suficiente para compensar a perda de um grau de liberdade?

Coeficiente de Determinação Ajustado, \bar{R}^2

(cont.)

- Usado para corrigir o fato de que adicionar variáveis independentes não relevantes ainda reduzirá a soma dos quadrados do erro


$$\bar{R}^2 = 1 - \frac{SQE/(n - K - 1)}{SQT/(n - 1)}$$

- (onde n = tamanho da amostra, K = número de variáveis independentes)
- R^2 ajustado fornece uma melhor comparação entre modelos de regressão múltipla com diferentes números de variáveis independentes
- Penalize o uso excessivo de variáveis independentes sem importância
- Menor que R^2

$$\bar{R}^2$$

<i>Regression Statistics</i>	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$\bar{R}^2 = .44172$



44.2% da variação nas vendas de tortas é explicada pela variação no preço e na publicidade, levando em consideração o tamanho da amostra e o número de variáveis independentes

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Coeficiente de Correlação Múltipla

- O coeficiente de correlação múltipla é a correlação entre o valor previsto e o valor observado da variável dependente

$$R = r(\hat{y}, y) = \sqrt{R^2}$$

- É a raiz quadrada do coeficiente de determinação múltiplo
- Usado como outra medida da relação linear entre a variável dependente e as variáveis independentes
- Comparável à correlação entre Y e X na regressão simples

Avaliando coeficientes individuais da Regressão

- Use testes t para coeficientes individuais
- Mostra se uma variável independente específica é condicionalmente importante

Hipóteses

- $H_0: \beta_j = 0$ (nenhuma relação linear)
- $H_1: \beta_j \neq 0$ (existe relação linear entre x_j e y)

Avaliando coeficientes individuais da Regressão

(cont.)

$H_0: \beta_j = 0$ (nenhuma relação linear)

$H_1: \beta_j \neq 0$ (existe relação linear entre x_j e y)

Estatística de Teste:

$$t = \frac{b_j - 0}{S_{b_j}}$$

(df = $n - k - 1$)

Avaliando coeficientes individuais da Regressão

(cont.)

<i>Regression Statistics</i>						
Multiple R	0.72213	<p>O valor de t para Preço é $t = -2.306$, com p-valor de 0.0398</p> <p>t-v O valor de t para Propaganda é $t = 2.855$, com p-valor de 0.0145</p>				
R Square	0.52148					
Adjusted R Square	0.44172					
Standard Error	47.46341					
Observations	15					
<i>ANOVA</i>		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	29460.027	14730.013	6.53861	0.01201	
Residual	12	27033.306	2252.776			
Total	14	56493.333				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888



Exemplo: Avaliando coeficientes individuais da Regressão

Da saída do Excel:

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

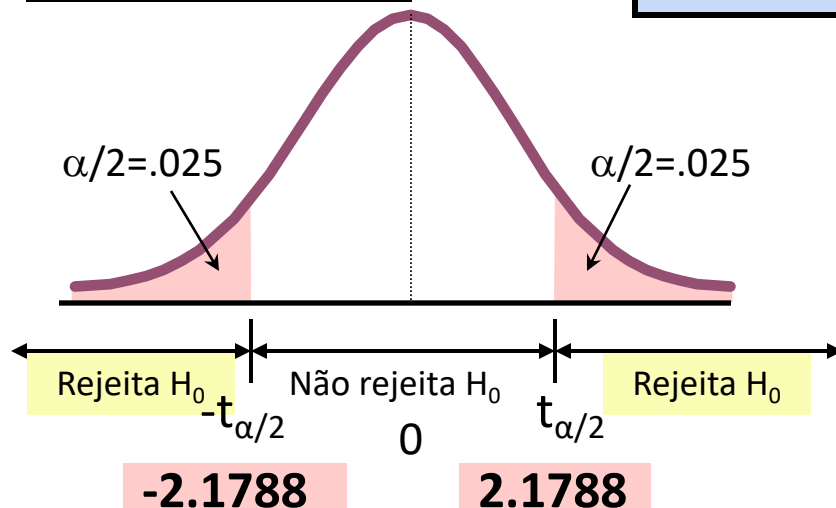
$$\text{d.f.} = 15 - 2 - 1 = 12$$

$$\alpha = .05$$

$$t_{12, .025} = 2.1788$$

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Price	-24.97509	10.83213	-2.30565	0.03979
Advertising	74.13096	25.96732	2.85478	0.01449

A estatística de teste para cada variável cai na região de rejeição (p-values < .05)



Decisão:

Rejeita H_0 para cada variável

Conclusão:

Há evidências de que tanto o preço quanto a publicidade afetam as vendas de tortas a $\alpha = .05$

Estimativa do intervalo de confiança para a inclinação

Estimativa do intervalo de confiança para a inclinação β_j

$$b_j \pm t_{n-K-1, \alpha/2} S_{b_j}$$

Sendo que t has
(n - K - 1) g.l.

	<i>Coefficients</i>	<i>Standard Error</i>
Intercept	306.52619	114.25389
Price	-24.97509	10.83213
Advertising	74.13096	25.96732

Here, t has
(15 - 2 - 1) = 12 d.f.

Exemplo: Construa um IC de 95% para efeitos de mudanças de preços (x_1) sobre a quantidade vendida de tortas:

$$-24.975 \pm (2.1788)(10.832)$$

Então o intervalo é $-48.576 < \beta_1 < -1.374$

Estimativa do intervalo de confiança para a inclinação

(cont.)

Estimativa do intervalo de confiança para a inclinação β_i

	<i>Coefficients</i>	<i>Standard Error</i>	...	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	...	57.58835	555.46404
Price	-24.97509	10.83213	...	-48.57626	-1.37392
Advertising	74.13096	25.96732	...	17.55303	130.70888

Exemplo: a saída do Excel também relata os valores do intervalo:

Estima-se que as vendas semanais sejam reduzidas entre 1,37 a 48,58 tortas para cada aumento de \$ 1 no preço de venda

Teste sobre todos os Coeficientes

- Teste F para a Significância global do Modelo
- Mostra se existe uma relação linear entre todas as variáveis X consideradas juntas e Y
- Use a estatística de teste F
- Hipóteses:

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ (não há relação linear)

$H_1: \text{at least one } \beta_i \neq 0$ (pelo menos uma variável independente afeta Y)

Teste F para significância geral

- Estatística de Teste:

$$F = \frac{MSR}{S_e^2} = \frac{SQR/K}{SQE/(n - K - 1)}$$

sendo que F tem k (numerador) e

$(n - K - 1)$ (denominador)

graus de liberdade


- A regra de decisão é

$$\text{Rejeita } H_0 \text{ se } F > F_{K, n-K-1, \alpha}$$

Teste F para significância geral

(cont.)

<i>Regression Statistics</i>	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15



$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom	P-value for the F-Test
----------------------------------	------------------------

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Teste F para significância geral

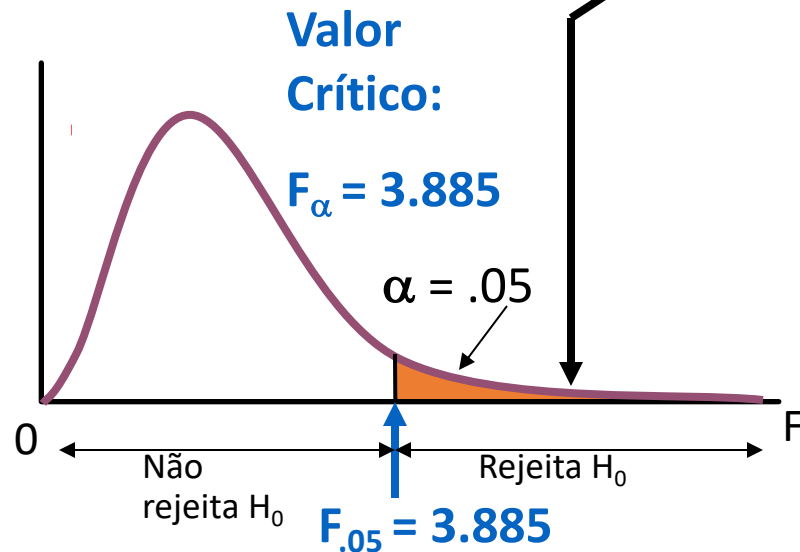
(cont.)

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_1 \text{ e } \beta_2 \text{ não são ambos } 0$$

$$\alpha = .05$$

$$gl_1 = 2 \quad gl_2 = 12$$



Estatística de Teste:

$$F = \frac{MSR}{MSE} = 6.5386$$

Decisão:

Como a estatística de teste F cai na região de rejeição ($p\text{-valor} < .05$), rejeita-se H_0

Conclusão:

Há evidências de que pelo menos uma variável independente afeta Y

Testes para um subconjunto de coeficientes de regressão

- Considere um modelo de regressão múltipla envolvendo as variáveis x_j e z_j , e a hipótese nula de que os coeficientes da variável z são todos zero:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \alpha_1 z_{1i} + \cdots + \alpha_r z_{ri} + \varepsilon_i$$

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

$$H_1: \text{pelo menos um } \alpha_j \neq 0. (j=1,2,\dots,r)$$

Testes para um subconjunto de coeficientes de regressão

(cont.)

- Objetivo: comparar a soma dos quadrados dos erros do modelo completo com a soma dos quadrados dos erros do modelo restrito
- Primeiro execute uma regressão para o modelo completo e obtenha SQE
- Em seguida, execute uma regressão restrita que exclui as variáveis z (o número de variáveis excluídas é r) e obtenha a soma do erro restrito dos quadrados SQE(r)
- Calcule a estatística F e aplique a regra de decisão para um nível de significância α

$$\text{Rejeite } H_0 \text{ se } F = \frac{(SQE(r) - SQE)/r}{S_e^2} > F_{r, n-K-r-1, \alpha}$$

Previsão

- Dado um modelo de regressão populacional

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

- Então, dado um novo ponto observado

$$(x_{1,n+1}, x_{2,n+1}, \dots, x_{K,n+1})$$

a melhor previsão linear não viesada de y_{n+1}^{\wedge} é

$$\hat{y}_{n+1} = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_K x_{K,n+1}$$

- É arriscado prever novos valores de X fora do intervalo dos dados usados para estimar os coeficientes do modelo, porque não temos dados para apoiar que o modelo linear se estende além do intervalo observado.

Usando a equação para fazer previsões

Preveja vendas para uma semana em que o preço de venda seja \$ 5,50 e a publicidade seja \$ 350:

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975 (5.50) + 74.131 (3.5) \\ &= 428.62\end{aligned}$$

A previsão de vendas é de 428,62 tortas

Observe que a publicidade está em \$ 100, então \$ 350 significa que $X_2 = 3.5$

Modelos de regressão não linear

- A relação entre a variável dependente e uma variável independente pode não ser linear
- É possível revisar o diagrama de dispersão para verificar relações não lineares
- Exemplo: modelo quadrático

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

- A segunda variável independente é o quadrado da primeira variável

Modelo de Regressão Quadrática

Forma do Modelos:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i$$

- sendo:

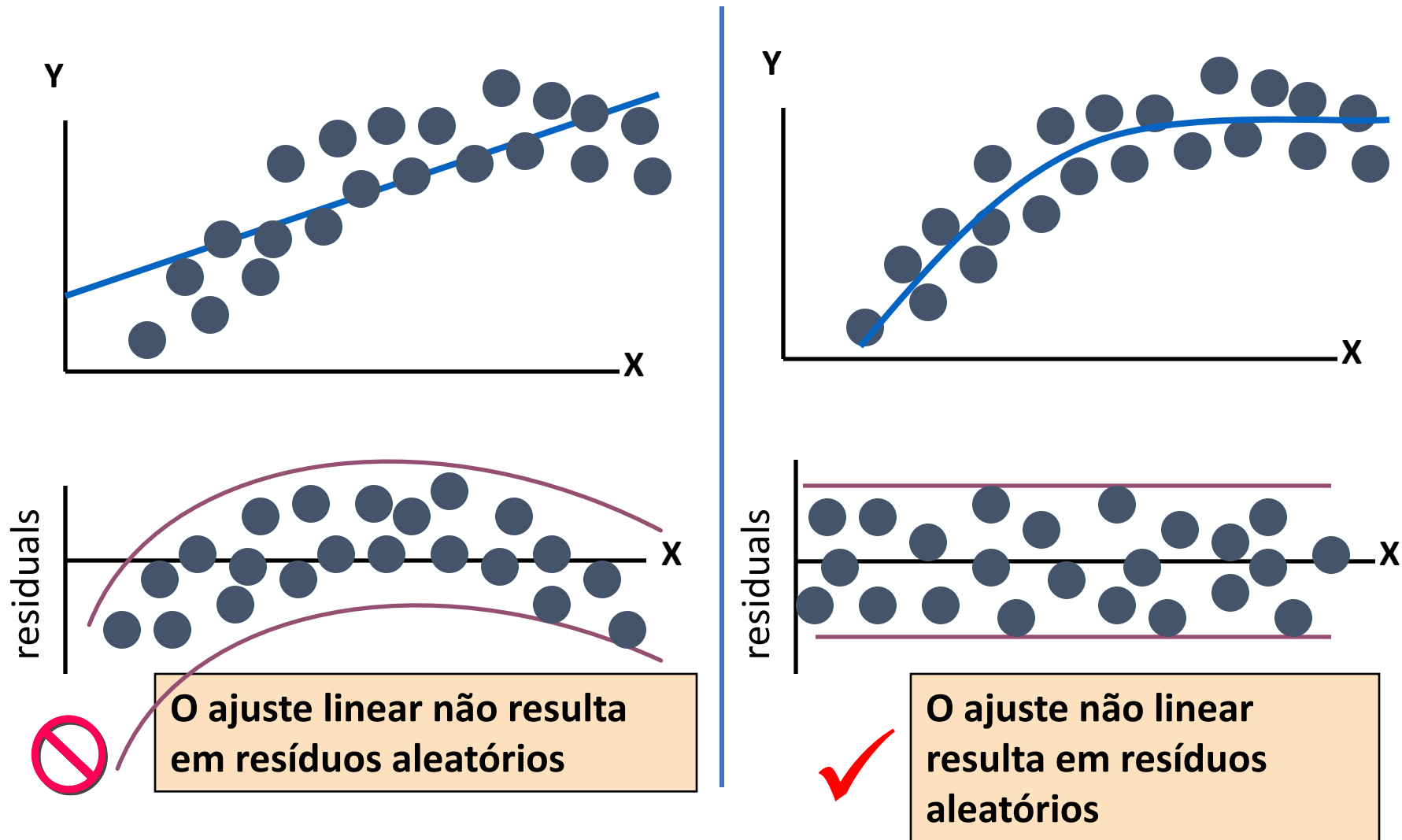
β_0 = o intercepto de Y

β_1 = coeficiente de regressão para efeito linear de X em Y

β_2 = coeficiente de regressão para efeito quadrático em Y

ε_i = erro aleatório em Y para observação i

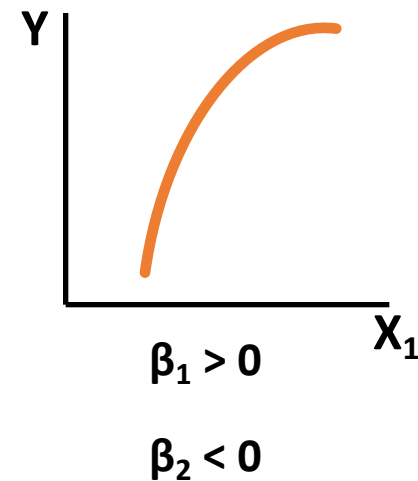
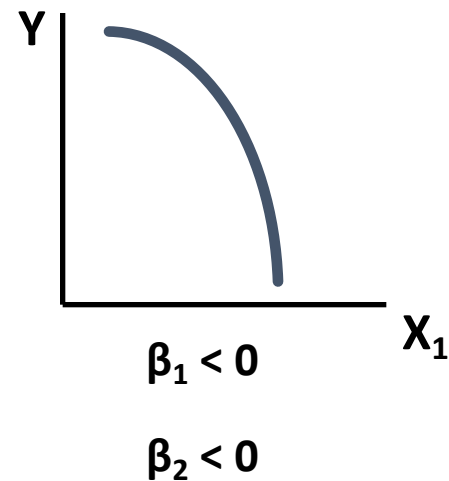
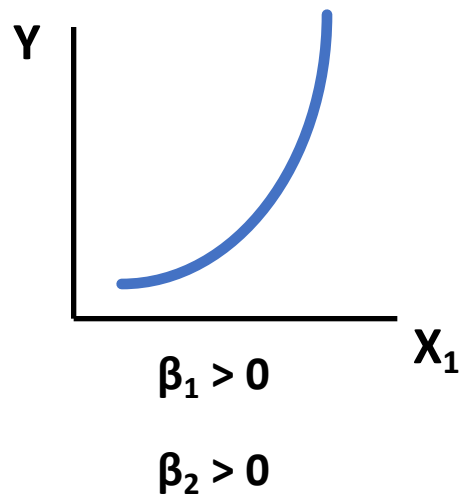
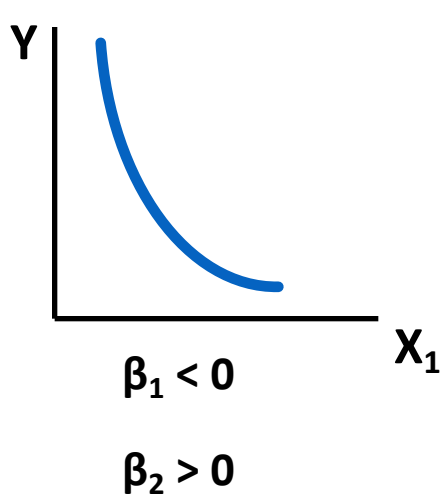
Ajustamento Linear vs. Não linear



Modelo de Regressão Quadrática

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Modelos quadráticos podem ser considerados quando o diagrama de dispersão assume uma das seguintes formas:



β_1 = o coeficiente do termo linear
 β_2 = th o coeficiente do termo quadrático

Teste de Significância: Efeito Quadrático

Testando o efeito quadrático

Compare a estimativa de regressão linear

$$\hat{y} = b_0 + b_1x_1$$

com a estimativa de regressão quadrática

$$\hat{y} = b_0 + b_1x_1 + b_2x_1^2$$

Hipóteses

$H_0: \beta_2 = 0$ (O termo quadrático não melhora o modelo)

$H_1: \beta_2 \neq 0$ (O termo quadrático melhora o modelo)

Teste de Significância: Efeito Quadrático

(cont.)

Testando o efeito quadrático

Hipóteses

$H_0: \beta_2 = 0$	(O termo quadrático não melhora o modelo)
$H_1: \beta_2 \neq 0$	(O termo quadrático melhora o modelo)

A estatística de teste é

$$t = \frac{b_2 - \beta_2}{S_{b_2}}$$

$$\text{d.f.} = n - 3$$

sendo:

b^2 = coeficiente de inclinação do termo quadrado

β_2 = inclinação hipotética (zero)

S_b^2 = erro padrão da inclinação

Teste de Significância: Efeito Quadrático

(cont.)

- Testando o Efeito Quadrático

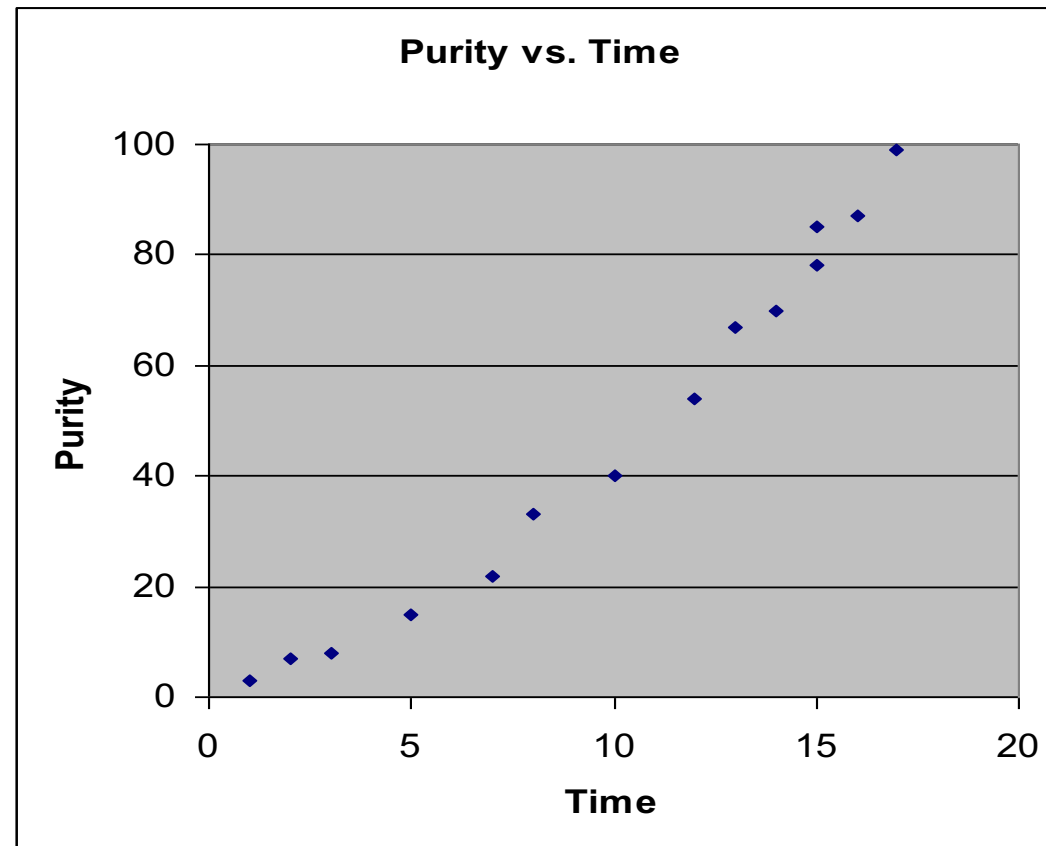
Comparar o \bar{R}^2 da regressão simples com o do modelo quadrático

- Se o \bar{R}^2 do modelo quadrático for maior que o R^2 do modelo de regressão simples, então o modelo quadrático é o melhor modelo.

Exemplo: modelo quadrático

Pureza	Tempo de Filtragem
3	1
7	2
8	3
15	5
22	7
33	8
40	10
54	12
67	13
70	14
78	15
85	15
87	16
99	17

A pureza aumenta conforme o tempo de filtragem aumenta:



Exemplo: modelo quadrático

(cont.)

- Resultados da Regressão Simples:

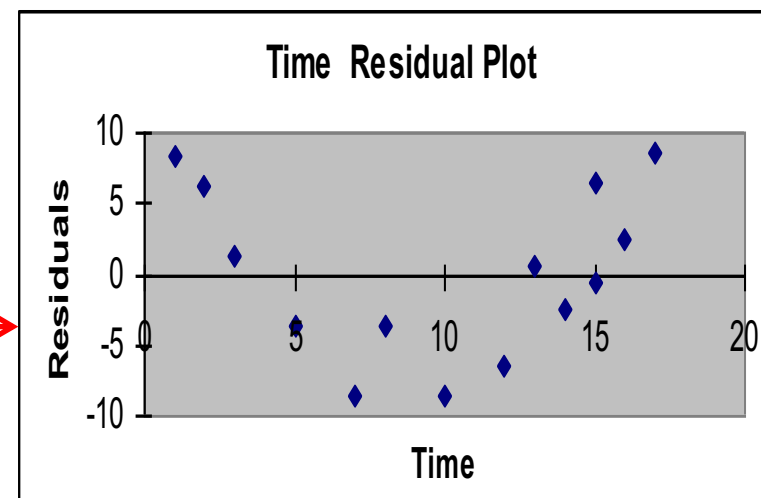
$$\hat{y} = -11.283 + 5.985 \text{ Time}$$

	Coefficients	Standard Error	t Stat	P-value
Intercept	-11.28267	3.46805	-3.25332	0.00691
Time	5.98520	0.30966	19.32819	2.078E-10

Regression Statistics	
R Square	0.96888
Adjusted R Square	0.96628
Standard Error	6.15997

F	Significance F
373.57904	2.0778E-10

Estatística t, Estatística F e R^2 são todos altos, mas os resíduos não são aleatórios:



Exemplo: modelo quadrático

(cont.)

■ Resultados da Regressão Quadrática:

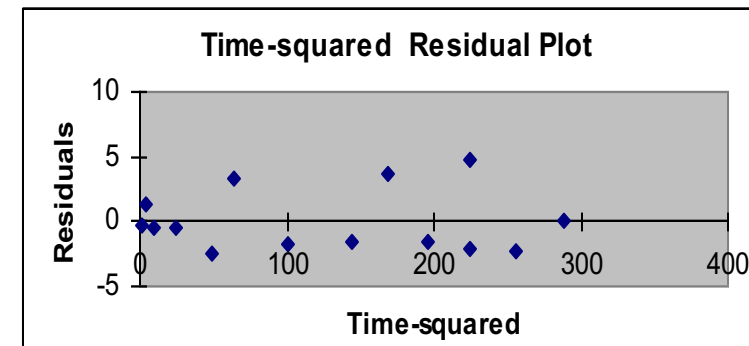
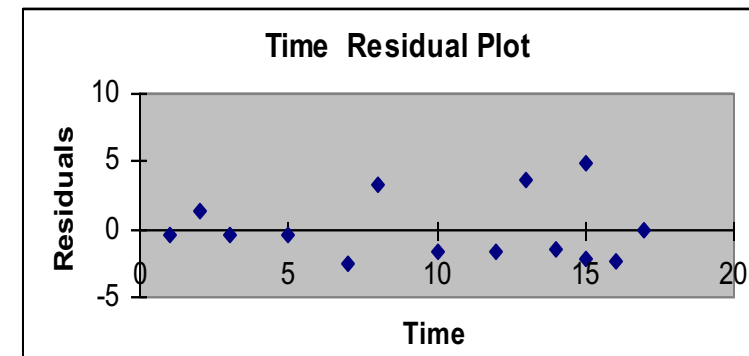
$$\hat{y} = 1.539 + 1.565 \text{ Time} + 0.245 (\text{Time})^2$$

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.53870	2.24465	0.68550	0.50722
Time	1.56496	0.60179	2.60052	0.02467
Time-squared	0.24516	0.03258	7.52406	1.165E-05

Regression Statistics	
R Square	0.99494
Adjusted R Square	0.99402
Standard Error	2.59513

F	Significance F
1080.7330	2.368E-13

O termo quadrático é significativo e melhora o modelo : R^2 é alto e s_e é menor, os resíduos agora são aleatórios



A transformação em Log

O modelo multiplicativo:

- Modelo multiplicativo original

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon$$

- Modelo multiplicativo transformado

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \log(\varepsilon)$$

Interpretação dos coeficientes

Para o modelo multiplicativo:

$$\log Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \log \varepsilon_i$$

Quando as variáveis dependentes e independentes são registradas:

O coeficiente da variável independente X_k pode ser interpretado como:

Uma mudança de 1% em X_k leva a uma mudança de b_k % na media do valor de Y

- b_k é a **elasticidade de** Y com respeito a uma mudança de X_k

Variáveis Dummy

- Uma variável dummy é uma variável independente categórica com dois níveis:
- sim ou não, ligado ou desligado, masculino ou feminino
- registrado como 0 ou 1
- Os interceptos de regressão são diferentes se a variável for significativa
- Assume inclinações iguais para outras variáveis
- Se houver mais de dois níveis, o número de variáveis dummies necessárias é (número de níveis - 1)

Exemplo Variáveis Dummies

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Seja:

y = Quantidade de tortas vendidas

x_1 = Preço

x_2 = Feriado ($x_2 = 1$ se um feriado ocorre durante a semana)

($x_2 = 0$ se não tem feriado na semana)



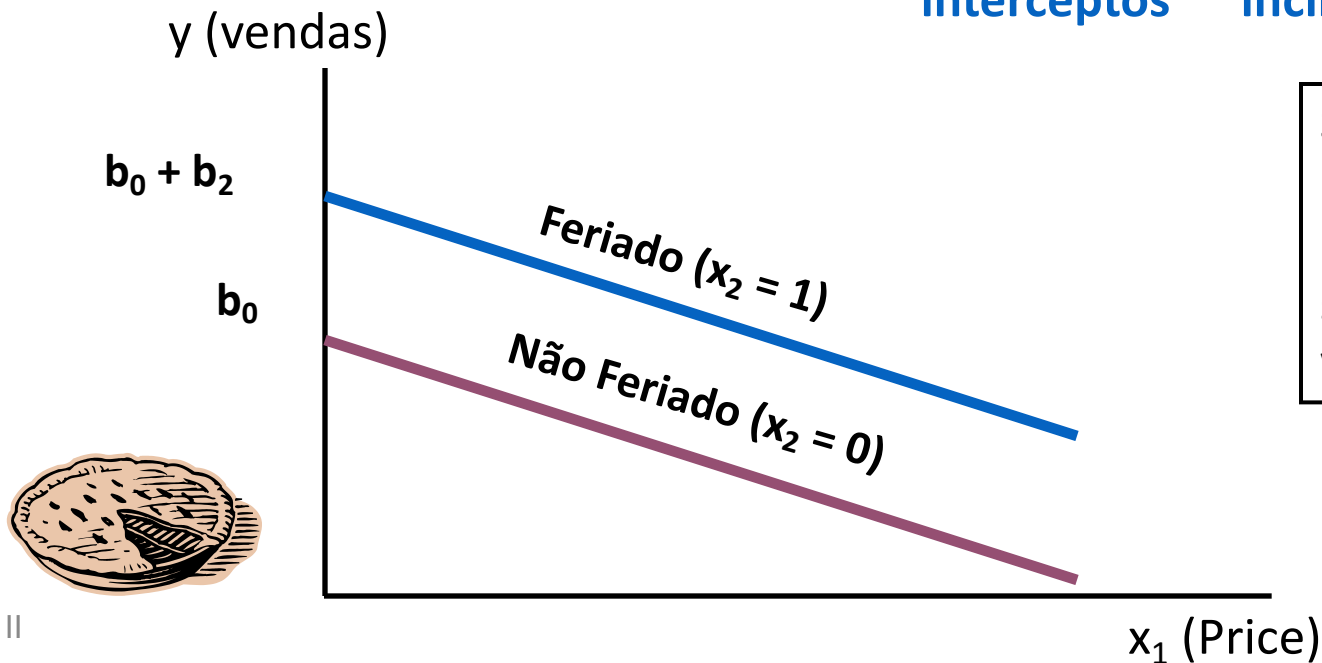
Exemplo Variáveis Dummies

(cont.)

$\hat{y} = b_0 + b_1 x_1 + b_2 (1) = (b_0 + b_2) + b_1 x_1$	Feriado
$\hat{y} = b_0 + b_1 x_1 + b_2 (0) = b_0 + b_1 x_1$	Não Feriado

Diferentes interceptos

Mesma inclinação



Se $H_0: \beta_2 = 0$ é rejeitada, então “Feriado” tem efeito significativo sobre as vendas de tortas

Interpretando o coeficiente da variável Dummy

Exemplo:

$$\widehat{Vendas} = 300 - 30 * Preço + 15 * Feriado$$

Vendas: número de tortas vendidas por semana

Preço: preço da torta em \$

Feriado: $\begin{cases} 1 & \text{se um feriado ocorre durante a semana} \\ 0 & \text{se um feriado não ocorre durante a semana} \end{cases}$

$b_2 = 15$: em média, as vendas foram 15 tortas maiores nas semanas com feriado do que nas semanas sem feriado, dado o mesmo preço



Interação entre variáveis explicativas

- Hipótese de interação entre um par de variáveis x
 - Resposta à uma variável x pode variar em diferentes níveis de outra variável x
- Contém o produto cruzado de dois termos

- $$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$
$$= b_0 + b_1x_1 + b_2x_2 + b_3(x_1x_2)$$

Efeito de Interação

- Dado:

$$\begin{aligned} Y &= \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1 \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \end{aligned}$$

Sem o termo de Interação, o efeito de X_1 sobre Y é medido por β_1

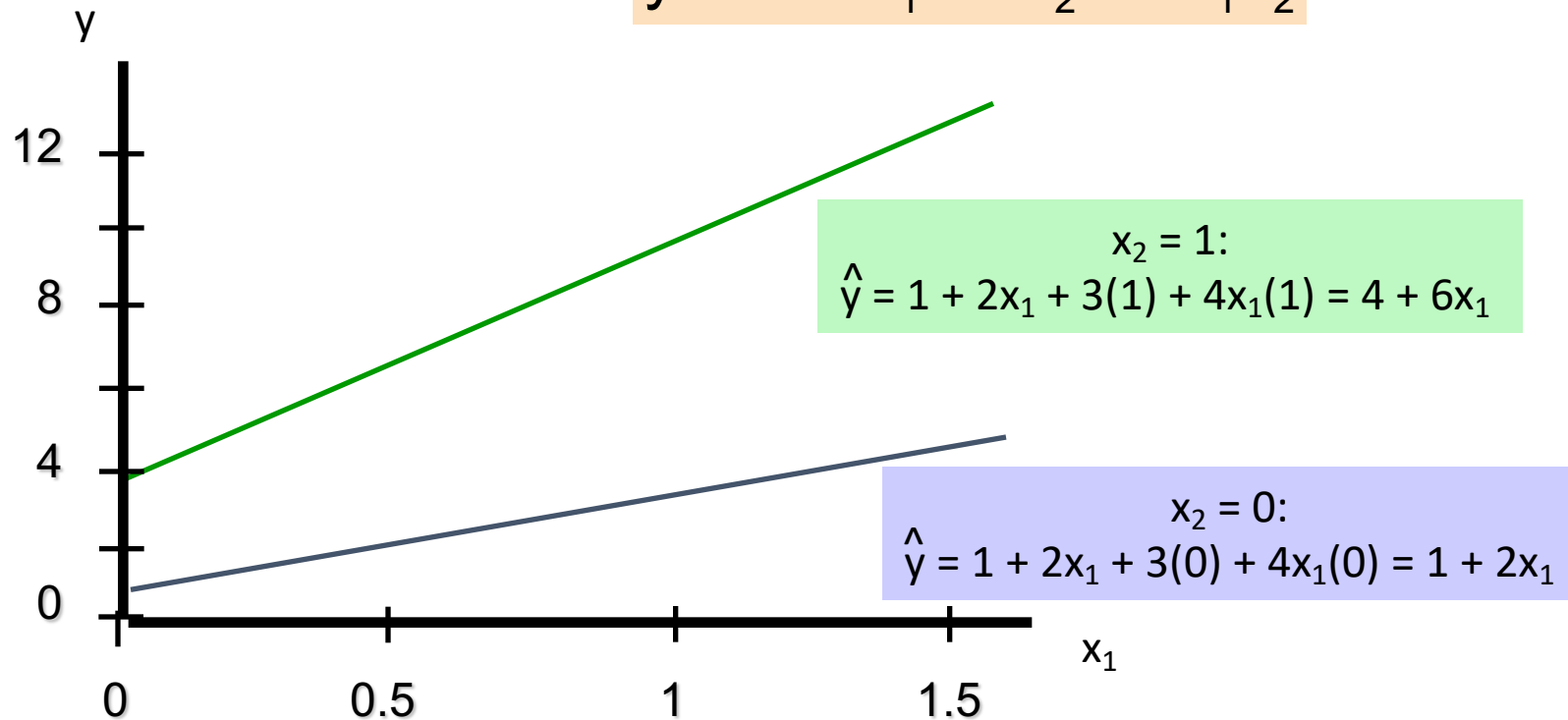
Com o termo de Interação, o efeito de X_1 sobre Y é medido por $\beta_1 + \beta_3 X_2$

Efeito das mudanças quando X_2 muda é $\beta_2 + \beta_3 X_1$

Exemplo de Interação

Suponha que x_2 é uma variável dummy e a equação da regressão estimada é

$$\hat{y} = 1 + 2x_1 + 3x_2 + 4x_1x_2$$



As inclinações são diferentes de o efeito de x_1 sobre y depende do valor de x_2

Significância do termo de interação

- O coeficiente b_3 é uma estimativa da diferença no coeficiente de x_1 quando $x_2 = 1$ comparado com quando $x_2 = 0$
- A estatística t para b_3 pode ser usada para testar a hipótese

$$H_0 : \beta_3 = 0 \mid \beta_1 \neq 0, \beta_2 \neq 0$$

$$H_1 : \beta_3 \neq 0 \mid \beta_1 \neq 0, \beta_2 \neq 0$$

- Se rejeitarmos a hipótese nula, concluímos que há uma diferença no coeficiente de inclinação para os dois subgrupos

Pressupostos do modelo de regressão múltipla

Erros (resíduos) do modelo de regressão múltipla

$$e_i = (y_i - \hat{y}_i)$$

Pressupostos:

- Os erros são normalmente distribuídos
- Os erros têm uma variação constante
- Os erros do modelo são independentes

Análise de resíduos em regressão múltipla

- Esses gráficos residuais são usados em regressão múltipla :
 - Resíduos vs. \hat{y}_i
 - Resíduos vs. x_{1i}
 - Resíduos vs. x_{2i}
 - Resíduos vs. tempo (se for dados de série de tempo)

Use os gráficos residuais para verificar as violações das suposições de regressão