Esse material foi extraído de Barbetta (2007 – cap 13)

Regressão linear simples

A análise de regressão é geralmente feita sob um referencial teórico que justifique a adoção de alguma relação matemática de causalidade.

Variável independente ou
Variável explicativa

Variável dependente ou
Variável resposta

- Predizer valores de uma variável dependente (Y) em função de uma variável independente (X).

- Conhecer o quanto variações de X podem afetar Y.

Exemplos de aplicações do modelo de regressão linear simples

Variável independente (*X*)	→	Variável dependente (*Y*)
Renda	→	Consumo (R$)
Gasto com o controle da qualidade (R$)	→	Número de defeitos nos produtos
Memória RAM do computador (Gb)	→	Tempo de resposta do sistema (segundos)
Área construída do imóvel (m^2)	→	Preço do imóvel (R$)

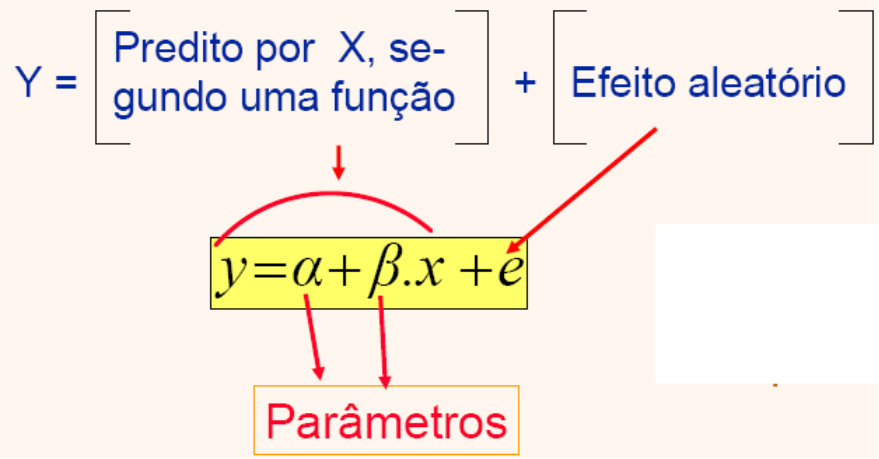
Análise de regressão

Amostra de observações (X, Y)

⇩

Conhecer o relacionamento entre X e Y

Modelo de regressão linear simples

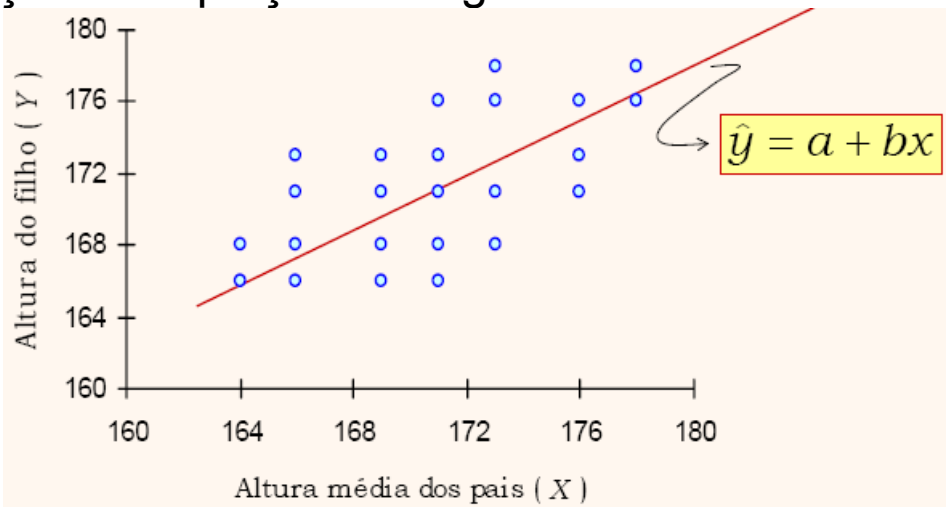


Pressupostos do modelo de regressão

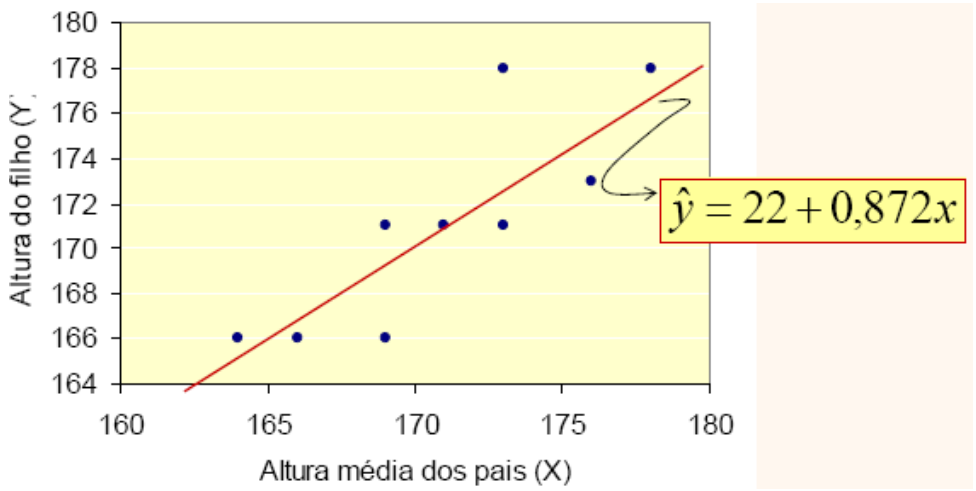
Os erros (e's) são independentes e variam aleatoriamente segundo a distribuição normal com $\mu = 0$ e σ^2 constante.

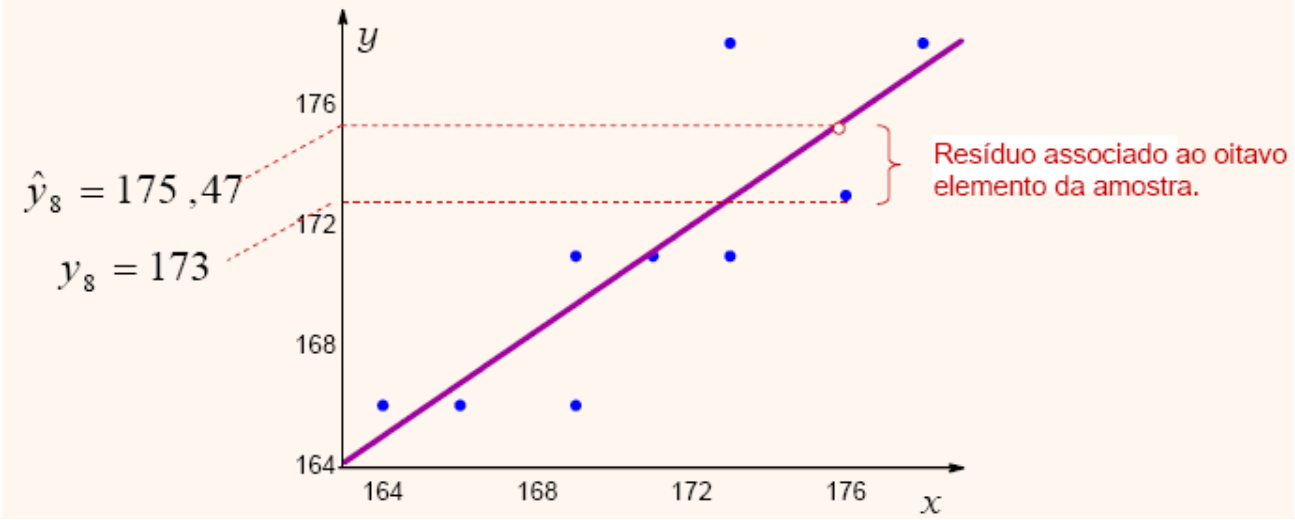
Estimativa dos parâmetros α e β

Construção da equação de regressão com base nos dados:

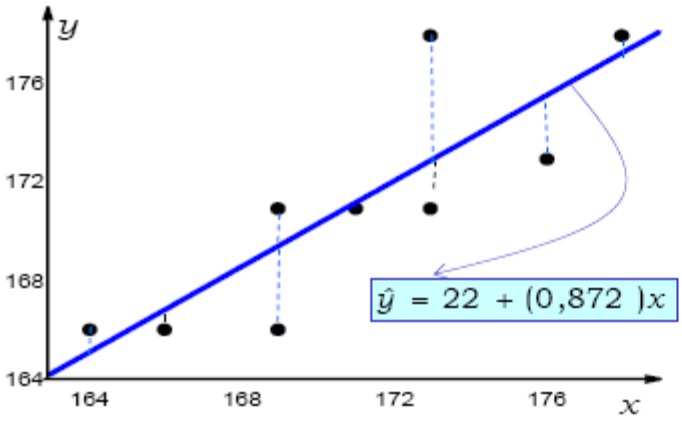


Estimativa dos parâmetros α e β





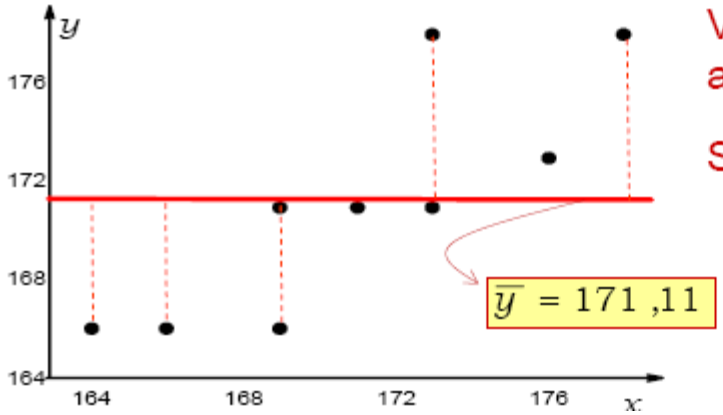
Variação explicada e não-explicada



Variação explicada
pelo modelo de regressão

Soma de quadrados devida ao erro aleatório:

$$SQE = \sum (y - \hat{y})^2$$



Variação em relação à média aritmética (variação total)

Soma de quadrado total:

$$SQT = \sum (y - \overline{y})^2$$

Exemplo 13.6 (BARBETTA, 2007) O anexo do capítulo 13 contém dados relativos a cinqüenta apartamentos da cidade de Criciúma – SC. Com o objetivo de construir um modelo para subsidiar a atualização dos valores dos tributos municipais, vamos realizar uma regressão entre valor (*Y*), em milhares de reais, e área privativa (*X*), em m^2.

```
Call:
lm(formula = Valor ~ Area, data = apto)

Residuals:
     Min       1Q   Median       3Q      Max
-132.936  -22.522   -2.683   16.082  140.491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.5736    14.6601  -4.405 5.91e-05 ***
Area          1.6658     0.1289  12.920  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.32 on 48 degrees of freedom
Multiple R-squared: 0.7767, Adjusted R-squared: 0.772
F-statistic: 166.9 on 1 and 48 DF,  p-value: < 2.2e-16
```

O R^2 (*R-quadrado*) igual a 0,777. Este resultado indica que na amostra, cerca de 78% da variação do valor de venda do apartamento pode ser *explicada* por uma relação linear com a área privativa. Os demais 22% são a parcela da variação provocada por outros fatores não incluídos no modelo de regressão. Essa parte aleatória tem erro padrão (*Se*) estimado igual a 43,32 mil reais.

A análise de variância (ANOVA) do modelo é o resultado de um teste estatístico para as hipóteses:
H_0: não existe relação linear entre *X* e *Y*; e
H_1: a relação linear entre *X* e *Y* é significativa (não é mero resultado do acaso).
O teste, conhecido como teste *F* do modelo, resultou em *F* = 166,9, com correspondente valor *p* = (o número de * é interpretado como legenda para os respectivos níveis de significância). Com esses valores de *p*, o teste estatístico rejeita H_0, indicando que a área privativa do apartamento (*X*) é significativa para *explicar* o seu preço (*Y*).

ANÁLISE DOS RESÍDUOS E TRANSFORMAÇÕES

Na seção anterior, estabelecemos um modelo para um conjunto de observações (*x, y*), relativo às variáveis *X* e *Y*, da forma

$$y = \alpha + \beta x + \varepsilon$$

Onde α e β são parâmetros estimados com os dados e ε representa o *erro aleatório*. Ou seja, estamos assumindo que *X* causa *Y* através de uma relação linear e toda a variação em torno dessa relação deve-se ao efeito do erro aleatório. Além disso, para a validade dos intervalos de confiança e testes estatísticos discutidos no Exemplo 13.6, é necessário supor que as observações de *Y* sejam independentes, e o termo de erro tenha distribuição aproximadamente normal com média nula e variância constante. Apresentaremos um processo gráfico para verificar se estas suposições podem ser válidas e, caso contrário, o que pode ser feito para adequar o modelo.

Um primeiro gráfico pode ser feito antes da análise de regressão. É o diagrama de dispersão, conforme discutido na Seção 13.1.

Tabela 13.1 Alguns dados, baseados no Censo Demográfico de 2000, de uma amostra aleatória de municípios brasileiros.

Município	DistCap	EspVida	MortInf	Alfab	Renda
Araruna (PR)	365	67,99	23,19	86,23	188,29
Nova Redenção (BA)	278	61,19	56,56	63,00	74,79
Monção (MA)	150	59,58	63,32	63,64	66,96
Porto Rico do Maranhão (MA)	78	58,96	66,05	79,33	65,34
Campo Erê (SC)	468	68,10	31,71	83,38	173,38
Lagoa do Piauí (PI)	40	63,65	47,08	65,81	60,00
São José das Palmeiras (PR)	486	71,01	16,62	77,54	150,67
Paraíba do Sul (RJ)	83	71,36	15,69	89,28	264,55
Malhada dos Bois (SE)	65	64,46	44,18	69,95	80,69
Jandaíra (BA)	175	62,45	51,57	59,72	58,68
Vespasiano (MG)	14	68,68	32,81	90,43	196,51
Ipaba (MG)	167	67,42	37,04	81,82	125,75

Fonte: Atlas de Desenvolvimento Humano (www.pnud.org.br/atlas).
Descrição das variáveis:
DistCap: distância da capital da respectiva Unidade da Federação.
EspVida: esperança de vida ao nascer
MortInf: mortalidade (número médio de mortes em 1.000) até um ano de idade.
Alfab: taxa de alfabetização (percentagem da população adulta alfabetizada).
Renda: renda *per capita* do município (R$).

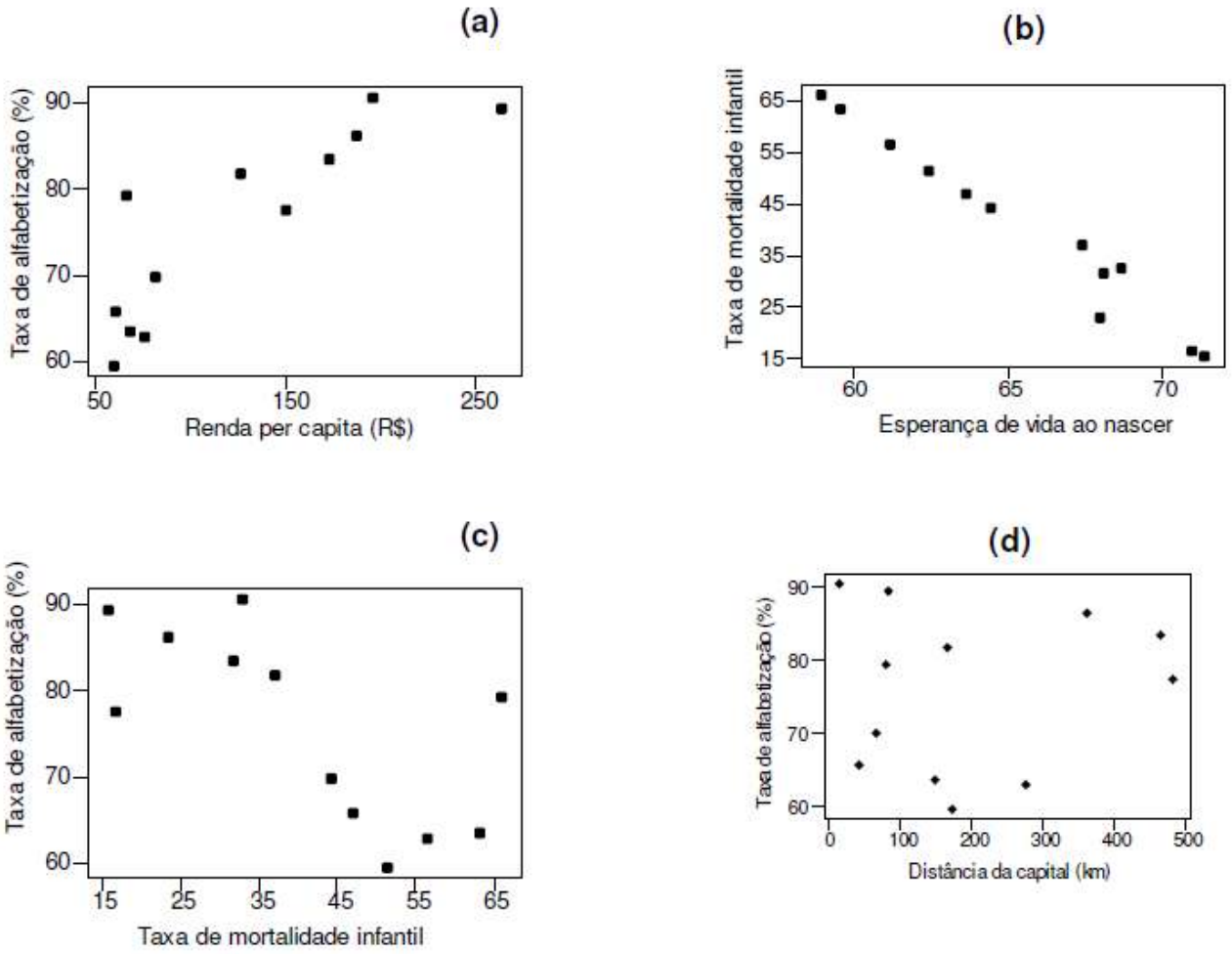


Figura 13.2 Alguns diagramas de dispersão, construídos com os dados Tabela 13.1.

Por esse gráfico, podemos verificar se a função linear é adequada para representar a forma estrutural entre *X* e *Y*. Veja o gráfico à esquerda da Figura 13.15.

Após a estimação dos parâmetros do modelo, podemos calcular os *resíduos* do modelo ajustado aos dados. O resíduo é calculado para cada observação, e definido como a diferença entre o valor observado *y* e o valor *predito* $\hat{y}$. Ou seja, *resíduo* = $y - \hat{y}$

Um gráfico apresentando os pares (*x*, *resíduo*) é bastante útil na avaliação do modelo de regressão. Veja o gráfico à direita da Figura 13.15.

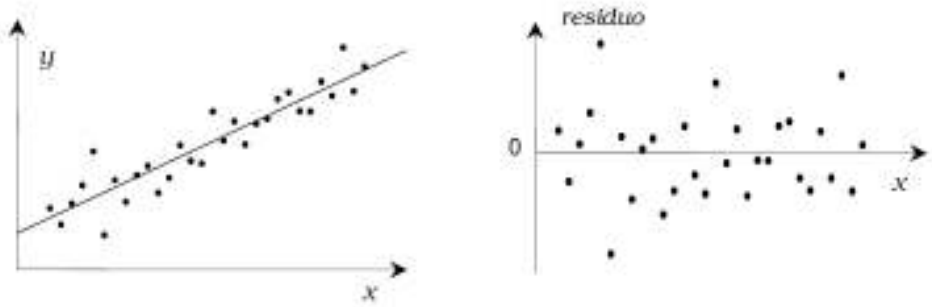


Figura 13.15 Gráficos para verificar a adequação do modelo.

Os gráficos da Figura 13.15 indicam uma situação em que as suposições do modelo estão aparentemente satisfeitas, pois os resíduos apresentam-se distribuídos de forma aleatória e razoavelmente simétrica em torno da reta de regressão. No gráfico dos resíduos, a reta de regressão corresponde à linha horizontal sobre o valor zero.

A Figura 13.16 apresenta uma situação em que temos um ponto discrepante. Esse ponto é visível nos dois gráficos, mas no gráfico dos resíduos ele aparece mais nitidamente. Seja:

$$resíduo\ padronizado = \frac{y - \hat{y}}{S_e}$$

Supostamente, os resíduos padronizados devem seguir uma distribuição normal padrão, pelo menos aproximadamente. Então, em torno de 95% dos valores devem estar entre 2 ou -2 (Capítulo 8). Fora deste intervalo, são casos suspeitos de serem discrepantes. Assim, o uso de resíduos padronizados é melhor para detectar pontos discrepantes.

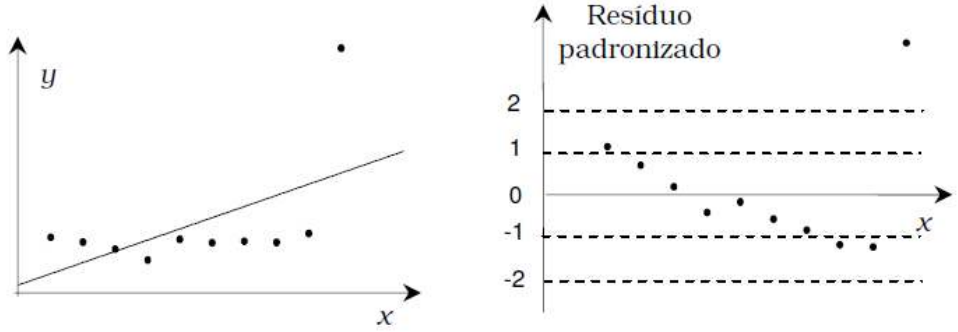


Figura 13.16 Gráficos indicando a presença de um valor discrepante.

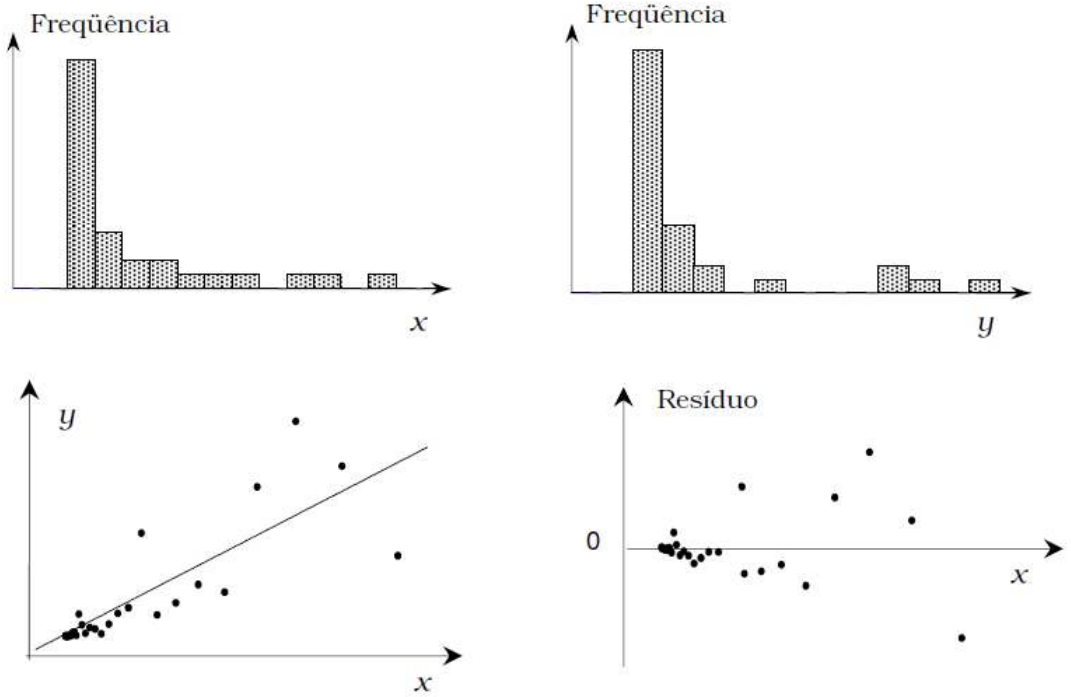


Figura 13.17 Gráficos indicando distribuições assimétricas de X e Y, além da variância de Y ser maior para valores maiores de X e Y.

Nesta situação, os valores grandes de X vão ter mais peso na determinação da inclinação da reta. Neste caso, recomendamos a aplicação da transformação logarítmica, tanto nos valores de X como nos valores de Y, estabelecendo o seguinte modelo: [12]

$$\log(y) = \alpha + \beta \cdot \log(x) + \varepsilon$$

A transformação logarítmica aumenta as distâncias entre os valores pequenos e reduz as distâncias entre os valores grandes, tornando distribuições assimétricas de cauda longa à direita em distribuições mais simétricas. Com isso, temos uma situação mais adequada para estabelecer a reta de regressão. Em termos computacionais, devemos:

a) calcular o logaritmo natural de cada valor x e de cada valor y;

b) aplicar a análise de regressão linear sobre os dados transformados [$\log(x)$, $\log(y)$]; e

c) construir novamente o gráfico de resíduos para verificar a adequação das suposições neste novo modelo.

---

[12] É comum usar o logaritmo natural ou na base 10. Outra transformação que se presta ao mesmo propósito é a raiz quadrada. Esta segunda transformação é usada nas situações em que a inadequação do modelo não aparece de forma tão forte como visto na Figura 13.17. Observamos que estas transformações são possíveis somente quando todos os valores são positivos.

A Figura 13.18 apresenta uma situação que sugere relação *não-linear*, com Y crescendo rapidamente para valores pequenos de X, e crescendo lentamente para valores grandes de X. É uma situação em que recomendamos uma transformação logarítmica (ou raiz quadrada) somente nos valores da variável X, ou seja, passamos a considerar o seguinte modelo para os dados:

$$y = \alpha + \beta \cdot \log(x) + \varepsilon$$

Note que esse modelo pode ser considerado linear em termos das variáveis $\log(x)$ e y (não mais entre x e y). Em termos computacionais, devemos:

a) calcular o logaritmo de cada valor x;

b) aplicar a análise de regressão linear sobre os dados [$\log(x)$, y]; e

c) construir novamente o gráfico de resíduos para verificar a adequação das suposições nesse novo modelo.

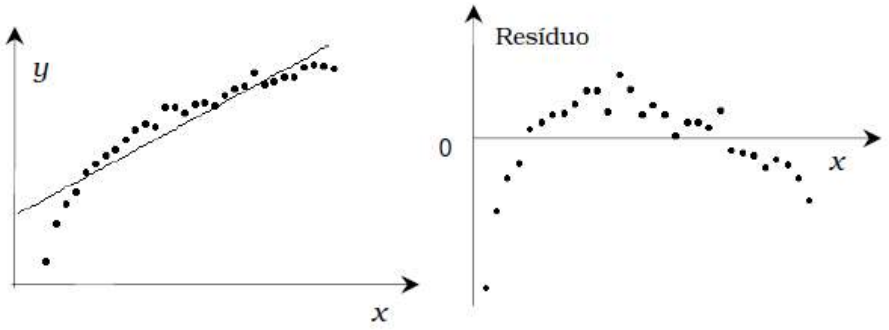


Figura 13.18 Gráficos indicando uma relação *não-linear*, aparentemente logarítmica.

A Figura 13.19 apresenta uma situação com os seguintes problemas: (1) relação *não-linear* para a parte estrutural do modelo e (2) aumento da variância à medida que X aumenta. Recomendamos uma transformação logarítmica nos valores da variável Y, ajustando o seguinte modelo aos dados:

$$\log(y) = \alpha + \beta x + \varepsilon$$

Para ajustar o modelo, devemos:

a) calcular o logaritmo de cada valor y;

b) aplicar a análise de regressão linear sobre os dados $[x, \log(y)]$; e

c) construir novamente o gráfico de resíduos para verificar se o novo modelo é mais adequado aos dados.

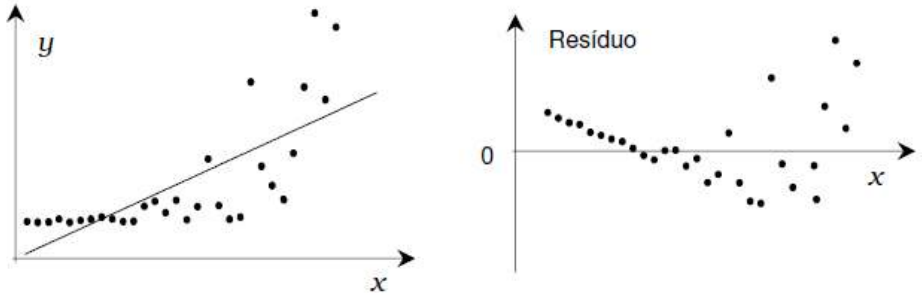


Figura 13.19 Gráficos indicando uma relação *não-linear* – aparentemente exponencial – e variância não-constante.

O uso de transformações auxilia o pesquisador a encontrar um modelo mais adequado para os dados, ainda que utilizando as expressões da regressão linear.

Exemplo 13.6 (continuação) Na seção anterior foi realizada uma regressão do valor de um imóvel (Y) com relação a sua área privativa (X), considerando uma amostra de cinqüenta apartamentos, apresentada no anexo deste capítulo. A Figura 13.20 apresenta a reta de regressão e o gráfico dos resíduos desse modelo.

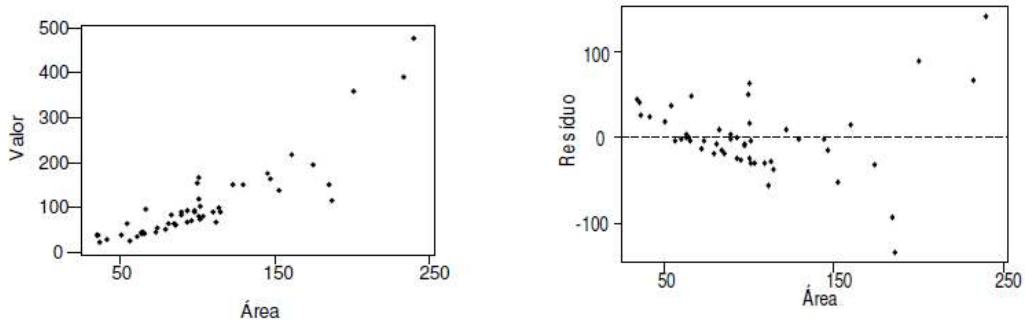


Figura 13.20 Gráficos de dispersão e dos resíduos (Exemplo 13.6).

Observamos na Figura 13.20 uma predominância de valores pequenos com respeito às duas variáveis. Isto era esperado porque são mais comuns apartamentos pequenos (área e preço pequenos) do que apartamentos grandes (área e preço grandes). Também podemos observar maior variabilidade nos apartamentos mais caros. Essas condições sugerem tentarmos uma transformação logarítmica em X e em Y. Assim, foi aplicado o logaritmo natural em cada um dos cinqüenta valores de X e Y. Por exemplo, o primeiro apartamento da amostra tem x = 96 m² e y = 69 mil reais. Aplicando o logaritmo natural, encontramos:

$$\log(x) = \log(96) = 4{,}56 \quad \text{e} \quad \log(y) = \log(69) = 4{,}23$$

A análise com os dados transformados produziu os gráficos de dispersão e de resíduos apresentados na Figura 13.21.

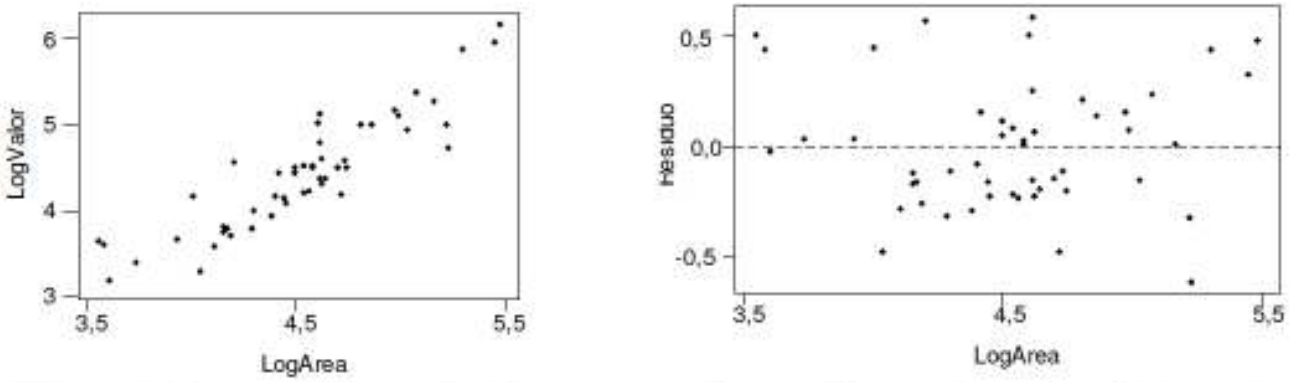


Figura 13.21 Gráficos de dispersão e dos resíduos (Exemplo 13.6), após transformações nas variáveis.

Verificamos pela Figura 13.21 que, após as transformações, as condições básicas do modelo estão aparentemente satisfeitas. A equação de regressão, obtida com apoio de um sistema computacional para análise estatística é:

$$Predição\ de\ \log(y) = -1{,}58 + (1{,}33) \cdot \log(x)$$

com $R^2 = 0{,}813$ e $S_e = 0{,}294$. Observar que o poder explicativo deste modelo é melhor que o anterior (81,3% contra 77,7%). Já o S_e não é comparável devido a transformação de escala.

Para predizer o valor de um apartamento com área privativa de 100 m², devemos, primeiramente, transformar este valor na escala logarítmica:

$$x = 100 \rightarrow \log(x) = 4{,}605$$

Aplicar o modelo de regressão:

$$Predição\ de\ \log(y) = -1{,}58 + (1{,}33) \cdot (4{,}605) = 4{,}545$$

Efetuar a transformação inversa do logaritmo:

$$\hat{y} = \exp\{4{,}545\} = 94{,}15$$

Assim, por este novo modelo, o apartamento valeria R$ 94.150,00. ▬

13.6 INTRODUÇÃO À REGRESSÃO MÚLTIPLA

Em geral, uma variável dependente (ou resposta) Y depende de várias variáveis independentes ou explicativas (X_1, X_2, ..., X_k). Na análise de regressão múltipla, vamos construir um modelo estatístico-matemático para se estudar, objetivamente, a relação entre as variáveis independentes e a variável dependente e, com o modelo construído, conhecer a influência de cada variável independente, como também, predizer a variável dependente em função do conhecimento das variáveis independentes. O Quadro 13.2 ilustra alguns exemplos.

Quadro 13.2 Aplicações do modelo de regressão múltipla.

Variáveis independentes (X_1, X_2, ..., X_k)	→	Variável dependente (Y)
X_1 = altura do pai (cm) X_2 = altura da mãe (cm) X_3 = sexo (1 = homem, 0 = mulher)	→	Y = altura de um indivíduo (cm)
X_1 = renda (R$) X_2 = poupança (R$) X_3 = taxa de juros (%)	→	Y = Consumo (R$)
X_1 = área construída do imóvel (m²) X_2 = idade (anos) X_3 = localização	→	Y = preço do imóvel (R$)
X_1 = memória RAM (Gb) X_2 = sistema operacional X_3 = tipo de processador	→	Y = tempo de resposta do sistema computacional (segundos)

Para estabelecer o modelo clássico de regressão múltipla, consideraremos que Y seja uma variável quantitativa contínua e X_1, X_2, ..., X_k sejam variáveis quantitativas ou indicadoras de certos atributos. A variável indicada deve ter valor 1 quando o atributo está presente; e 0 quando não está presente. Por exemplo, a variável X_3 = *localização do imóvel* pode ter valor 1 quando o imóvel estiver numa área valorizada, e 0 quando estiver numa área pouco valorizada. Também será considerado que Y é uma variável aleatória, isto é, somente será conhecida após a observação do elemento (indivíduo, imóvel, etc.), enquanto X_1, X_2, ..., X_k também podem provir de observação ou serem estabelecidas *a priori*.

A análise de regressão múltipla parte de um conjunto de observações (x_1, x_2,..., x_k, y), relativas às variáveis X_1, X_2, ..., X_k e Y. Diremos que um dado valor y depende dos correspondentes valores x_1, x_2,..., x_k, mas também de uma infinidade de outros fatores não incluídos no modelo, que serão representados por ε (*erro aleatório*). Mais especificamente, supomos o seguinte modelo para as observações:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

onde α, β_1 , β_2 , ... , β_k são parâmetros a serem estimados com os dados e ε representa o *erro aleatório*, cujo desvio padrão também pode ser estimado pelos dados. As suposições são análogas às suposições da regressão simples, acrescentando que as variáveis independentes X_1, X_2, ..., X_k não devem ter correlações altas entre si.

Exemplo 13.7 Voltando à questão de construir um modelo para o valor de um apartamento (Y) com os dados do anexo deste capítulo. Sejam as variáveis independentes:

X_1 = área comum do apartamento (m²);

X_2 = idade (anos);

X_3 = consumo de energia elétrica do morador (Kw/mês) e

X_4 = localização (1= área valorizada; 0 = área pouco valorizada).

Como discutimos no Exemplo 13.6, as variáveis Y e X_1 serão analisadas na escala logarítmica. A variável X_3 está sendo usada como uma *proxi* do padrão de vida do morador do apartamento e, por sua vez, da qualidade do apartamento. Temos o seguinte modelo teórico para os dados:

$$\log(y) = \alpha + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

```
lm(formula = ln_Valor ~ Energia + Idade + ln_Area + Local, data = ZX)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37240 -0.18794 -0.03707  0.16261  0.51185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.206678   0.376375  -3.206  0.00248 **
Energia      0.002419   0.001610   1.503  0.13982
Idade       -0.024816   0.005358  -4.632  3.1e-05 ***
ln_Area      1.195395   0.083900  14.248  < 2e-16 ***
Local        0.077575   0.075708   1.025  0.31100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2341 on 45 degrees of freedom
Multiple R-squared: 0.8888, Adjusted R-squared: 0.8789
F-statistic: 89.94 on 4 and 45 DF,  p-value: < 2.2e-16
```

Observamos o valor de R^2 (*Rquadrado*) igual a 0,8888 e o erro padrão (*Se*)= 0,2341. Comparando com os resultados do Exemplo 13.6 (R^2 = 0,813 e *Se* = 0,294), vemos melhora no modelo com a inclusão das variáveis: idade, gasto de energia elétrica e localização. O valor R^2 = 0,889, indica quase 90% da variação do logaritmo do valor de um apartamento pode ser *explicado* por uma relação linear que

envolve logaritmo da área comum (X_1), idade (X_2), consumo de energia elétrica do morador (X_3) e dois níveis de localização (X_4).

O *teste F do modelo* resultou na estatística F = 89,94, com correspondente valor *p* extremamente pequeno (menor que um milésimo). Assim, o teste estatístico rejeita H_0, indicando que as variáveis independentes escolhidas são significativas para *explicar* a variável dependente.
A primeira coluna apresenta as estimativas dos coeficientes, de onde podemos extrair a seguinte equação:

$$\textit{Predição de } \log(y) = -1{,}208 + 1{,}195 \cdot \log(x_1) - 0{,}025x_2 + 0{,}0024x_3 + 0{,}076x_4$$

Assim, tendo a área do apartamento (x_1), a idade (x_2), o consumo de energia elétrica (x_3) e a localização (x_4) podemos obter uma predição de seu valor. Por exemplo, um apartamento com 100 m², que tenha 5 anos de uso, morador consumindo 200 Kw e localização em área valorizada, temos:

---

14 Cabe observar que o teste estatístico refere-se à população, ou seja, quando se tem uma amostra muito pequena, podemos obter um valor alto de R^2 e o teste aceitar H_0.

$$\text{Predição de } \log(y) = -1{,}208 + 1{,}195 \cdot \log(100) - (0{,}025) \cdot 5 + (0{,}0024) \cdot 200 + (0{,}076) \cdot 1$$

ou: Predição de $\log(y) = 4{,}726$. Portanto: $\hat{y} = \exp(4{,}726) = 112{,}84$

ou, seja, valor estimado de R$ 112.840,00.

Devemos observar que os sinais dos coeficientes do modelo construído estão coerentes. Coeficiente de X_1 positivo, isto é, quanto maior o apartamento, maior deverá ser o seu valor; coeficiente de X_2 negativo (quanto mais velho, menor o valor); coeficiente de X_3 positivo (quanto maior o consumo de energia do morador, maior o valor); e coeficiente de X_4 positivo (em área valorizada, maior o valor).