

Diagnóstico em Regressão

Rejane Sobrino Pinheiro
Tania Guillén de Torres

Diagnósticos em Regressão

Introdução

- ◆ Conjunto de ferramentas para análise dos resíduos, detecção de *valores extremos (outliers)*, de *pontos influentes (alavanca)* e para avaliação de colinearidade.

Análises iniciais

- ◆ Importante conhecer algumas características básicas dos dados:
 - ✓ O tipo de unidade de análise (homens idosos, lâminas etc.)
 - ✓ O procedimento de coleta dos dados
 - ✓ A unidade de medida de cada variável
 - ✓ O intervalo razoável e o valor típico de cada variável
- ◆ Estas informações, juntamente com a análise exploratória e com as estatísticas correspondentes, podem ser usadas para detecção de erros nos dados e de potenciais violações dos pressupostos.

Análises iniciais (cont...)

- ◆ **Estatística descritiva** - depende do tipo de variável (proporção, médias, medianas e outros percentis, etc.).
- ◆ Nas variáveis quantitativas, recomenda-se listar os **5 maiores e 5 menores** valores de cada variável. Embora simples, juntamente com o conhecimento dos 4 itens do primeiro parágrafo, favorece a detecção de erros nos dados ou da presença de *outliers*.
- ◆ Pode-se calcular também as estatísticas para alguns grupos importantes (ex: homens e mulheres) de dados de interesse.
- ◆ Interessante comparar os resultados obtidos com o que é esperado, dado o conhecimento científico sobre as diferentes variáveis.
- ◆ Estatísticas mais elaboradas também são úteis nesta etapa, incluindo correlação entre pares de variáveis e entre a variável resposta e as variáveis independentes. Exame de colinearidade.

Análises iniciais - Estatística descritiva

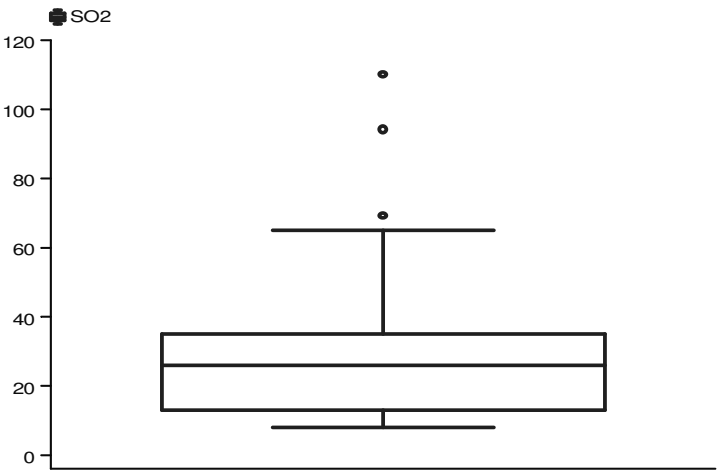
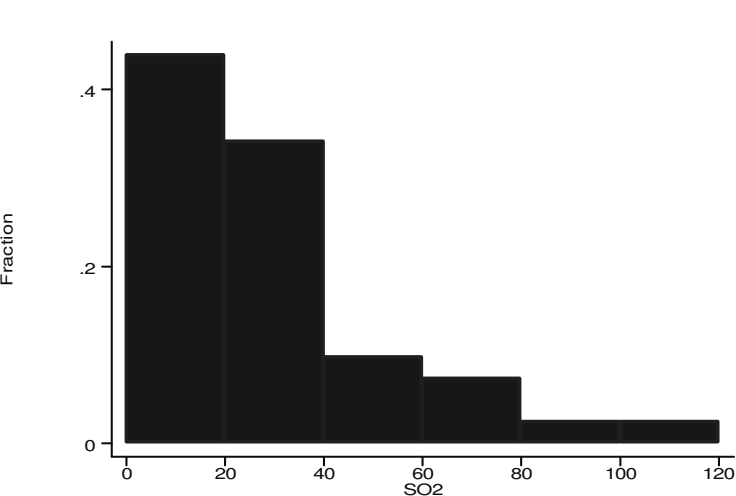
Variáveis Contínuas

. sum SO2- days, detail

SO2

Percentiles			Smallest			
1%	8		8		Obs	41
5%	9		9		Sum of Wgt.	41
10%	10		9			
25%	13		10			
50%	26				Mean	30.04878
			Largest		Std. Dev.	23.47227
75%	35		65			
90%	61		69		Variance	550.9476
95%	69		94		Skewness	1.643887
99%	110		110		Kurtosis	5.521466

. graph SO2, xlab(0,20,40,60,80,100,120) ylab bin(6)



Análise exploratória de dados

- ❖ Diagramas de espalhamento entre variável resposta (dependente) e as independentes. Por exemplo Peso vs. Altura.

PESO vs. ALTURA

A afastado da nuvem com relação ao peso, mas no intervalo de plausibilidade de altura → potencial outlier

A = 1 observation, B = 2 observations, etc.

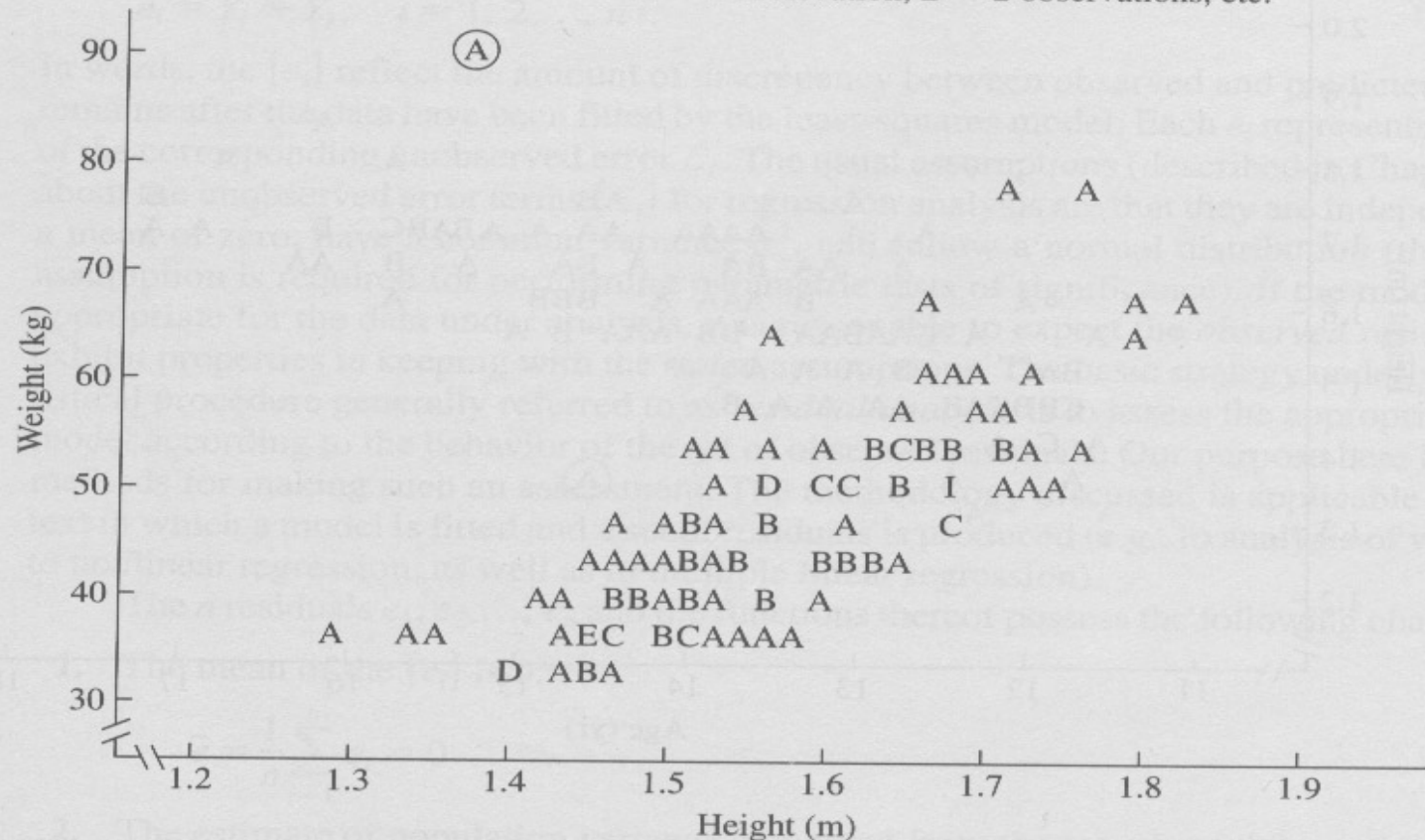
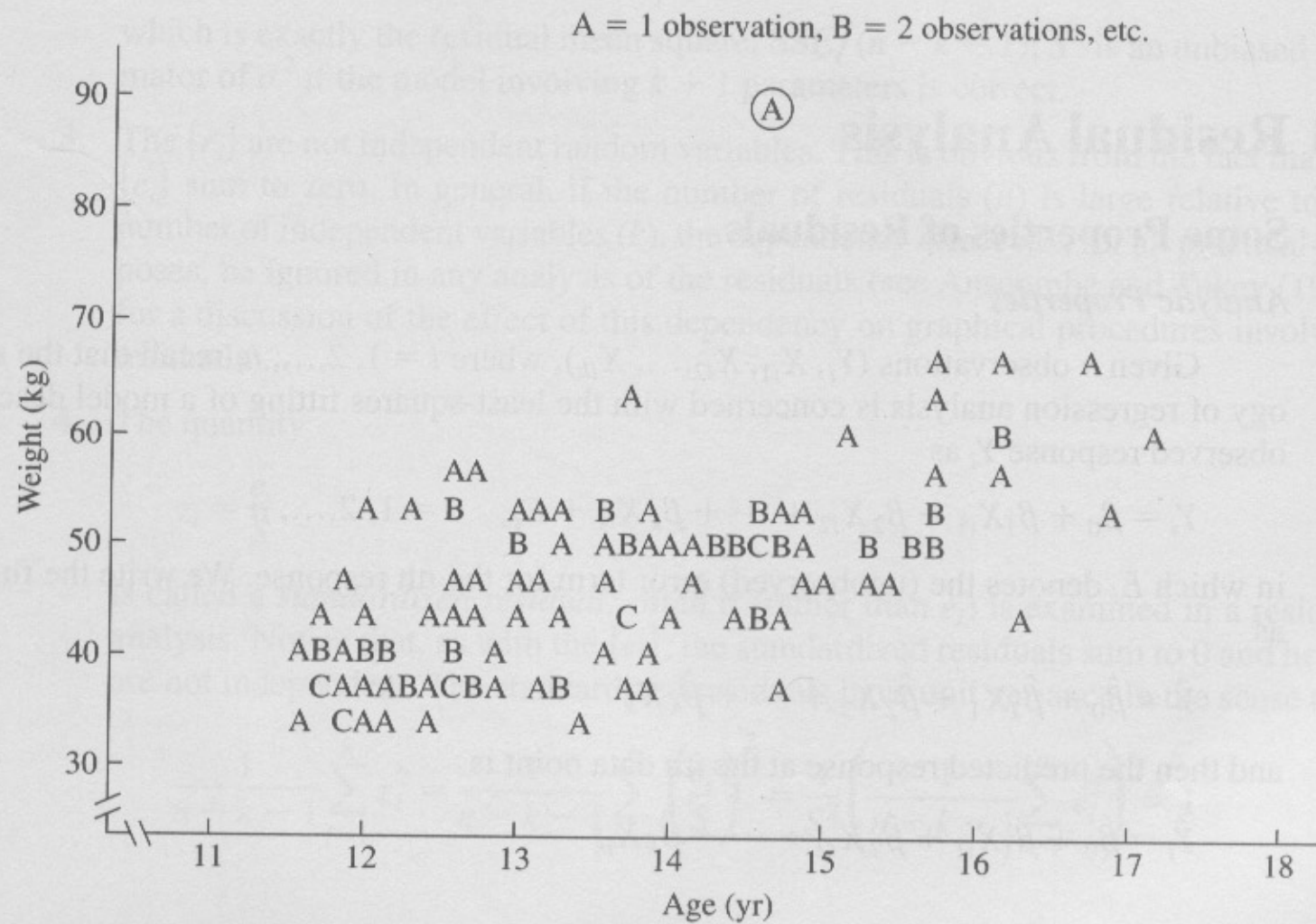


FIGURE 12-2 Children's body weight as a function of age; data from Lewis and Taylor (1967) ($n = 127$).

PESO vs. IDADE



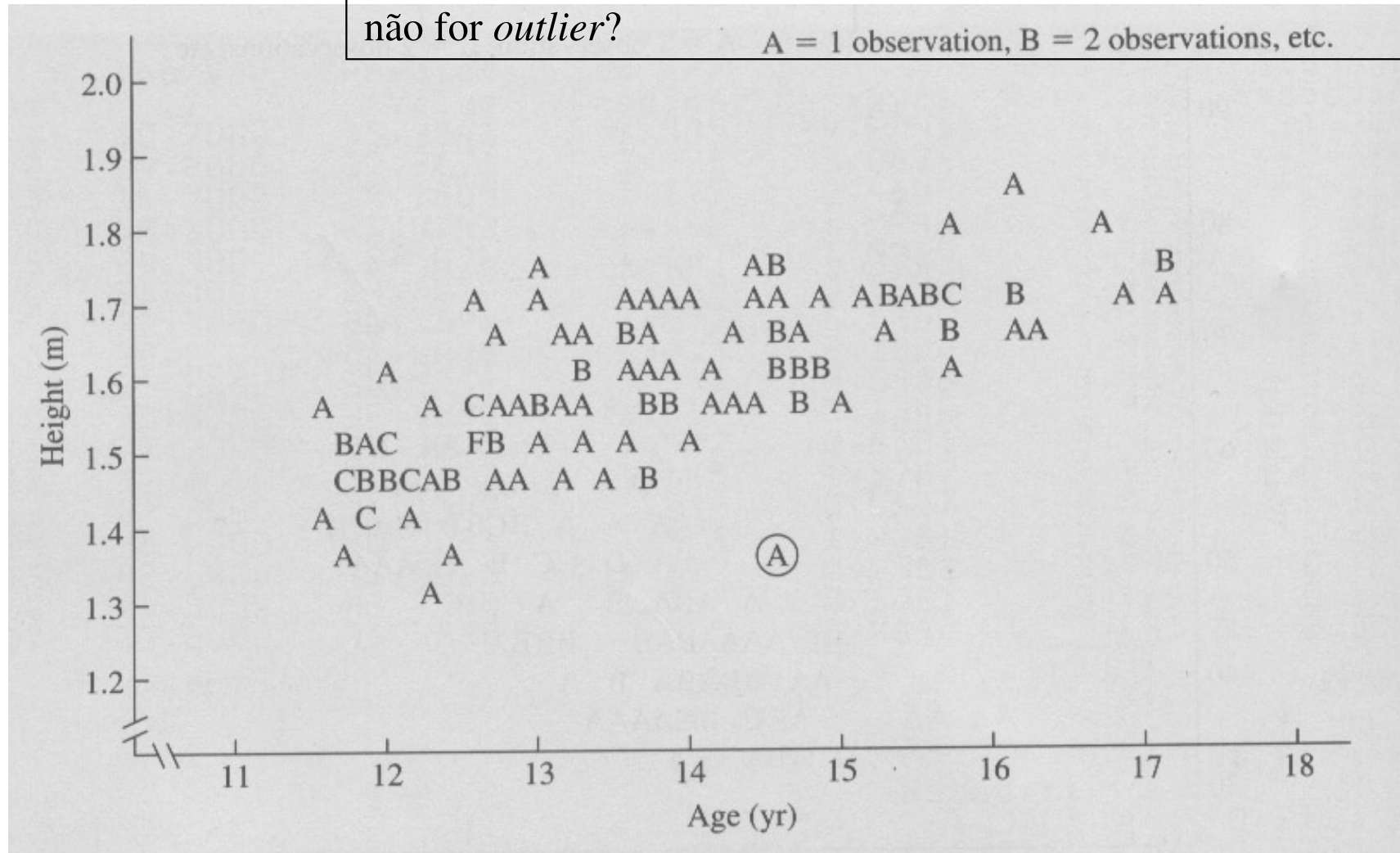
Análise exploratória de dados

cont.

- ❖ Diagrama de dispersão das variáveis independentes, para avaliar colinearidade (forte associação entre as variáveis independentes).

ALTURA vs. IDADE

A afastado da nuvem → potencial *outlier* ou pq está afastado, se não for *outlier*?
A = 1 observation, B = 2 observations, etc.



- ◆ Nos 3 gráficos aparece uma observação destoante.
- ◆ Duas interpretações possíveis para ela:
 - ✓ Erro de medida ou digitação
 - ✓ Valor está correto e o seu efeito na relação deve ser analisado.
- ◆ Não quer dizer que deva ser retirada (a princípio, mas deve-se ficar “de olho” nela).

Análise dos resíduos

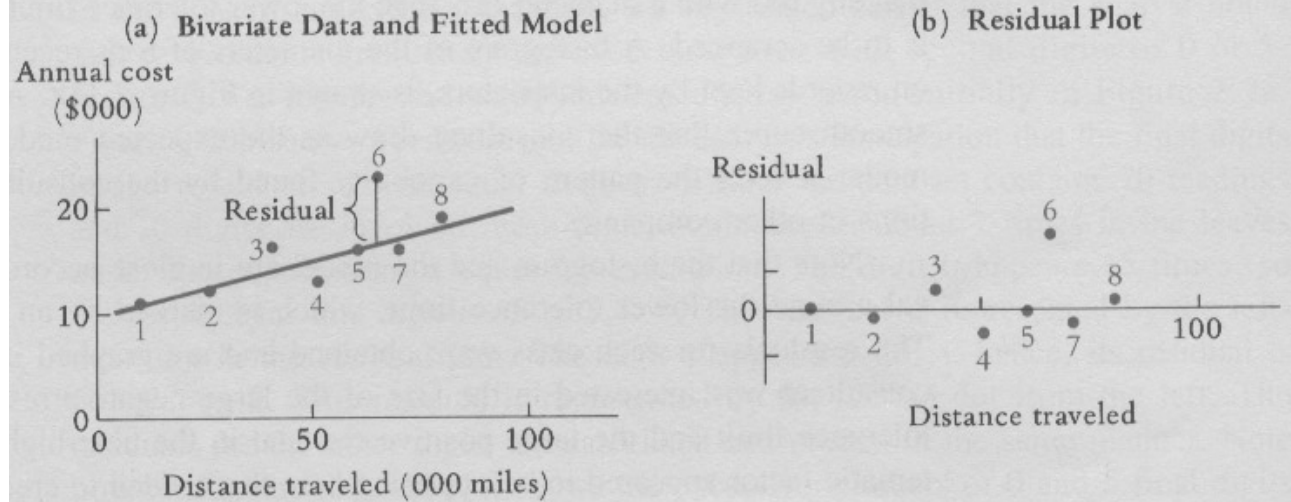
◆ Resíduo:

- ✓ diferença entre o valor observado (Y_i) e o valor predito pela regressão ($\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$)
- ✓ discrepância que permanece após os dados terem sido ajustados pelo modelo de mínimos quadrados.

$$\varepsilon_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \dots, n$$

Posição dos Pontos em torno da reta vs. Resíduos

FIGURE 2.12 Fitting a model to bivariate data, and the associated residual plot for the delivery fleet example



O ponto 6 está mais afastado do modelo ajustado (reta) → o resíduo é maior para este ponto.

❖ Análise dos resíduos pode ser usada para estudar a adequação do modelo linear, verificando se:

- ✓ A função de regressão é linear
- ✓ A distribuição de Y possui variância constante para todos os valores de X (homocedasticidade)
- ✓ Distribuição de Y é normal $\rightarrow \varepsilon_i$ é normal
- ✓ Os termos de erro ε_i são independentes.

Os pressupostos para o erro (resíduo) são:

- Distribuição normal (necessária para a execução de testes de hipóteses paramétricos)

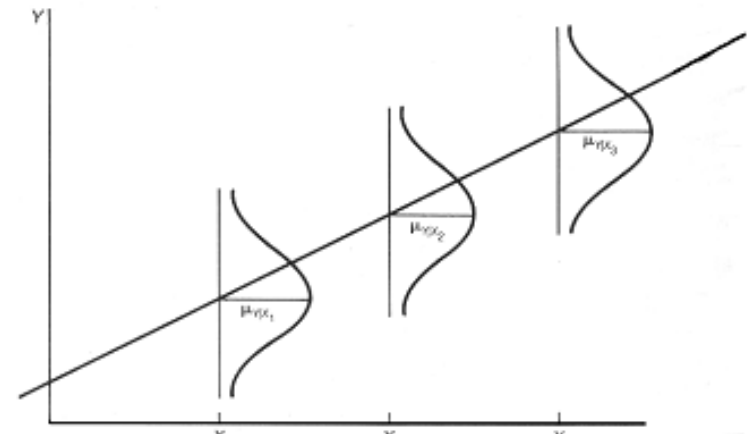
$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

- Média zero

$$\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i = 0$$

- Variância σ^2 constante

$$S_\varepsilon^2 = \frac{1}{n - k - 1} \sum_{i=1}^n \varepsilon_i^2$$

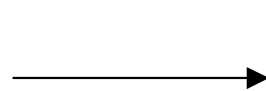


Padronização dos resíduos

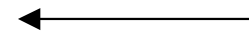
Qtos DP está afastado

Não subtraís μ , pois é 0

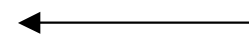
$$z_i = \frac{e_i - \mu}{S_e}$$



$$z_i = \frac{e_i}{S_e}$$



Resíduo original



Desvio padrão dos resíduos

- ◆ Resíduo padronizado tem variância 1 e média 0.

$$z_i \sim N(0,1)$$

$$\frac{1}{n-k-1} \sum_{i=1}^n Z_i^2 = \frac{1}{n-k-1} \sum_{i=1}^n \left(\frac{\varepsilon_i}{S} \right)^2 = \frac{1}{S^2} \left(\frac{1}{n-k-1} \sum_{i=1}^n \varepsilon_i^2 \right) = 1$$

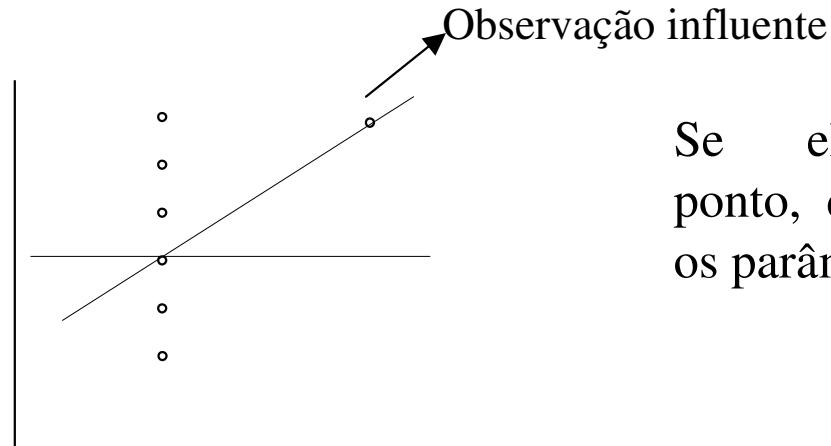
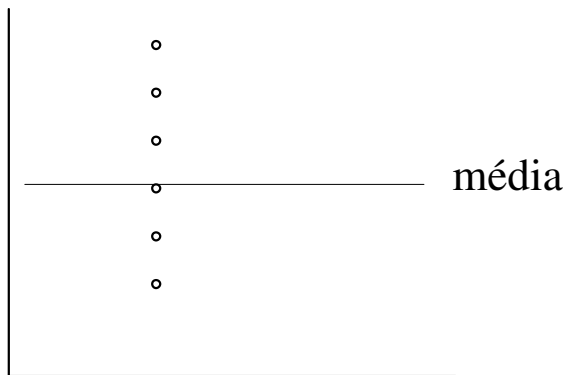
Resíduos studentizados

$$r_i = \frac{e_i}{S \sqrt{1 - h_i}} = \frac{z_i}{\sqrt{1 - h_i}}$$

$$0 \leq h_i \leq 1$$

$h_i \rightarrow$ alavanca - medida da importância da i -ésima observação em determinar o ajuste do modelo.

À medida que $h_i \rightarrow 1$, denominador $\rightarrow 0$ e $r_i \rightarrow \infty$



Se eliminarmos o ponto, qto que alteram os parâmetros?

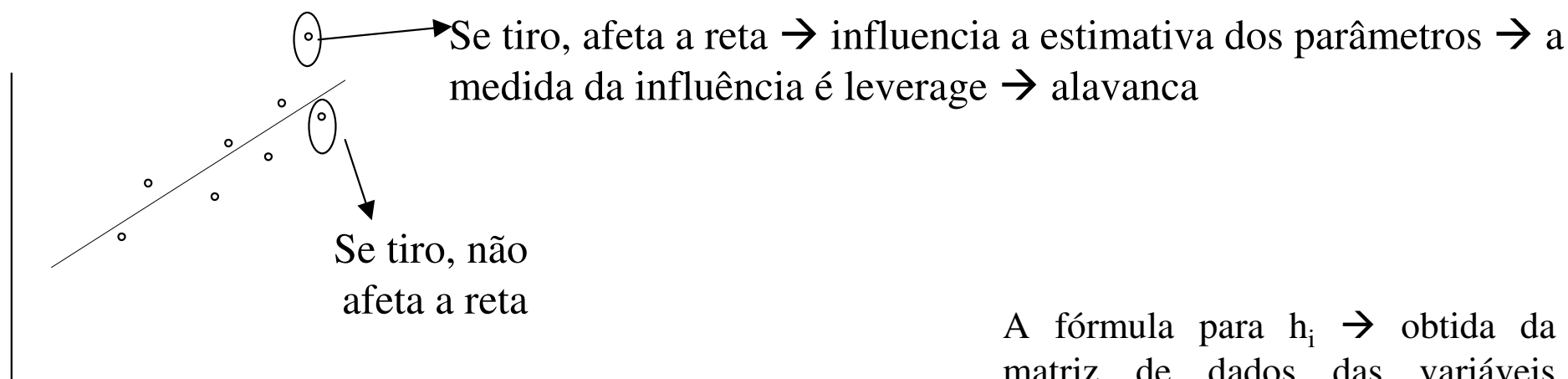
Resíduo Jackknife

$$r_{(-i)} = r_i \sqrt{\frac{S^2}{S_{(-i)}^2}} = \frac{e_i}{\sqrt{S_{(-i)}^2(1-h_i)}} = r_i \sqrt{\frac{(n-k-1)-1}{(n-k-1)-r_i^2}}$$

Variância usando todos os pontos

Variância tirando 1 ponto → se a diferença dos modelos, das variâncias (razão grande), o ponto é influente

- ❖ Valores altos de h_i ressaltam $r_{(-i)}$
- ❖ r_i segue aproximadamente uma distribuição de Student
- ❖ $r_{(-i)}$ segue uma distribuição de Student



A fórmula para h_i → obtida da matriz de dados das variáveis independentes → transposta

Análise Univariada dos Resíduos Padronizados , studentizados ou do tipo jackknife :

- Análise Descritiva: Utilizando os resíduos padronizados, espera-se que, caso sigam uma distribuição normal:
 - Metade deve ser negativa e metade positiva .
 - A média, mediana e a moda sejam “0”.
 - A variância é aproximadamente 1?
 - Aproximadamente 68% deles caiam entre os valores -1 e +1.
 - Aproximadamente 95% deles caiam entre os valores -2 e +2.
 - Aproximadamente 99% deles caiam entre os valores -3 e +3.
 - Comparar os maiores valores com os percentis p_{95} ou p_{99} e os menores valores com os percentis p_1 ou p_5 .
 - Valores absolutos maiores do que 2,5 ou 3 indicariam a presença de um possível *outlier*.

Coeficiente de assimetria:

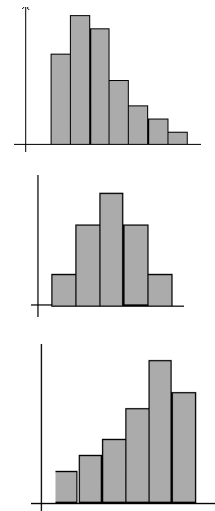
O **coeficiente de assimetria** (*skewness*) → descreve o alongamento horizontal da distribuição de frequência para um lado ou outro, de modo que uma cauda de observações é maior e tem mais observações do que a outra. Se uma distribuição é assimétrica, a média se desloca em direção da cauda alongada, mais do que a mediana, porque a média é mais fortemente influenciada por valores extremos.

O **coeficiente de assimetria** assume o valor zero quando a distribuição é simétrica, por exemplo a normal. Valores negativos para distribuições que apresentam uma cauda mais prolongada no lado esquerdo da distribuição, e valores positivos quando a distribuição apresenta uma cauda mais prolongada no lado direito da distribuição.

Coeficiente de assimetria (*skewness*)

- Desvios médios ao cubo (em relação à média)
- Mede o grau de assimetria
- Distribuição simétrica → **skewness = 0**

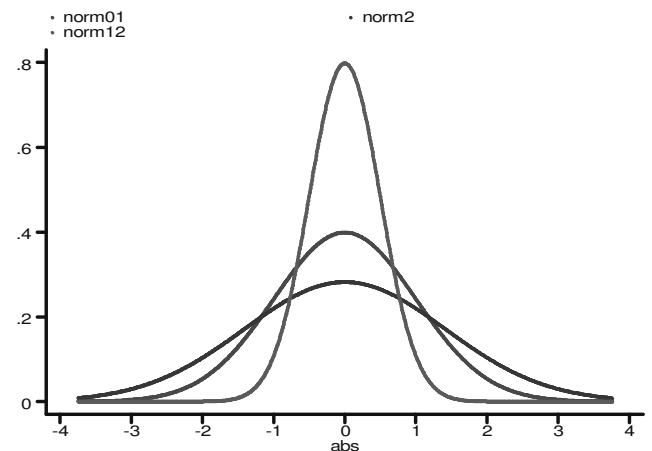
$$sk(e) = \left(\frac{n}{n-2} \right) \left(\frac{1}{n-1} \right) \sum_{i=1}^n \left(\frac{e_i - \bar{e}}{S_e} \right)^3, \quad sk(e) = \begin{cases} > 0 & \text{assimetria +} \\ = 0 & \text{simétrico} \\ < 0 & \text{assimetria -} \end{cases}$$



Kurtose:

A **kurtose** é caracterizada por um alongamento vertical da distribuição de frequência. Quanto menor for o coeficiente de kurtose, mais achatada será a distribuição. A distribuição normal tem um coeficiente de kurtose igual a 3.

$$kurt(e) = \left(\frac{n(n+1)}{(n-2)(n-3)} \right) \left(\frac{1}{n-1} \right) \sum_{i=1}^n \left(\frac{e_i - \bar{e}}{S_e} \right)^4$$
$$kurt(e) = \begin{cases} > 3 & \text{caudas - pequenas} \\ = 3 & \text{distrib - normal} \\ < 3 & \text{caudas - longas} \end{cases}$$



Em curvas bem achatadas, semelhança com a normal é mais problemática. Provavelmente, entre +/- 2 desvios há < 95% das observações. → > 5% além destes limites.

Análise dos Resíduos Padronizados , studentizados ou do tipo jackknife :

- Avaliar a **simetria** através de medidas que descrevem a forma de uma distribuição, como é o caso do **coeficiente de assimetria** (*skewness*) e do coeficiente de **kurtose**.

. sum respad, detail				
Standardized residuals				

	Percentiles	Smallest		
1%	-1.719021	-1.719021		
5%	-1.719021	-1.081339		
10%	-1.081339	-.6255674	Obs	12
25%	-.5719923	-.5184171	Sum of Wgt.	12
50%	.0925361		Mean	-.0234159
		Largest	Std. Dev.	.9982075
75%	.3692959	.3150306		
90%	.4842856	.4235612	Variance	.9964182
95%	2.338388	.4842856	Skewness	.6720124
99%	2.338388	2.338388	Kurtosis	4.091408

Média um pouco negativa

Não esperado para distr simétricas (=0)
Mais concentrado de um lado que de outro

Referência para distr normal = 3

- ❖ Pequenos desvios da distribuição normal não produzem efeitos importantes no modelo de regressão. Porém, as assimetrias muito acentuadas influenciam na estimação dos intervalos de confiança e nos testes de hipóteses.

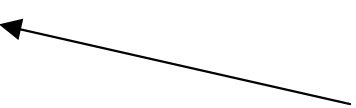
Análise gráfica dos resíduos

Dois tipos de gráficos são básicos:

- ❖ Análise **unidimensional** dos resíduos

- Histogramas,
- box plot etc.

- ❖ Análise **bidimensional** dos resíduos

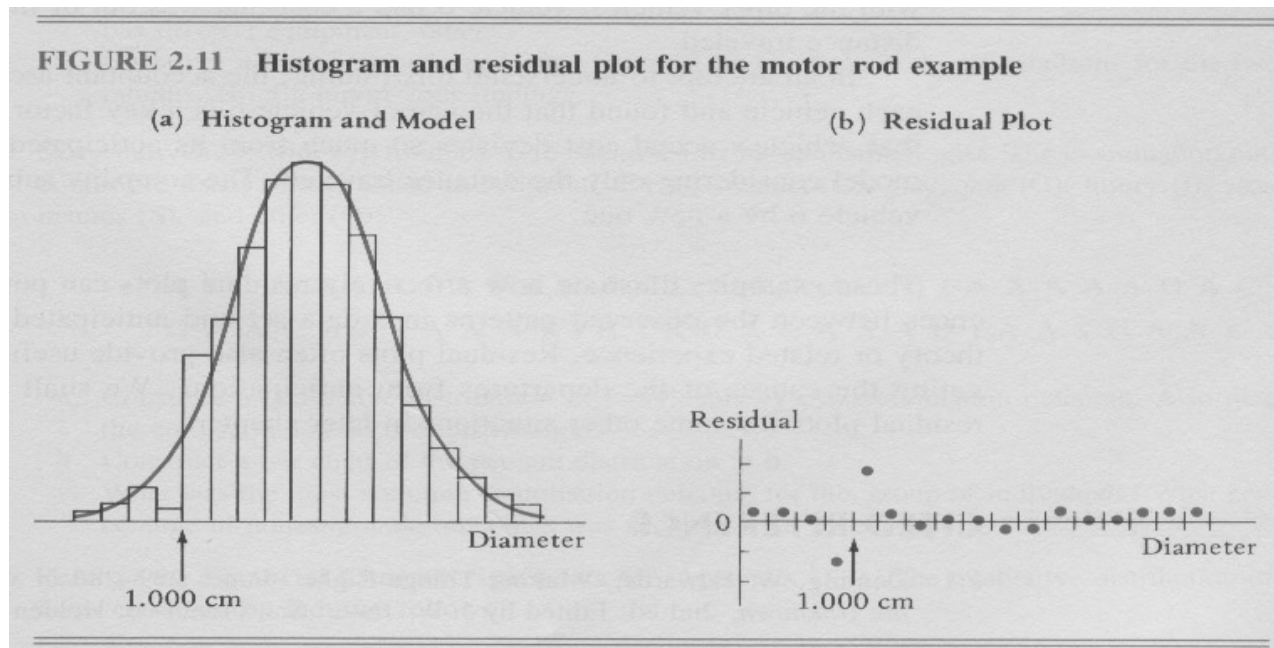
- em relação à variável resposta.
- em relação às variáveis independentes.
- em relação ao tempo. 

Se tivermos anotados data, hora da coleta, pode ser que existam erros sistemáticos com relação ao horário da coleta, data em que foi feita etc.

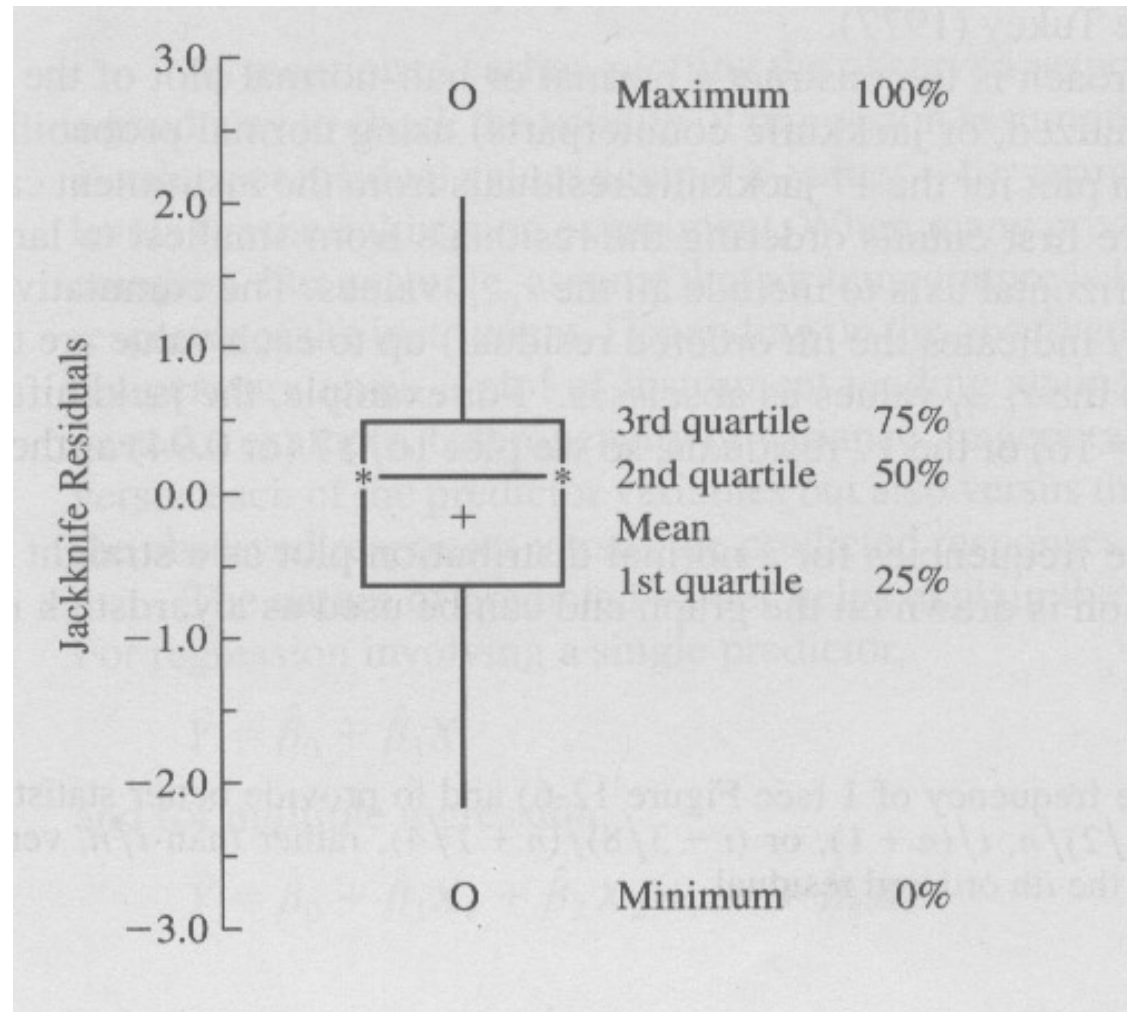
Gráficos Unidimensionais dos Resíduos

Histograma

- ❖ Podemos fazer um histograma para avaliar simetria do gráfico.
- ❖ Podemos comparar os resíduos observados com os valores que seriam esperados no caso de normalidade, calculando alguns pontos onde deveriam ser encontrados os percentis 5%, 25%, 95% etc., usando a distribuição $t(\text{percentil}; n-2 \text{ graus de liberdade})$.



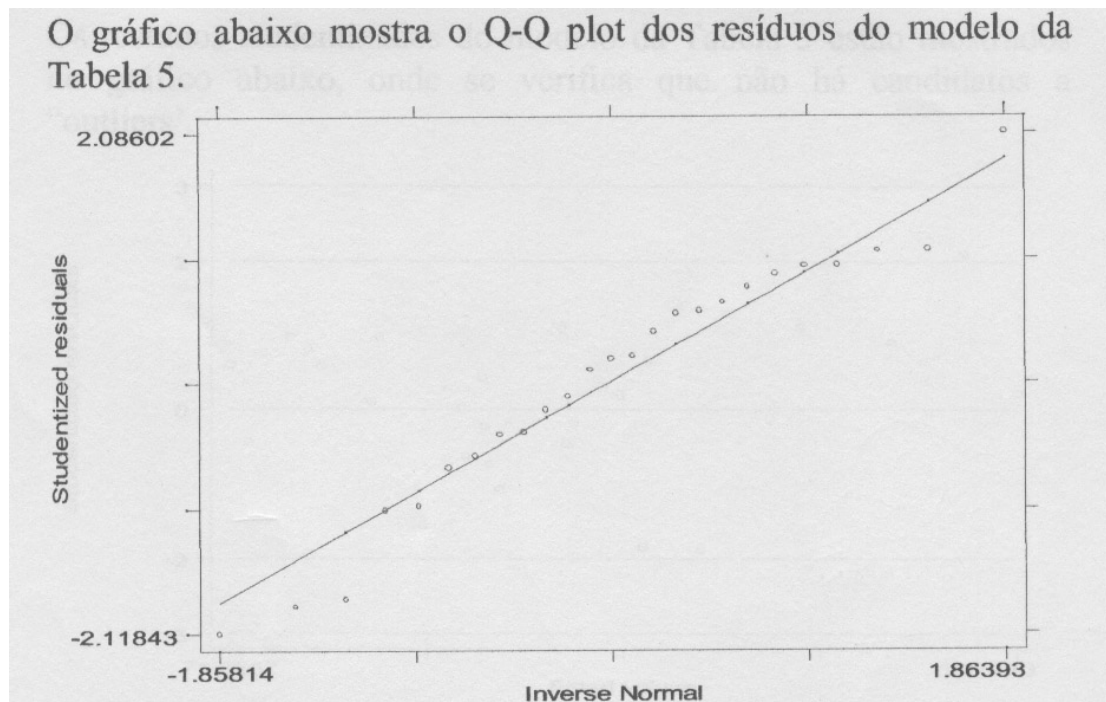
Boxplot dos resíduos Jackknife



- ◆ O gráfico mostra uma distribuição "quase simétrica".
- ◆ Não são apresentados valores muito afastados (3σ).

Normalidade dos resíduos

- ❖ Podemos fazer esta análise utilizando o gráfico Q-Q plot
- ❖ Gráfico dos resíduos padronizados contra os percentis de uma distribuição normal.
- ❖ Se os resíduos tiverem uma distribuição normal, o gráfico Q-Qplot terá os pontos caindo sobre uma reta de 45°



observado

esperado

Gráficos bidimensionais

O ideal seria uma reta no zero, ou uma nuvem de pontos em torno dela

- ❖ Das respostas observadas *vs.* as respostas preditas.
- ❖ Resíduos *vs.* as respostas preditas. → Normalidade, variância cte, *outliers*
- ❖ Resíduos *vs.* as variáveis independentes X . → Outliers - pontos de alavanca
- ❖ Resíduos *vs.* o tempo. → Desvios sistemáticos no tempo
- ❖ Resíduos *vs.* variáveis não incluídas no modelo.


→ Se houver tendência, a nova variável deve ser incluída.

→ *Ex: PAS = idade + sexo*

→ *Resíduos x IMC → se apresentar tendência, incluir*

Gráficos bidimensionais

Ideal que a nuvem seja bem próxima
à diagonal



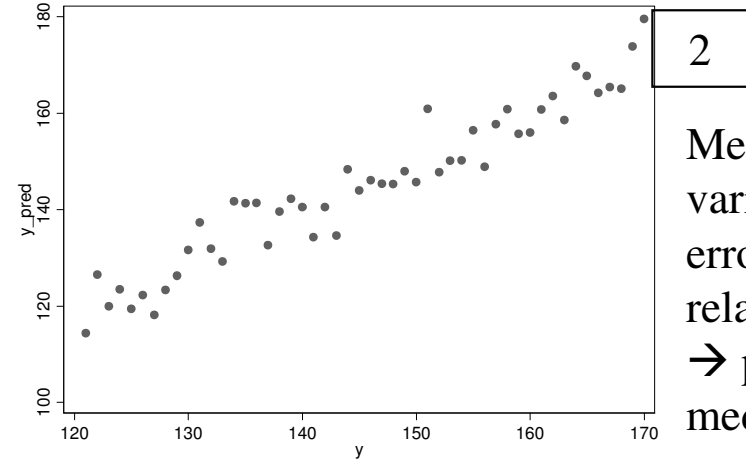
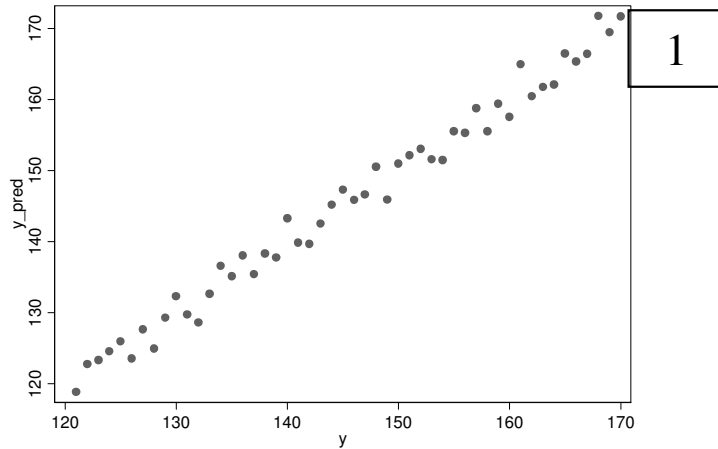
- ❖ Gráfico da variável dependente (Y_i) vs. os valores preditos (\hat{Y}) → permite avaliar a qualidade do ajuste e a força da associação.
- ❖ Gráfico dos erros (ϵ) com os valores preditos (\hat{Y}) → Permite avaliar a hipótese de variância constante, de linearidade, ideal que o gráfico apresente uma distribuição aleatória, nuvem de pontos sem qualquer padrão sistemático.
- ❖ Gráfico dos erros (ϵ) com cada uma das variáveis regressoras ou independentes (X): o ajuste será bom se o diagrama tiver um padrão aleatório em torno de “zero”, no eixo das ordenadas.
- ❖ Modelos inadequados mostrarão algum padrão sistemático. A não-linearidade se tornará evidente quando estes gráficos sugerirem a necessidade de incluir no modelo termos de maior ordem.

- ❖ Gráfico dos erros com o tempo: para avaliar independência das observações. Ideal que não apresente tendência.
- ❖ Se o gráfico dos resíduos vs. variáveis não incluídas no modelo apresentarem algum padrão sistemático quer dizer que devem ser adicionadas ao modelo.

Valores observados vs. valores preditos

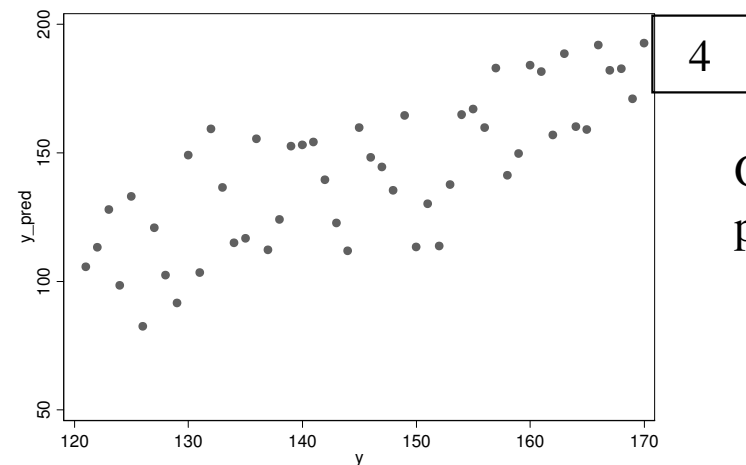
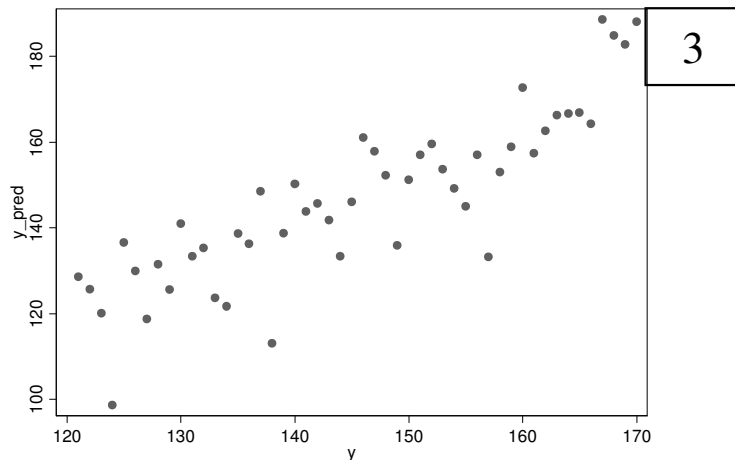
- ❖ Gráfico da variável dependente (Y_i) vs os valores preditos (\hat{Y}) permite avaliar a qualidade do ajuste e a força da associação.

Melhor
ajuste



Menor
variância dos
erros 1 em
relação ao 2
→ piora à
medida que
avança p/ 4

$$Pas = \beta_0 + \beta_1 \text{ Idade} + e$$

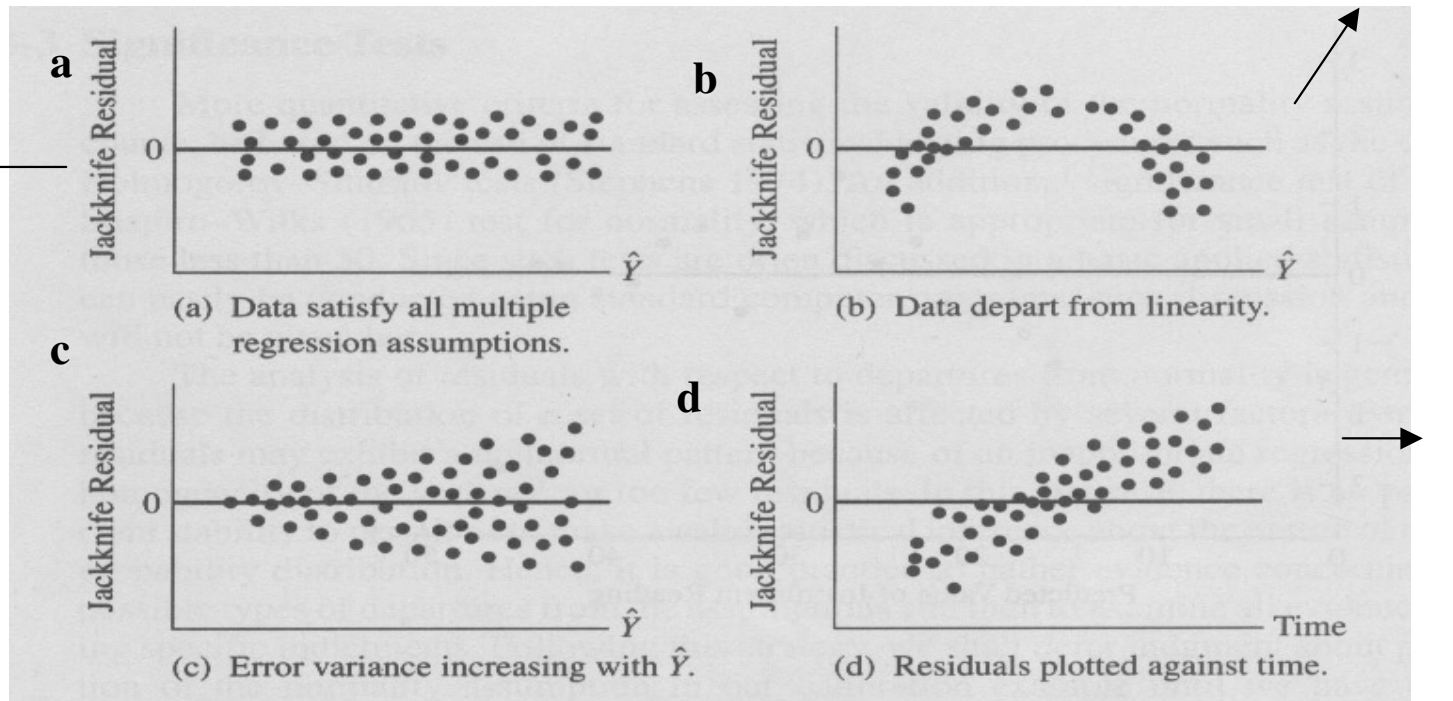


Qualidade
pior

Fonte: Dados simulados com $\sigma = 2, 5, 10$ e $20 \rightarrow$ quanto maior σ , menor a precisão de \hat{Y}

Resíduos vs valor predito (\hat{Y})

Elipse em torno de zero



Usar modelos que incluem estrutura de dependência.

Ex: entrevistas ou medidas cujos resultados são dependentes no tempo. Pode

a → Padrão esperado quando todos os pressupostos são atendidos (faixa de pontos em torno do zero) - ausência de tendências.

b → Padrão típico de quando os dados se afastam do padrão linear.

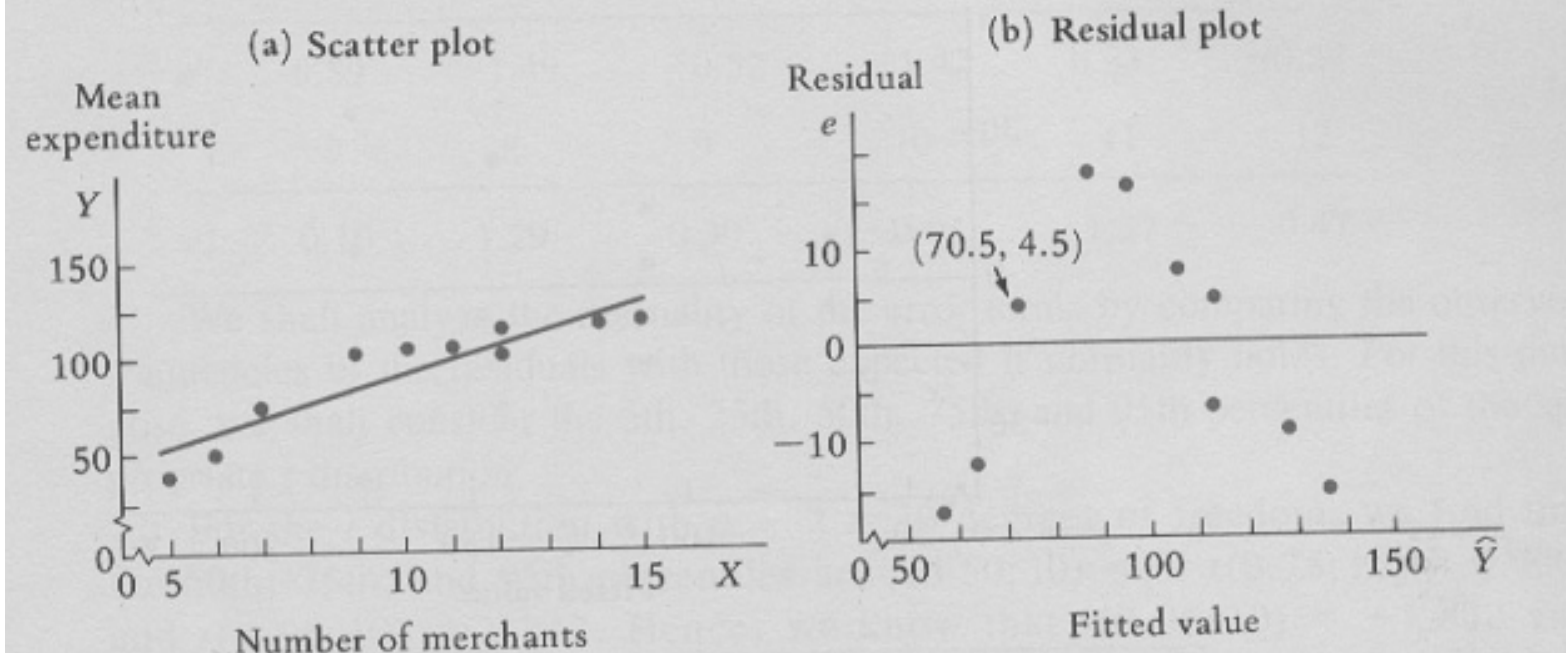
c → Variância não homogênea (cresce com o aumento de \hat{Y})

d → muitas vezes, os dados são coletados em função do tempo; neste exemplo, mostra-se uma clara correlação dos resíduos ao longo do tempo, ou seja, no tempo t_0 , se um resíduo é positivo, no tempo t_0+1 o resíduo também é positivo e assim por diante.

ser que os indivíduos sejam da mesma família entrevistados em tempos diferentes → talvez incluir termo quadrático (olhando o gráfico b)

Exemplo de Modelo não linear

FIGURE 19.4 Scatter and residual plots illustrating a nonlinear regression relation — Discount card use example



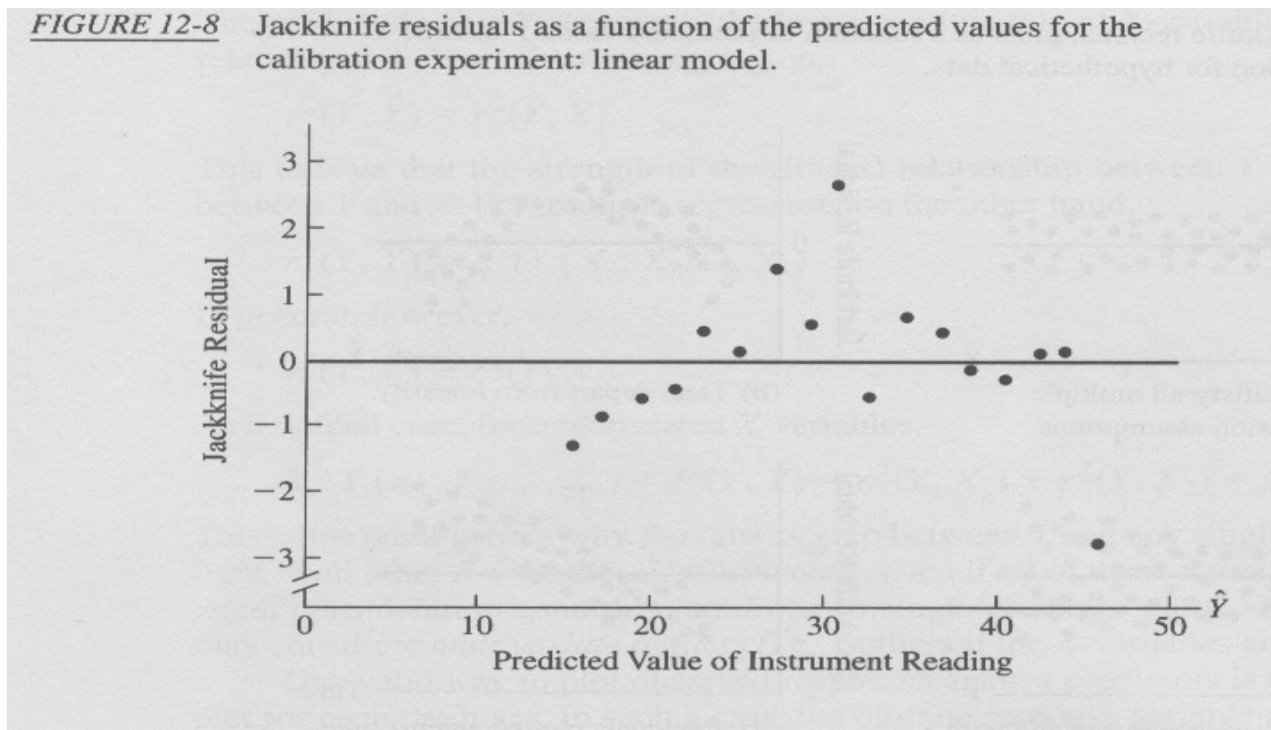
- Relação não linear.
- Pode ser investigada também com o diagrama de espalhamento.
- Se a relação fosse linear, os resíduos formariam uma faixa em torno do valor zero.

Resíduos vs. valores preditos

cont.

Linearidade:

O gráfico dos **resíduos** vs. **valores preditos** [leitura instrumento
f(poluente)] $Y = 15.39 + 3.79X$



Não
linearidade

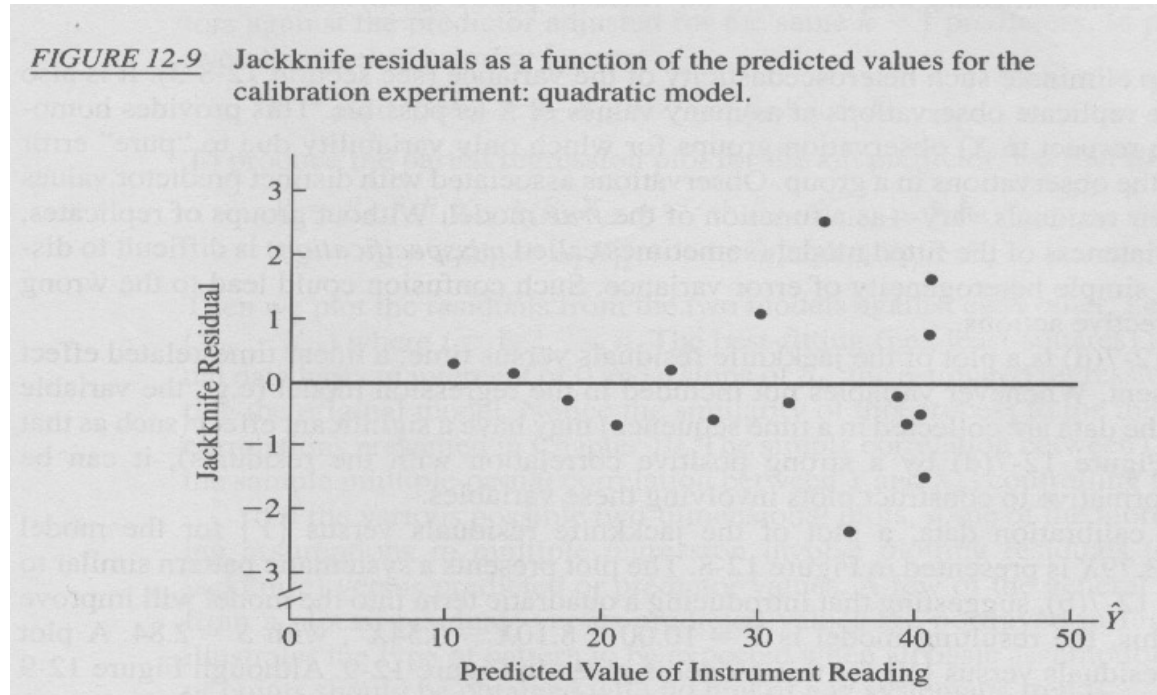
→ Violação do
pressuposto
de linearidade

→ Incorporar
termo
quadrático para
melhorar
ajuste?

- ❖ O padrão semelhante ao da figura b.
- ❖ Sugestão que a introdução de um termo quadrático no modelo iria melhorá-lo.

Linearidade:

- ❖ O gráfico dos **resíduos** vs. **valores preditos** - inclusão termo quadrático [leitura instrumento $f(\text{poluente})$] $Y = 10.00 + 8.10X - 0.54X^2$



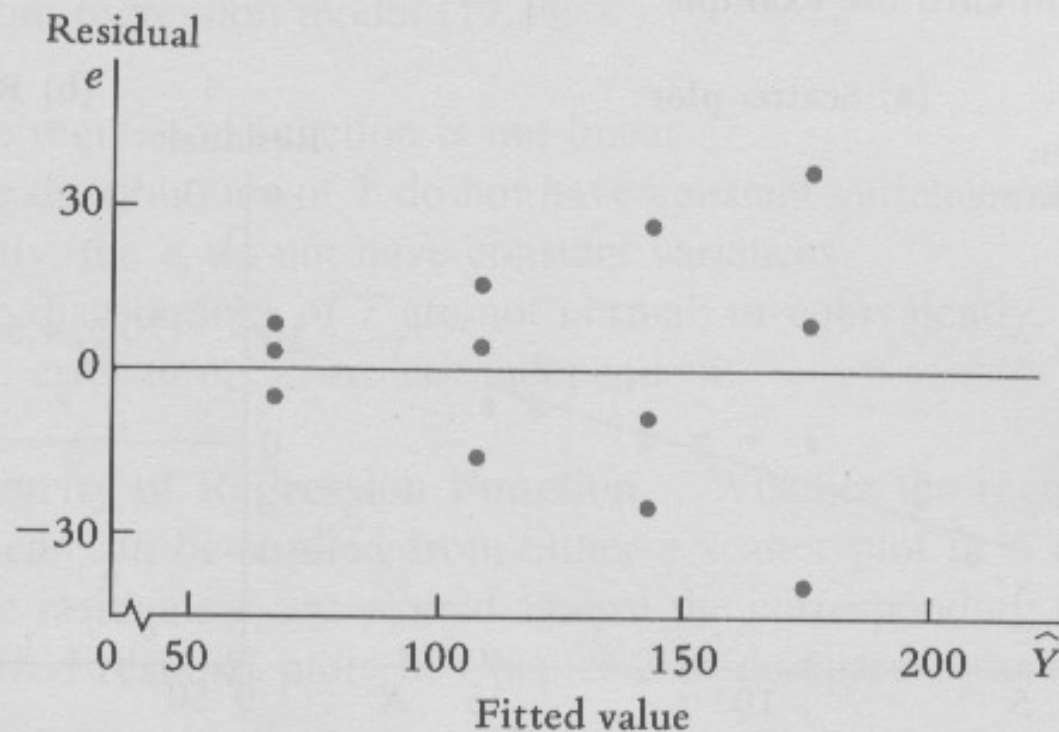
- ❖ Embora haja grande espalhamento dos pontos (em parte em função do tamanho da amostra), a figura mostra um padrão de faixa horizontal em torno do zero.
- ❖ Podemos também admitir que o padrão se assemelha ao da figura c (heterocedasticidade).

Resíduos vs. valores preditos

cont.

Homogeneidade das variâncias

FIGURE 19.6 Residual plot illustrating nonconstant error variance

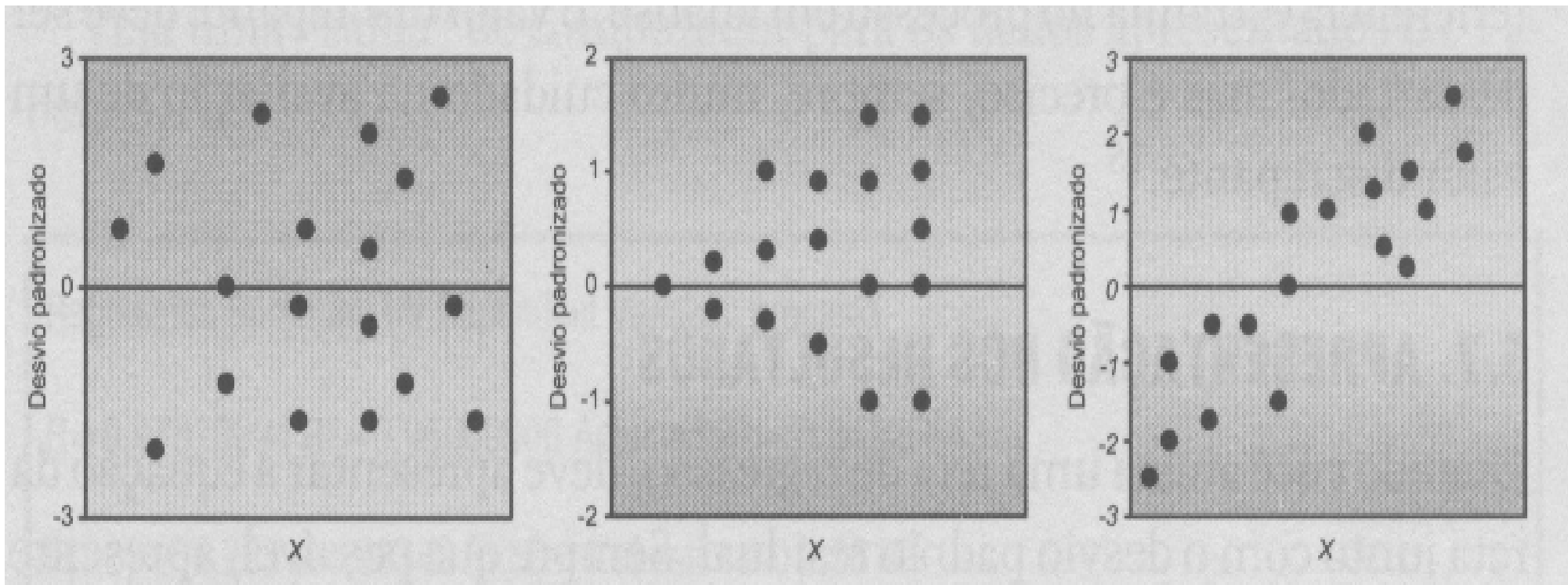


Variância não homogênea

→ Pensar em transformação da variável, como termo de maior potência

→ Ex: Incluir x^2 ?

Resíduos vs. variáveis independentes X



- O 1º. gráfico mostra que os resíduos estão aleatoriamente distribuídos → não há qualquer tendência aparente.
- O 2º. gráfico mostra que maiores resíduos estão ocorrendo para maiores valores de X → o pressuposto de variância constante não foi obedecido. → Transformar a variável
- O 3º. gráfico mostra que o modelo não é linear. → Incorporar termo quadrático?

Correlação entre os erros

- ❖ Quando os dados são ordenados no tempo
- ❖ Construir gráfico dos resíduos ordenados no tempo também auxilia na visualização de correlação entre estes erros.

Detecção de *outliers*

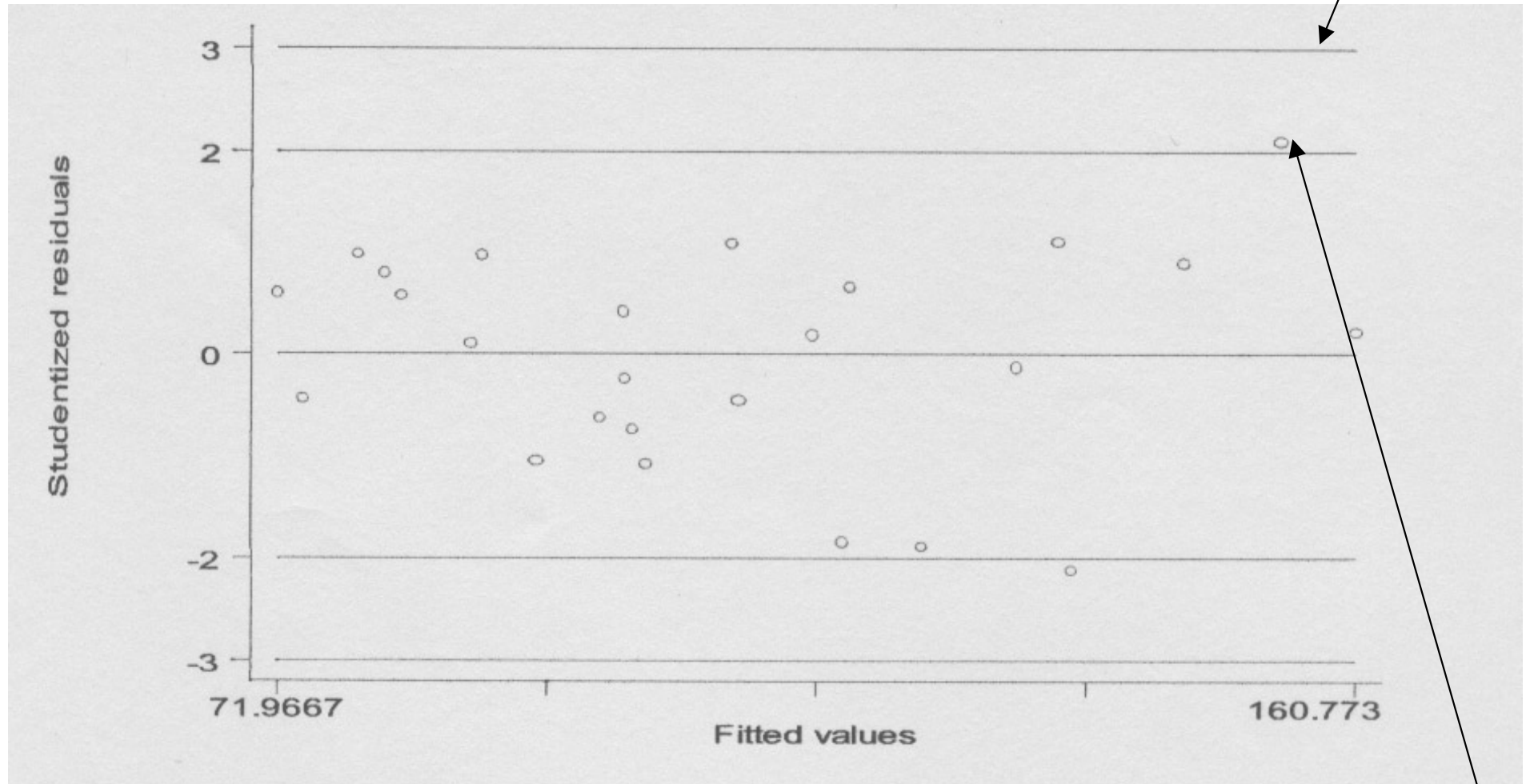
- ❖ Um *outlier* de um conjunto de resíduos é um valor atípico.
- ❖ Pode cair acima de 2,5 a 3 desvios padrão além da média do conjunto de resíduos padronizados
- ❖ A presença de tal valor pode afetar o ajuste pelos mínimos quadrados
- ❖ *Outliers* podem causar um impacto importante nas conclusões de um estudo.
- ❖ É de interesse saber em que medida este ponto afeta o ajuste.
- ❖ Não se recomenda excluí-lo, somente nos casos de certeza de ser um valor errado.
- ❖ No mínimo, realizar análises com e sem a presença dos *outliers*.

- ❖ Cálculo DSE (Desvio Studentizado Extremo):
- ❖ Existem diversas formas de analisar se uma observação é um *outlier*. Iremos apresentar um modo simples
- ❖ Padronizam-se os valores, para saber a quantos desvios padrões da média eles estão.
- ❖ Estatística DSE = $\max_{i=1,2,\dots,n} |X_i - \bar{X}| / S$
- ❖ Os que se afastam **muito**, podem ser considerados *outliers*.
- ❖ O que é "muito"?

Detecção de *outliers*

Arredondado do 2,58 que equivale a 99% da normal (\approx da de Student \rightarrow g.l.?);

2 (1,96) \rightarrow 95%



Está acima de 2, mas muito próximo. Há outra de 3,44 (não aparece no gráfico) que seria a mais importante para nos preocuparmos.

Observações influentes (Leverage ou alavanca)

- ◆ Uma observação influente é definida como aquela que, por alguma razão, causa grandes mudanças em alguns ou em todos os parâmetros do modelo, quando ela é omitida do conjunto de dados.

h_i – medida da importância da i -ésima observação no ajuste do modelo

h_i = i -ésimo elemento da matriz $H = X(X'X)^{-1}X'$

Diagonal do produto da matriz das variáveis independentes

$0 \leq h_i \leq 1$ e $h_i > 2(k+1)/n$ indicaria a presença de uma observação influente.

No. de parâmetros

Distância de Cook:

- ❖ Ajuda a descobrir possíveis *outliers*
- ❖ Ela quantifica o peso da observação no modelo
- ❖ É uma medida da mudança dos coeficientes de regressão, quando se retira do modelo essa observação.
- ❖ Pontos com valores acima de 1 são valores suspeitos.
- ❖ Os maiores que 2 sinalizam sérios problemas.

$$d_i = \left(\frac{1}{k+1} \right) r_i^2 \left(\frac{h_i}{1-h_i} \right) = \left(\frac{e_i^2 h_i}{(k+1) S_i^2 (1-h_i)^2} \right)$$

Resíduo Jackknife

Resíduo

DFbeta:

- ❖ Permite avaliar o efeito de cada observação nas estimativas de cada um dos parâmetros do modelo ajustado.
- ❖ São calculados os coeficientes Dfbeta para cada variável.
- ❖ Uma observação é considerada influente se:
 - ❖ $|Dfbeta_k| > 1$ se $n \leq 30$
 - ❖ $|Dfbeta_k| > 2/\sqrt{n}$ se $n > 30$

Colinearidade

- ❖ Colinearidade → Forte relação entre variáveis independentes
- ❖ Pode gerar problemas numéricos de modo a gerar estimativas inacuradas dos coeficientes da regressão, variabilidade e no valor-P.
- ❖ Supondo a regressão com 2 variáveis independentes X_1 e X_2 .

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

- ❖ Pode-se demonstrar que

$$\hat{\beta}_j = c_j \left[\frac{1}{1 - r^2(X_1, X_2)} \right]$$

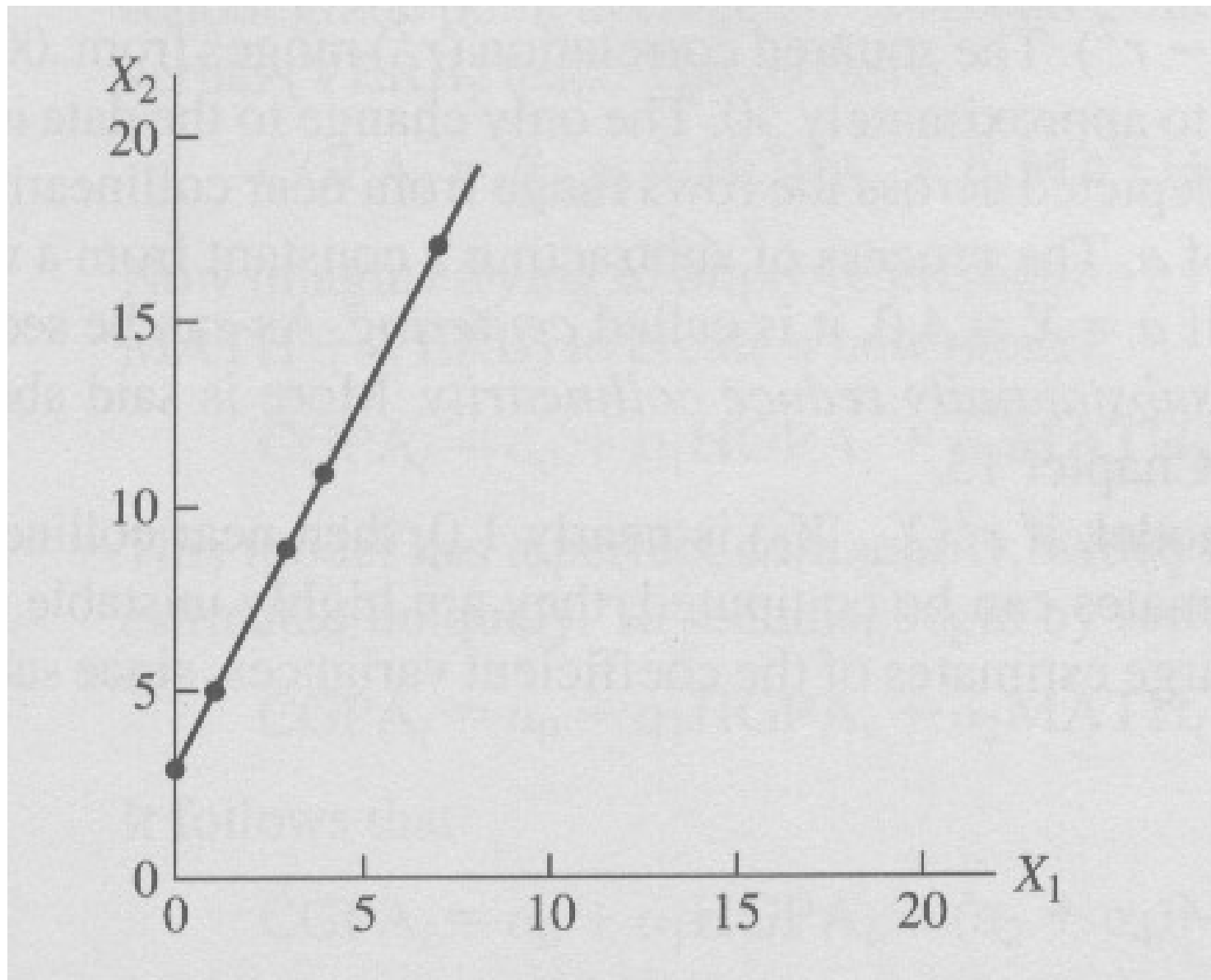
B depende das observações (c_j), mas tb da correlação

para $j = 1$ ou $j = 2$

$c_j \rightarrow$ valor que depende dos dados.

$r^2(X_1, X_2)$ é ao quadrado da correlação entre X_1 e X_2 .

Colinearidade



Colinearidade

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

Então:

$$\bar{Y} - \hat{\beta}_0 = \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 = (c_1 \bar{X}_1 + c_2 \bar{X}_2) \left[\frac{1}{1 - r^2(X_1, X_2)} \right]$$

$\bar{Y} - \hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ são proporcionais a $\left[\frac{1}{1 - r^2(X_1, X_2)} \right]$ (VIF)

Se $r^2(X_1, X_2) \rightarrow 1$ então :

$$[1 - r^2(X_1, X_2)] \rightarrow 0 \quad \text{e} \quad \left[\frac{1}{1 - r^2(X_1, X_2)} \right] \rightarrow \infty$$

- Superestima os coeficientes, a variância (que também é proporcional à parcela) e os testes que têm por base o valor do coeficiente e a variância (p-valor)

Colinearidade

❖ $r^2 > 0.90$ merece atenção

❖ existe uma medida que verifica se a correlação pode causar problema de colinearidade

❖ VIF - Variance inflation factor

Se não detectou colinearidade na exploratória, O VIF ajuda depois do ajuste, na fase de diagnóstico

$$VIF = \left[\frac{1}{1 - r^2(X_1, X_2)} \right]$$

❖ $VIF \geq 10$

0,8 e 0,85 por ex. tb afetam as estimativas. O ponto de corte é controverso

❖ Regra prática: $VIF \geq 10 \rightarrow r^2 > 0.90$ ou $r > 0.95$

Estratégias alternativas

Algumas estratégias podem ser adotadas quando os pressupostos básicos não são atendidos.

Transformações

Existem 3 razões básicas pra usar transformações matemáticas dos dados:

Ou variável resposta ou a independente

1. Estabilizar a variância no caso de heterocedasticidade
2. Normalizar a variável dependente Y.
3. Linearizar o modelo de regressão → caso os dados não sugiram uma relação linear.

Mais que normalizar Y, normalizar a distribuição dos resíduos

A distr Y → normalidade de Y é condicionada à X (distr Y para idade=20, para idade=30 etc.)

As transformações mais usadas são:

Log ($Y' = \log Y$)

- ❖ Para estabilizar variância, quando ela cresce acentuadamente com o aumento de $Y \longrightarrow$ Cauda à direita
- ❖ Normalizar a distribuição da variável dependente Y (caso a distribuição dos resíduos seja marcadamente assimétrica à direita)
- ❖ Para linearizar a relação de Y e X , caso a relação sugira uma inclinação consistentemente crescente.

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

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

$$X_2 - X_1 \rightarrow 1 \text{ unidade}$$

$$\log Y_2 - \log Y_1 = \beta_0 + \beta_1 X_2 - \beta_0 - \beta_1 X_1$$

$$\log Y_2 - \log Y_1 = \beta_1 (X_2 - X_1)$$

$$\log Y_2 - \log Y_1 = \beta_1$$

$$\log \left(\frac{Y_2}{Y_1} \right) = \beta_1$$

$$\frac{Y_2}{Y_1} = 10^{\beta_1} \Rightarrow Y_2 = 10^{\beta_1} Y_1$$

As transformações mais usadas são:

Raiz quadrada ($Y' = \sqrt{Y}$):

- ❖ Estabiliza a variância nos casos em que a variância é proporcional à média de Y .
- ❖ Em especial se a variável dependente tem uma distribuição de Poisson.

As transformações mais usadas são:

Inverso ($Y' = 1/Y$)

- ❖ Estabiliza a variância nos casos em que a variância é proporcional à 4a. potência da média de Y ($\text{var} \propto \bar{Y}^4$).
- ❖ Indica que um aumento abrupto ocorre a partir de um determinado limiar de Y .
- ❖ Esta transformação minimiza o efeito de valores elevados de Y , uma vez que a transformação os trará para próximo de zero.
- ❖ Aumentos grandes em Y ocasionarão aumentos pequenos em Y' (Y transformada)

As transformações mais usadas são:

Quadrado ($Y' = Y^2$)

- ❖ Estabiliza a variância quando a variância diminui com a média de Y
- ❖ Para normalizar a variável dependente Y , se a distribuição dos resíduos é assimétrica à esquerda
- ❖ Linearizar o modelo se a relação original for curvilínea para baixo (se a inclinação consistentemente decresce com o aumento de X).

As transformações podem ser realizadas também na variável independente X

FIGURE 19.7 Use of a reciprocal transformation on the independent variable to linearize a regression relation

