CE071 - Análise de Regressão Linear

Cesar Augusto Taconeli

20 de maio, 2018

Aula 5 - Diagnóstico do modelo de regressão linear

Introdução

- A especificação de um modelo de regressão depende de várias suposições;
- A verificação das suposições assumidas é necessária para a validade do modelo ajustado e das consequentes inferências;
- Após o ajuste do modelo, devemos avaliar a validade dessas suposições, bem como checar outros possíveis problemas de ajuste.
- Esta etapa da análise de é comumente denominada *diagnóstico da regressão* ou simplesmente *análise de diagnóstico*.

Introdução

- Os potenciais problemas quanto à especificação de um modelo de regressão linear são:

1. A média de y, condicional a $\boldsymbol{x}$, foi especificada como $E(y|\boldsymbol{x}) = \boldsymbol{x}'\beta$;
2. Assumimos que os erros são independentes e têm variância constante (σ^2);
3. Assumimos que os erros têm distribuição normal;
4. A presença de observações atípicas pode ter considerável impacto nos resultados.

Resíduos em regressão linear

- Os resíduos, conforme definido anteriormente, são dados por:

$$r_i = y_i - \hat{y}_i, \quad i = 1, 2, ..., n. \tag{1}$$

- O vetor de resíduos, $\boldsymbol{r'} = (r_1, r_2, ..., r_n)$, pode ser expresso na seguinte forma:

$$\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}, \tag{2}$$

em que $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$, I é a matriz identidade $n \times n$ e $\boldsymbol{\epsilon}$ o vetor de erros.

Resíduos em regressão linear

- Decorre, da definição dos resíduos, que:

1. $E(\boldsymbol{r}) = \boldsymbol{0}$;
2. $Var(\boldsymbol{r}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$;
3. Os resíduos têm distribuição normal, uma vez que são combinações lineares dos $\epsilon's$.

- Podemos descrever a distribuição dos resíduos, de forma resumida, por:

$$\begin{aligned} & r_i \sim Normal(0, \sigma^2(1 - h_{ii})); \\ & Cov(r_i, r_i') = -\sigma^2(h_{ii'}), \quad i, i' = 1, 2, ..., n; i \neq i'. \end{aligned} \tag{3}$$

Resíduos padronizados

- Resíduos escalonados são úteis para a identificação de valores extremos (outliers).
- Uma primeira versão de resíduos escalonados são os **resíduos padronizados**, definidos por:

$$e_i = \frac{r_i}{QM_{Res}}, \quad i = 1, 2, ..., n. \tag{4}$$

- Neste caso, QM_{Res} serve como estimativa para as variâncias dos resíduos.
- Observações com $|e_i| > 3$ são potenciais outliers e devem ser investigadas.

Resíduos studentizados

- Os **resíduos studentizados** têm como vantagem adicional incorporar as variâncias individuais dos resíduos no escalonamento, sendo definidos por:

$$t_i = \frac{r_i}{\sqrt{QM_{Res}(1 - h_{ii})}}, \quad i = 1, 2, ..., n. \tag{5}$$

- Por sua construção, os resíduos studentizados têm variância igual a um qualquer que seja a locação da observação ($\boldsymbol{x}_i$) se o modelo especificado se ajustar aos dados.
- Resíduos studentizados são recomendados por facilitar a identificação de **observações influentes**.

Resíduos studentizados externamente

- **Resíduos studentizados externamente** fazem uso da estratégia *leave one out* na estimação de σ^2:

$$t_{(i)} = \frac{r_i}{\sqrt{QM_{Res_{(i)}}(1 - h_{ii})}}, \quad i = 1, 2, ..., n, \tag{6}$$

em que $QM_{Res_{(i)}}$ é a estimativa de σ^2 gerada pelo modelo ajustado com $n - 1$ observações (exceto a $i-$ésima).

- Pode-se mostrar que o ajuste de n modelos não é necessário para o cômputo de $QM_{Res_{(i)}}$, uma vez que:

$$QM_{Res_{(i)}} = \frac{(n - p)QM_{Res} - r_i^2/(1 - h_{ii})}{n - p - 1}. \tag{7}$$

Resíduos parciais

- **Resíduos parciais** permitem avaliar a relação entre a resposta e uma particular covariável **ajustado o efeito das demais covariáveis**.

- Suponha que o modelo ajustado contenha as covariáveis $x_1, x_2, ..., x_k$. O resíduo parcial associado à variável x_j é definido por:

$$r_i^*(y|x_j) = r_i + \hat{\beta}_j x_{ij}, \quad i = 1, 2, ..., n. \tag{8}$$

- Observe que o resíduo parcial *desconta* do resíduo original o efeito de x_j.

Análise de resíduos

- Diversos gráficos podem ser construídos para checar o ajuste de modelos de regressão com base nos resíduos, dentre os quais:

1. Resíduos *vs* valores ajustados:
   - Verificar padrões sistemáticos que podem indicar especificação incorreta do preditor do modelo;
   - Avaliar se os erros tê variância constante;
   - Identificar *outliers*.
2. Gráfico quantil-quantil:
   - Checar se os erros tem distribuição (aproximadamente) normal;
   - Identificar outliers.

Análise de resíduos

3. Resíduos *vs* ordem de coleta:
   - Analisar possível correlação nos dados induzida pela ordem de coleta (caso se aplique);
4. Resíduos *vs* variável incluída no modelo
   - Verificar tendência não linear, indicativo de que o efeito da variável na resposta não é bem explicado pelo modelo.
   - Avaliar variância não constante.
5. Resíduos parciais *vs* correspondente variável explicativa
   - Analisar a relação entre a resposta e a variável sob investigação ajustado o efeito das demais variáveis.
6. Resíduos *vs* variáveis não incluídas no modelo
   - Objetivos similares ao gráfico de resíduos parciais.

Padrões em gráficos de resíduos

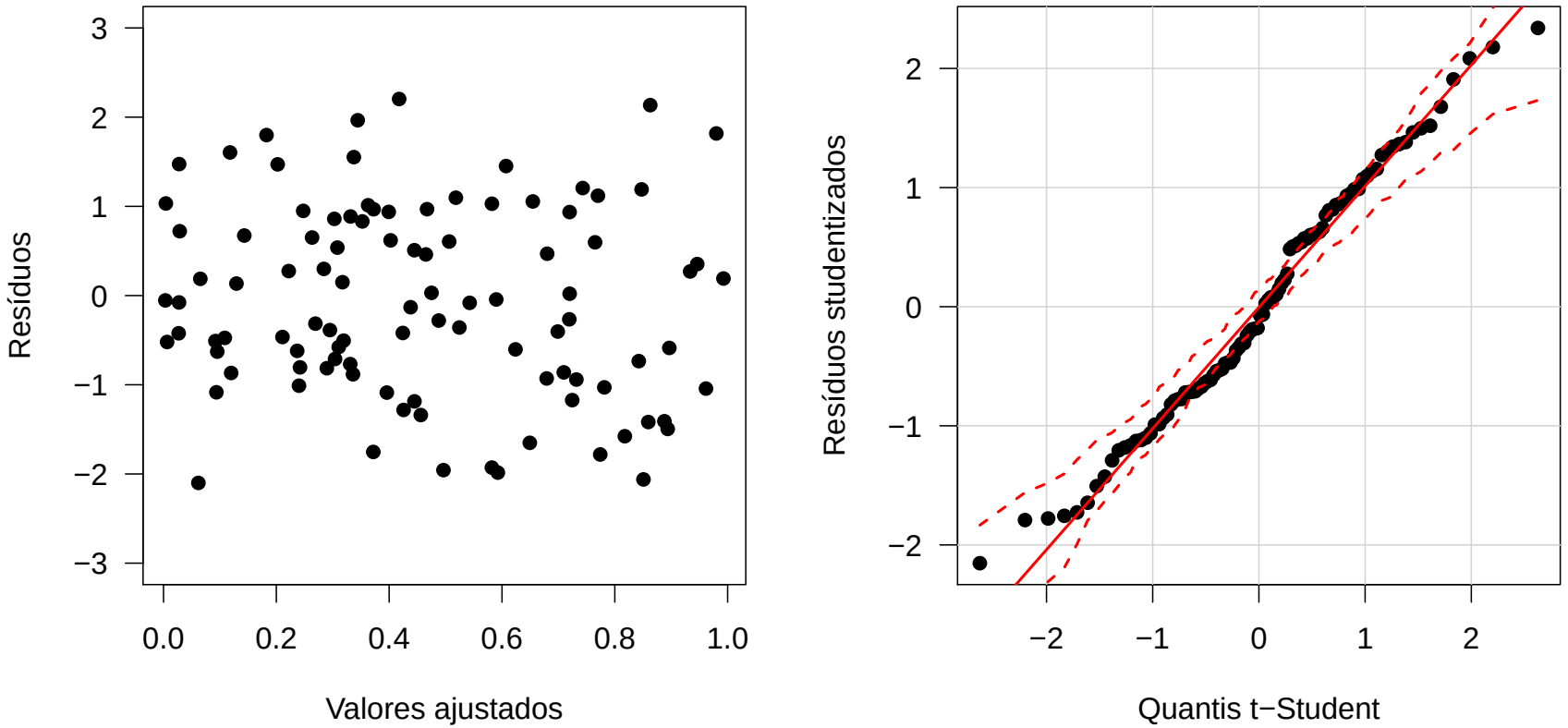

Figura 1: Ajuste satisfatório

Padrões em gráficos de resíduos

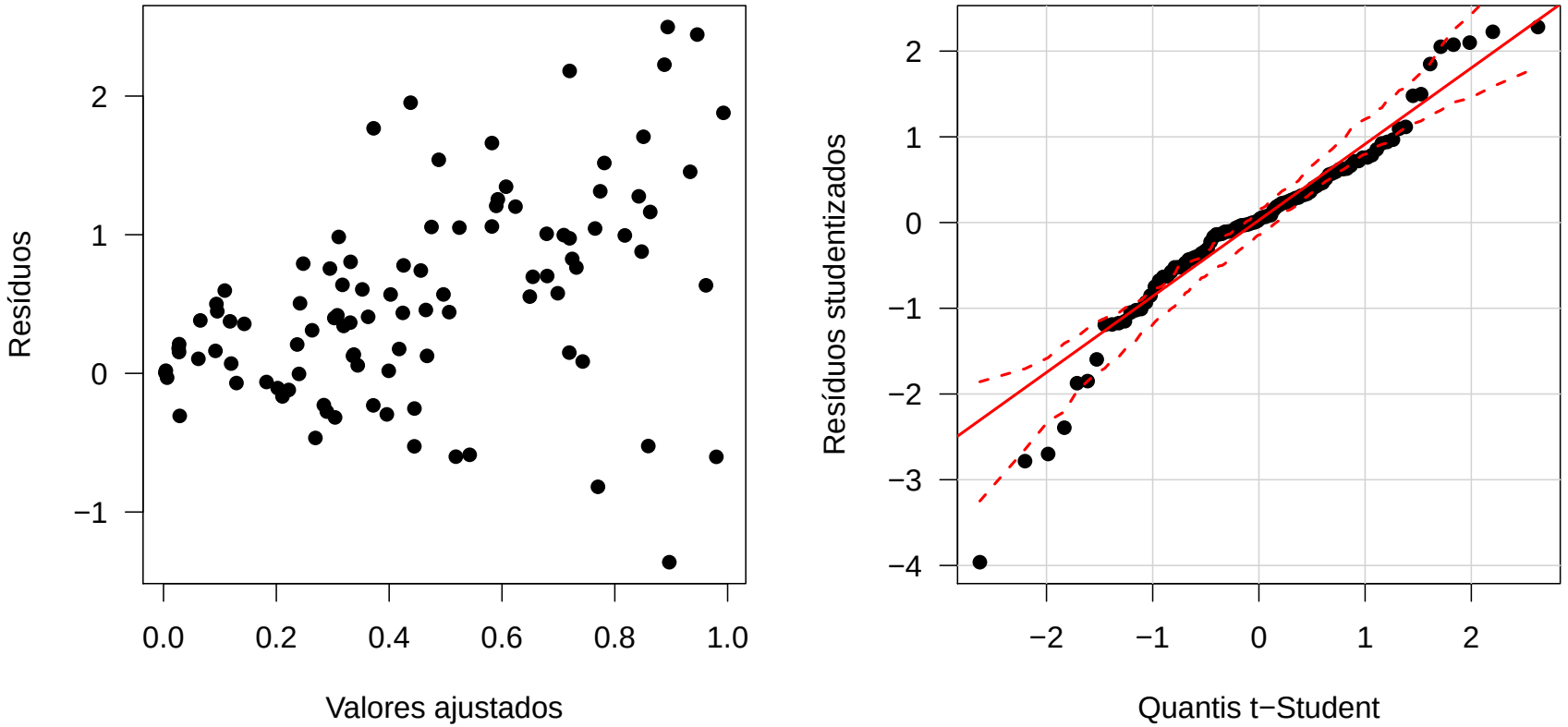

Figura 2: Variância não constante

Padrões em gráficos de resíduos

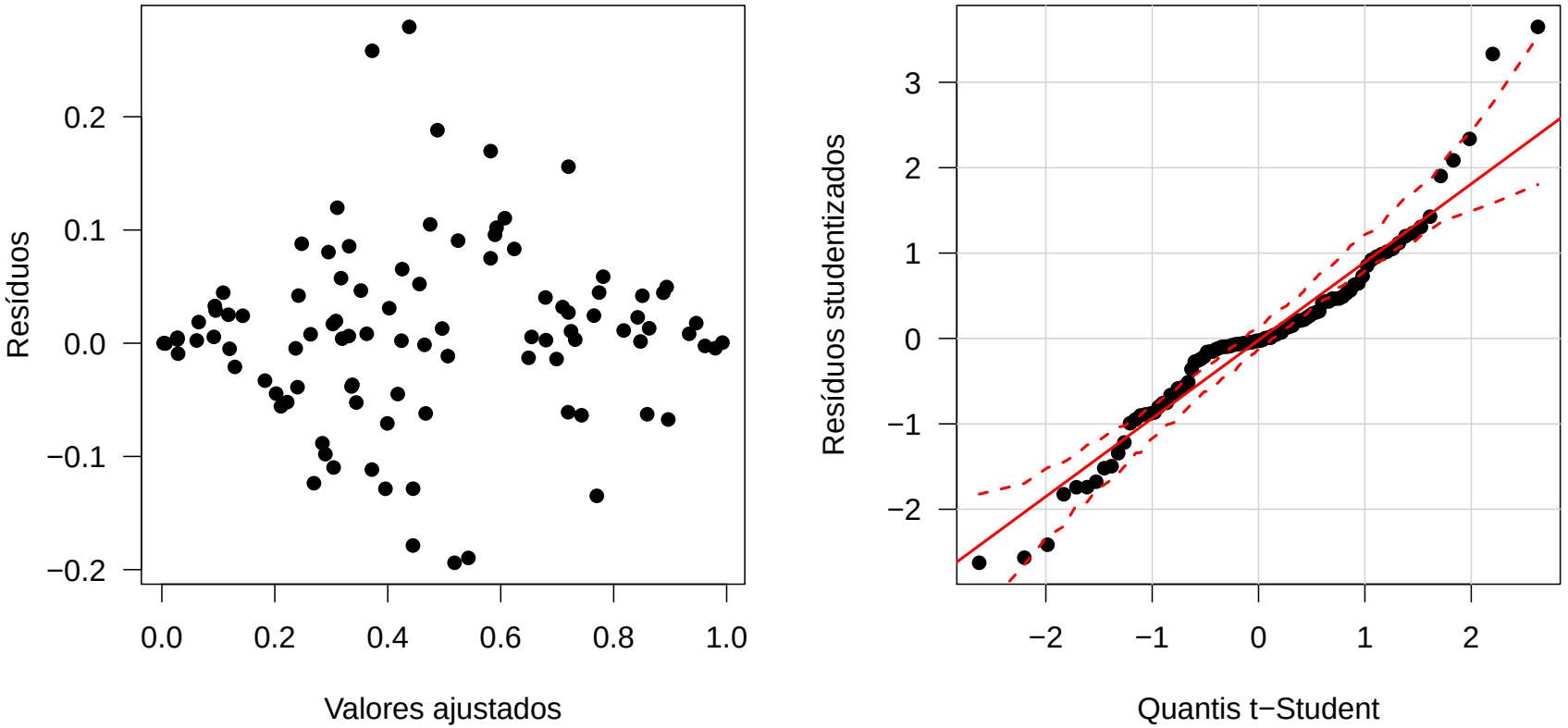

Figura 3: Variância não constante (2)

Padrões em gráficos de resíduos

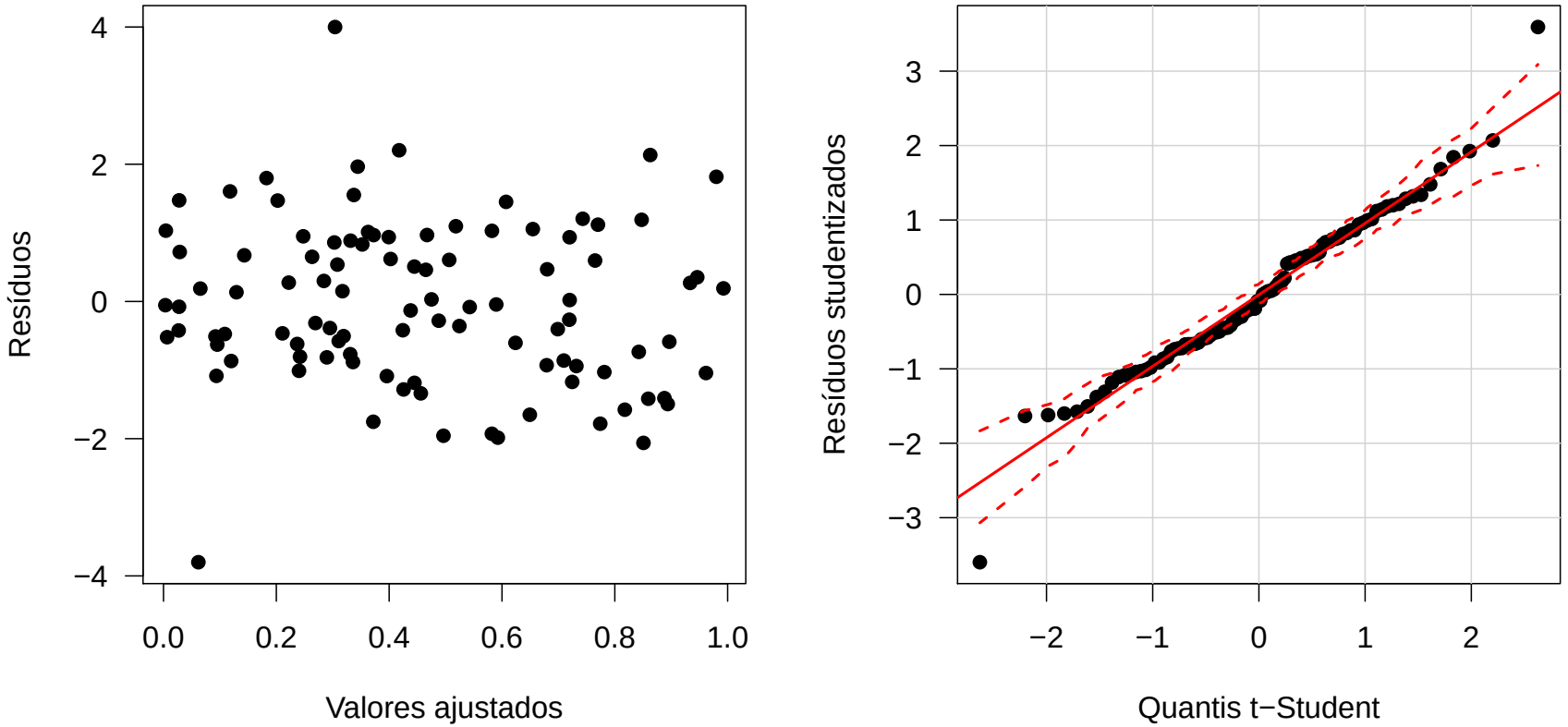

Figura 4: Presença de outliers

Padrões em gráficos de resíduos

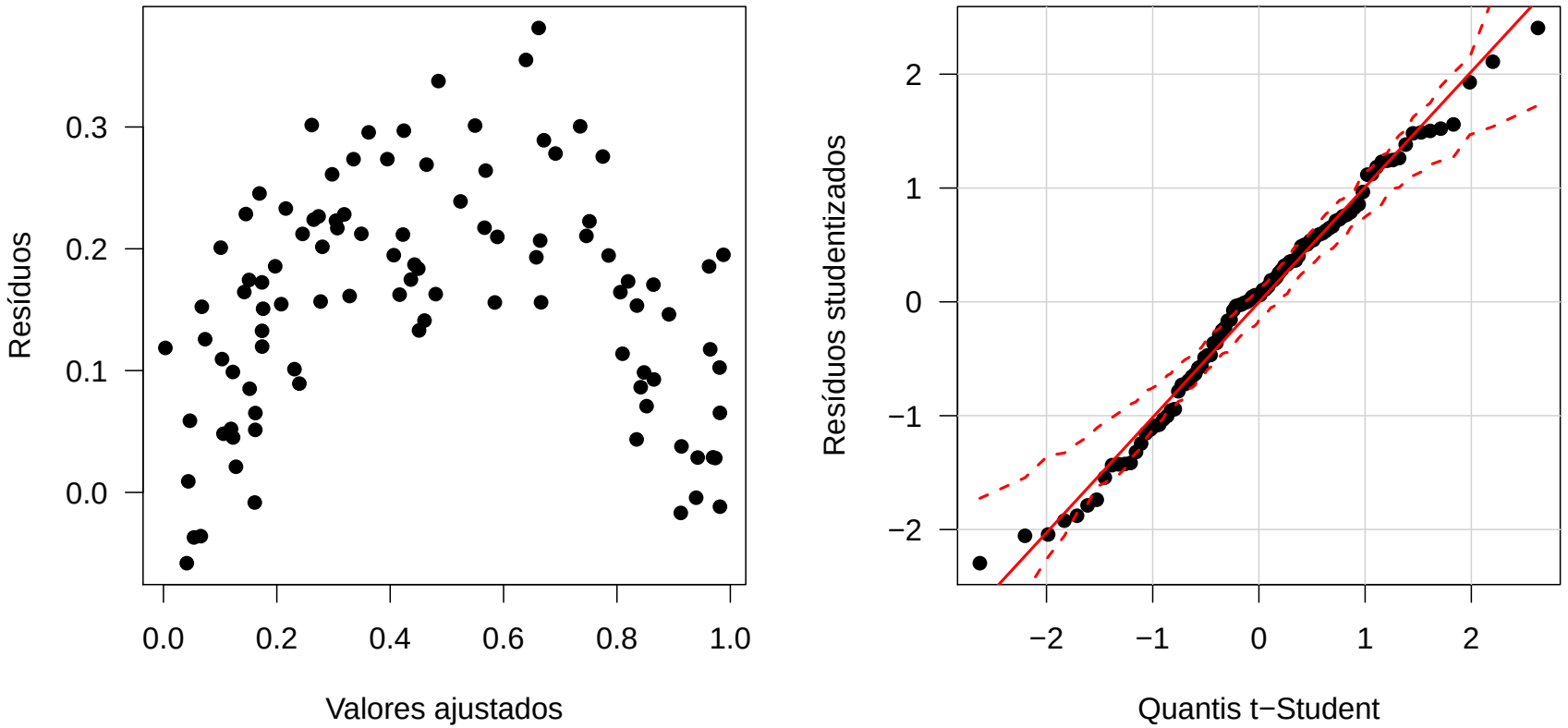

Figura 5: Não linearidade

Padrões em gráficos de resíduos

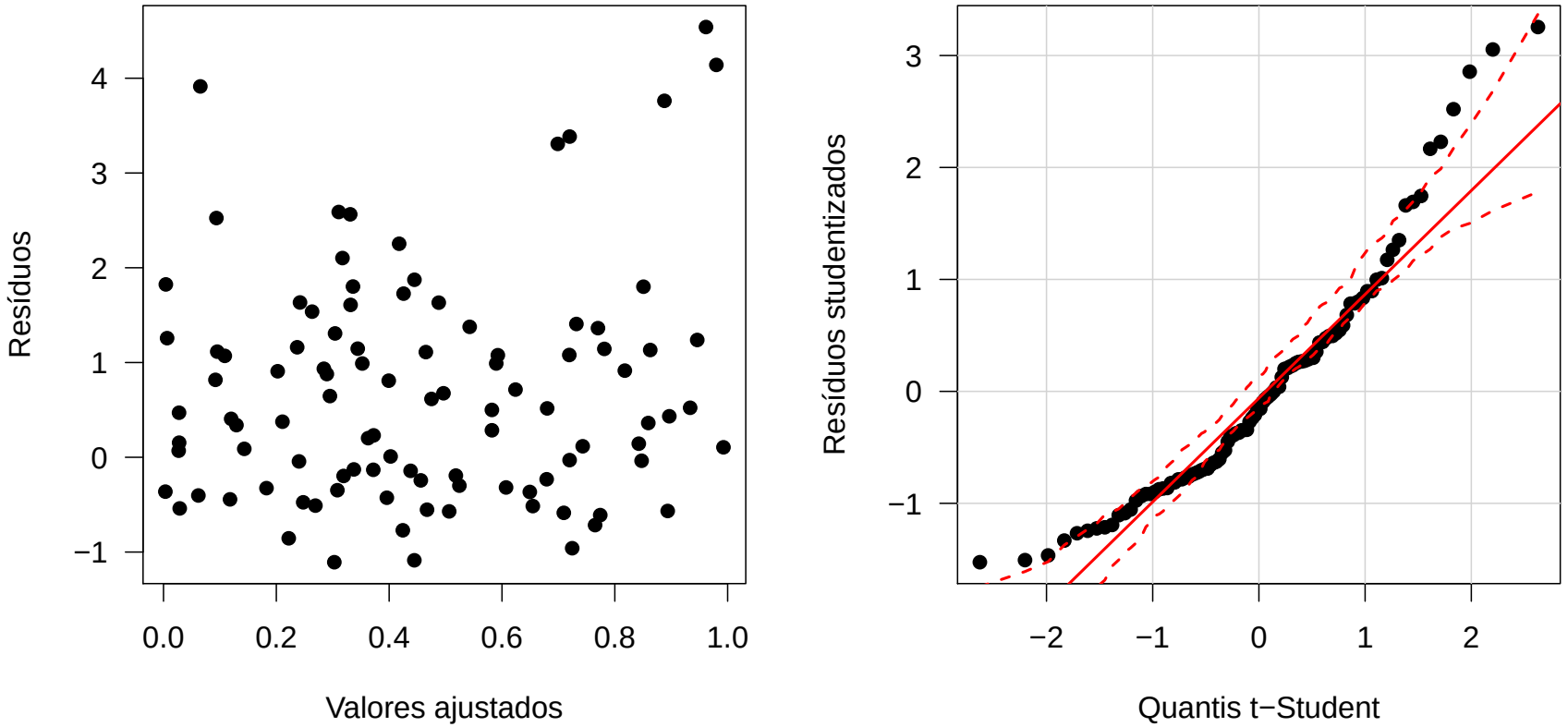

Figura 6: Erros com distribuição assimétrica

Padrões em gráficos de resíduos

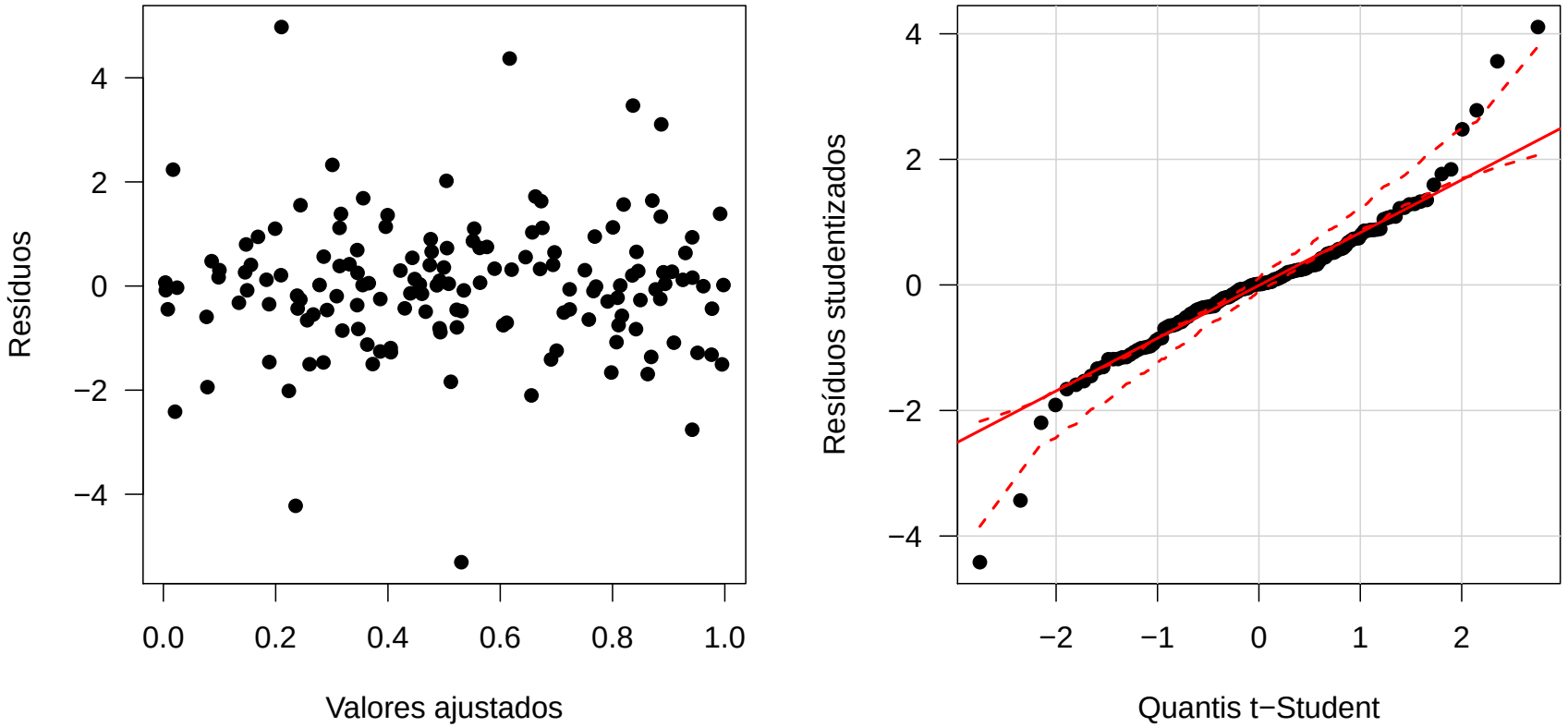

Figura 7: Erros com distribuição simétrica - caudas pesadas

Padrões em gráficos de resíduos

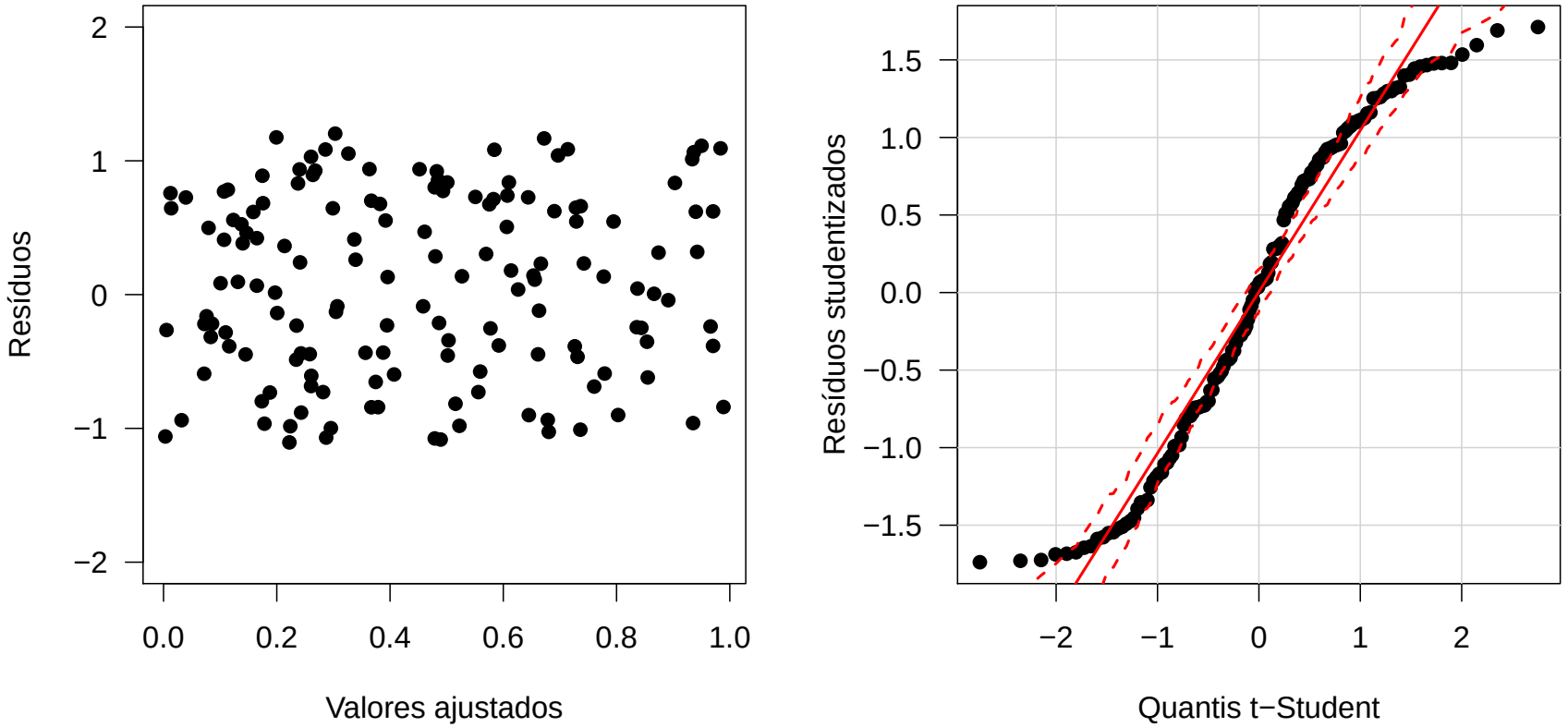

Figura 8: Erros com distribuição simétrica - caudas leves

Padrões em gráficos de resíduos

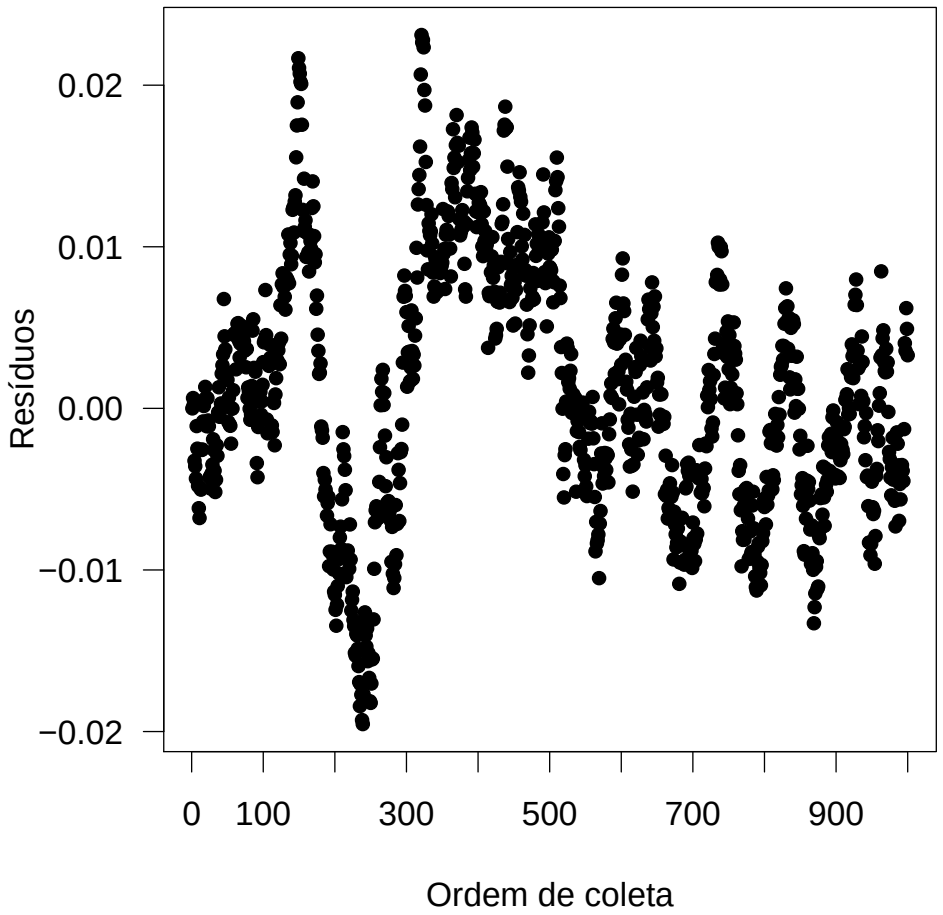

Figura 9: Erros auto-correlacionados

Padrões em gráficos de resíduos

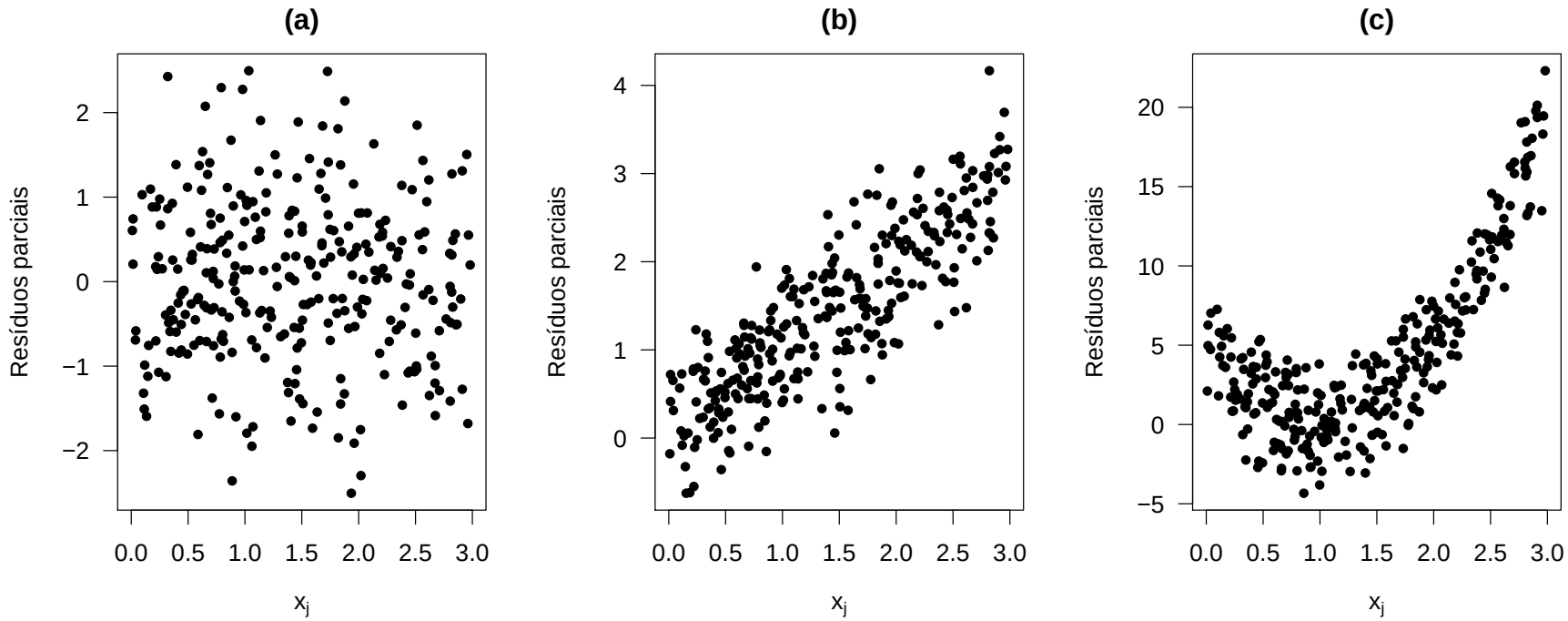

Figura 10: Gráficos de resíduos parciais: (a) Não efeito da variável (ajustado pelo efeito das demais); (b) Efeito linear; (c) Efeito não linear

Padrões em gráficos de resíduos

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

1. Teste de Shapiro-Wilk: A hipótese nula é que os resíduos têm distribuição normal;

2. Teste de Bartlett: Aplicado ao teste da hipótese nula de que os resíduos têm variância constante em k grupos (ou níveis de um fator);

3. Teste de Durbin-Watson: Serve para testar a hipótese nula de que os resíduos não apresentam autocorrelação.

Padrões em gráficos de resíduos

- O uso dos testes em substituição à análise gráfica é **altamente desaconselhável**, porque:

1. Testes de hipóteses não fornecem informações necessárias para avaliar adequadamente o desajuste e identificar medidas corretivas;

2. Desvios moderados (e aceitáveis) das suposições dos modelos podem produzir evidências significativas de desajuste caso a amostra seja suficientemente grande;

3. Para amostras pequenas, os testes podem não ter poder suficiente para indicar desvios consideráveis (e não aceitáveis) das suposições assumidas.

Identificando observações não usuais

- Neste ponto vamos tratar de observações que apresentam comportamento atípico numa análise de regressão:

1. **Outliers:** Observações que não são bem ajustadas pelo modelo;
2. **Observações influentes:** Observações que afetam alguma propriedade do modelo ajustado de maneira substancial;
3. **Ponto de alavanca:** É um ponto extremo no espaço das variáveis explicativas.

- Uma mesma observação pode apresentar duas ou mesmo as três características simultaneamente.

Identificando observações não usuais

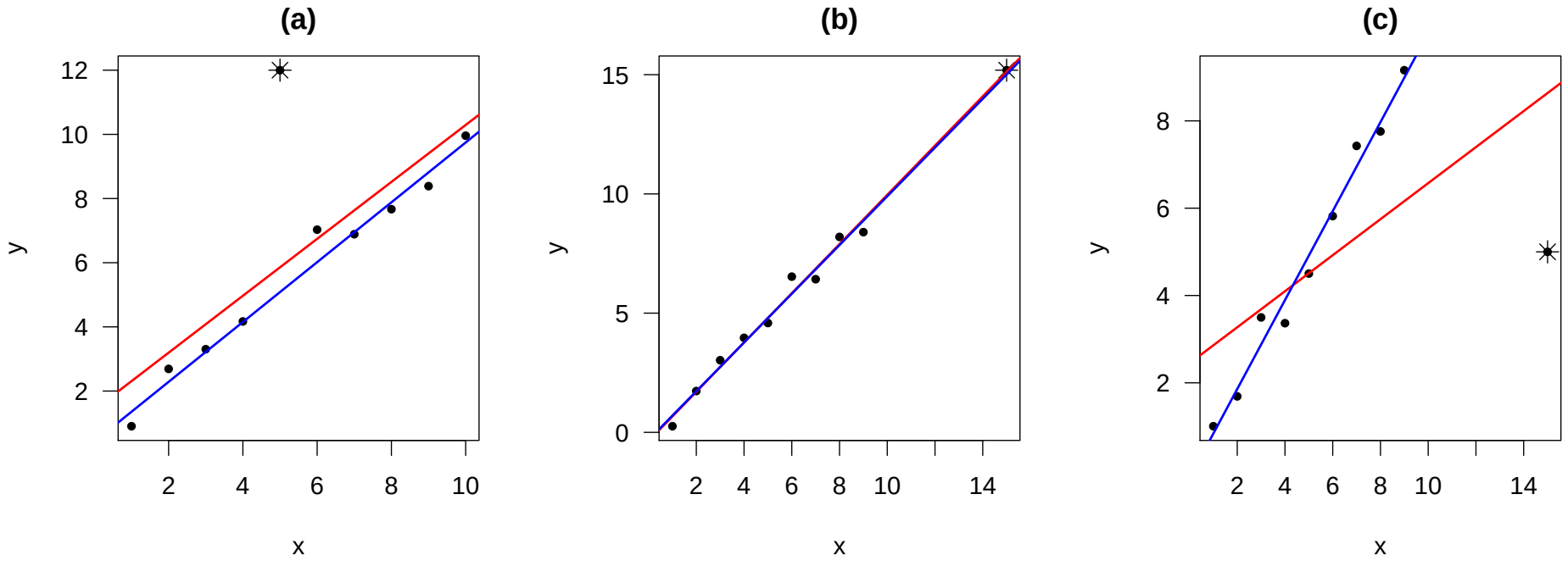

Figura 11: Observações atípicas - as retas em vermelho são ajustadas com todos os pontos e as azuis excluindo as respectivas observações atípicas.

Identificando observações não usuais

- As observações atípicas apresentadas na Figura 11 podem ser classificadas como:

1. Outlier (a): trata-se de um valor extremo de y para o seu particular valor de x. No entanto, não pode ser classificado como ponto de alavanca ou influente;

2. Ponto de alavanca (b): trata-se de um ponto com valor extremo de x. No entanto não é um valor mal ajustado pelo modelo, nem tem grande influência no ajuste;

3. A observação em (c) apresenta as três características atípicas: é um ponto extremo quanto a x, claramente influente e mal ajustado pela reta de regressão (extremo quanto a y).

Outliers

- A maneira mais eficaz de identificar outliers é através da análise dos resíduos escalonados (por exemplo os resíduos studentizados);
- Resíduos escalonados com valor absoluto maior que 3 são potenciais indicadores de outliers.
- Importante ter em mente que a existência de um "grande número de outliers" deve ser resultado da má especificação do modelo, e não propriamente indicador de observações atípicas.
- Outliers devem ser cuidadosamente avaliados, e a causa dos correspondentes valores investigada.

Outliers

- Dependendo da origem do outlier, a observação pode (e deve) ser excluída da análise.
- Algumas causas que justificam a exclusão da observação são a coleta ou o registro incorreto do dado (se possível, ele deverá ser corrigido) e problemas nos instrumentos de medida, dentre outros.
- Em outros casos, não há uma justificativa de ordem operacional para excluir o outlier (a observação é atípica mas sua ocorrência é plausível).
- Nesses casos **não se deve eliminar a observação da análise** simplesmente com o objetivo de obter um melhor ajuste.

Outliers

- Um procedimento recomendável para a análise de regressão na presença de outliers é checar o efeito desses dados nos principais resultados do ajuste.

- Para isso, pode-se ajustar um novo modelo para a base sem outliers e comparar os resultados aos obtidos com o uso da base completa.

- Alterações substanciais nas estimativas, como trocas de sinais, ou mudanças nas significâncias dos parâmetros devem ser relatadas, complementando a análise.

Pontos de alavanca

- Pontos de alavanca correspondem a observações com valores atípicos (extremos) no espaço das variáveis explicativas.

- Pontos remotos no espaço das covariáveis são potencialmente (mas não necessariamente) pontos influentes, podendo alterar de maneira substancial as estimativas e correspondentes erros padrões, dentre outros.

- A forma mais eficiente de detectar pontos de alavanca é através da matriz chapéu:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'. \tag{9}$$

Pontos de alavanca

- Já vimos que $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. Desta forma:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + ... + h_{ii}y_i + ... + h_{in}y_n, \quad i = 1, 2, ..., n. \tag{10}$$

- Assim, h_{ii} pode ser interpretado como o peso exercido por y_i em seu próprio ajuste ($\hat{y}_i$).

- Observações com valores extremos para h_{ii} são pontos de alavancagem.

Pontos de alavanca

- Adicionalmente, pode-se mostrar que os elementos h_{ii} estão relacionados à distância de Mahalanobis da i-ésima observação ao centroide de $\boldsymbol{x}$ ($\bar{\boldsymbol{x}}$).

- A distância de Mahalanobis entre $\boldsymbol{x}_i$ e $\bar{\boldsymbol{x}}$ é dada por:

$$D(\boldsymbol{x_i}, \bar{\boldsymbol{x}}) = (\boldsymbol{x_i} - \bar{\boldsymbol{x}})' \hat{\Sigma}^{-1} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}), \tag{11}$$

em que $\hat{\Sigma}$ é a matriz de covariâncias estimada de $\boldsymbol{x}$.

- Assim, quanto mais afastada estiver $\boldsymbol{x}_i$ do centroide de $\boldsymbol{x}$, maior o valor de h_{ii} (e maior o potencial de alavancagem da observação).

Pontos de alavanca

- Outra propriedade importante de $\boldsymbol{H}$ é que seu traço é igual a p, sendo p o rank de $\boldsymbol{X}$.

- Assim, se cada observação contribuir igualmente para o seu próprio "auto ajuste", teremos um h_{ii} médio, para cada observação, igual a p/n.

- É convencional classificar uma observação i como sendo de alavanca caso o correspondente h_{ii} seja maior que $2p/n$.

Nota: Observações com elevado h_{ii} e elevado resíduo studentizado são potenciais pontos influentes.

Observações influentes

- Observações influentes são aquelas que, quando removidas da base de dados, produzem expressiva mudança no ajuste do modelo.

- As estratégias usadas para identificação de observações influentes fazem uso da estratégia *leave one out*.

- Neste caso, determinada propriedade (estimativa de parâmetros, predições,...) do modelo é avaliada para os modelos ajustados considerando toda a base e mediante exclusão de cada observação da base.

- Na prática, não há necessidade de proceder os ajustes de todos os n modelos, havendo expressões para o cálculo das medidas de interesse usando apenas o ajuste baseado na base completa.

Observações influentes

- Uma das principais medidas de influência é a **distância de Cook**, definida como a distância das estimativas de mínimos quadrados obtidas com as n observações $(\hat{\beta})$ para as estimativas obtidas mediante exclusão da base da observação i $(\hat{\beta}_{(i)})$:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' \boldsymbol{X}' \boldsymbol{X} (\hat{\beta}_{(i)} - \hat{\beta})}{pQM_{Res}}, \quad i = 1, 2, ..., n. \tag{12}$$

- Uma regra usual é classificar como influentes observações tais que $D_i > 1$.
- Uma forma equivalente de calcular D_i é dada por:

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, ..., n. \tag{13}$$

Observações influentes

Algumas outras medidas de influência são:

- **DFBetas:** Medem a alteração na estimativa de um particular β_j resultante da deleção da i-ésima observação:

$$DFBetas_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QM_{Res_{(i)}} C_{jj}}}, \quad i = 1, 2, ..., n, \tag{14}$$

em que $\hat{\beta}_{j(i)}$ e $QM_{Res_{(i)}}$ são calculados mediante exclusão da i-ésima observação e C_{jj} é o j-ésimo elemento da diagonal de $(\boldsymbol{X'X})^{-1}$.

- Recomenda-se investigar observações para as quais $|DFBetas_{j,i}| > 2/\sqrt{n}$.

Observações influentes

- **DFFITS:** Mede a alteração na predição ou valor ajustado de uma observação resultante de sua deleção:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{QM_{Res_{(i)}} h_{ii}}}. \tag{15}$$

- Recomenda-se investigar observações para as quais $|DFFITS_i| > 2/\sqrt{p/n}$.

Observações influentes

- Uma vez detectada uma ou mais observações influentes, é necessário avaliar adequadamente o impacto dessas observações nos principais resultados da análise;
- Quanto a deletar tais observações, as mesmas orientações apresentadas quanto ao tratamento de outliers se aplicam aqui;
- Novamente, deve-se avaliar criteriosamente se a presença de múltiplos outliers e obsevações influentes não se deve à má especificação do modelo;
- Uma alternativa para análise na presença de observações atípicas é usar métodos robustos, que atribuam menor peso a tais observações no ajuste do modelo.

Multicolinearidade

- A multicolinearidade se caracteriza por uma quase dependência linear entre as variáveis regressoras.

- Se as colunas da matriz $\boldsymbol{X}$ $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ forem exatamente colineares, ou seja, se houver um conjunto de constantes $c_1, c_2, ..., c_n$ nem todas nulas, tal que:

$$\sum_{j=1}^{p} c_j \boldsymbol{X}_j = \boldsymbol{0}, \tag{16}$$

segue que $(\boldsymbol{X}'\boldsymbol{X})$ é singular, não havendo solução única na estimação por mínimos quadrados.

Efeitos da multicolinearidade

- Nos casos em que as colunas da matriz $\boldsymbol{X}$ exibem uma quase dependência linear, como resultado tem-se baixa precisão (elevado erro) na estimação dos parâmetros do modelo.

- Para o modelo de regressão linear múltipla, a variância de $\hat{\beta}_j$, estimador de um particular parâmetro do modelo, pode ser expressa por:

$$Var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}, \tag{17}$$

em que R_j^2 é o coeficiente de determinação da regressão de x_j nas demais variáveis.

- É fácil observar que $Var(\hat{\beta}_j) \to \infty$ quando $R_j^2 \to 1$.

Diagnóstico de multicolinearidade

- O termo $VIF_j = 1/(1 - R_j^2)$ é chamado **fator de inflação da variância** e pode ser utilizado para diagnóstico de multicolinearidade.
- Se as colunas de $\boldsymbol{X}$ forem ortogonais, então $VIF_j = 1$ para todo j.
- Quanto mais próximos de 1 os valores de VIF_j, menor a preocupação com a multicolinearidade e seus efeitos;
- Uma regra prática, mas não formal, para indicação de multicolinearidade é a identificação de qualquer $VIF_j > 10$.

Como lidar com a multicolinearidade

- Alguns procedimentos podem ser adotados para contornar o problema da multicolinearidade, dentre eles:

1. Coleta de dados adicionais: coletar dados em regiões do espaço de covariáveis não amostradas (ou amostradas com baixa frequência);

2. Reespecificação do modelo: por exemplo, se as variáveis x_1, x_2 e x_3 exibirem multicolinearidade, pode-se optar por:
   - Substituí-las por alguma função que preserve a informação original mas reduza a colinearidade (ex: $z = (x_1 + x_2 + x_3)/3$ ou $w = x_1 x_2 / x_3$ ou...);
   - Eliminar uma ou duas das variáveis pode ser uma alternativa, embora isso possa reduzir o poder preditivo do modelo;

Como lidar com a multicolinearidade

3. Regressão Ridge - O método ridge consiste em encontrar um estimador $\hat{\beta}^*$ que seja viciado mas com menor variância que $\hat{\beta}$, o estimador de mínimos quadrados.

4. Regressão com componentes principais - O método de componentes principais permite identificar um conjunto de $q < p$ combinações lineares ortogonais das variáveis regressoras originais que expliquem a maior parcela possível da variação original presente em $\boldsymbol{X}$.

- Após identificadas as novas variáveis (componentes), as p variáveis originais podem ser substituídas pelos q componentes principais no ajuste do modelo de regressão.