Análise dos resíduos e *Outlier*, Alavancagem e Influência

Diagnóstico na análise de regressão

Usadas para detectar problemas com o ajuste do modelo de regressão.

- Presença de observações mal ajustadas (pontos aberrantes).
- Presença de pontos de alavanca
- Presença de pontos influentes
- Adequação das suposições iniciais para os erros
  - Normalidade dos erros
  - Independência dos erros
  - Variância constante (homogeneidade) dos erros
  - Relação linear entre as variáveis X e Y

Qualidade do modelo	Métodos gráficos Testes estatísticos

Matriz de projeção H

$$H = X(X^T X)^{-1} X^T$$

- h_{ii} – mede o quão distante a observação y_i está das demais n-1 observações no espaço definido pelas variáveis explicativas.
- h_{ii} – só depende do valor das variáveis explicativas.
- Esse elemento mede a influência da í-ésima resposta sobre seu valor ajustado.
- Representa uma medida de alavancagem
- Se h_{ii} é grande (próximo de 1), existem valores atípicos das variáveis explicativas. Alavancagem

$$\hat{\boldsymbol{y}} = H\boldsymbol{y}$$

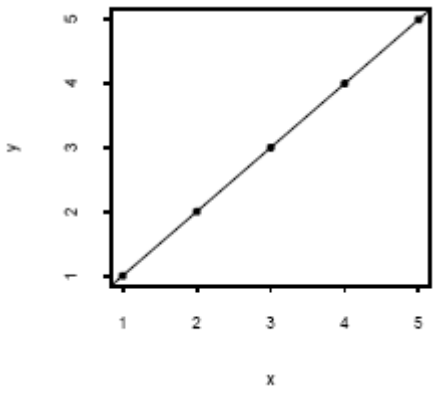
$$h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}} \right)$$

$$h_{ij} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{XX}} \right)$$

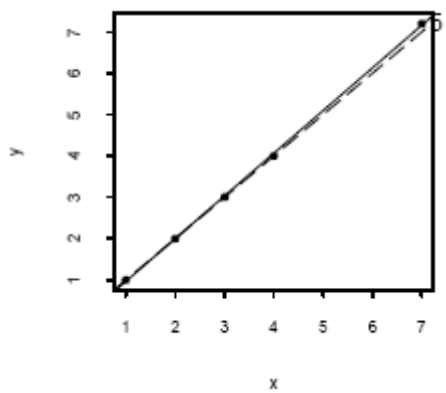
$$\frac{1}{n} \leq h_{ii} \leq 1 \qquad \sum_i h_{ii} = p$$

Alavancagem

- O elemento h_{ii} representa uma medida de alavancagem;
- Heurística: Se $h_{ii} > 2p/n$ => y_i é pto de alavanca.



Pontos sem nenhuma pertubação



Ponto 5 é um ponto de alavanca

Resíduos

Diagnóstico para a variável resposta é realizado através de uma análise de resíduos. Os resíduos são definidos como:

$$r_i = Y_i - \hat{Y}_i$$

Os resíduos podem ser considerados como erros observados, para distingui-los do erro verdadeiro desconhecido ε_i no modelo de regressão:

$$\varepsilon_i = Y_i - E(Y_i)$$

Para o modelo de regressão, temos:

suposição $\longrightarrow$ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Resíduos: Detecção de *outliers*

Violações de suposições é comum em situações em que existem pontos aberrantes

Se o modelo é adequado para os dados, os resíduos observados devem refletir essas suposições.

1. **A relação entre y e os regressores é aproximadamente linear**
2. **O erro tem média zero e variância constante**
3. **Os erros são não correlacionados**
4. **Os erros são normalmente distribuídos**

Violação das suposições acima podem produzir instabilidade no modelo.

Quando o tamanho da amostra é grande em comparação com o número de parâmetros no modelo de regressão, o efeito de dependência entre os resíduos e_i é relativamente sem importância e pode ser ignorado.

Resíduos

1. Resíduos ordinários

$$r_i = y_i - \hat{\mu}_i$$

- Não é muito informativo;
- Não permite detectar pontos remotos.
- Variância não é constante (não propicia comparação). Necessita ser padronizado.
- $r_i \sim N(0, \sigma^2(1-h_{ii}))$.
- Quanto maior for h_{ii} (alavancagem) menor será a variabilidade do resíduo. O valor estimado de y é praticamente determinado pelo valor de y.

Resíduos

2. Resíduos padronizados usando MS_R para estimar a variância. Os resíduos usualmente são menores.

$$r_i^{**} = \frac{r_i - \bar{r}}{\sqrt{MS_R}} = \frac{r_i}{\sqrt{MS_R}}$$

Um valor r>3 indica um ponto aberrante

Vantagens:
Se o modelo está correto, todos resíduos tem a mesma variância.
Apropriado para verificar normalidade dos erros e homogeneidade.

Desvantagem: É difícil detectar violações do modelo usando resíduos ordinários ou padronizados pois esses resíduos são menores.

Resíduos

3. Resíduos padronizados (semi-studentizados)

$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{MS_R(1-h_{ii})}}$$

Um valor $|r^*| > 2$ indica um ponto aberrante

- Variância constante (propicia comparação)
- Usados na verificação de normalidade dos erros e homogeneidade das variâncias.
- Dependência entre r^*_i e $\hat{\sigma}^2$
- Pode-e contornar a dependência obtendo uma estimativa para σ^2 removendo a i-ésima obervação. Com isso y_i e $\hat{y}_i$ são independentes.
- Desvantgaem: r_i não tem distribuição t-student.

Quando n é grande os resíduos semi-studentizados não diferem muito dos resíduos padronizados.
Quando o h_{ii} é alto (próximo de 1) e o resíduo é grande, o ponto tem grande chance de ser um ponto influente. Nesse caso, é recomendável o uso dos resíduos semi-studentizados.

Resíduos

4. Resíduos Studentizado $$t_i = \left(\sqrt{\frac{n-p-1}{n-p-r_i^{*2}}} \right) r_i^*$$

•É baseada na estimativa de variância para os erros removendo a i-ésima observação.

• t_i tem distibuição t-student com n-p-1 graus de liberdade.

• É um teste de hipótese para ponto aberrante. Compara-se o valor absoluto de ti com t(α/2,n-p-1). Se o ti observado for maior, o ponto é um *outilier*.

•Uma abordagem usando um valor *cutoff* é mais usual do que comparar cada valor de tt com t(α/2,n-p-1).

• Em muitas situações, resíduos studentizados diferem dos

•resíduos semi-studentizados. Contudo, se uma observação é um ponto influencial a estatística será mais sensível para esse ponto.

• Detecção de outilier necessita de investigação simultânea com detecção de ponto influente.

Influência

- Conhecer o grau de dependência entre o modelo ajustado e o vetor de observações (y);
- As técnicas são baseadas na exclusão de uma única observação;
- Procuram medir o impacto dessa perturbação nas estimativas dos parâmetros.
- Se não existem pontos influentes, pode-se confiar mais no modelo proposto.

Influência

- Distância de Cook (D_i) introduzida por Cook (1977)

$$D_i = \frac{h_{ii}}{p(1-h_{ii})} r_i^*$$

- Combina o resíduo padronizado (ri*) com a medida de alavancagem (h_{ii});
- Se Di > Fp, n-p (0.5) => yi será influente.
- Como na distribuição F o quantil 50 é aproximadamente 1, recomenda-se na prática que se a distância de Cook for muito menor que 1, então a eliminação do ponto não irá influenciar as estimativas dos parâmetros.
- Para investigar melhor a influencia com maiores valores de D_i, deve-se eliminar essas observações e re-computar as estimativas dos parâmetros.

Não é adequada quando h_{ii} é proximo de zero e o resíduo é studentizado é grande.

Influência

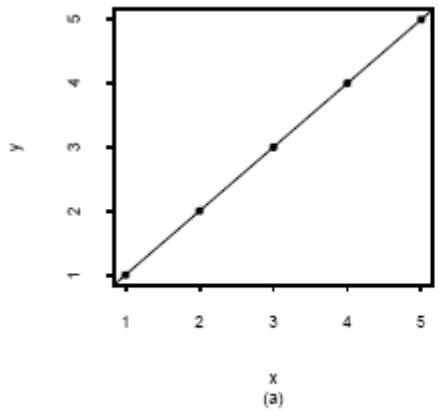
- DFFITS introduzida por Belsley et al. (1980)

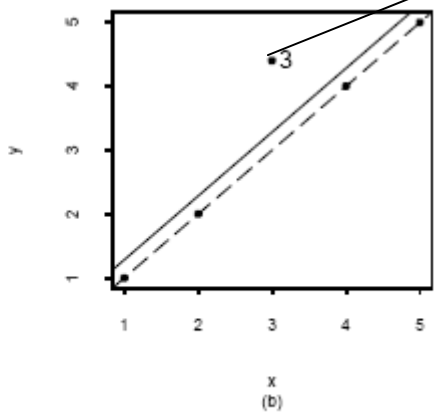
$$DFFITS_i = t_i \left(\frac{h_{ii}}{p(1-h_{ii})} \right)^{1/2}$$

- Medida alternativa a Di
- Combina o resíduo studentizado (t_i) com a medida de alavancagem (hii);
- Se DFFITSi > 2{p/(n-p)}1/2 => yi será influente.
- **Não é adequada quando h_{ii} é proximo de zero e o resíduo é studentizado é grande.**

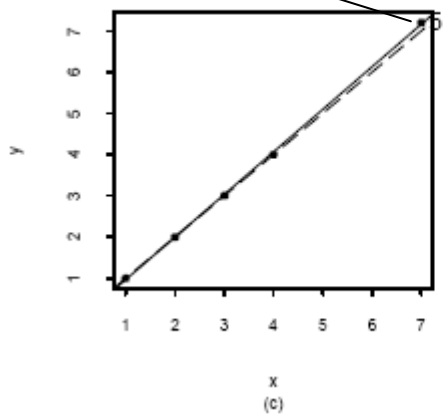
Tratamento de observações influenciais : Recomenda-se o uso de técnicas de regressão robusta (mínimos quadrados reponderados) ou regressão com suposição de distribuição de caudas pesadas para os erros.

Outlier

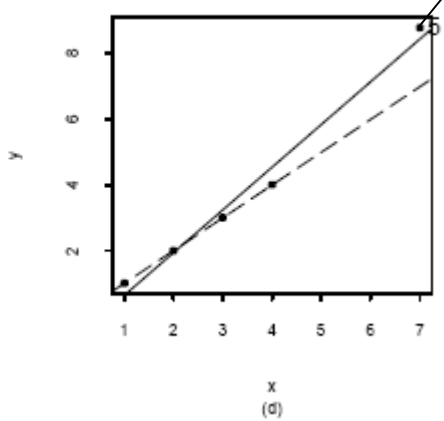




Ponto de alavanca



Outlier e ponto de alavanca



Diagnóstico

Técnicas gráficas são mais informativas do que os testes de hipóteses. Testes não são largamente usados.

1. Gráfico dos resíduos padronizados versus a ordem das observações para observações aberrantes.
2. Gráfico dos resíduos padronizados versus os valores ajustados para verificar aleatoriedade dos resíduos
3. Gráfico de probabilidade dos resíduos padronizados ordenados versus os quantis da distribuição normal padrão para verificar suposição de normalidade.
4. Gráfico dos resíduos padronizados versus regressor para verificar verificar heterocedasticidade e relação não linear
5. Gráfico de h_{ii} versus ordem das observações para verificar pontos de alavancagem.
6. Gráfico de D_i ou DFITTS versus ordem das observações para verificar pontos de influência.

Não linearidade da função de regressão ou heterocedasticidade da variância:

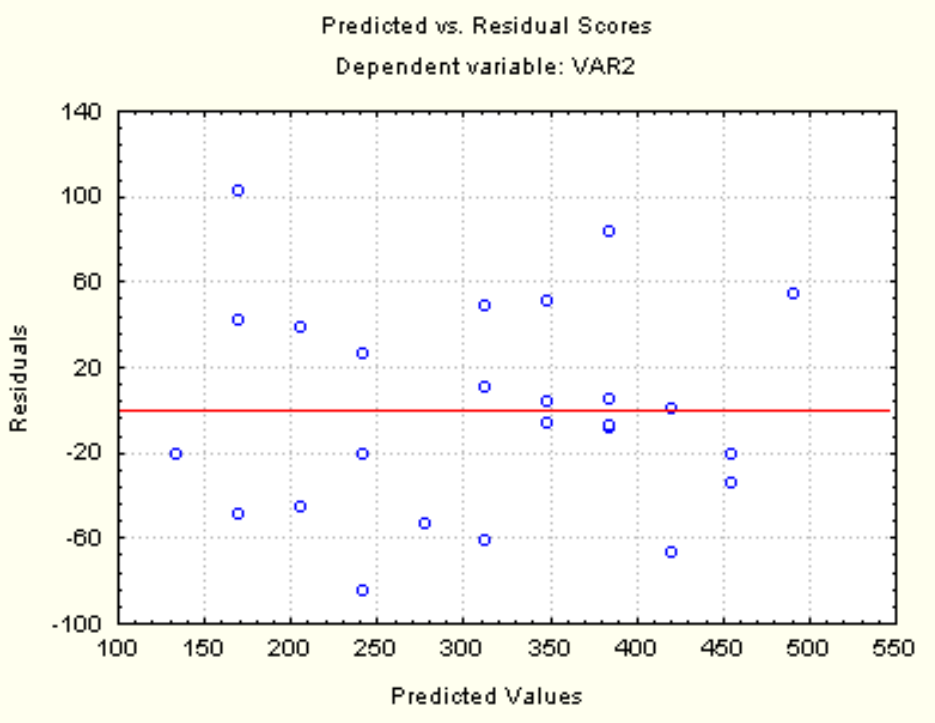
A verificação de que a função de regressão é adequada aos dados pode ser feita através do gráfico dos resíduos versus valores ajustados ou dos resíduos versus variáveis preditoras.

Os resíduos devem estar aleatoriamente distribuídos em uma faixa de valores.

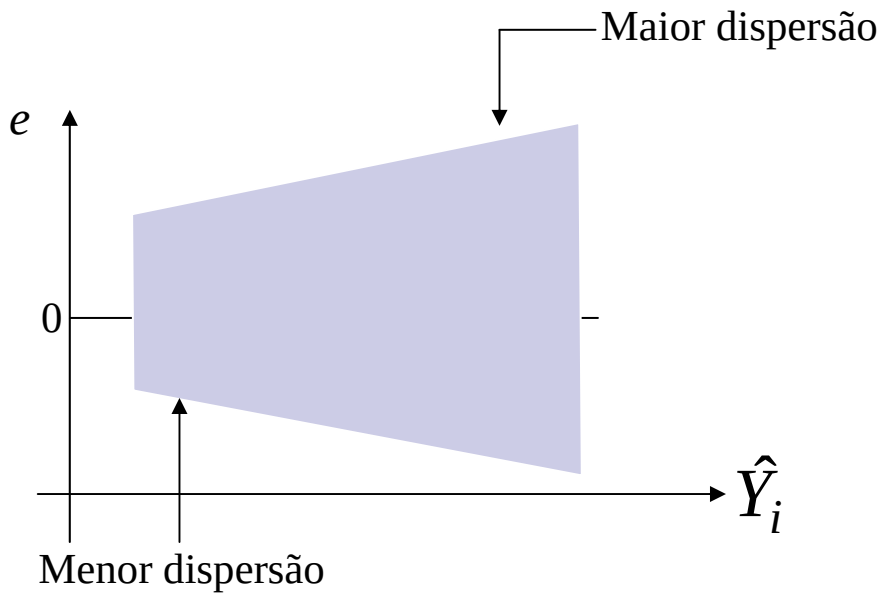
Caso verificar-se um comportamento sistemático, termos adicionais ou alternativos devem ser incluídos no modelo. Por exemplo, termo quadrático no regressor ou uma transformação deve ser considerada.

Não linearidade da função de regressão ou heterocedasticidade da variância:

Nesta figura temos um protótipo da situação em que um modelo de regressão linear é adequado. Observe que os resíduos se distribuem aleatoriamente em torno da média zero.



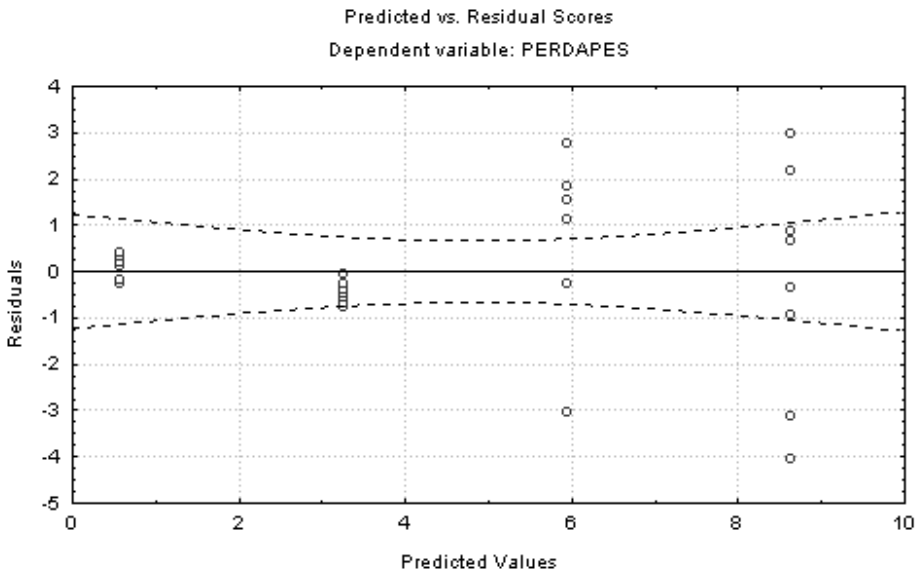
Heterocedasticidade da variância:



Uma abordagem usual para tratar com desigualdade de variância é aplicar transformação na variável resposta ou no regressor ou ainda usar o método de mínimos quadrados ponderados.

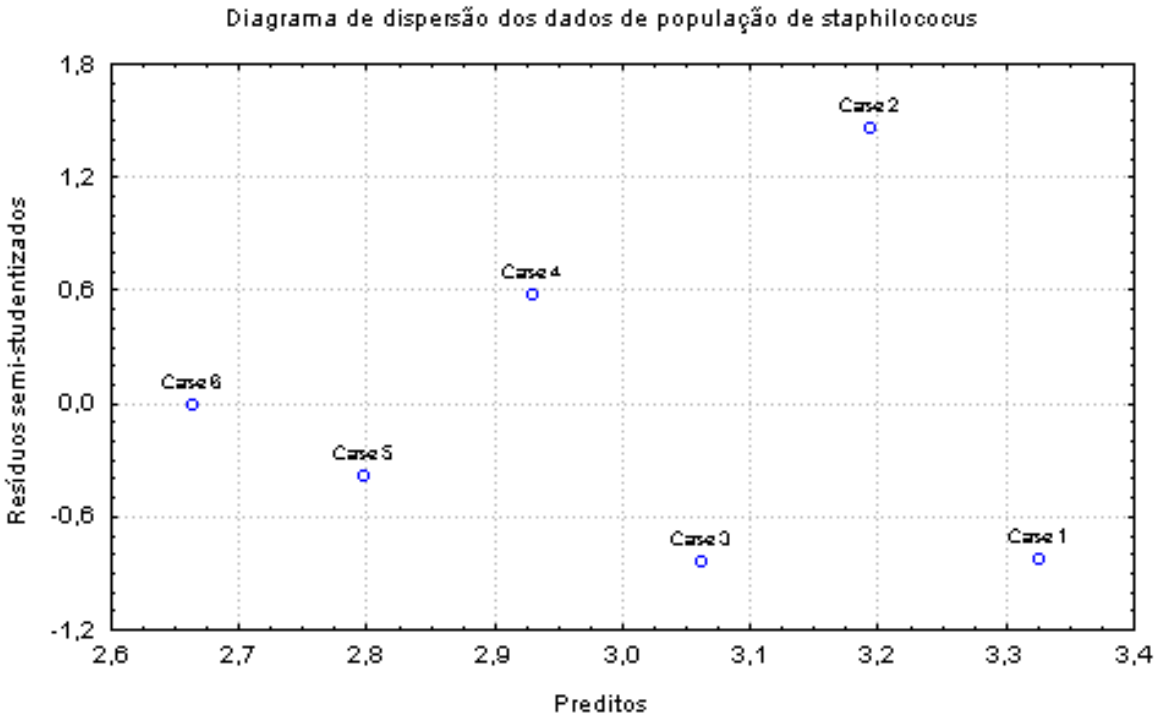
Heterocedasticidade da variância:

O gráfico dos resíduos versus valores preditos (ajustados) mostra que quanto maiores são os valores preditos maior é a dispersão dos resíduos.



Presença de outliers

Outliers são valores extremos, atípicos, ou seja, são observações que não são bem ajustadas pelo modelo. Resíduos que são outliers podem ser identificados a partir de um gráfico dos resíduos versus a variável preditora ou valores ajustados. Pode-se usar também o box-plot ou ramo-e-folhas. O uso dos resíduos padronizados são particularmente úteis, pois é fácil identificar resíduos que estão muitos desvios padrões a partir de zero. *Regra:* considera-se outliers os resíduos que estão 4 ou mais desvios padrões a partir de zero.



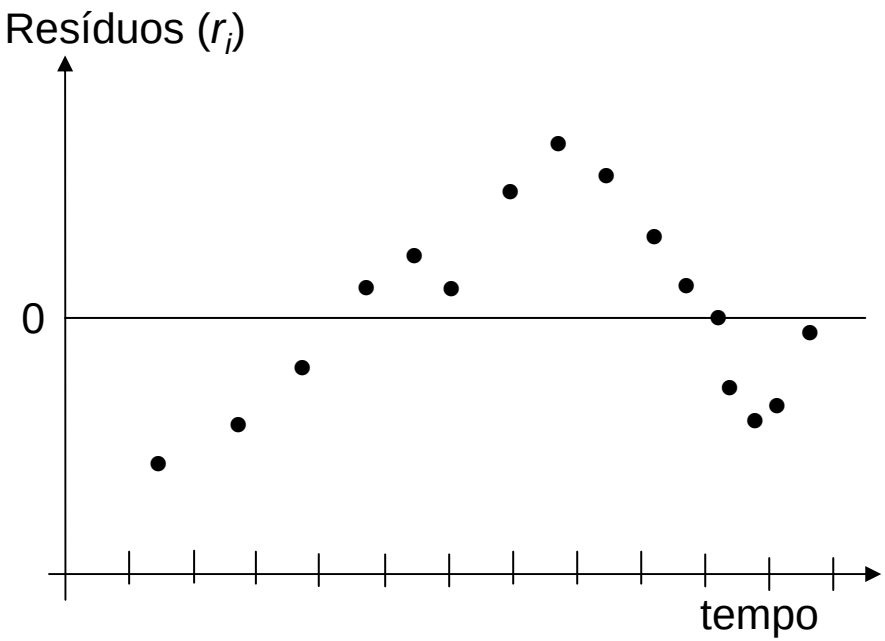
O gráfico ao lado apresenta os resíduos padronizados e não contém outliers.

Outliers podem introduzir grandes dificuldades na análise estatística. Deve-se descartar um outlier se ele representa um erro de registro, erro de medida, falha de equipamento ou algum outro problema similar.

Recomenda-se usar técnicas de regressão robusta quando existem *ouliers*.

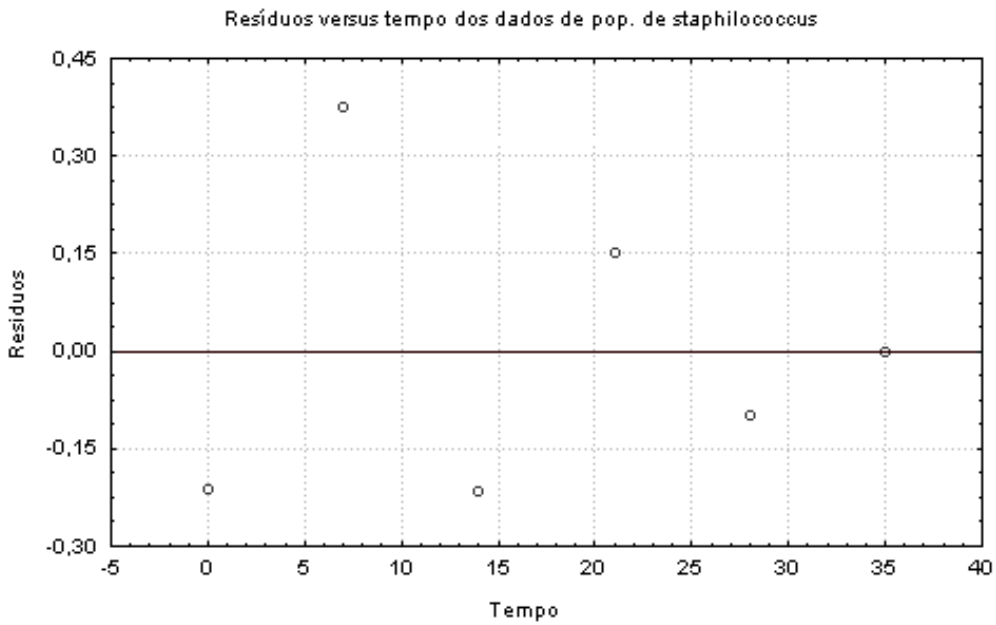
Falta de independência dos erros

Se a seqüência de tempo em que os dados foram coletados é conhecida, deve-se fazer um gráfico dos resíduos versus seqüência.



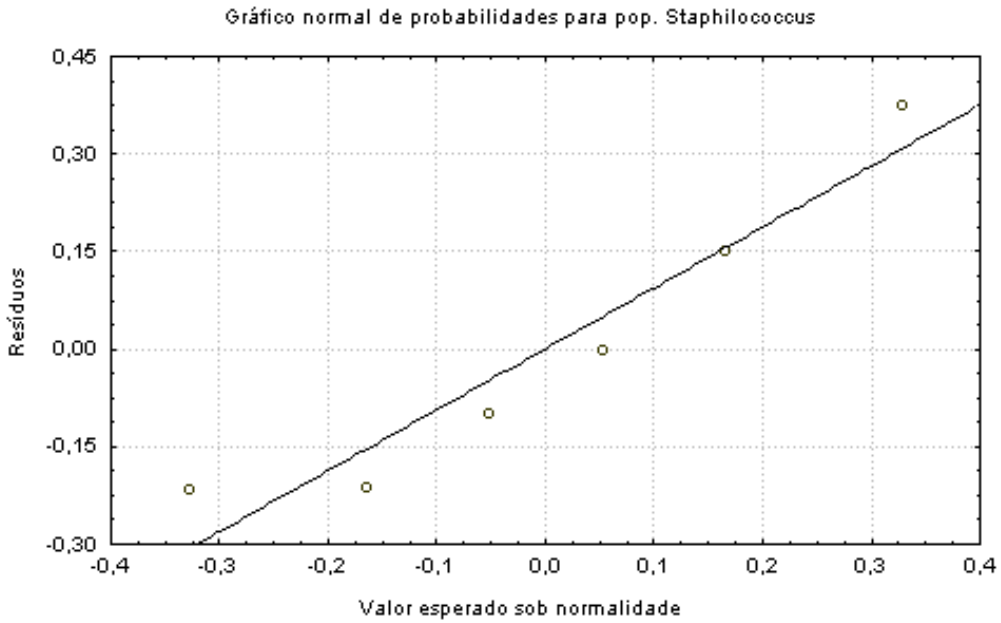
Quando os resíduos são independentes, eles devem se distribuir aleatoriamente em torno de zero. Deve alternar os pontos em torno de zero.

Independência dos erros

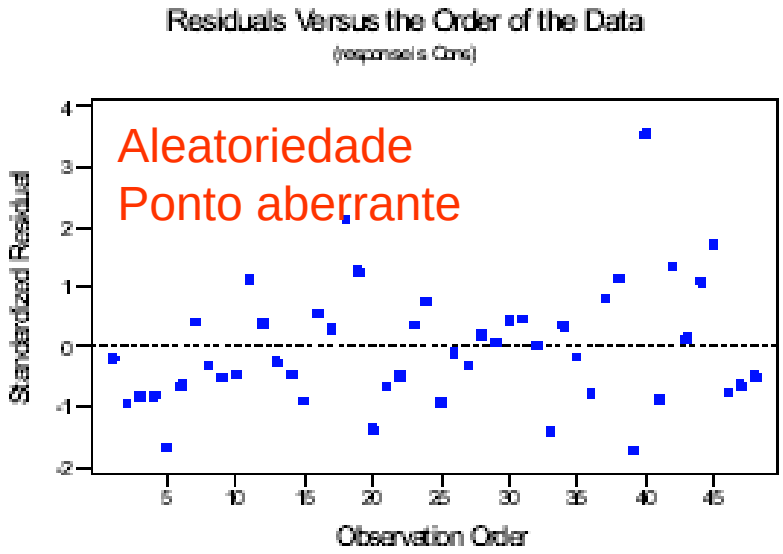


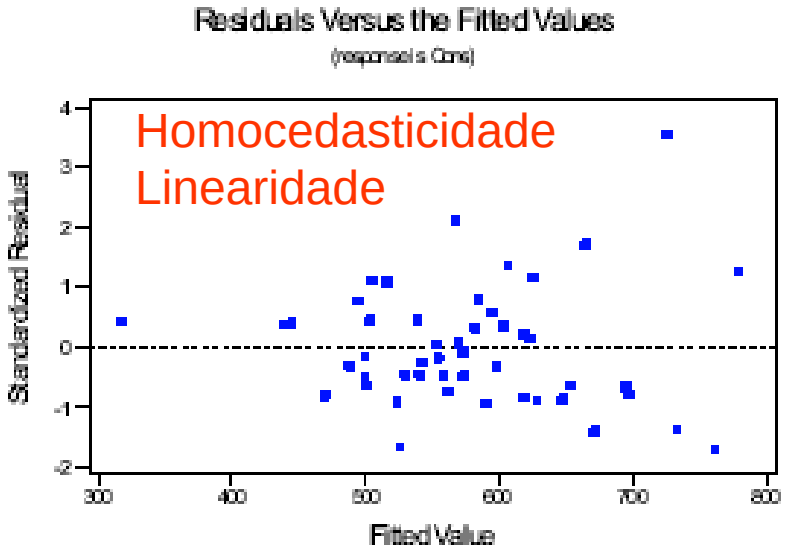
Normalidade dos erros

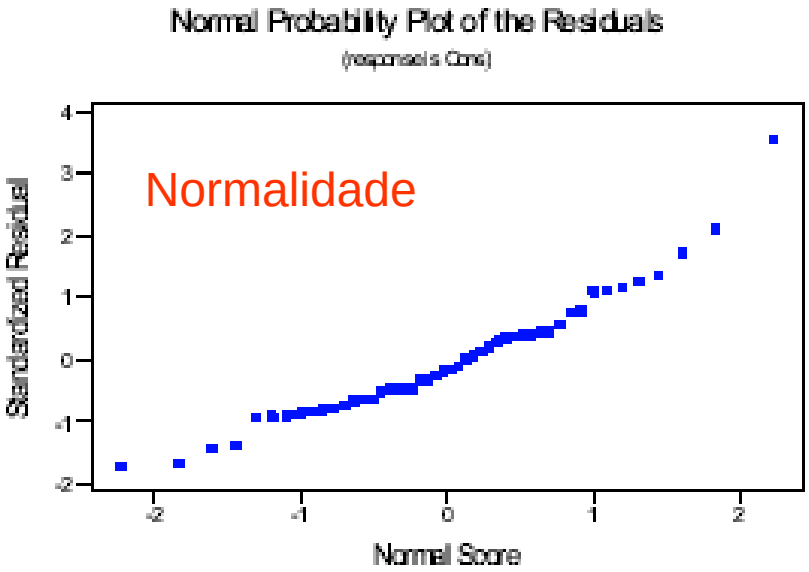
Situação: normalidade dos erros.

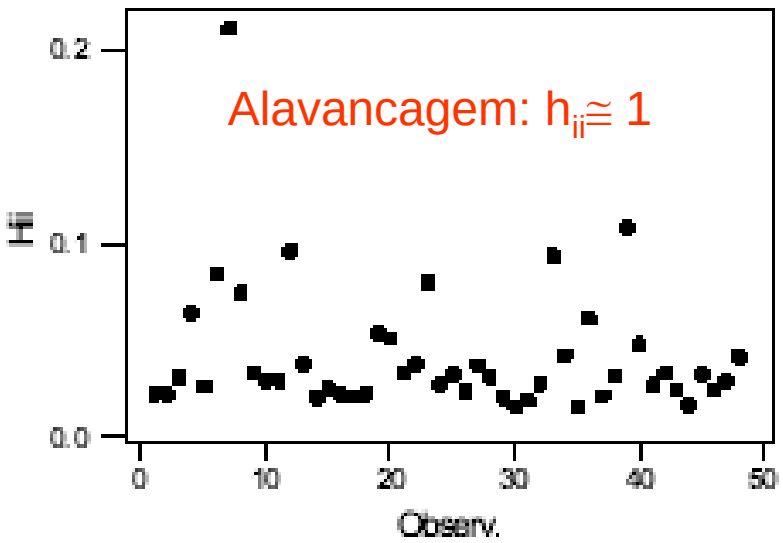


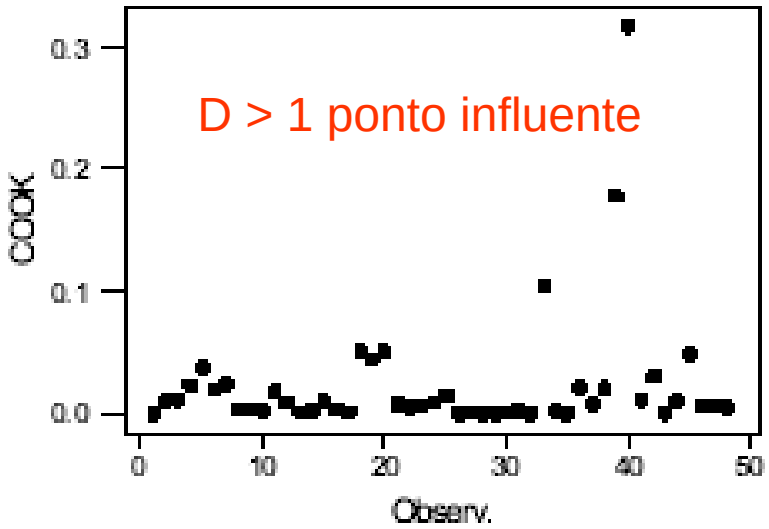
Resíduos muito afastados da linha indicam não normalidade para os erros ou presença de *ouliers* (distribuição de caudas pesadas).

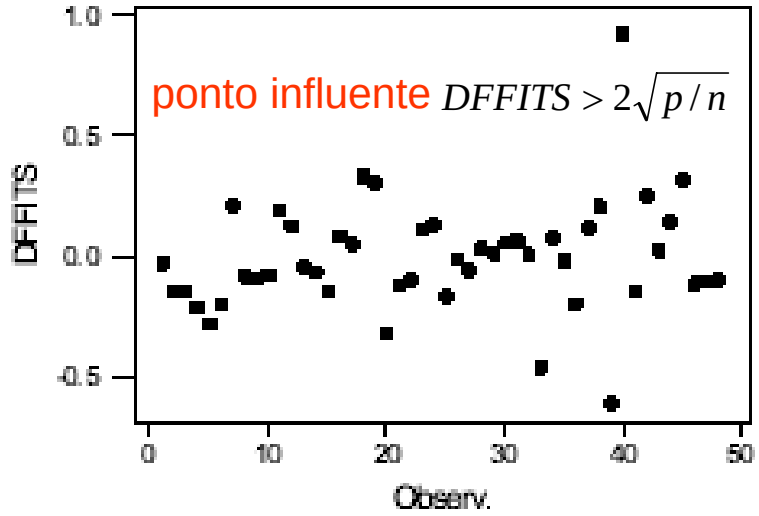

Residuals Versus the Order of the Data
(response is Conc)
Aleatoriedade
Ponto aberrante
Standardized Residual
Observation Order


Residuals Versus the Fitted Values
(response is Conc)
Homocedasticidade
Linearidade
Standardized Residual
Fitted Value


Normal Probability Plot of the Residuals
(response is Conc)
Normalidade
Standardized Residual
Normal Score


Alavancagem: $h_{ii} \cong 1$
Hii
0.2
0.1
0.0
0
10
20
30
40
50
Observ.


D > 1 ponto influente
COOK
0.3
0.2
0.1
0.0
0
10
20
30
40
50
Observ.


ponto influente $DFFITS > 2\sqrt{p/n}$
DFFITS
1.0
0.5
0.0
-0.5
0
10
20
30
40
50
Observ.