Modelos de Regressão Múltipla - Parte VIII

Erica Castilho Rodrigues

21 de Outubro de 2014

Observações não usuais

As observações não usuais podem ser:

- **Outliers:** não se ajustam bem ao modelo (resíduo grande);
- **Alavancas:** têm valor não usual da variável explicativa;
- **Influentes:** quando presentes, mudam o ajuste do modelo substancialmente.

- Nem todo outlier é influente.
- Assim como nem toda alavanca é influente.
- Mas um ponto influente é um outlier e/ou uma alavanca.
- Os resíduos exercem importante papel na análise de observações não usuais.

A matriz H

- Lembre que no modelo de regressão

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  o estimador de mínimos quadrados é dado por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- Os valores estimados para $\mathbf{Y}$ são dados por

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{H}\,\mathbf{Y}\,.$$

- A matriz

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

  é chamada de matriz **hat** (chapéu), pois liga $\mathbf{Y}$ a $\hat{\mathbf{Y}}$

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}\,.$$

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & & & \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix}$$

- A matriz **H** é simétrica? Sim, pois

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

e

$$\mathbf{H}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \, .$$

- A matriz **H** é idempotente

$$\mathbf{H}^p = \mathbf{H}, \text{ para qualquer potência } p.$$

- Vejamos porque isso é verdade

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$$

- O vetor de resíduos pode ser escrito em função da matriz **H**

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

- A matriz de covariâncias do vetor de resíduos é dada por

$$Var(\mathbf{e}) = (\mathbf{I} - \mathbf{H})Var(\mathbf{Y})(\mathbf{I} - \mathbf{H})'$$

$$= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})'$$
$$= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})'$$

- Mas temos que

$$(\mathbf{I} - \mathbf{H})' = \mathbf{I} - \mathbf{H}' = \mathbf{I} - \mathbf{H}$$

pois $\mathbf{H}' = \mathbf{H}$.

- Além disso

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}^2 = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}$$

pois $\mathbf{H}$ é indepotente, logo

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' = I - \mathbf{H}\,.$$

- E portanto

$$Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$

- Os termos da diagonal nos fornecem as variâncias dos resíduos

$$Var(e_i) = \sigma^2(1 - h_{ii})\,.$$

- Não é a variância dos erros!
- Quanto vale a variância dos erros? σ^2

- Os termos fora da diagonal nos fornecem as covariâncias dos resíduos

$$Cov(e_i, e_j) = -\sigma^2(h_{ij}) \text{ para } i \neq j \ .$$

- Iremos mostrar agora que

$$\sum_i h_{ii} = p \ .$$

- Temos que

$$\sum_i h_{ii} = \text{traço}(\mathbf{H}) \ .$$

- Vamos usar a seguinte propriedade

$$\text{traço}(\mathbf{AB}) = \text{traço}(\mathbf{BA}) \ .$$

- Temos então que

$$\text{traço}(\mathbf{H}) = \text{traço}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \text{traço}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})$$

$$= \text{traço}(\mathbf{I}_p) = p\,.$$

- Mostramos assim que

$$\sum_i h_{ii} = \text{traço}(\mathbf{H}) = p\,.$$

- Temos ainda que

$$h_{ii} < 1, \text{ pois } Var(e_i) = \sigma^2(1 - h_{ii}) > 0\,.$$

Tipos de Resíduos

- Temos três principais tipos de resíduos.

Resíduo ordinário

- Definido por

$$e_i = Y_i - \hat{Y}_i \, .$$

- Se as suposições do modelo estão corretas

$$Var(e_i) = \sigma^2(1 - h_{ii}) \quad Cov(e_i, e_j) = -\sigma^2 h_{ij}$$

Resíduo “estudentizado” internamente

- É definido por

$$r_i = \frac{e_i}{\sqrt{S^2(1 - h_{ii})}}$$

onde S^2 é a estimativa de σ^2 do modelo ajustado com todas observações.

- Se as suposições do modelo estão corretas

$$Var(r_i) = 1 \quad Cov(r_i, r_j) \text{ tende a ser pequena.}$$

Resíduo “estudentizado” externamente

- É definido por

$$t_i = \frac{e_i}{\sqrt{S_i^2(1 - h_{ii})}}$$

  onde S_i^2 é a estimativa de σ^2 do modelo ajustado sem a i-ésima observação.
- Se a i-ésima observação infla a variância do modelo:
  - S_i^2 será bem menor que S^2;
  - assim não causa a falsa impressão que o resíduo é pequeno.

- Não é necessário reajustar o modelo retirando cada uma das observações, pois

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}$$

- Sob a suposição de normalidade dos erros, temos que

$$t_i \sim t_{n-p} \, .$$

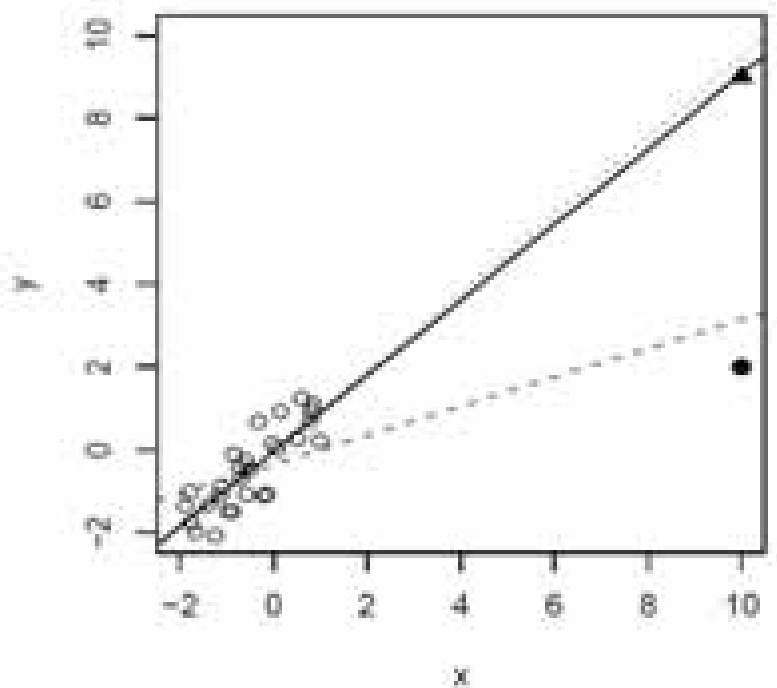
- Isso pode ser usado para detecção de observações atípicas.

Observações Discrepantes (*Outliers*)

Outlier

- Observação com valor discrepante de Y.
- É um ponto que não se ajusta bem ao modelo.
- O valor ajustado $\hat{Y}_i$ é muito distante de Y_i.
- O resíduo e_i é grande.
- Podem afetar o ajuste do modelo.
- Quando retirados, o modelo muda substancialmente.

- A figura abaixo exemplifica a presença de um *outlier* afetando o ajuste do modelo.



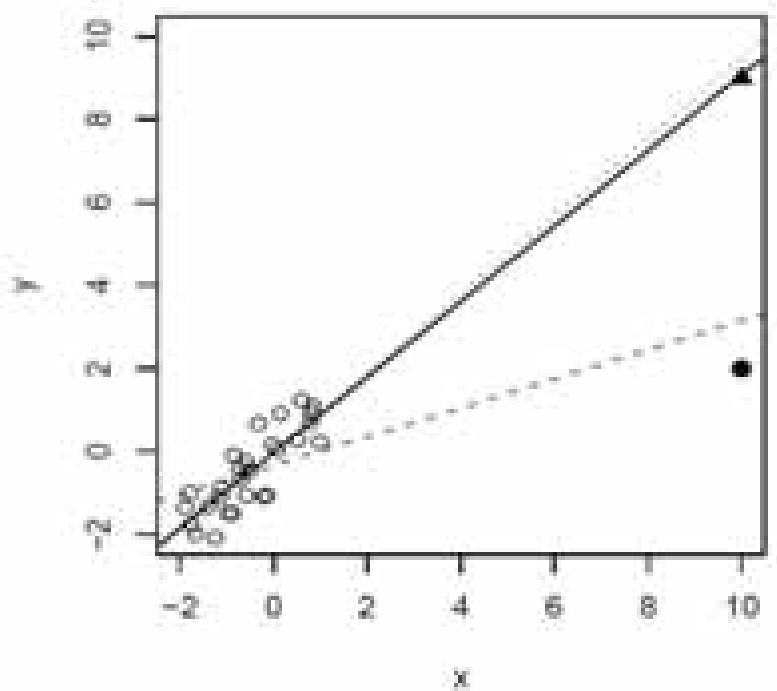
Linha pontilhada: regressão ajustada com os pontos ○;
Linha sólida: regressão ajustada com os pontos ○ e ▲;
Linha tracejada: regressão ajustada com os pontos ○ e ●.

- Nem todo ponto que possui um valor alto de e_i é um *outlier*.
- Como identificar pontos que são de fato *outliers*?
- Valores extremos de t_i na distribuição t_{n-p} indicam que a i-ésima observação é um *outlier*.
- Um *outlier* só deve ser excluído se um erro é identificado.
- Caso isso ocorra deve-se destacar essa exclusão nas conclusões finais do trabalho.

There should be strong non-statistical evidence before we discard any of the outliers. Otherwise, it can be dangerous as "deleting a desirable response" may give the user a false sense of precision in estimation or prediction.

Exemplo

- Vamos analisar o *outlier* presente no seguinte gráfico.



Linha pontilhada: regressão ajustada com os pontos ○;
Linha sólida: regressão ajustada com os pontos ○ e ▲;
Linha tracejada: regressão ajustada com os pontos ○ e ●.

Exemplo (continuação)

- Temos nesse caso que

$$n = 30 \quad p = 2\,.$$

- Devemos comparar t_i com o valor crítico da distribuição t_{28}.
- Temos que

$$t_{28;0.95} = 1,70\,.$$

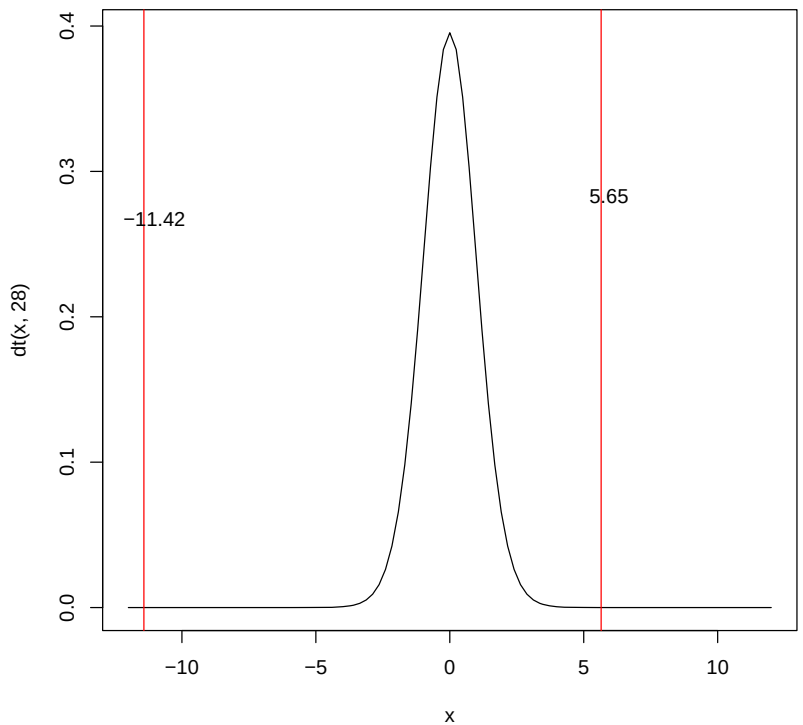
- Os pontos marcados com bola vazia são tais que

$$-1.12 < t_i < 1.12\,.$$

- O ponto marcado com triângulo cheio possui $t_i = 5.65$.
- O ponto marcado com bola cheia possui $t_i = -11.42$.
- Esses dois últimos pontos são disprepantes na distribuição t_{28}.

Exemplo (continuação)

- A figura a seguir mostra onde estão localizados os dois pontos na distribuição t_{28}.



Observações Alavanca (*Leverage*)

Observações alavanca

- Observações extremas em X.
- Observações tais que o valor de X está muito afastado dos demais.
- Elas não necessariamente afetam a reta de regressão.

- Seja a matriz

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

com elementos da diagonal $h_{ii} < 1$.

- Sabe-se que o resíduo da i-ésima observação tem variância

$$Var(e_i) = \sigma^2(1 - h_{ii}) \, .$$

- Se h_{ii} é alto ($h_{ii} \approx 1$) $\Rightarrow Var(e_i) << \sigma^2$.
- Como consequência o ajuste é forçado a se aproximar de Y_i.
- Ou seja, $Y_i \approx \hat{Y}_i$, pois

$$\hat{Y} = \mathbf{HY} \text{ portanto } \hat{Y}_i = \sum_{j=1}^{n} H_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j \, .$$

- A matriz **H** depende de **Y**? Não, apenas dos valores **X**.
- Então h_{ii} é alto por causa de um valor atípico de X e não de Y.
- Já vimos que

$$\sum_{i=1}^{n} h_{ii} = p$$

  onde p é a dimensão do vetor de coeficientes β.
- Se nenhum ponto "força" o ajuste, devemos ter
  - todos h_{ii} aproximadamente iguais, nenhum "puxa" mais;

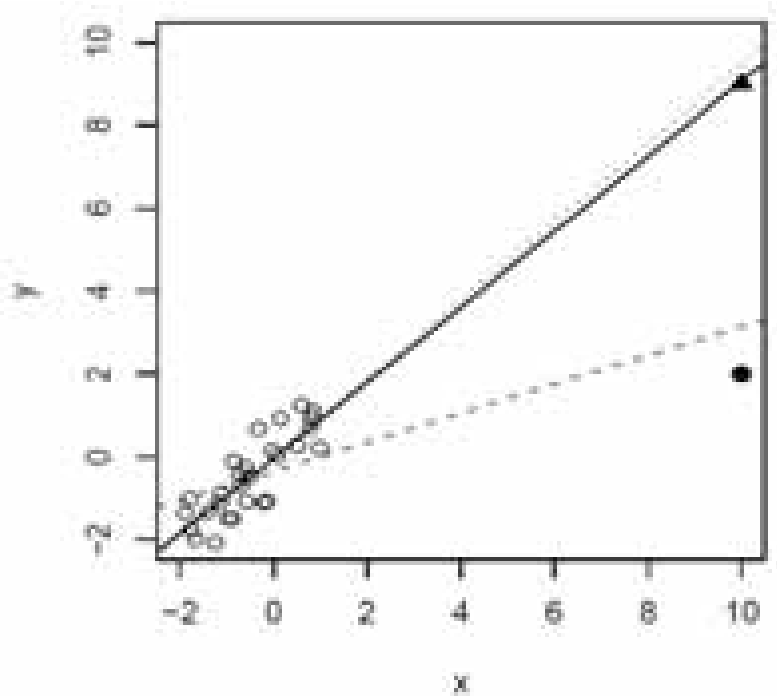
$$h_{ii} \approx \frac{p}{n} .$$

- Uma regra empírica é olhar com mais atenção pontos tais que

$$h_{ii} > \frac{2p}{n} .$$

Exemplo

- Vamos considerar novamente o gráfico do exemplo anterior.



Linha pontilhada: regressão ajustada com os pontos ○;
Linha sólida: regressão ajustada com os pontos ○ e ▲;
Linha tracejada: regressão ajustada com os pontos ○ e ●.

Exemplo (continuação)

- Nesse caso temos que

$$n = 30 \quad p = 2 \quad p/n = 0.07 \quad 2p/n = 0.14\,.$$

- Os pontos marcados com bola vazia são tais que

$$0.03 < h_{ii} < 0.05\,.$$

- Os pontos marcados com bola e triângulo cheios são tais que

$$h_{ii} = 0.46$$

  portanto, pela regra empírica, são considerados pontos alavanca.

Observações Influentes

Ponto Influente

- É aquele cuja remoção do banco de dados causa grande mudança no ajuste.
- Pode ou não ser um *outlier*.
- Pode ou não ser uma *alavanca*.
- Tende a ter pelo menos um dessas duas propriedades.
- A medida mais usada para detectar pontos influentes é a **Distância de Cook**.

- Vejamos como é definida a **Distância de Cook**.
- Sejam $\hat{\beta}$ e $\hat{\mathbf{Y}}$ os estimadores de β e $\mathbf{Y}$.
- Esses estimadores são calculados usando as n observações.
- Retiramos a i-ésima observação do banco de dados.
- Recalculamos os estimadores.
- Vamos denotá-los por $\hat{\beta}_{i}$ e $\hat{\mathbf{Y}}_{i}$.

- A Ditância de Cook mensura:
  - o quanto muda as estimativas dos coeficientes de regressão com a retirada da observação *i*

$$\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)} \, .$$

- Observe que

$$\hat{\mathbf{Y}} = \mathbf{X}'\hat{\boldsymbol{\beta}} \quad \text{e} \quad \hat{\mathbf{Y}}_{(i)} = \mathbf{X}'\hat{\boldsymbol{\beta}}_{(i)}$$

portanto

$$\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)} = \mathbf{X}'(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

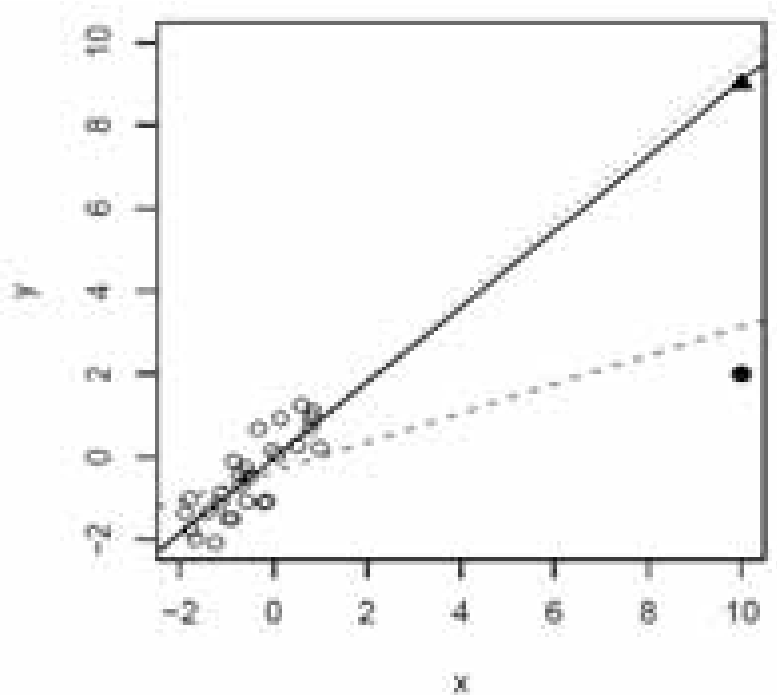
- A distância Cook da i-ésima observação é dada por:

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p\hat{\sigma}^2} = \frac{r_i h_{ii}}{p(1 - h_{ii})}$$

  onde
  - $\hat{\sigma}^2$ é a estimativa de σ^2 (baseada no modelo com as n observações);
  - p é a dimensão dao vetor β;
  - r_i é o resíduo padronizado;
  - h_{ii} é a alavanca.
- Um valor alto de D_i (em relação às demais)
  - indica que a observação i é influente, os valores estimados com e sem a observação são bem próximos.
- A influência pode ser consequência de:
  - um valor extremo do resíduo padronizado (o ponto é *outlier*);
  - um valor alto (próximo de 1) de h_{ii} (o ponto é uma alavanca).

Exemplo

- Vamos considerar novamente o gráfico do exemplo anterior.



Linha pontilhada: regressão ajustada com os pontos ○;
Linha sólida: regressão ajustada com os pontos ○ e ▲;
Linha tracejada: regressão ajustada com os pontos ○ e ●.

Exemplo (continuação)

- Os pontos marcados com bola vazia são tais que

$$5.7 \times 10^{-6} < D_i < 0.027\,.$$

- Os pontos marcados com bola e triângulo cheios são tais que

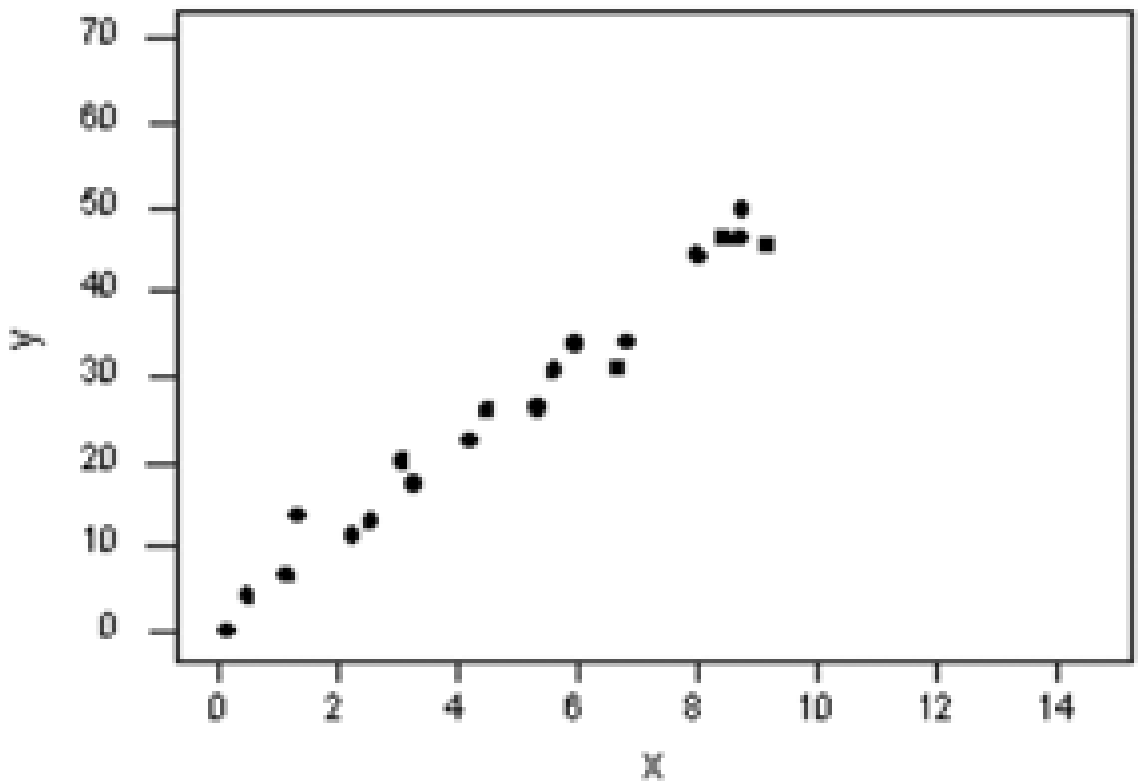
$$D_i = 6.4 \quad \text{e} \quad D_i = 9.8\,.$$

Como lidar com *outlier*?

- Caso tenha alguma justificativa teórica,
  - remover a observação.
- Coletar mais dados para verificar se os pontos identificados são, de fato, *outliers*.
- Aplicar uma transformação em Y ou ajustar um modelo não linear.
- Tratar os *outliers* separadamente.
- Usar métodos de regressão robusta.
- Apresentar os resultados com e sem os *outliers*.

Exemplo:

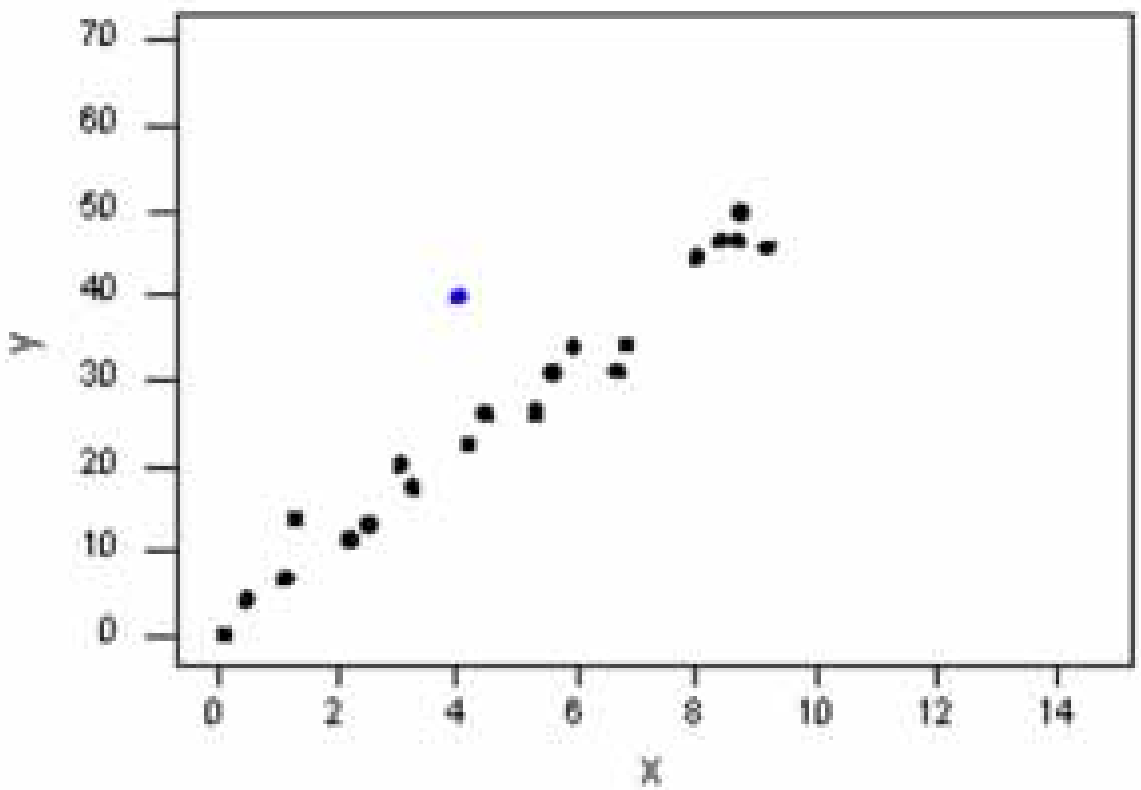
- O gráfico a seguir apresenta algum *outlier*, pontos de alavanca ou pontos influentes?



- Não, todos os dados seguem um padrão geral.

Exemplo:

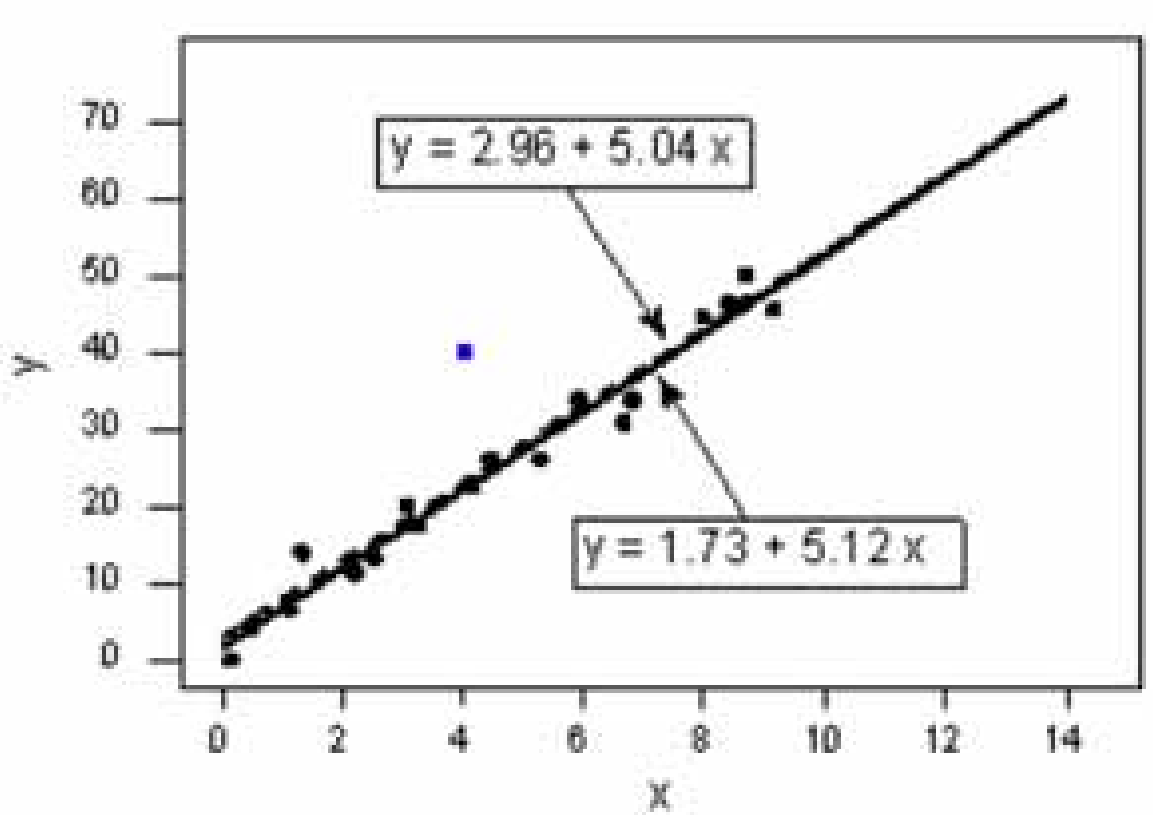
- O gráfico a seguir apresenta algum *outlier*, pontos de alavanca ou pontos influentes?



- O ponto azul pode ser considerado um *outlier*.

Exemplo:

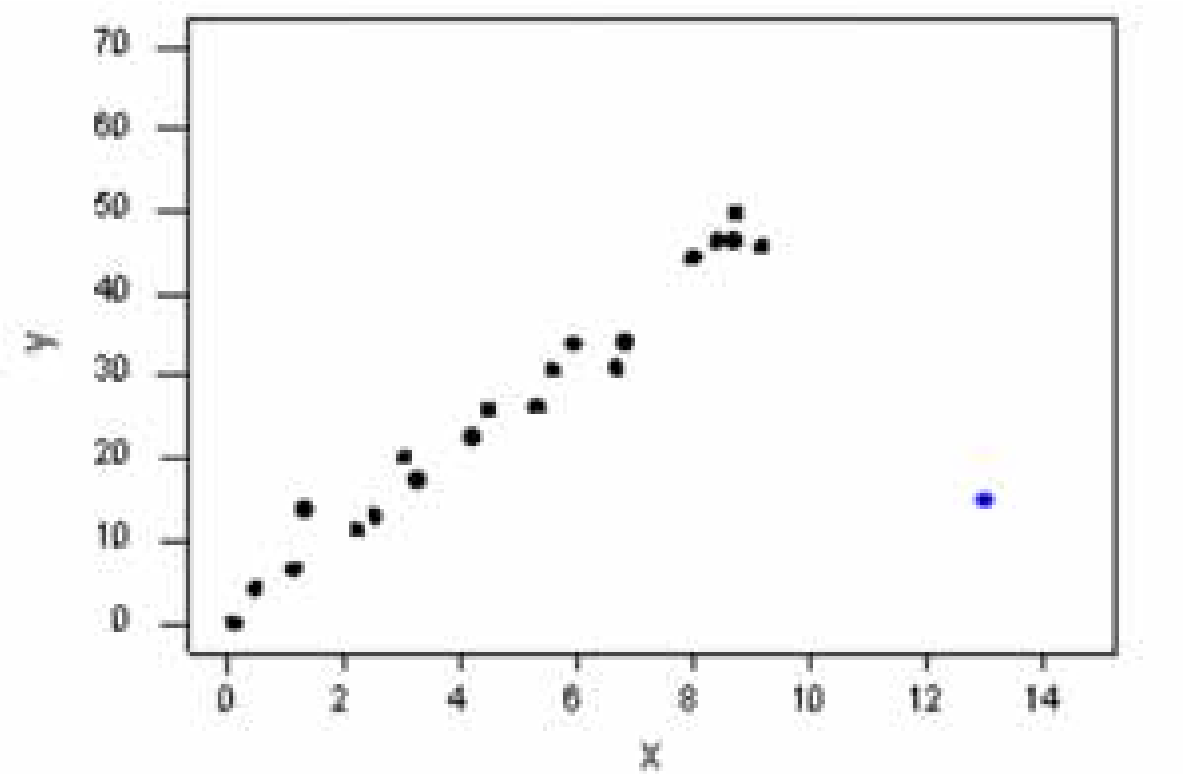
- Queremos agora verificar se é um ponto influente.
- O gráfico apresenta as retas ajustadas com e sem o ponto.



- A reta não muda muito, o ponto não é influente.

Exemplo:

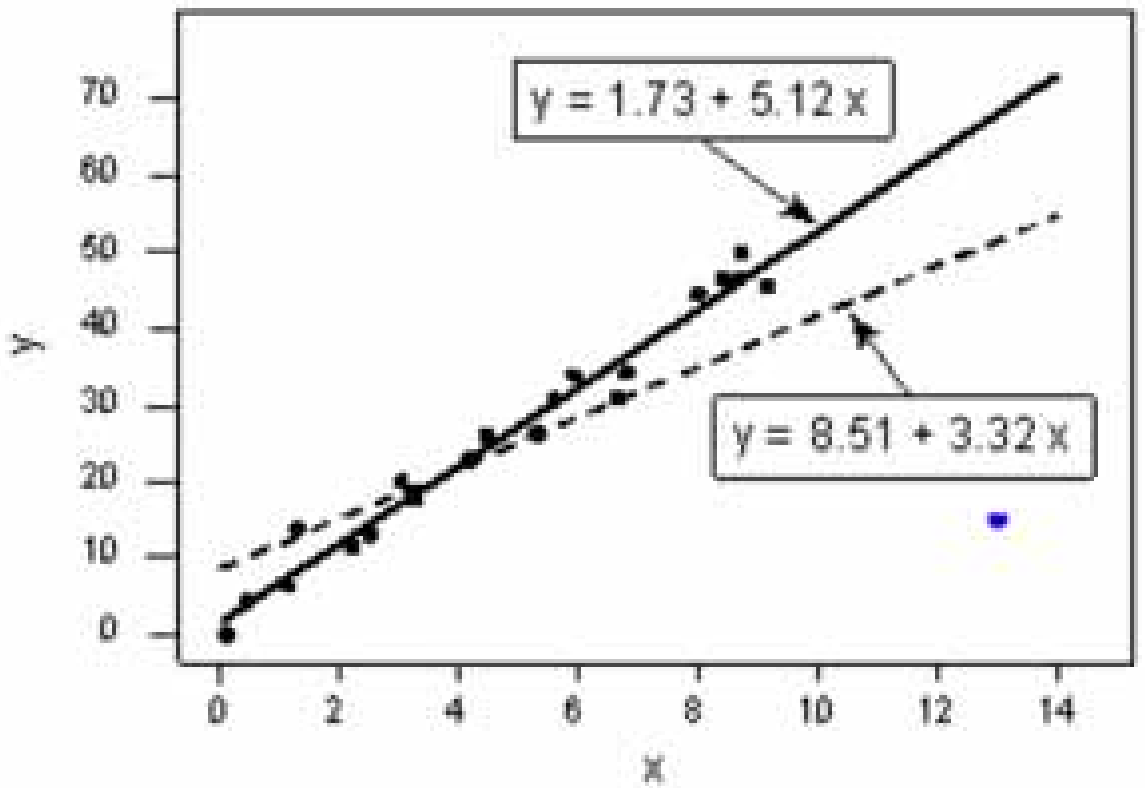
- O gráfico a seguir apresenta algum *outlier*, pontos de alavanca ou pontos influentes?



- Esse ponto é *outlier* e muito provavelmente será um ponto influente.

Exemplo:

- Queremos agora verificar se é um ponto influente.
- O gráfico apresenta as retas ajustadas com e sem o ponto.



Exemplo:

- A reta muda muito, o ponto é influente.