Métodos de Diagnóstico

Gilberto A. Paula

Departamento de Estatística
IME-USP, Brasil
giapaula@ime.usp.br

2º Semestre 2023

Sumário

Objetivos

Neste material serão discutidos alguns conceitos relacionados com Métodos de Diagnóstico em regressão linear múltipla:

- Solução de Mínimos Quadrados
- Pontos de Alavanca
- Análise de Resíduos
- Análise de Influência

Objetivos

Os principais objetivos da análise de diagnóstico são:

- Avaliar se há afastamentos importantes das suposições feitas para o modelo
- Avaliar se há presença de observações atípicas ou discrepantes.

As observações atípicas podem ser classificadas em três categorias: (i) pontos de alavanca (ii) pontos aberrantes e (iii) pontos influentes.

Pontos de Alavanca

São observações em que o vetor $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ está remoto no subespaço gerado pelas colunas da matriz $\mathbf{X}$. Essas observações têm influência desproporcional no próprio valor ajustado.

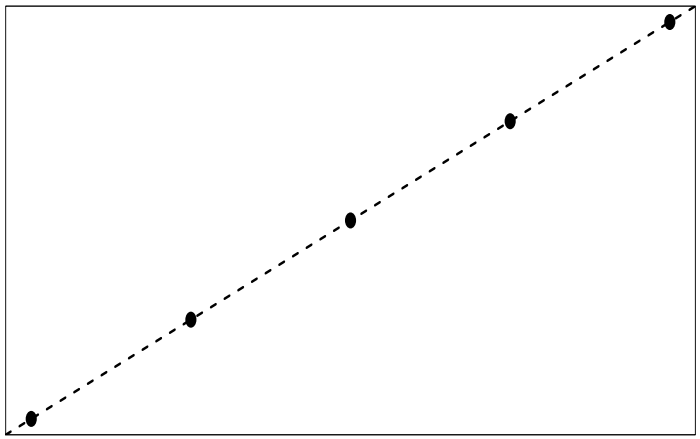
Pontos Aberrantes

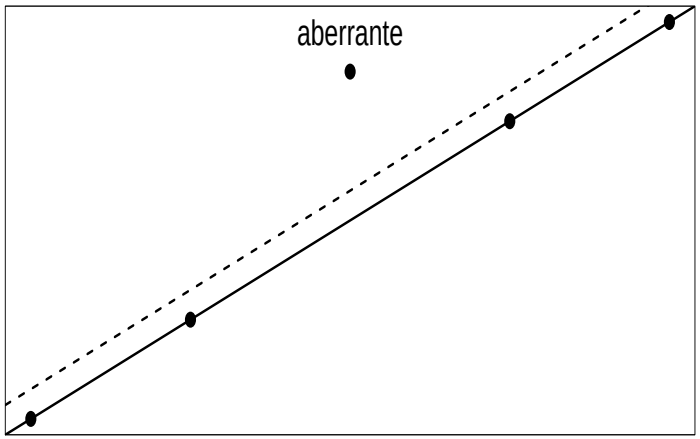
São observações com resíduo alto, posicionadas fora da banda de confiança. Ou seja, observações mal ajustadas pelo modelo. Em geral essas observações têm influência desproporcional na predição das respostas.

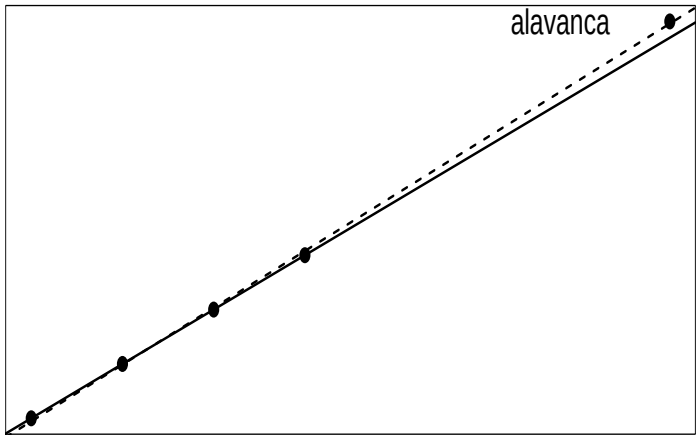
Pontos Influentes

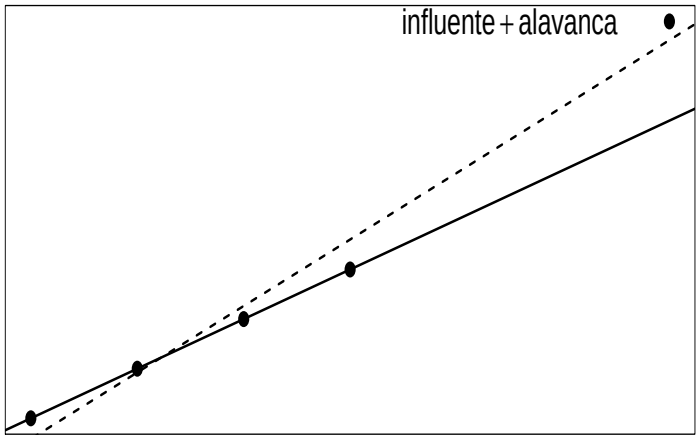
São observações com peso desproporcional nas estimativas dos coeficientes do componente sistemático do modelo. Em geral são pontos de alavanca mas a recíproca nem sempre é verdadeira.

Ilustração Observações Discrepantes









Sumário

Valor Predito

Tem-se pela solução de mínimos quadrados que

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y},$$

em que $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ é o projetor linear que projeta $\mathbf{y} \in R^n$ no subespaço $C(\mathbf{X})$ gerado pelas colunas da matriz $\mathbf{X}$.

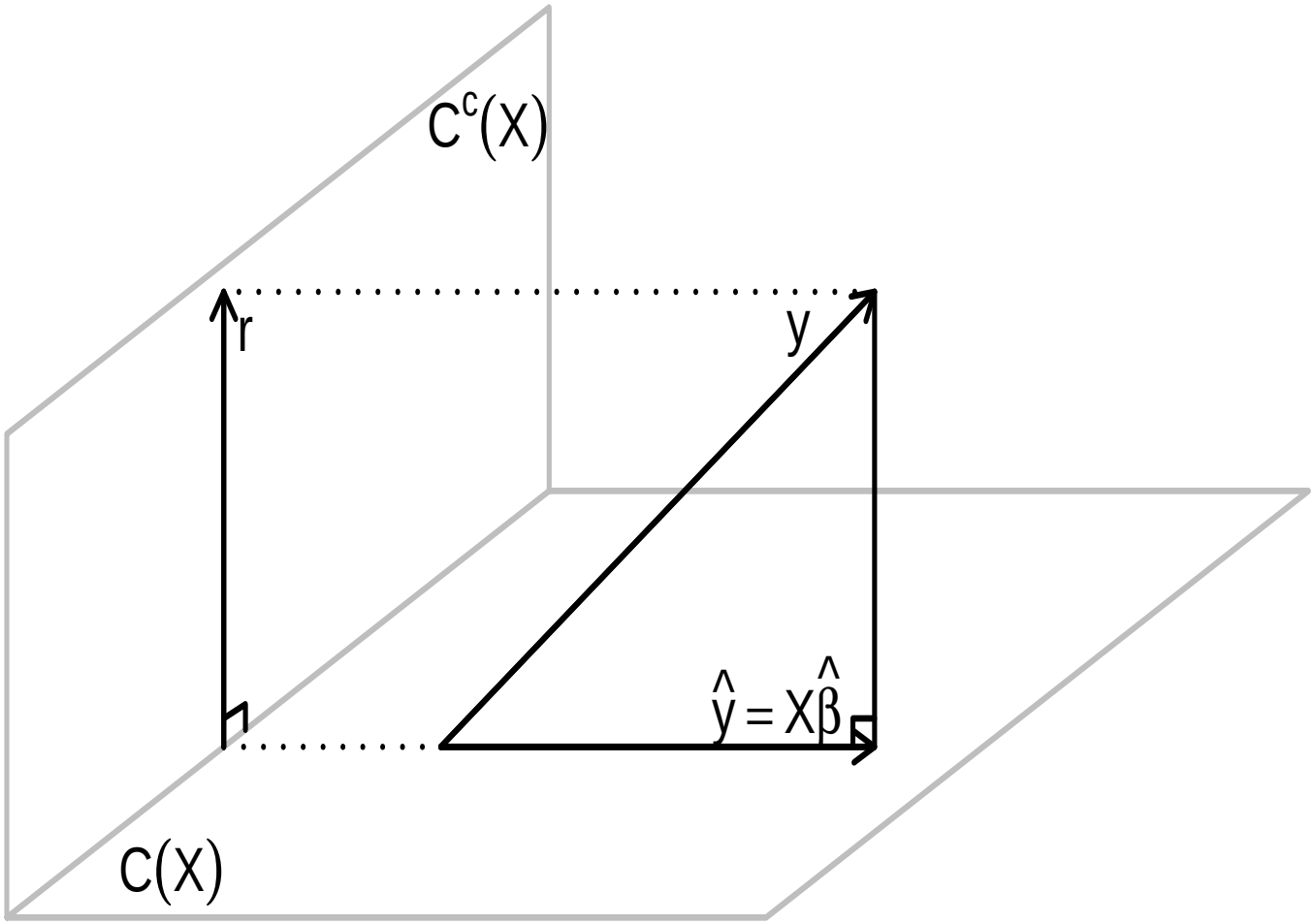
Resíduo Ordinário

Similarmente, segue pela solução de mínimos quadrados que

$$\mathbf{r} = (\mathbf{y} - \widehat{\mathbf{y}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{y},$$

em que $(\mathbf{I}_n - \mathbf{H})$ é o projetor linear que projeta $\mathbf{y} \in R^n$ no ortocomplemento de $C(\mathbf{X})$, $\mathbf{r}$ é o vetor de resíduos ordinários e $\mathbf{I}_n$ denota a matriz identidade de ordem n.

Ilustração Solução de Mínimos Quadrados



Sumário

Pontos de Alavanca

Definição

Uma observação é definida como ponto de alavanca se tem uma alta influência no próprio valor ajustado. Essa influência é medida através da derivada $\partial\widehat{y}/\partial y$. Ou seja, mede o impacto que uma variação infinitesimal na resposta causa no valor ajustado.

Definição

Da relação $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ obtém-se

$$\widehat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j,$$

em que h_{ij} denota o elemento (i, j) da matriz simétrica $\mathbf{H}$ de dimensão $n \times n$. Daí segue que $\partial \widehat{y}_i / \partial y_i = h_{ii}$ e ainda pode-se mostrar que

$$h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i,$$

em que $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$, para $i = 1, \ldots, n$.

Definição

Como a matriz **H** é idempotente $\mathbf{H} = \mathbf{HH}$ segue que

$$\sum_{j=1}^{n} h_{ij}^2 = h_{ii} \rightarrow \sum_{j \neq i} h_{ij}^2 = h_{ii} - h_{ii}^2 = h_{ii}(1 - h_{ii}),$$

então $h_{ii} \geq 0$ e $h_{ii}(1 - h_{ii}) \geq 0$ e portanto

$$0 \leq h_{ii} \leq 1.$$

Note que se $h_{ii} = 1$ então $h_{ij} = 0 \ \ \forall j \neq i$ e logo $\widehat{y}_i = y_i$.

Ponto de Corte

Proposta para classificar pontos de alavanca

$$h_{ii} \geq 2\bar{h},$$

em que $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n}$. Tem-se que

$$\sum_{i=1}^{n} h_{ii} = \text{tr}(\mathbf{H}) = \text{tr}\{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\} = \text{tr}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\} = \text{tr}(\mathbf{I}_p) = p.$$

Portanto, deve-se destacar observações tais que

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

Para amostras grande sugere-se $h_{ii} \geq \frac{3p}{n}$.

Pontos de Alavanca - Exemplo Telhados $h_{ii} \geq 2p/n$

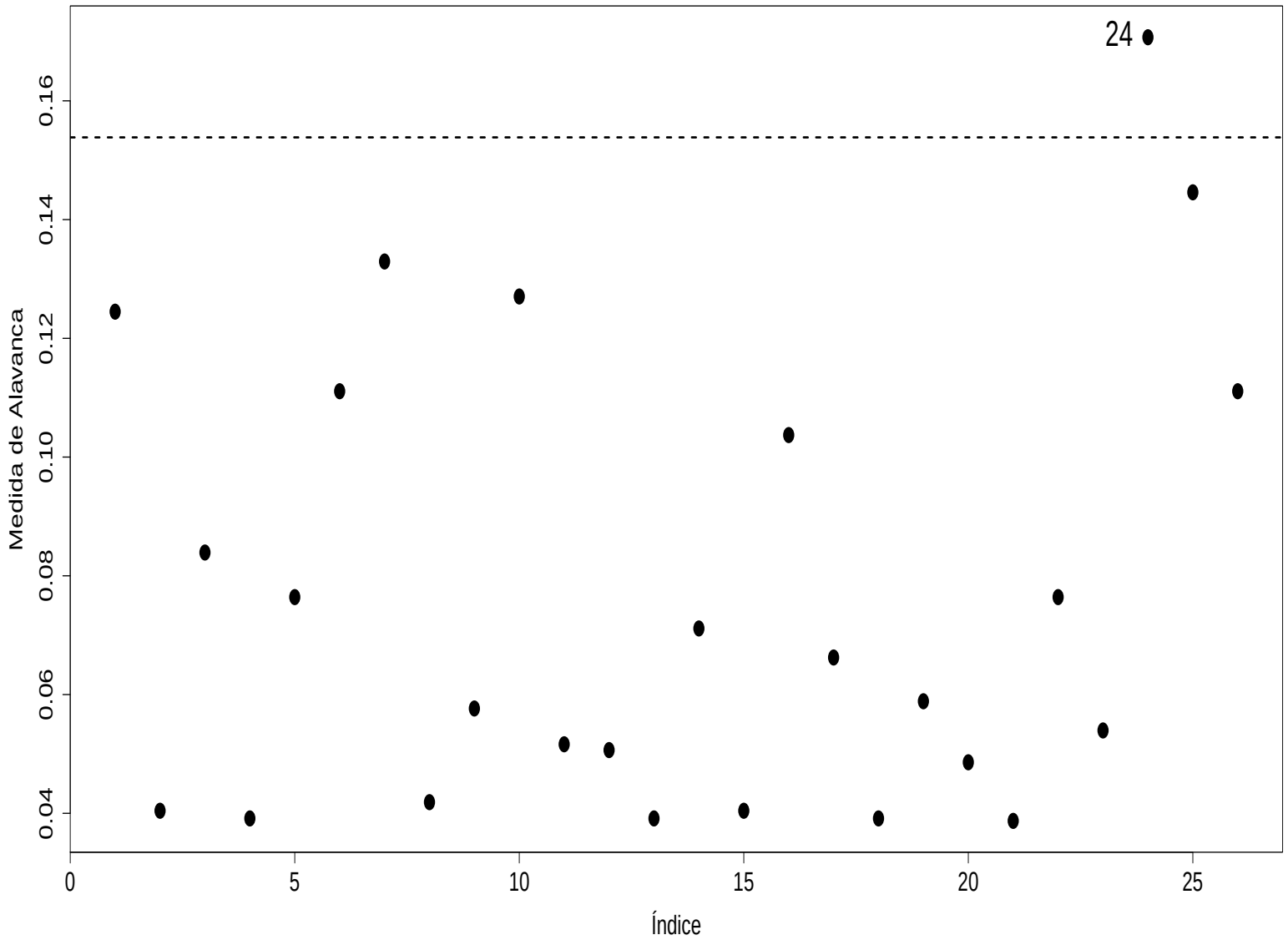
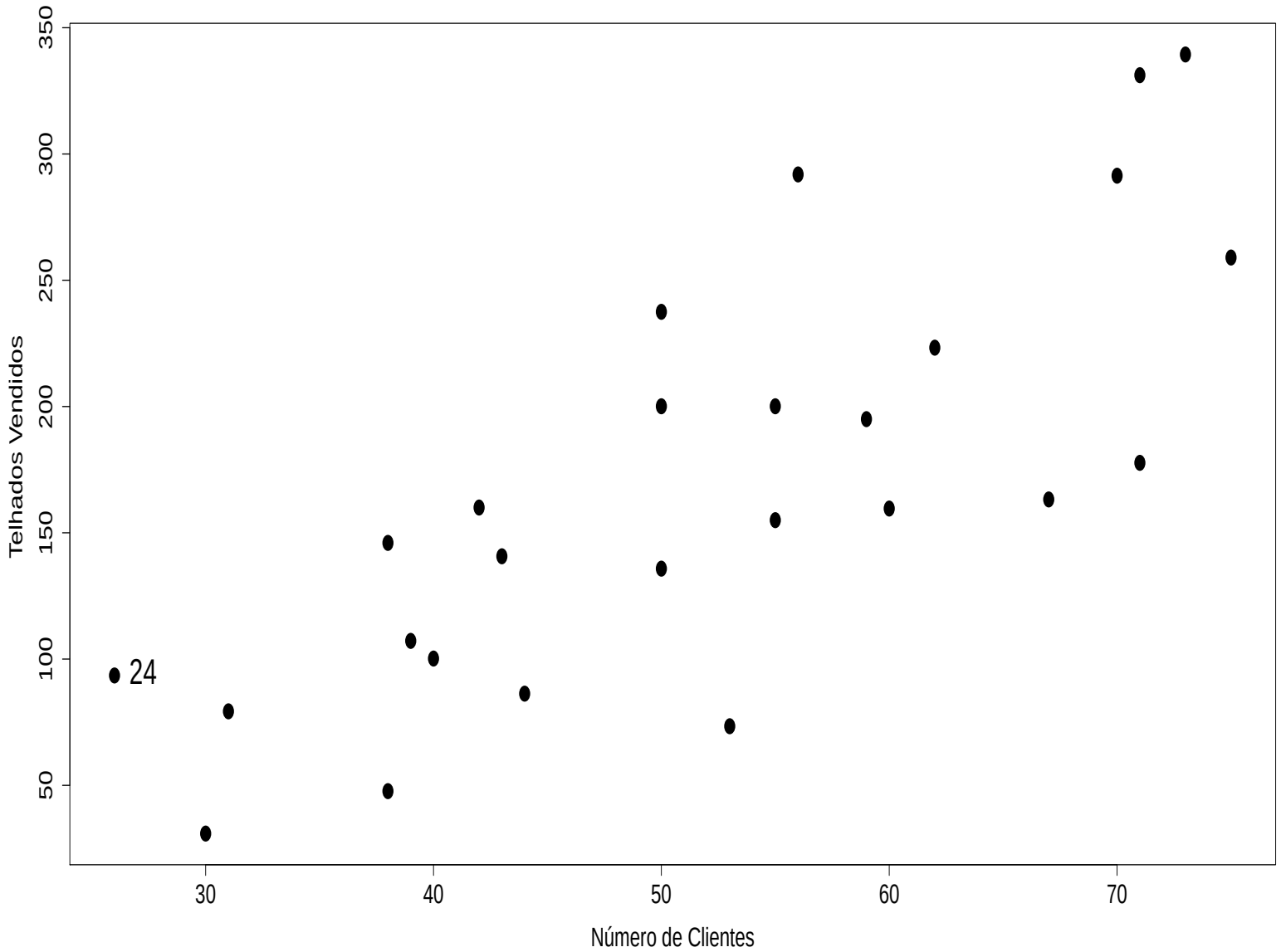


Diagrama de Dispersão - Exemplo Telhados



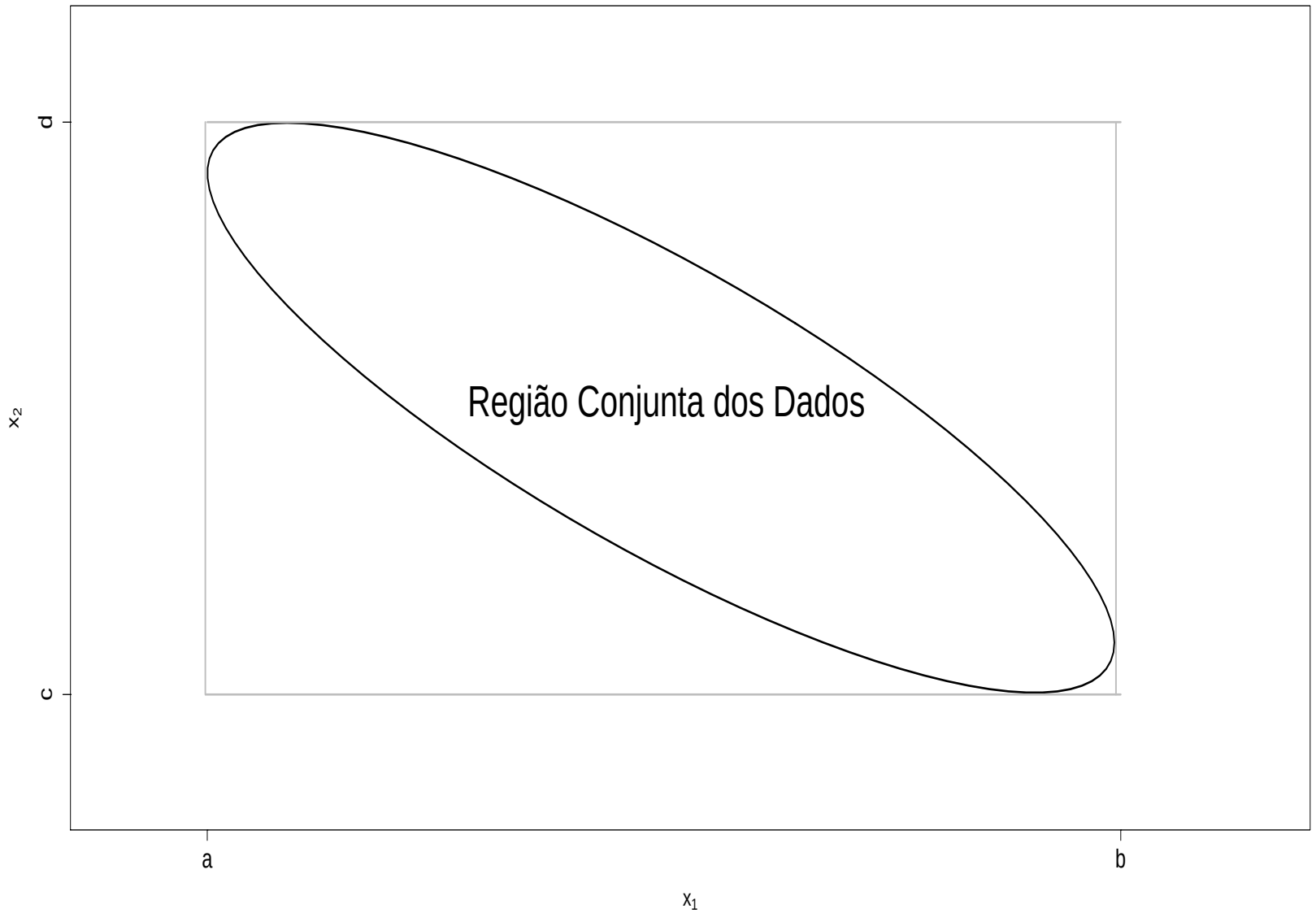
Limites para a Predição

Supor uma nova observação com valores para as variáveis explicativas representados por $\mathbf{z} = (z_{i1}, z_{i2}, \ldots, z_{ip})^\top$. Qual a condição para obter $\widehat{y}(\mathbf{z})$?
Segundo Montgomery et al.(2021) pode-se fazer predição (interpolação) no modelo de regressão linear múltipla com segurança se a condição abaixo for satisfeita

$$\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} \leq h_{\max} \; \forall \mathbf{x} \in R^p,$$

em que $h_{\max} = \max\{h_{11}, \ldots, h_{nn}\}$. Logo, uma condição para predição de $y(\mathbf{z})$ é que $\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{z} \leq h_{\max}$.

Ilustração Região Conjunta dos Dados $\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} \leq h_{\max}$



Sumário

Resíduo Ordinário

O vetor de resíduos ordinários é definido por

$$\mathbf{r} = (\mathbf{I}_n - \mathbf{H})\mathbf{y},$$

em que $\mathbf{r} = (r_1, \ldots, r_n)^\top$ com $r_i = y_i - \widehat{y}_i$, para $i = 1, \ldots, n$. Tem-se que

$$\begin{aligned} \mathrm{E}(\mathbf{r}) &= \mathrm{E}(\mathbf{Y}|\mathbf{X}) - \mathbf{H}\mathrm{E}(\mathbf{Y}|\mathbf{X}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{0}. \end{aligned}$$

Resíduo Ordinário

A matriz de variância-covariância de $\mathbf{r}$ fica dada por

$$\begin{aligned} \text{Var}(\mathbf{r}) &= \text{Var}\{(\mathbf{I}_n - \mathbf{H})\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{I}_n - \mathbf{H})\text{Var}(\mathbf{Y}|\mathbf{X})(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H}). \end{aligned}$$

Resíduo Ordinário

Portanto, segue que

$$\mathbf{r} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H})).$$

e consequentemente

- $r_i \sim \mathrm{N}(0, \sigma^2(1 - h_{ii}))$
- $\mathrm{Cov}(r_i, r_j) = -\sigma^2 h_{ij},\ i \neq j$
- $\mathrm{Corr}(r_i, r_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}$, para $i, j = 1, \ldots, n$.

Portanto, os resíduos têm distribuição normal de média zero, variâncias não constantes e são correlacionados.

Resíduo Padronizado

Para que os resíduos sejam comparáveis é preciso padronizá-los. Uma padronização natural seria o resíduo normalizado

$$t_{r_i} = \frac{r_i}{\sigma\sqrt{1-h_{ii}}} \sim \mathrm{N}(0,1),\ i = 1,\ldots,n.$$

Porém, é preciso estimar σ^2.

Distribuição t-Student

A estatística t-Student é construída da seguinte forma:

$$t = \frac{Z}{\sqrt{U/\nu}} \sim t_\nu,$$

em que $Z \sim \mathrm{N}(0,1)$, $U \sim \chi^2_\nu$ e Z e U são variáveis aleatórias independentes.

Resíduo Studentizado

Tem-se que $t_{r_i} \sim \mathrm{N}(0,1)$ e que $(n-p)s^2/\sigma^2 \sim \chi^2_{(n-p)}$, porém t_{r_i} e s^2 não são independentes. Logo o resíduo

$$t_i = \frac{r_i}{s\sqrt{1-h_{ii}}}$$

não segue distribuição $t_{(n-p)}$.

Resíduo Studentizado

Sugestão é substituir s^2 por $s^2_{(i)}$ que denota o erro quadrático médio do modelo sem a i-ésima observação. Agora, tem-se que

- $t_{r_i} \sim \mathrm{N}(0,1)$,
- $(n-p-1)s^2_{(i)}/\sigma^2 \sim \chi^2_{(n-p-1)}$
- t_{r_i} e $s^2_{(i)}$ são independentes.

Logo, o resíduo

$$t_i^* = \frac{r_i}{s_{(i)}\sqrt{1-h_{ii}}} \sim t_{(n-p-1)},$$

para $i = 1, \ldots, n$.

Resíduo Studentizado

É possível obter $s^2_{(i)}$ sem a necessidade de reajustar o modelo retirando cada observação. Mostra-se que

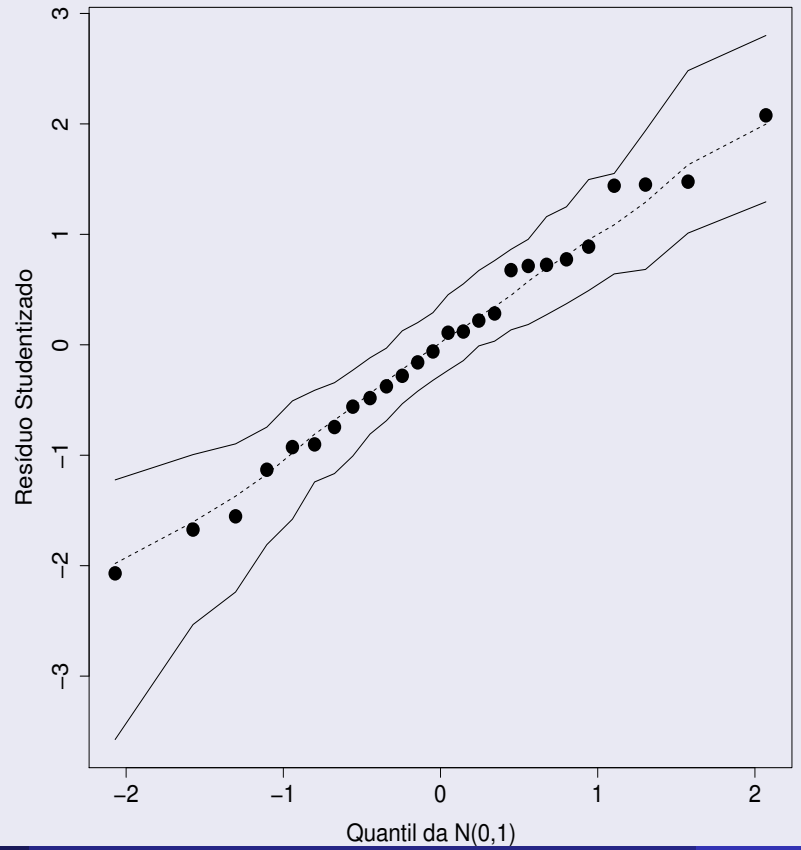
$$s^2_{(i)} = s^2 \left(\frac{n - p - t_i^2}{n - p - 1} \right).$$

para $i = 1, \ldots, n$.

Resíduo Studentizado

Gráficos sugeridos com o resíduo t_i^*:

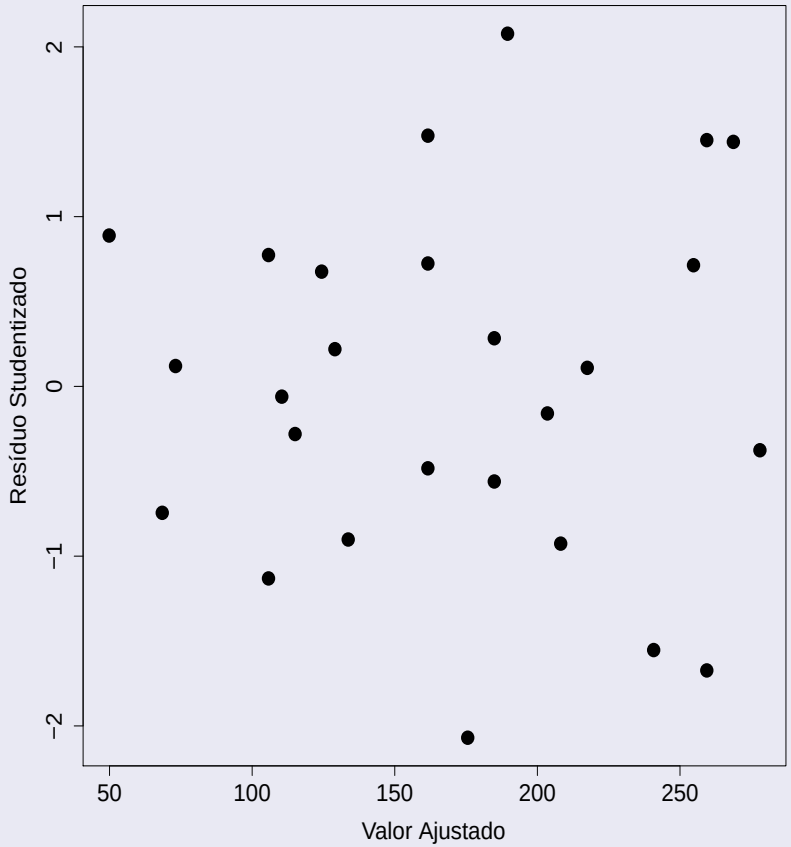
- Gráfico normal de probabilidades com banda de confiança empírica. Espera-se os pontos de forma aleatória dentro da banda de confiança:



Resíduo Studentizado

Gráficos sugeridos com o resíduo t_i^*:

- Gráfico de t_i^* contra valores ajustados $\widehat{y}_i$. Desde que $\text{Cov}(\mathbf{r}, \widehat{\mathbf{y}}) = \mathbf{0}$ espera-se distribuição uniforme dos pontos:



Resíduo Studentizado

Gráficos sugeridos com o resíduo t_i^*:

- Gráfico de t_i^* contra a ordem das observações para detectar (quando fizer sentido) correlação temporal dos dados.
- Gráfico de t_i^* contra valores de variáveis explicativas contínuas para avaliar se há algum termo que não foi incluído no componente sistemático do modelo.

Sumário

Região de Confiança

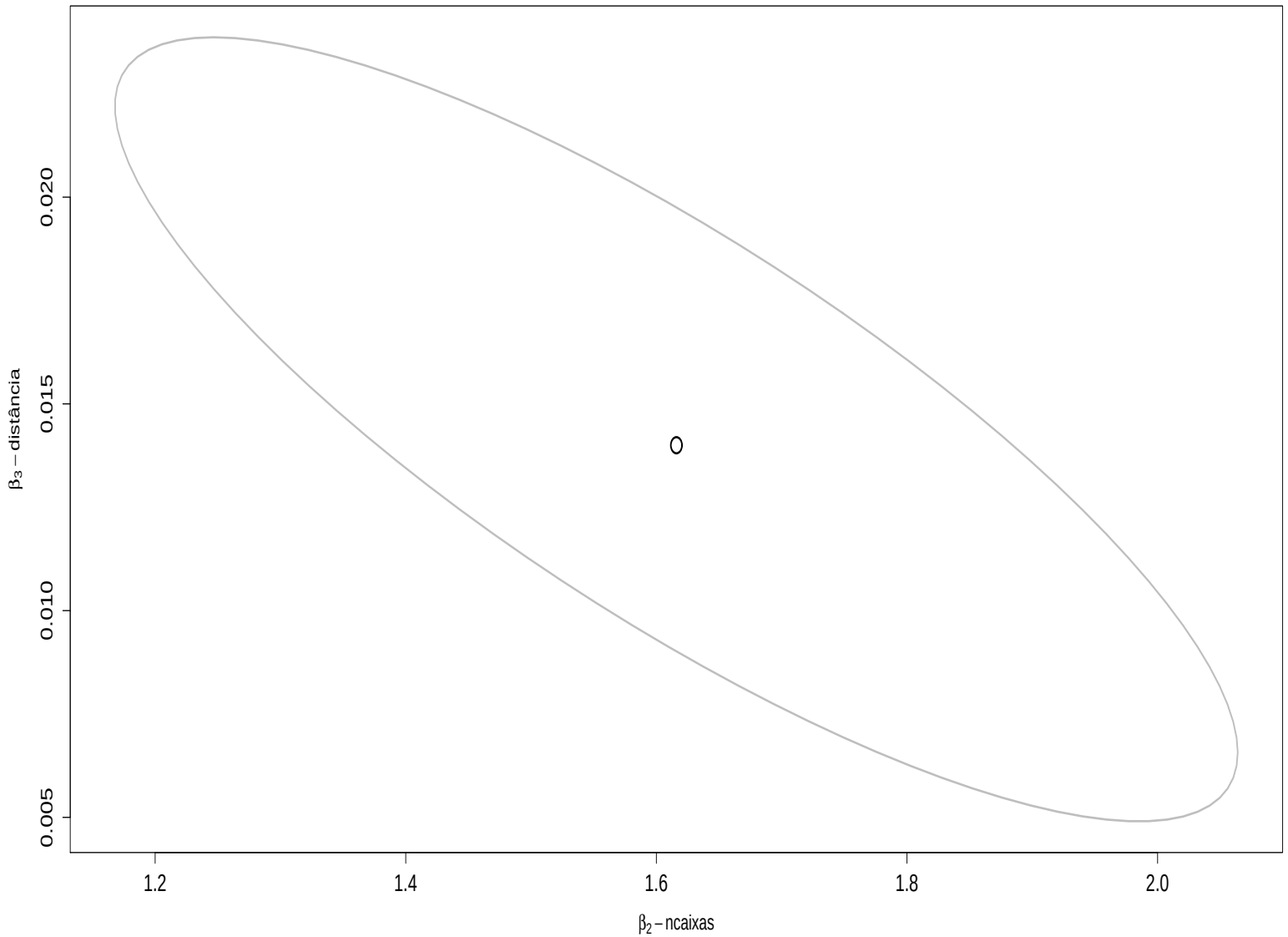
Uma região de confiança de coeficiente $(1-\alpha)$ para β é dada por

$$\frac{(\widehat{\beta}-\beta)^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\beta}-\beta)}{ps^2} \leq F_{p,(n-p),(1-\alpha)},$$

em que $F_{p,(n-p),(1-\alpha)}$ denota o quantil $(1-\alpha)$ de uma distribuição F com p e $(n-p)$ graus de liberdade. Essa região de confiança é construída usando o resultado abaixo

$$P\left\{\frac{(\widehat{\beta}-\beta)^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\beta}-\beta)}{ps^2} \leq F_{p,(n-p),(1-\alpha)}\right\} = 1-\alpha.$$

Ilustração Região Conjunta (β_2, β_3) Exemplo Delivery



Distância de Cook

A distância de Cook é definida por

$$\mathrm{D}_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{ps^2},$$

em que $\widehat{\boldsymbol{\beta}}_{(i)}$ denota a estimativa de mínimos quadrados quando a i-ésima observação não é considerada no modelo.

Distância de Cook

Tem-se que

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{(i)} &= \{\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\}^{-1}\mathbf{X}_{(i)}^\top \mathbf{y}_{(i)} \\
&= \{\mathbf{X}^\top\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^\top\}^{-1}\{\mathbf{X}^\top\mathbf{y} - \mathbf{x}_i y_i\} \\
&= \left\{(\mathbf{X}^\top\mathbf{X})^{-1} + \frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}}{1-h_{ii}}\right\}\{\mathbf{X}^\top\mathbf{y} - \mathbf{x}_i y_i\} \\
&= \widehat{\boldsymbol{\beta}} - \frac{r_i}{(1-h_{ii})}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i,
\end{aligned}
$$

em que $r_i = y_i - \widehat{y}_i$ e $h_{ii} = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$, para $i = 1, \ldots, n$.

Distância de Cook

Portanto, obtém-se

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{r_i}{(1-h_{ii})}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i.$$

E a distância de Cook fica dada por

$$\begin{aligned} D_i &= \frac{1}{ps^2}\frac{h_{ii}r_i^2}{(1-h_{ii})^2} \\ &= \left\{\frac{r_i}{s\sqrt{1-h_{ii}}}\right\}^2 \frac{h_{ii}}{(1-h_{ii})}\frac{1}{p} \\ &= \frac{1}{p}t_i^2\frac{h_{ii}}{(1-h_{ii})}. \end{aligned}$$

Distância de Cook

Portanto, a distância de Cook fica dada por

$$\mathrm{D}_i = \frac{1}{p} t_i^2 \frac{h_{ii}}{(1 - h_{ii})}.$$

Como $h_{ii}/(1 - h_{ii})$ é uma função crescente de h_{ii}, então D_i será grande se $|t_i|$ e/ou h_{ii} forem (for) grande(s). Uma proposta de pontos suspeitos de serem influentes é olhar aqueles pontos tais que

$$\mathrm{D}_i \geq F_{p,(n-p),(1-\alpha)}.$$

Outras sugestões: $\mathrm{D}_i \geq \bar{\mathrm{D}} + k\mathrm{DP}(\mathrm{D}_i)$, para $k = 2, 3, 4$.

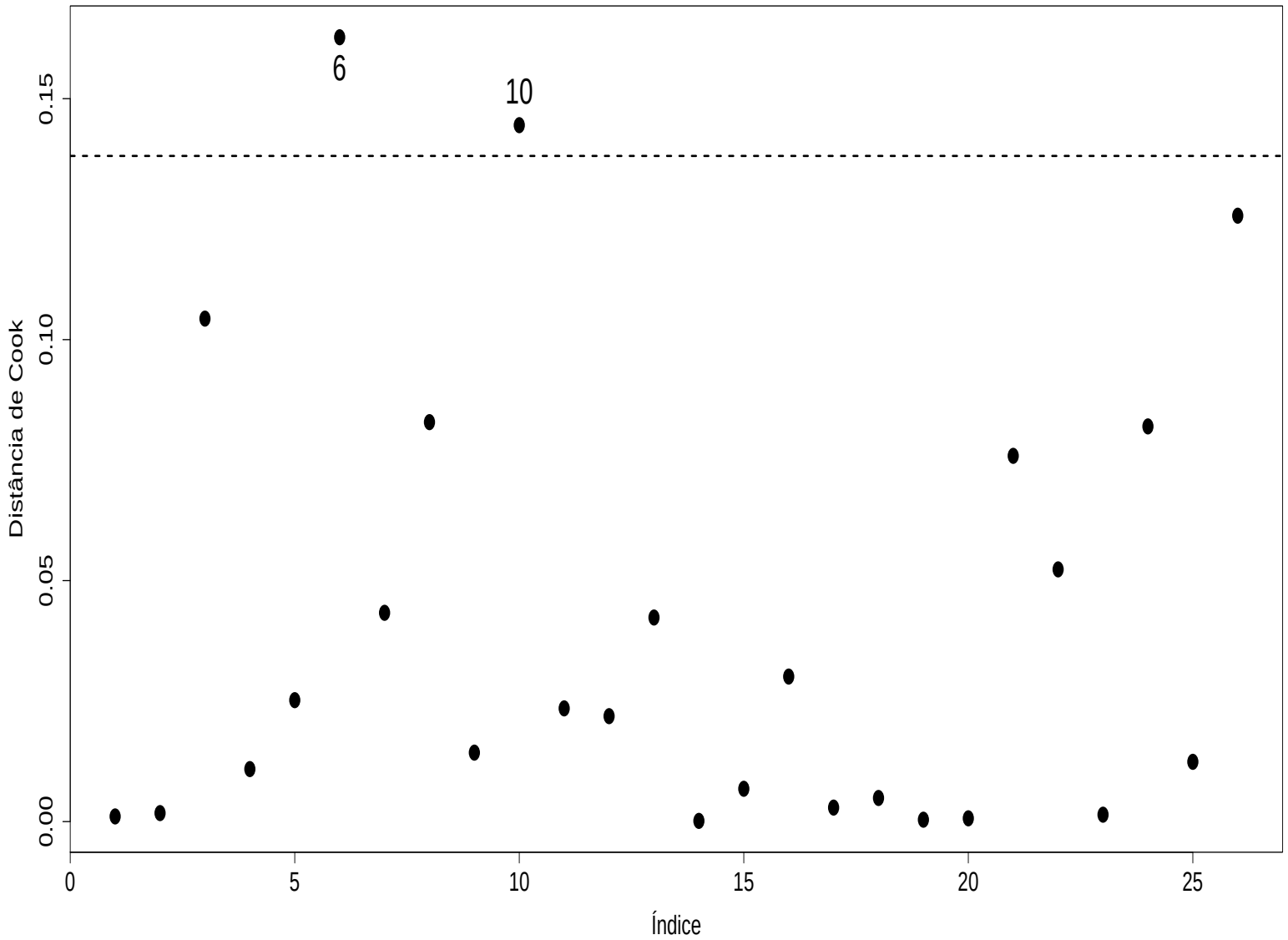
Distância DFFITS$_i$

Outra medida de influência derivada da distância de Cook substituindo s^2 por $s^2_{(i)}$ é dada por

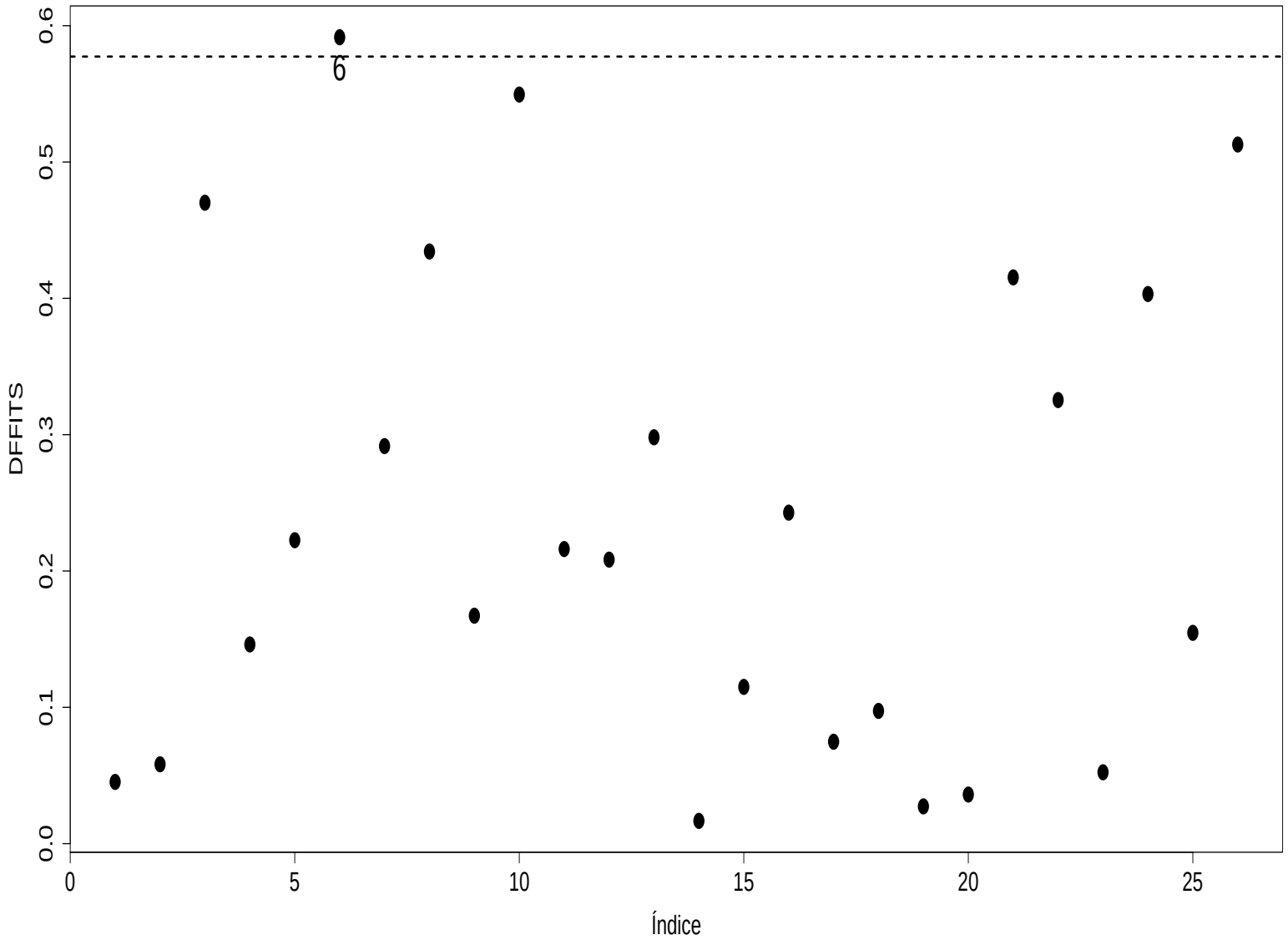
$$\begin{aligned} \text{DFFITS}_i &= \frac{|r_i|}{s_{(i)}\sqrt{1-h_{ii}}}\left\{\frac{h_{ii}}{1-h_{ii}}\right\}^{1/2} \\ &= |t_i^*|\left\{\frac{h_{ii}}{1-h_{ii}}\right\}^{1/2}. \end{aligned}$$

Deve-se olhar as observações tais que $\text{DFFITS}_i \geq 2\{p/(n-p)\}^{1/2}$.

Distância de Cook - Exemplo Telhados $D_i \geq \bar{\mathbf{D}} + 2\mathbf{DP}(D_i)$



DFFITSi - Exemplo Telhados $\text{DFFITS}_i \geq 2\{p/(n-p)\}^{1/2}$



Medida DFBETAS$_{ji}$

Define-se a seguinte medida para avaliar a influência da eliminação da i-ésima observação no j-ésimo coeficiente estimado da regressão:

$$\begin{aligned}
\text{DFBETAS}_{ji} &= \frac{(\widehat{\beta}_j - \widehat{\beta}_{j(i)})}{s_{(i)}\sqrt{C_{jj}}} \\
&= \frac{\mathbf{C}_j^\top \mathbf{x}_i r_i}{s_{(i)}(1-h_{ii})\sqrt{C_{jj}}} \\
&= \frac{p_{ji}}{\sqrt{\mathbf{p}_j^\top \mathbf{p}_j}} \frac{t_i^*}{\sqrt{1-h_{ii}}},
\end{aligned}$$

em que $\mathbf{C} = (\mathbf{X}^\top\mathbf{X})^{-1}$, $\mathbf{C}_j$ denota a j-ésima coluna de $\mathbf{C}$, p_{ji} e $\mathbf{p}_j^\top$ denotam, respectivamente, o (j,i)-ésimo elemento e a j-ésima linha de $\mathbf{P} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, $i = 1,\ldots,n$ e $j = 1,\ldots,p$. $\text{DFBETAS}_{ji} > \frac{2}{\sqrt{n}}$.

Gráfico da Variável Adicionada

Considere o seguinte modelo de regressão linear múltipla:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + z_i\gamma,$$

$i = 1, \ldots, n$, em que Z é uma variável contínua adicionada no modelo linear. Pode-se mostrar, após algumas manipulaçõea algébricas, que

$$\hat{\gamma} = \frac{\mathbf{z}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}}{\mathbf{z}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{z}},$$

em que $\mathbf{z} = (z_1, \ldots, z_n)^\top$ e as demais quantidades como definido anteriormente.

Gráfico da Variável Adicionada

Sejam os resíduos

- $\mathbf{r} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ (da regressão linear de $\mathbf{y}$ contra as colunas de $\mathbf{X}$)
- $\mathbf{v} = (\mathbf{I}_n - \mathbf{H})\mathbf{z}$ (da regressão linear de $\mathbf{z}$ contra as colunas de $\mathbf{X}$).

Então tem-se que

$$\hat{\gamma} = (\mathbf{v}^\top \mathbf{v})^{-1}\mathbf{v}^\top \mathbf{r}.$$

Ou seja, $\hat{\gamma}$ é coeficiente angular da regressão passando pela origem de $\mathbf{r}$ contra $\mathbf{v}$. Assim, o diagrama de dispersão entre $\mathbf{r}$ e $\mathbf{v}$ pode revelar a forma mais apropriada da covariável Z entrar no modelo.

Sumário

Deleção de Observações Suspeitas

O procedimento mais tradicional de verificação das observações suspeitas é através da deleção individual de cada observação suspeita, computando-se a variação percentual de cada coeficiente da regressão e o respectivo valor-P. Denota-se o conjunto das m observações supeitas por $S = \{S_1, \ldots, S_m\}$.

Variação Percentual

A variação percentual do j-ésimo coeficiente da regressão quando a i-ésima observação não é considerada no ajuste é definido por

$$\Delta_{ij} = \left| \frac{\hat{\beta}_{(i)j} - \hat{\beta}_j}{\hat{\beta}_j} \right| \times 100\%$$

para $j = 1, \ldots, p$ e $i \in$ S. Deve-se associar a cada observação deletada o novo valor-P de cada coeficiente. Variações percentuais desproporcionais (muito acima de $(1/n) \times 100\%$) são esperadas, porém deve-se dar atenção quando ocorrerem mudanças inferenciais.

Comparação entre Observações Suspeitas e não Suspeitas

Um outro procedimento usual é comparar alguma medida resumo das observações suspeitas com a mesma medida resumo obtida de r amostras aleatórias de tamanho m das observações não suspeitas. Por exemplo, pode-se computar a medida

$$\text{MRC}_S = \max_{1 \leq j \leq p} \left| \frac{\hat{\beta}_{(S)j} - \hat{\beta}_j}{\hat{\beta}_j} \right| .$$

Comparar MRC_S com as r medidas das r amostras aleatórias de tamanho m extraídas do grupo de observações não suspeitas. Se MRC_S for muito maior que $\max_{1 \leq j \leq r} \text{MRC}_{NS_j}$ é um indício de que as observações em S são discrepantes. Sugere-se utilizar que $r \geq 10$.

Procedimentos Mais Usuais

Os seguintes procedimentos são usuais para acomodar pontos discrepantes:

- Aplicar transformações nas variáveis explicativas, por exemplo padronização, raiz quadrada e logarítmica
- Incluir termos não lineares em variáveis explicativas contínuas
- Incluir (ou retirar) interações
- Regressão ponderada
- Aplicar método robusto
- Mudar a distribuição dos erros

Sumário

Referências

- Belsley, D. A.; Kuh, E. e Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2021). *Introduction to Linear Regression Analysis, 6th Edition*. Hoboken: Wiley.