CAP 4 - Análise de Resíduos e Diagnósticos

No modelo $\underset{\sim}{Y} = \mathbb{X}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}$, os elementos ε_i do vetor $\underset{\sim}{\varepsilon}$ são as diferenças $\varepsilon_i = Y_i - \hat{Y}_i$ e representam a variabilidade natural dos dados. Assumimos que os erros ε_i são independentes e $\varepsilon_i \sim N(0, \sigma^2)$.

Se as suposições são violadas, temos:

- falhas sistemáticas: não linearidade; heterocedasticidade; dependência dos resíduos, não normalidade, etc
- falhas isoladas: presença de pontos atípicos

Ajustado um modelo a um conjunto de dados, o material básico que será utilizado para verificar as pressuposições, são:

- valores estimados $\hat{Y}_i$;
- os resíduos;
- estimativa da variância residual $\hat{\sigma}^2 = QMRes$
- elementos da diagonal da matriz H (Hat matrix)

$$H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \qquad h_{ii} \Rightarrow \text{leverage}$$

Em relação às técnicas usadas para verificação do ajuste do modelo, estas podem ser:

- formais: aninhar um modelo sob pesquisa em uma classe maior de modelos → trabalhar a idéia de deleção de pontos
- informais: inspeção de gráficos.

Tipos de Resíduos

① Resíduos Ordinários (r_i ou $\hat{\varepsilon}_i$)

Dado o ajuste do modelo $\underset{\sim}{Y} = X\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}$ e $\underset{\sim}{\hat{Y}} = H\underset{\sim}{Y}$, o resíduo ordinário é dado por: $\underset{\sim}{\hat{\varepsilon}} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} = (I-H)\underset{\sim}{Y}$.

Enquanto que os erros $\underset{\sim}{\varepsilon} \sim N(\phi, I\sigma^2)$ $\begin{cases} E(\varepsilon_i) = 0 \\ cov(\varepsilon_i, \varepsilon_{i'}) = 0 \\ Var(\varepsilon_i) = \sigma^2 \end{cases}$

isso não ocorre com $\underset{\sim}{\hat{\varepsilon}}$, ou seja,

$$Var(\underset{\sim}{\hat{\varepsilon}}) = Var[(I-H)\underset{\sim}{Y}] = (I-H)\,Var(\underset{\sim}{Y})(I-H)^T =$$

$$= (I-H)(I-H)^T\sigma^2 = (I-H)\sigma^2$$

$$Var(\underset{\sim}{\hat{\varepsilon}}) = \begin{bmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ & 1-h_{22} & \cdots & -h_{2n} \\ & & \ddots & \vdots \\ \text{SIMÉTRICA} & & & 1-h_{nn} \end{bmatrix} \sigma^2$$

$\Rightarrow \begin{cases} Var(\hat{\varepsilon}_i) = (1-h_{ii})\sigma^2 \\ Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_{i'}) = -h_{ii'}\sigma^2 \end{cases}$

$Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_{i'})$

$Var(\hat{\varepsilon}_i)$

② Resíduos Padronizados (r_{pi})

Por definição os r_{pi} são obtidos por:

$$r_{pi} = \frac{\hat{\varepsilon}_i}{s}, \text{ em que: } s = \sqrt{QMRes}$$

→ permitem uma avaliação de valores discrepantes pois não são esperados valores maiores (em módulo) do que 2 ou 3;

→ como a média de $\hat{\varepsilon}_i$ é nula e s^2 é uma estimativa aproximada de sua variância espera-se que r_{pi} tenha uma distribuição t aproximada
  → $\hat{\varepsilon}_i$ e s não são independentes
  → $Var(\hat{\varepsilon}_i) = (1-h_{ii})\sigma^2$ → não considerado na dist t.

③ Resíduo Estudentizado Internamente (rSi)

$$rSi = \frac{\hat{\varepsilon}_i}{\sqrt{Var(\hat{\varepsilon}_i)}} = \frac{\hat{\varepsilon}_i}{\sqrt{(1-hii)QMRes}}$$

→ são mais sensíveis que os anteriores por considerarem variâncias distintas

→ valores discrepantes podem afetar profundamente a variância residual

→ Numerador ($\hat{\varepsilon}_i$) e denominador ($QMRes$) são variáveis dependentes → $Cov(\hat{\underset{\sim}{\varepsilon}}, QMRes) \neq 0$.

④ Resíduos Estudentizados Externamente ($rS(i)$)

$$rS(i) = \frac{\hat{\varepsilon}_i}{S(i)\sqrt{1-hii}}$$ em que: $S(i)$ é o desvio-padrão residual sem a i-ésima observação.

↳ livre da influência da i-ésima obs.

Sob normalidade, $rS(i)$ tem distribuição t-Student com $(n-p-1)$ gl.

ATENÇÃO: $|rS(i)| > 100\left(1-\frac{r}{2n}\right)$ percentil da $t_{(n-p-1)}$

ESTATÍSTICAS PARA DIAGNÓSTICOS

As observações podem ser classificadas, basicalle, como:

1) OBS INCONSISTENTES ⇒ resíduo grande

$|rS(i)| \geq t_{(n-p-1)}$ percentil $\frac{r}{2n}$ casos (a) e (c) GRÁFICO APOSTILA

OBS: Dentro dos inconsistentes, se o leverage (hii) for pequeno ⇒ observação é um OUTLIER (caso a)

2) PONTO DE ALAVANCA: Observações com leverage grande.

grande $\Rightarrow h_{ii} \geq \frac{2p}{n}$

visualmente: h_{ii} é grande quando a observação está relacionada a um x_i distante de $\bar{x}$.

Os pontos de alavanca podem ser:

→ BOM (se consistente) → RS(i) peq.

→ RUIM (se inconsistente) → RS(i) gde.

3) INFLUENTE: uma observação é considerada influente quando sua omissão (no conjunto de dados) resulta em mudanças significativas em alguns aspectos do modelo

→ no ajuste geral

→ nas estimativas dos parâmetros

As estatísticas:

① LEVERAGE (h_{ii}): mede a distância de uma observação em relação às demais, em termos da variável X.

No caso de uma RLS e $x_i = X_i - \bar{X}$, temos:

$$H = \underset{n\ 2}{\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} \underset{2\ 2}{\begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum x_i^2 \end{bmatrix}} \underset{2\ n}{\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}} = \underset{n\ 2}{\begin{bmatrix} 1/n & x_1/\sum x_i^2 \\ \vdots & \vdots \\ 1/n & x_n/\sum x_i^2 \end{bmatrix}} \underset{2\ n}{\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}} =$$

$$= \begin{bmatrix} \frac{1}{n} + \frac{x_1^2}{\sum x_i^2} & \frac{1}{n} + \frac{x_1 x_2}{\sum x_i^2} & \cdots & \frac{1}{n} + \frac{x_1 x_n}{\sum x_i^2} \\ & \frac{1}{n} + \frac{x_2^2}{\sum x_i^2} & \cdots & \frac{1}{n} + \frac{x_2 x_n}{\sum x_i^2} \\ & & \cdots & \end{bmatrix}$$

Portanto,

$$h_{ii} = \frac{1}{n} + \frac{x_i^2}{\sum x_i^2} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum x_i^2}$$

x_i longe $\bar{x} \Rightarrow h_{ii} \uparrow$

$x_i = \bar{x} \Rightarrow h_{ii} = \frac{1}{n}$

↳ qdo tem β_0

e

$$h_{ii'} = \frac{1}{n} + \frac{x_i x_{i'}}{\sum x_i^2} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_{i'} - \bar{x})}{\sum x_i^2}$$

Para um modelo com intercepto: $\frac{1}{n} \leq h_{ii} \leq 1$

Para um modelo para pela origem (sem β_0): $0 \leq h_{ii} \leq 1$

Além disso, $tr(H) = \sum_{i=1}^{n} h_{ii} = p \Rightarrow \bar{h} = p/n$ (valor médio)

↳ são consideradas observações influentes aquelas cujo h_{ii} supera em duas ou três vezes a sua média.

Lembrando que,

$$\underset{\sim}{\hat{y}} = H \underset{\sim}{y} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ \vdots & h_{22} & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$$\hat{y}_1 = h_{11} y_1 + h_{12} y_2 + \cdots + h_{nn} y_n$$

↳ $\hat{y}_1$ estimado é uma média ponderada de todas as observações $\Rightarrow h_{ii'}$ são os pesos de ponderação

↳ observações com $h_i \geq 2p/n \Rightarrow$ merecem atenção

↳ NO R: $h_i \geq 3p/n$

② DFBeta: mede a alteração no vetor $\hat{\underset{\sim}{\theta}}$ ao se retirar a i-ésima observação na análise.

$$DFBeta_{(i)} = \hat{\underset{\sim}{\theta}} - \hat{\underset{\sim}{\theta}}_{(i)}$$

no R: influente se $|DFBetaS_{(i)}| > 1$

③ DFFitS: mede a alteração provocada no valor ajustado $\hat{Y}_i$ ao retirar a i-ésima observação da análise.

$$DFFitS = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{h_{ii}\, S^2_{(i)}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} rS_{(i)}$$

↳ potencial de influência

merecem atenção as obs com $|DFFitS| > 2\sqrt{\frac{p}{n}}$

No R: $|DFFitS| > 3\sqrt{\frac{p}{n-p}}$

④ DISTANCIA DE COOK: também mede o afastamento do vetor de estimativas provocado pela retirada da i-esima observação

$$D_{(i)} = \frac{h_{ii}}{(1-h_{ii})} \cdot \frac{1}{p} \cdot (rS_i)^2$$

No R: $D_{(i)} > F_{\{0.5;\, p,\, n-p\}}$

⑤ COV-RATIO: mede a influencia da i-esima observação na $Var(\hat{\underset{\sim}{\theta}})$

$$COV\text{-}RATIO = \frac{S^2_{(i)}\, det\left|(X'_{(i)}X_{(i)})^{-1}\right|}{S^2\, det\left|(X'X)^{-1}\right|}$$

No R: $\left|COV\text{-}RATIO - \vec{1}\right| > \frac{3p}{n-p}$ ou $\frac{3p}{n}$

GRÁFICOS

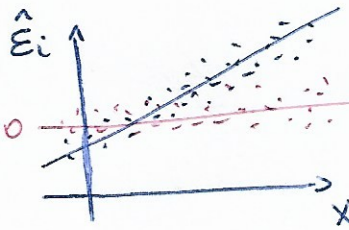
(1) RESÍDUOS VERSUS X fora

Relação entre os resíduos do modelo ajustado e uma variável ainda não incluída no modelo.

EXEMPLO: Y ↝ variável resposta

X_1 e X_2 ↝ variáveis regressoras

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \Rightarrow \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$



↝ ideal

↳ não incluída no modelo

ALTERNATIVA MELHOR: gráfico da variável adicionada

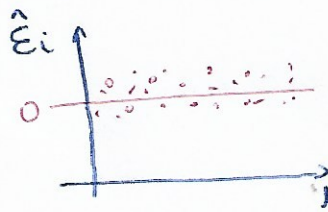
EXEMPLO: Y ~ variável resposta

X_1 e X_2 ↝ variáveis regressoras

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \Rightarrow \hat{\varepsilon}_i \text{ (resíduo)}$$

$$X_{2i} = \beta_0 + \beta_1 X_{1i} + \mu_i \Rightarrow \hat{\mu}_i \text{ (resíduo)}$$

$\hat{\varepsilon}_i \sim \hat{\mu}_i$



↝ ideal: coef. angular não significativo ($\beta_1 = 0$) de uma reta que passa pela origem ($\beta_0 = 0$)

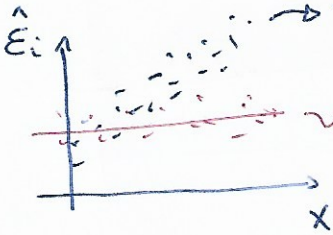
(2) RESÍDUOS VERSUS X_{dentro}

Verificar se ainda existe uma relação sistemática entre os resíduos de um modelo e a X_j já incluída no modelo ⇒ inclusão de termos de maior ordem X^2)

EXEMPLO: Y → variável resposta

X_1 e X_2 → variáveis regressoras

$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \Rightarrow \hat{\varepsilon}_i$



→ necessidade de maior ordem

→ ideal

→ já incluída no modelo

ALTERNATIVA MELHOR: gráfico de resíduo + componente

EXEMPLO: Y → variável resposta

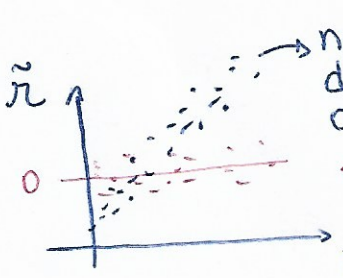
X_1 e X_2 → variáveis regressoras

$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \Rightarrow \hat{\varepsilon}_i$

↑ adiciona o componente $\hat{\beta}_1 X_{1i}$

$\tilde{r} = \hat{\varepsilon}_i + \hat{\beta}_1 X_{1i}$

↳ $\tilde{r}$ VERSUS X_1



→ necessidade de maior ordem

→ ideal: coef angular não significativo ($\beta_1 = 0$) e intercepto também não significativo ($\beta_0 = 0$)