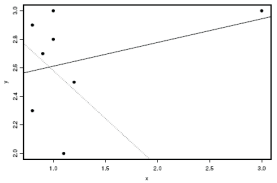
Análise de Diagnóstico

- A hipótese que todas as observações têm igual influência na estimação dos coeficientes do modelo de regressão linear nem sempre se verifica na prática.
- Uma observação pode substancialmente alterar os resultados obtidos.
- Mas por vezes as observações discordantes podem passar despercebidas e ter um efeito diminuto sobre a análise da regressão, noutros casos podem exercer uma influência grande sobre os parâmetros estimados, provocando estimativas desastrosas.
- Há muitas situações em que um único ponto pode ser determinante para uma recta de regressão estimada.



Análise de Diagnóstico

Em geral, faz-se a distinção entre três tipos de observações discordantes (ver Hadi):

- Outliers: observações que têm resíduos de valor elevado quando comparado com as outras observações.
- High-leverage points: observações que estão afastadas, no espaço das variáveis explicativas, da maioria das outras observações.
- Influentes: observações que, individualmente ou colectivamente, influenciam o modelo de regressão linear estimado. Eliminar estas observações na estimação dos modelos de regressão conduz a grandes mudanças nas estimativas dos coeficientes.

Nota:

- **Pequenos valores nos resíduos de certas observações não significa que não possam ser observações discordantes.**
- **Outliers não são necessariamente observações influentes nem observações influentes são necessariamente outliers;**
- **High-leverage points não são necessariamente observações influentes nem observações influentes são necessariamente**

Exemplo:

Resíduos

Para identificar os outliers, pode-se analisar os resíduos:

- Unstandardized residual:
$$e_i = y_i - \hat{y}_i$$
- Standardized residual:
$$r_i = \frac{e_i}{s(e_i)}$$
onde $s(e_i)$ representa a estimativa do desvio pardão de e_i
- Studentized residual :
$$r_{(i)} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$
onde $\hat{\sigma}_{(i)}$ é a estimativa da variância dos erros no modelo de regressão quando a observação i é eliminada.
Nota: Estes resíduos seguem uma distribuição t-student com $(n-p-2)$ graus de liberdade.

Pode-se mostrar que:

$$r_{(i)} = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}}$$

Resíduos

- Alguns autores sugerem que observações com $|r_i| \geq 3$ são identificadas como outliers.
- Alguns autores desenvolveram alguns testes para identificar observações outliers.
- No entanto, representações gráficas dos resíduos são aconselhadas para identificar estas observações:
  - QQ-plot dos resíduos;
  - gráfico dos resíduos em função dos valores estimados da variável dependente;
  - gráfico dos resíduos versus número de observação.

High-leverage points

Para identificar high-leverage points pode-se avaliar:

- leverage values da observação i:

$$h_{ii} = x_i(X^TX)^{-1}x_i^T$$

  - Alguns autores sugerem que observações podem ser declaradas como high-leverage points se $h_{ii} > \frac{2(p+1)}{n}$.
  - No entanto, o gráfico dos versus número de observação claramente identifica os high-leverage points.

  Propriedades:

  - $0 \leq h_{ii} \leq 1$
  - Média dos elementos da diagonal H é $\frac{p+1}{n}$

Exercício: Mostrar que na RLS: $h_{ii} = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum(x_i-\overline{x})^2}$

High-leverage points

Para identificar high-leverage points pode-se também avaliar:

- Mahalanobis distance: mede a distância, no espaço das variáveis explicativas, a que uma observação se encontra da média das outras observações:

$$M_i = \frac{n(n-2)}{n-1}\frac{h_{ii} - 1/n}{1-h_{ii}}$$

  Valores elevados de M_i identificam os high-leverage points.

- Weighted Squared Standardized Distance

Observações Influentes

Para identificar observações influentes pode-se avaliar:

- Distância de Cook: mede a influência que a observação tem nos coeficientes:

$$C_i = \frac{\sum_{i=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)} = \frac{r_i^2}{p+1}\frac{h_{ii}}{1-h_{ii}}$$

  - Valores elevados de C_i identificam as observações influentes.
  - Alguns autores sugerem que observações podem ser declaradas influentes se $C_i > F(0.5, p+1, n-p-1)$.
  - O gráfico C_i versus número de observação claramente identifica as observações influentes.

Observações Influentes

Para identificar observações influentes pode-se avaliar:

- DFFITS mede a influência que a observação tem nos valores estimados da variável dependente:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)})^2}{\hat{\sigma}^2_{(i)}(h_{ii})} = \frac{r_i^2}{p+1}\frac{h_{ii}}{1-h_{ii}}$$

  onde $\hat{y}_{i(i)}$ é o valor estimado da observação i no modelo de regressão estimado sem essa observação.

  - Valores elevados de $DFFITS_i$ identificam as observações influentes.
  - Alguns autores sugerem que $|DFFITS_i| > 2\sqrt{\frac{p+1}{n-p-1}}$ observações podem ser declaradas influentes.
  - O gráfico $DFFITS_i$ versus número de observação claramente identifica as observações influentes. ”

Observações Influentes

- Quando os dados contêm apenas um outlier, a sua identificação é um problema simples.
- No entanto, se os dados contém mais que uma observação outlier, a sua identificação pode ser complicada. Alguns problemas podem surgir:
  - Os resíduos e_i e os valores leverage estão relacionados por:

$$h_{ii} + \frac{e_i^2}{SSE} \leq 1$$

  Esta desigualdade indica que valores elevados de h_{ii} tendem a ter resíduos pequenos.
  - Masking : Ocorre quando não se detectam os outliers pois são "mascarados"por outras observações.
  - Swamping: Ocorre quando observações são incorrectamente identificadas como outliers.

Análise de Diagnóstico

O FAZER COM OS OUTLIERS?
Outliers, high -leverage points e observações influentes não devem ser automaticamente eliminadas do modelo, pois podem não ser necessariamente "más"observações. Pelo contrário podem indicar informação importante sobre os dados. Devem por isso ser analisados para identificar possíveis causas do aparecimento dessas observações.

Exercício:

O aumento dos preços na Tailândia, durante o período de 1940-1946, é expresso na tabela seguinte:

Ano (x)	40	41	42	43	44	45	46
Aumento (y)	1.62	1.63	1.90	2.64	2.05	2.13	1.94

(a) Apresente um diagrama de dispersão.

(b) Estime a recta de regressão e represente-a no diagrama efectuado na alínea anterior.

(c) Para cada valor do regressor, calcule as seguintes medidas de diagnóstico: *leverages*, resíduos estandardizados, resíduos *Studentizados*, $DFITS$ e distâncias de Cook.