# Análise de Influência e Alavancagem

Prof. Caio Azevedo

(grande parte do material apresentado foi extraído do livro Modelos de regressão com apoio computacional do Prof. Gilberto A. Paula)

link

# Motivação

- Estando ou não um modelo de regressão bem ajustado, é possível que uma ou mais observações possam influenciar as estimativas dos parâmetros (consequentemente, tudo o que depender delas).
- Por "influenciar" vamos entender que a presença e/ou a ausência de uma ou mais observações pode mudar, significativamente, os resultados (conclusões) inferenciais estatisticamente e/ou em termos do problema.
- No caso da classe dos MLG vista neste curso, os parâmetros em questão são $(\boldsymbol{\beta}', \phi)'$.

# Objetivos

- Para evitar discussões mais conceituais (também porque melhorar o ajuste de um modelo pode evitar que observações sejam "influentes", vamos considerar que o modelo em análise está bem ajustado).
- Assim, as técnicas que veremos não se prestam, **em princípio**, a avaliar se o modelo se ajusta bem ou não aos dados, mas "somente" em identificar as observações (possivelmente) influentes, segundo algum critério.
- Além disso, vamos nos focar no conceito de influência com base na mudança na previsão das observações (alavancagem) e no afastamento da verossimilhança (distância de Cook).

# Objetivos

- Estudaremos duas medidas: uma de alavancagem ("leverage"), e a distância de Cook (veja também aqui e referências nela contidas).
- A medida de alavancagem busca medir a influência em relação à eventuais mudanças na previsão das observações, enquanto que a distância de Cook visa medir a influência com base em eventuais mudanças na verossimilhança. Contudo, em ambos os casos, verifica-se se a retirada de uma (ou mais de uma) observação(ões) afeta as estimativas dos parâmetros.
- Para a classe dos MRNLH sugerimos a leitura de link e link

# Medida de alavancagem (pontos alavancas) “h”

- A ideia principal subjacente ao conceito de ponto de alavanca é a de avaliar a influência de cada $Yi(y_i)$ sobre o próprio valor ajustado (predito) $\widehat{Y}_i(\widetilde{Y}_i)$.
- Definem-se: $\widehat{Y}_i = g^{-1}\left(\sum_{j=1}^{p} X_{ji}\widehat{\beta}\right)$ e $\widetilde{Y}_i = g^{-1}\left(\sum_{j=1}^{p} X_{ji}\widetilde{\beta}\right)$, em que $\widehat{\beta}(\widetilde{\beta})$ é algum estimador (estimativa) de $\beta$, por exemplo, o EMV.
- Usualmente, essa influência é medida (apropriadamente) através de $\frac{\partial \widetilde{Y}_i}{\partial y_i}$.
- Assim, para os MRNLH, uma vez que $\widetilde{Y}_i = \boldsymbol{H}_i y_i$, em que $\boldsymbol{H}_i = \boldsymbol{X}_i' \left(\boldsymbol{X}\boldsymbol{X}\right)^{-1} \boldsymbol{X}_i$, tal cálculo é bem simples.

# Medida de alavancagem (pontos alavancas) “h”

- Com efeito, para os MRNLH os elementos da diagonal principal da matriz de projeção (ou matriz “hat”) $\boldsymbol{H} = \boldsymbol{X}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}$, representam a medida de alavancagem, i.e.,

$$\boldsymbol{h} = \text{diag}(\boldsymbol{H}) \rightarrow h_{ii} = [\boldsymbol{h}]_{ii}.$$

- No entanto, para os MLG's, a menos que se considere a função de ligação identidade $\left(g(\widetilde{Y}_i) = \sum_{j=1}^{p} X_{ji}\widetilde{\boldsymbol{\beta}}\right)$, tal cálculo pode ser bastante complicado.

# Medida de alavancagem (pontos alavancas) “h”

- Uma definição de pontos de alavanca que tem sido utilizada na classe dos MLG's, proposta por Pregibon (1981), é construída fazendo-se uma analogia entre o estimador de máxima verossimilhança para $\beta$ para um dado MLG (aqui) e a solução de mínimos quadrados de uma regressão normal linear ponderada (aqui).

link

# Medida de alavancagem (pontos alavancas) “h”

- Considerando a expressão para $\widehat{\boldsymbol{\beta}}$ obtida na convergência do processo iterativo (algoritmo Escore de Fisher, aqui), tem-se que:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{z}},$$

em que $\widehat{\boldsymbol{z}} = \widehat{\boldsymbol{\eta}} + \widehat{\boldsymbol{W}}^{-1/2}\widehat{\boldsymbol{V}}^{-1/2}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$.

- Portanto, $\widehat{\boldsymbol{\beta}}$ pode ser visto como a solução de mínimos quadrados da regressão linear de $\widehat{\boldsymbol{W}}^{1/2}\widehat{\boldsymbol{z}}$ contra as colunas de $(\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X})$ (solução de mínimos quadrados ponderados).

link

# Medida de alavancagem (pontos alavancas) “h”

- Portanto, a matriz de projeção da solução de mínimos quadrados da regressão linear de $\widehat{z}$ contra $\boldsymbol{X}$, com pesos $\widehat{\boldsymbol{W}}$, fica dada por

$$\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{1/2}.$$

- O resultado acima sugere a utilização dos elementos $\widehat{h}_{ii}$ da diagonal principal de $\widehat{\boldsymbol{H}}$ para detectar a presença de pontos de alavanca nesse modelo de regressão normal linear ponderada (MLG).
- Quanto maior o valor de $\widehat{h}_{ii}$, mais indicações pode-se ter de que o ponto $i$ é (de) alavanca.

# Distância de Cook

- Supondo $\phi$ conhecido, o afastamento pela verossimilhança quando elimina-se a i-ésima observação é denotado por

$$LD_i = 2\left(l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_{(i)})\right),$$

em que $l(\boldsymbol{\beta})$ é a log-verossimilhança de um MLG com $\phi$ conhecido (aqui).

- Uma aproximação da expressão acima é dada por:

$$LD_i \approx \phi\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)'\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right).$$

link

# Distância de Cook

- Conforme Cook, R. D. and Weisberg, S. (1982), uma aproximação para $\widehat{\boldsymbol{\beta}}_{(i)}$ é dada por:

$$\widehat{\boldsymbol{\beta}}^1_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{RP_i\sqrt{\widehat{\omega}_i\phi^{-1}}}{\left(1-\widehat{h}_{ii}\right)}\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i,$$

em que (resíduo de Pearson) $RP_i = \frac{\sqrt{\phi}(Y_i-\widehat{\mu}_i)}{\sqrt{\widehat{V}_i}}$.

# Distância de Cook

- Portanto:

$$LD_i \approx \left( \frac{\widehat{h}_{ii}}{1 - \widehat{h}_{ii}} \right) T_{S_i}^2,$$

em que (resíduo padronizado) $T_{S_i} = \frac{\sqrt{\phi}(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{V}_i\left(1 - \widehat{h}_{ii}\right)}}$.

- Quanto maior o valor de $LD_i$, mais indicações pode-se ter de que o ponto $i$ é influente.

# Observações

- Diferentemente do que ocorre para os MRNLH, não existe um "ponto de corte apropriado", a partir do qual considera-se um ponto como (candidato à) alavanca e/ou (candidato à) influente.
- Uma abordagem (heurística) comumente usada é analisar, mais detalhadamente, as observações com maiores valores (em relação as demais) para cada uma das medidas. Isso se aplica a ambas as medidas.
- Posteriormente, pode-se avaliar (individual ou conjuntamente) o impacto de tais observações, retirando-as e avaliando o quanto a(s) estimativa(s) de $\beta$, se modifica(m).