MLG

Curso de Modelos Lineares Generalizado - DEST/UFMG
Marcos Oliveira Prates

23 de outubro de 2017

Técnicas para Diagnósticos em GLM

- Assim como para modelos lineares técnicas de diagnósticos foram desenvolvidas para GLM.
- Um passo inicial é extender as analises feitas para modelos lineares e achar uma medida equivalente na classe de GLM.
- Dessa forma, seria interessante que essas extensões equivalentes também preservem a interpretação anterior.

Pontos de alavanca

- A matriz H para modelos lineares é obtida por

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

  e vimos que os h_{ii} podem ser utilizados para detectar pontos de alavanca.
- Porém pode se verificar que h_{ii} é dado por

$$\frac{\partial \hat{y}_i}{\partial y_i}$$

- Assim, Wei, Hu e Fung (1998) propuseram uma forma geral para obter a matriz

$$\left(\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}'}\right)_{n \times n}$$

  quando as respostas são continuas.

- Especificamente para GLM pode se mostrar que

$$\hat{G}L = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}'} = D_\beta(-\ddot{l}_{\beta\beta})^{-1}\ddot{l}_{\beta y}|_{\hat{\beta}}$$

onde $D_\beta = \frac{\partial \mu}{\partial \beta}$, $\ddot{l}_{\beta\beta} = \frac{\partial^2 l(\beta)}{\partial \beta \beta'}$ e $\ddot{l}_{\beta y} = \frac{\partial^2 l(\beta)}{\partial \beta \mathbf{y}'}$.

- Em modelos GLM temos que $D_\beta = \frac{\partial \mu}{\partial \beta} = N\mathbf{X}$ e $\ddot{l}_{\beta y} = \frac{\partial^2 l(\beta)}{\partial \beta \mathbf{y}'} = \phi^{-1}\mathbf{X}' W^{-1} N$, onde $N = \text{diag}(d\mu_1/d\eta_1, \ldots, d\mu_n/d\eta_n)$.
- Substituindo $\ddot{l}_{\beta\beta} = -\phi^{-1}(\mathbf{X}' W\mathbf{X})$, seu valor esperado. Obtemos:

$$\hat{G}L = \hat{N}\mathbf{X}(\mathbf{X}' W\mathbf{X})^{-1}\mathbf{X}' W^{-1}\hat{N}$$

- Selecionando a diagonal $\hat{G}L_{ii}$ temos uma alternativa para representar pontos de alavanca no caso GLM.
- Note que
$$\hat{G}L_{ii} = \hat{\omega}_i \mathbf{x}_i'(\mathbf{X}'W\mathbf{X})^{-1}\mathbf{x}_i,$$
onde $\omega_i = \frac{(d\mu_i/d\eta_i)^2}{W_i}$.
- $\hat{G}L_{ii}$ apesar de análoga a definição de obtida no casos de modelos lineares essa estimativa não é única.

- Pontos de alavanca também podem ser obtidos fazendo uma analogia entre o estimador de de máxima verosimilhança para $\hat{\beta}$ no MLG e a solução de mínimos quadrados para modelos lineares ponderados.
- Vimos que a estimativa para o $\hat{\beta}$ para modelos GLM pode ser pensada como a solução de um modelo linear ponderado com a seguinte forma:

$$\hat{\beta}^{(t+1)} = (\mathbf{X}' W^{(t)} \mathbf{X})^{-1} \mathbf{X}' W^{(t)} z^{(t)},$$

onde $\hat{z} = \mathbf{X}\beta^{(t)} + W^{-1}(\mathbf{Y} - \mu^{(t)})$

- Logo, após convergência $\hat{\beta}$ pode ser interpretado como a solução de mínimos quadrados de $\hat{W}\hat{z}$ contra as colunas de $\hat{W}^{1/2}X$.

- No caso da solução de modelos lineares com mínimos quadrados a matriz de projeção é dada por:

$$\hat{H} = \hat{W}^{1/2}\mathbf{X}(\mathbf{X}'\hat{W}\mathbf{X})^{-1}\mathbf{X}W^{1/2}.$$

- Assim, sugere a utilização do $\hat{h}_{ii}$ como a medida de ponto de alavanca (Pregbon, (1981)).
- Para ligações canônicas temos que $\hat{GL}_{ii} = \hat{h}_{ii}$.
- Com ligações não canônicas pode se mostrar que $\hat{GL}_{ii} = \hat{h}_{ii}$ para um tamanho de amostra grande.
- Como $\hat{h}_{ii}$ depende de $\hat{\mu}_i$ sugere para detectar pontos de alavanca fazer o gráfico $\hat{h}_{ii} \times \hat{\mu}_i$.

Resíduos para GLM

- Uma definição análoga aos resíduos studentizados para modelos lineares pode ser feitas para modelos GLM.
- Todavia isso não garante que as propriedades continuem valendo.
- Assim pode se pensar em resíduos que preservem as propriedades desejadas.

- Uma primeira proposta seria considerar os resíduos ordinários da solução de mínimos quadrados da regressão de $\hat{z}$ em $\mathbf{X}$, definido por $r^\star = \hat{W}^{-1/2}(\mathbf{y} - \hat{\mu})$
- Se assumirmos que a variância de $\hat{z} \approx W^{-1}\phi$, então $\mathrm{Var}(r^\star) \approx \phi(I - \hat{H})$.
- Logo, podemos pensar o resíduo padronizado como

$$t_{S_i} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\phi \hat{W}_i (1 - \hat{h}_{ii})}}$$

- Como na prática $\hat{\mu}$ não é conhecido e nem $\hat{z}$ é normalmente distribuído, as propriedades de t_i não são verificadas para t_{S_i}.
- Williams (1984) mostra através de simulação que na prática a distribuição de t_{S_i} são assimétricas.
- O resíduo mais utilizado em GLM é o seguinte:

$$t_{D_i} = \frac{\sqrt{D^*(y_i;\hat{\mu}_i)}}{\sqrt{1-\hat{h}_{ii}}} = \frac{\sqrt{D(y_i;\hat{\mu}_i)}}{\sqrt{\phi(1-\hat{h}_{ii})}}.$$

- Williams (1984) verificou através de simulações que a distribuição de t_{D_i} é mais próxima da normalidade.

Identificando Observações Influentes

- Observação Influente
  Como para modelos lineares, após identificar observações que são outliers com respeito aos valores de Y e/ou valores de X, o próximo passo é determinar se essas observações são ou não pontos influentes.
  Continuamos a considerar uma observação influente se a exclusão dessa observação causa uma grande mudança no ajuste da função regressão.

Cook's Distance para GLM

- Supondo ϕ conhecido vamos definir a influência LD_i como

$$LD_i = 2(l(\hat{\beta}) - l(\hat{\beta}_{(i)}))$$

onde $\hat{\beta}_{(i)}$ denota o valor estimado para $\hat{\beta}$ sem a i-ésima observação nos dados.

- O calculo de LD_i não possui forma analítica. Assim, utiliza-se uma expansão de Taylor de segunda ordem em LD_i para se obter

$$LD_i \approx \phi^{-1}(\hat{\beta} - \hat{\beta}_{(i)})'(\mathbf{X}'W\mathbf{X})(\hat{\beta} - \hat{\beta}_{(i)})$$

onde $\phi^{-1}\mathbf{X}'W\mathbf{X}$ é o valor esperado de $-\ddot{l}(\hat{\beta})_{\beta\beta}$.

- De forma geral não é possível encontrar de forma fechada $\hat{\beta}_{(i)}$. Portanto, a aproximação (Pregbon, (1981)) é utilizada:

$$\hat{\beta}^1_{(i)} = \hat{\beta} + [-\ddot{l}(\hat{\beta})_{\beta\beta}]^{-1} l_{(i)}(\hat{\beta})$$

  onde $l_{(i)}(\hat{\beta})$ é a função de log-verosimilhança sem a i-ésima observação.

- Substituindo, $\ddot{l}(\hat{\beta})_{\beta\beta}$ e $l_{(i)}(\hat{\beta})$ por seus valores esperados temos que

$$\hat{\beta}^1_{(i)} = \hat{\beta} + \frac{\hat{r}_{P_i}\sqrt{\hat{\omega}_i}}{\phi^{1/2}(1 - \hat{h}_{ii})} (\mathbf{X}'\hat{W}\mathbf{X})^{-1}\mathbf{x}_i$$

- Ao determinar $\hat{\beta}^1_{(i)}$ podemos substituir na aproximação de segunda ordem e obtemos

$$LD_i \approx \left(\frac{\hat{h}_{ii}}{1 - \hat{h}_{ii}} \right) t^2_{S_i}$$

- A validade dessa aproximação ainda esta sendo investigada por pesquisadores. Até o momento acredita-se que a mesma subestima o verdadeiro valor de LD_i, porém é suficiente para chamar a atenção de pontos influentes.

Diagnósticos de Influência Local

- Diagnósticos de influência local também podem ser feitos para GLM. Porém, seus cálculos não são tao simples.
- Influência local tem sido estudada por diversos pesquisadores. A ideia consiste basicamente em perturbar o vetor de covariáveis e verificar como essa perturbação influencia as estimativas.
- Para isso, Cook (1986) utilizou o conceito de curvatura normal

$$C_a(\theta) = 2|a'\Delta'(\ddot{l}_{\hat{\theta}\hat{\theta}})^{-1}\Delta a|$$

onde a é uma direção unitária que se deseja analisar a influência.

- Uma sugestão é utilizar $C_{a_{\max}}$, ou seja, a direção a de maior curvatura. Isso implica, que observações sob pequena perturbação exerce influência desproporcional em LD
- Para GLM se consideramos a função de perturbação da forma

$$l(\beta|\delta) = \sum_{i=1}^{n} \delta_i l_i(\beta),$$

com $0 \leq \delta_i \leq 1$, pode se mostrar que

$$\Delta = \phi^{-1/2} X' \hat{W}^{1/2} D(\hat{r}_P)$$

onde $D(\hat{r}_P) = \mathrm{diag}(\hat{r}_{P_1}, \ldots, \hat{r}_{P_n})$ e $r_{P_i} = \frac{\sqrt{\phi}(y_i - \hat{\mu}_i)}{\sqrt{\hat{W}_i}}$.

- Substituindo $\ddot{l}_{\hat{\theta}\hat{\theta}}^{-1}$ por seu valor esperado obtemos

$$C_a(\theta) = 2|a' D(\hat{r}_P) \hat{H} D(\hat{r}_P) a|$$

- Se escolhemos a para ser na direção da i-ésima observação $\mathbf{X}_i$ temos que

$$C_i = 2\hat{h}_{ii}\hat{r}_{P_i}$$

- Uma sugestão para detectar observações influentes é verificar se $C_i > \bar{C} \pm 2\text{sd}(C)$
- Em particular, o vetor a_{max}, ou seja, a direção de maior influência é dado pelo autovetor correspondente ao maior autovalor da matriz

$$B = D(\hat{r}_P)\hat{H}D(\hat{r}_P)$$

- O gráfico de a_{max} contra a ordem das observações pode ser usado para detectar observações influentes.

- Se desejamos detectar observações influentes na estimativa de um coeficiente em particular, associado a variável X_i, podemos reescrever o vetor $a_{\max}$ como

$$a'_{\max} = \left(\frac{\nu_1 \hat{r}_{P_1}}{\sqrt{C_{a_{\max}}}}, \ldots \frac{\nu_n \hat{r}_{P_n}}{\sqrt{C_{a_{\max}}}} \right)$$

  onde $\nu_1, \ldots, \nu_n$ sao obtidos da regressao linear de $\mathbf{X}_i$ nas colunas de $\mathbf{X}_{-i}$ com pesos $\hat{W}$, ou seja, $\nu = \hat{W}^{1/2}\mathbf{X}_i - W^{1/2}\mathbf{X}_{-i}(\mathbf{X}'_{-i}\hat{W}\mathbf{X}_{-i})^{-1}\mathbf{X}'_{-i}\hat{W}\mathbf{X}_i$.
- O gráfico do novo $a_{\max}$ contra a ordem das observações pode ser usado para detectar observações influentes.