PROCEDIMENTOS GRÁFICOS PARA IDENTIFICAÇÃO DE PONTOS INFLUENTES NO AJUSTE DE UM MODELO LOGISTICO

Giulia Orsi Gonçalves
A.C. Nielsen
Pedro Ferreira Filho
Departamento de Estatística - UFSCar

1. Introdução:

Atualmente a regressão logística é um procedimento estatístico amplamente utilizado em diferentes áreas. Destaca-se a sua utilização em áreas como finanças, *marketing*, saúde entre outras. A resposta a problemas nestas áreas, na grande parte dos casos, é feita a partir ajuste de modelos de regressão logística, uma vez que a variável resposta de interesse é dicotômica.

Do ponto de vista de metodologia estatística, a regressão logística é um caso particular de regressão não linear onde a variável resposta assume apenas dois valores se dicotômica, ou mais de dois valores no caso de politômica.

Por outro lado, de forma complementar, a utilização de métodos gráficos em estatística é de extrema importância; a partir deles, podemos averiguar o comportamento de variáveis, tendências de comportamento e verificar a adequabilidade do ajuste de um modelo. Em muitas situações, a utilização de procedimentos gráficos é de extrema importância para uma melhor compreensão dos resultados obtidos. No ajuste de um modelo de regressão logística não é diferente; embora pouco conhecidos e conseqüentemente, pouco utilizados, diferentes procedimentos gráficos podem ser utilizado de forma a facilitar a análise dos dados observados, a verificação da qualidade do ajuste e a interpretação dos resultados obtidos.

A partir deste princípio, Friendly[4] publicou o livro “Visualizing Categorical Data”. Neste livro, o autor propõe métodos gráficos específicos para diferentes procedimentos para análise de dados categóricos. Especificamente, Friendly propõe procedimentos gráficos para verificação da qualidade de ajuste e interpretação dos resultados de um modelo de regressão logística.O objetivo desse trabalho é o de apresentar alguns destes procedimentos gráficos .

2. Material e Métodos

2.1. Modelo Logístico:

O ajuste de modelos de regressão, em geral, procura verificar o quanto um conjunto de covariáveis X (também chamadas variáveis preditoras) tem a capacidade de “explicar” uma dada variável resposta Y. Nas situações mais usuais esta resposta Y é uma medida quantitativa e neste caso o ajuste de um modelo de regressão linear geral pode ser utilizado. Por outro lado, em diferentes áreas do conhecimento, a variável resposta é categórica e em grande parte dos casos, dicotômica. A regressão logística constitui-se então em uma alternativa para análise da relação entre a resposta categórica Y e as covariáveis X no estudo. O caso mais simples é aquele onde temos a presença de uma única covariável X, chamado modelo de regressão simples. Na presença de duas ou mais covariáveis temos um modelo de regressão múltipla. É importante também destacar que as

covariáveis podem ser medidas quantitativas ou qualitativas ou também produtos de medidas quantitativas por uma variável qualitativa.

Considerando agora um conjunto de p variáveis preditoras x, um modelo para $E(Y=1/x)$ é dado pelo Modelo Logístico:

$$E(Y=1/x_i)=\frac{\exp\left\{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right\}}{1+\exp\left\{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right\}}=p_{x_i}=p_i\ .$$

2.2. Diagnóstico do Modela:

A qualidade de ajuste do modelo ajustado engloba medidas de influência de pontos, afim de verificar se existe alguma observação que apresente impacto acima do esperado no modelo ajustado de regressão. Para o modelo usual de regressão, usa-se o *leverage*, Distância de Cook, DFBETAs, entre outros; na regressão logística existem para a maioria dos casos, medidas análogas as utilizadas para a regressão usual. Pregibon [10] apresentou a base teórica para a fundamentação desses métodos, explorando a relação entre os modelos logísticos e mínimos quadrados ponderados.

Por exemplo, sabe-se que a influência de pontos depende multiplicativamente do resíduo (diferença entre y e $\hat{y}$) e de seus *leverage* (quão fora do comum x_i está no espaço das variáveis explicativas). Na regressão logística o resíduo puro é definido por: $e_i = y_i - \hat{p}_i$, em que

$$\hat{p}_i=\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}.$$

Os resíduos de Pearson e *Deviance* são mais utilizados para verificar observações mal ajustadas e também são componentes da qualidade de ajuste. Os resíduos de Pearson é dado por:

$$r_i=\frac{e_i}{\sqrt{p_i(1-p_i)}}$$

E a estatística de teste de Qui-Quadrado de Pearson é dada pela soma dos resíduos ao quadrado: $\chi^2=\sum r_i^2$, já o resíduo *Deviance* é dado por:

$$g_i \equiv \pm -2\left[y_i \log p_i+(1-y_i)\log(1-p_i)\right]^{1/2}$$

Em que o sinal de gi é o mesmo do resíduo ei. De forma similar ao resíduos de Pearson, a soma dos quadrados dos resíduos *Deviance* também criam um *Deviance* "geral"; $G^2=\sum g_i^2$

É importante notar que quando y_i é binomial baseado em n tentativas, o resíduo de Pearson é então denotado por:

$$r_i=\frac{y_i-n_i p_i}{\sqrt{n_i p_i(1-p_i)}}$$

Seguindo o mesmo pensamento, temos que para o mesmo caso, o resíduos *Deviance* será:

$$g_i \equiv \pm -2\left[y_i \log n_i p_i+(1-y_i)\log(1-n_i p_i)\right]^{1/2}$$

A medida *Leverage* mensura aproximadamente o potencial impacto que um caso individual possa ter no resultado como um todo. O *Leverage* é calculado a partir dos elementos diagonais h_{ii} da matriz *hat*, ou matriz chapéu H:

$$H = X^*\left(X^{*\prime}X^*\right)^{-1}X^{*\prime}, \qquad (4.11)$$

Em que $X^* = V^{1/2}X$ e $V = diag\left[\hat{p}(1-\hat{p})\right]$. No caso de uma regressão usual, normalmente observamos que os valores dos *leverage*s estão entre 0 e 1, e que se $h_{ii} > 2(k+1)/n$, então é considerado "grande". Entretanto, ao trabalharmos com uma regressão usual, estimada por mínimos quadrados ordinários, estamos trabalhando com uma matriz H dependente somente dos x's; enquanto que, quando estamos trabalhando com a regressão logística, essa matriz também depende dos valores da variável dependente e das probabilidades ajustadas.

O resultado disso é que uma observação pode ser extremamente fora do comum nos preditores, mas ainda assim, não possuir um grande valor h_{ii}, se a probabilidade ajustada estiver próxima de 0 ou 1.

As medidas de influência verificam até que ponto excluir uma observação, normalmente considerada fora do padrão terá impacto nos parâmetros da regressão, os valores ajustados e as estatísticas de qualidade de ajuste. Numa regressão comum, essas medidas podem ser calculadas exatamente de apenas uma regressão, enquanto que na regressão logística é necessário um maior cuidado; é necessário que as observações sejam excluidas uma a uma e que o modelo seja reajustado, de forma a verificar o efeito exato de cada observação dcluida. Isso ocorre uma vez que estamos trabalhando com estimação de equações que são não lineares. Dessa forma, Pregibon [10] mostrou quão análogo o diagnóstico de exclusão pode ser aproximado, performando um passo adicional do procedimento iterativo.

A medida mais simples para verificar a influência da observação i na diferença padronizada dos coeficientes para cada variável, devido à omitização dessa observação é o DFBETA, que é dado por:

$$b - b_{(-i)} = \left(X'VX\right)^{-1} x_i \frac{(y_i - p_i)}{(1-h_{ii})}$$

Dessa forma, a diferença estimada padronizada no coeficiente para a variável j é

$$DFBETAi_j \equiv \frac{b_{(-i)j} - b_j}{\hat{\sigma}\left(b_j\right)}$$

Em que $\hat{\sigma}\left(b_j\right)$ é o erro estimado de b_j. Se existe k variáveis regressoras, então teremos k+1 conjuntos de DFBETAs, o que faz com que a verificação seja cansativa. Representações gráficas fazem com que essa verificação seja mais simples, assim como várias estatísticas descritivas consideradas abaixo.

A influência geral da observação i nas estimativas dos coeficientes de regressão é auxiliada por uma medida análoga à distância de Cook (Neter *et al* [9]), que mede a diferença de b para todo conjunto de dados $b_{(-i)}$ estimado sem a observação i. Uma medida, C_i, é definida por:

$$C_i \equiv \left(b - b_{(-i)}\right)\left(X'VX\right)^{-1}\left(b - b_{(-i)}\right)$$

E calculada por:

$$C_i = \frac{r_i^2 h_{ii}}{(1-h_{ii})^2}$$

Uma segunda medida, denotada por $\overline{C_i}$, é calculada por:

$$\overline{C_i} = \frac{r_i^2 h_{ii}}{(1-h_{ii})} = (1-h_{ii})C_i$$

Ambas medidas acima podem ser obtidas no output do proc logistic do *software* SAS. É interessante notar que $\overline{C_i}$ nunca será maior que C_i, uma vez que $0 < h_{ii} < 1$. Os resíduos de Pearson e

deviance definidos acima não possuem variância igual, mas sim aproximadamente 1- h_{ii}. As versões *Studentizadas* de ambos resíduos possuem variância igual uma vez se divididas por $\sqrt{1-h_{ii}}$. O resíduo *deviance* studentizado é então obtido por:

$$g_i* = \frac{g_i}{\sqrt{1-h_{ii}}}.$$

Entretanto, nota-se que normalmente é utilizado o quadrado desses resíduos associado com a observação i deletada:

$$\Delta G^2_{(-i)} = \frac{g^2{}_i}{\sqrt{1-h_{ii}}},$$

E para o resíduo studentizado de Pearson:

$$\Delta \chi^2_{(-i)} = \frac{r^2{}_i}{\sqrt{1-h_{ii}}}.$$

Ambas estatísticas descritas acima possuem assintoticamente uma distribuição χ^2_1, então um valor que exceda o valor de 4, que é o valor crítico aproximado da χ^2_1, vale a pena verificar.

Representar graficamente a mudança no χ^2 contra cada *leverage* ou probabilidade predita é particularmente muito útil para detectar casos indevidamente influentes. Esses são gráficos análogos aos que são recomendados para modelos lineares por Fox [3] e Friendly [4]. A influência geral estimada para cada caso nos coeficientes estimados ($\overline{C}_i$ ou C_i) podem ser mostradas em um gráfico de bolha em que os símbolos são círculos proporcionais ao $\overline{C}_i$ ou C_i.

Como já foi dito anteriormente, uma outra medida muito utilizada para verificar a influência dos pontos é o DFBETA. Essa medida verifica a influência da observação i na diferença padronizada nos coeficientes para cada variável, devido à omitização dessa observação. Apesar de ser a medida mais simples utilizada, torna-se cansativa de ser observada, quando temos um grande número de variáveis explicativas.

3. Aplicação

Consideremos o problema apresentado por Colosimo [2]. O estudo considera que ao constatar que um paciente desenvolveu câncer de próstata, é fundamental, para se decidir qual tratamento utilizar, saber se o câncer já se espalhou para os linfonodos próximos.O objetivo do trabalho foi medir a capacidade de predição para o envolvimento nodal de cinco variáveis pré-operatórias cuja coleta é menos invasiva que uma cirurgia. As variáveis são: resultado da radiografia, sendo 0 negativo e 1 positivo, grau – resultado da biopsia, idade (à época do diagnóstico, medida em anos), ácido - nível de fosfatase ácida (x100) e medida do estádio do tumor, sendo codificado por 1 o estado mais grave do câncer e 0 o estado menos grave.

A partir dos resultados do ajuste obtidos por Colosimo[2], podemos verificar a existência de pontos influentes a partir de diferentes representações gráficas da mudança no χ^2. Essas representação são dadas nas figuras a seguir:

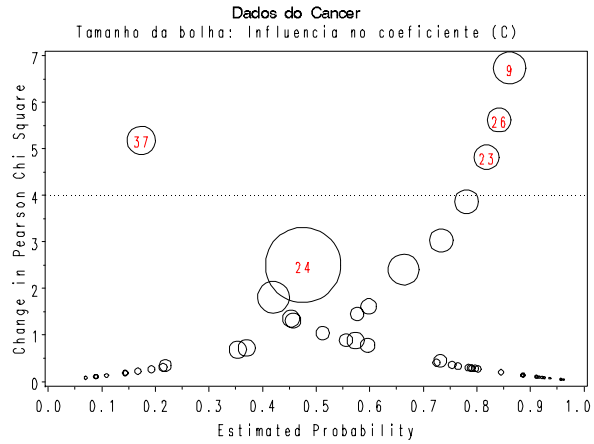

Figura 1: Probabilidade estimada pela diferença na Qui-Quadrado de Pearson.

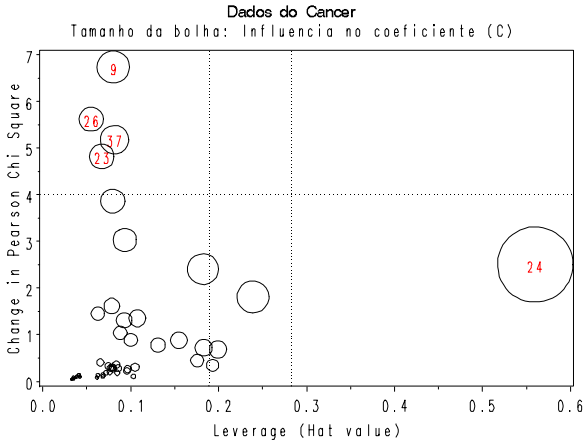

Figura 2: *Leverage* pela diferença na Qui-Quadrado de Pearson.

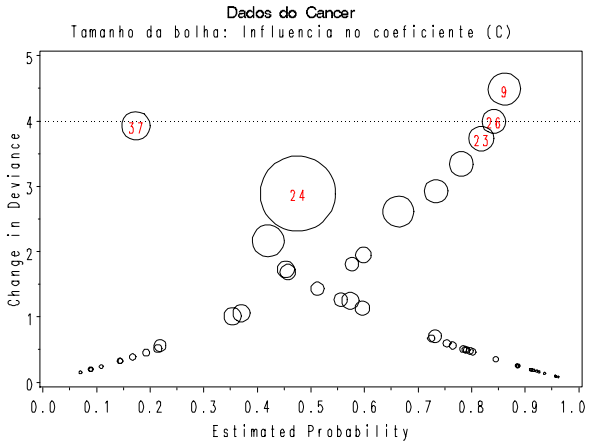

Figura 3.: Probabilidade estimada pela diferença no *Deviance*.

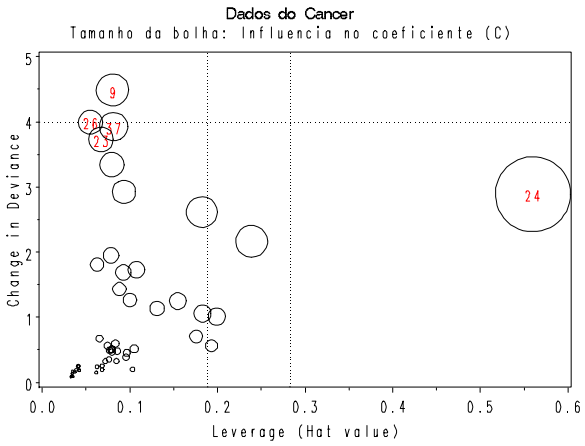

Figura 4.: *Leverage* pela diferença no *Deviance*.

Nas Figuras 1 a 4 é possível observar que as observações 9 e 24 são as aqueles que geram maiores mudanças no qui-quadrado de Pearson e no *Deviance*, quando em relação ao *Leverage* e à probabilidade estimada.

Por outro lado, para verificarmos o impacto nas estimativas dos parâmetros do modelo ajustado podemos verificar a seguinte representação gráfica de DFBETA para cada uma das variáveis preditoras:

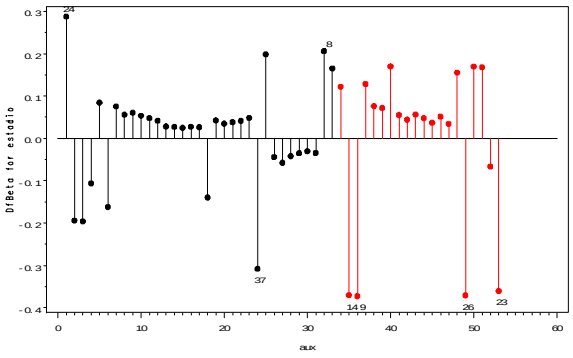

Figura 5.: Gráfico do DFBeta para Estádio.

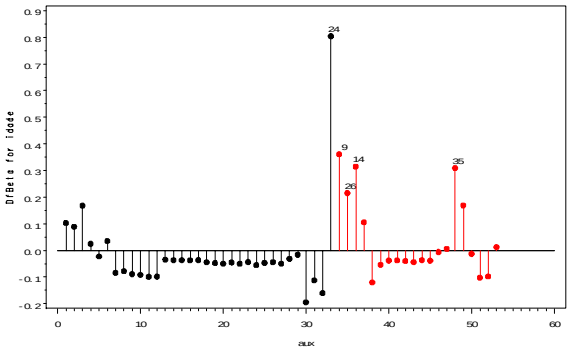

Figura 6.: Gráfico do DFBeta para Idade.

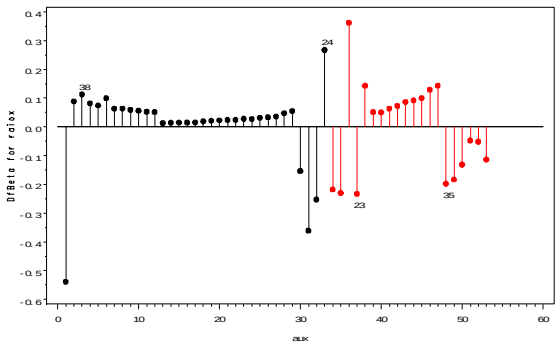

Figura 7.: Gráfico do DFBeta para Raio-x .

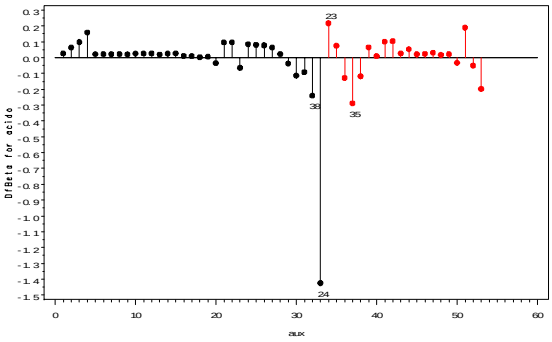

Figura 8.: Gráfico do DFBeta para Ácido.

As figuras, 5 a 8 acima, ratificam os resultados observados nas figuras 1 a 4. O impacto da observação 24 nos valores estimados das diferentes variáveis no modelo é extremamente significativo. Observe-se que esse indivíduo não apresentou câncer na próstata, apesar do alto nível de acido, o que torna essa observação consideravelmente influente. É importante destacar que esses gráficos foram feitos ordenando a variável resposta de forma a facilitar a visualização e auxiliar em entender o porquê de uma observação ser influente.

Conclusão:

Com esse trabalho, procura-se mostrar a utilização de alguns procedimentos gráficos para a identificação de pontos influentes no ajuste de um modelo de regressão logística. Os gráficos apresentados permitem de forma bastante simples a identificação de pontos influentes mesmo na presença de uma grande quantidade de observações

Referências Bibliográficas

[1] Allison, P. D., *Logistic Regression using SAS: Theory and Application*, SAS, 1999.
[2] Colosimo, E.A; Soares, J.F., *Métodos Estatísticos na Pesquisa Clínica*, Sociedade Brasileira de Biometria, 1995.
[3] Fox, J., Regression Diagnostics: And Introduction. Sage Pulblications, Beverlly Hills, CA, 1991.
[4] Friendly, M., *SAS Systems for Statistical Graphics.* SAS Institute Inc., Cary, NC, Primeira Edição, 1991.
[5] Friendly, M., *Visualizing Categorical Data*, SAS, 2000.
[6] Giolo, S. R., *Análise de Dados Categóricos*, UFPR, 2006.
[7] Hosmer, D. W.; Lemeshow, S., *Applied Logistic Regression*, Editora Wiley, 1989.
[8] Joiner, B., Lurking variables: Some examples. *The American Statistician*, 35: 227-233, 1981.
[9] Neter, J.; Kutner, M.H.; Nachtsheim, C.J., Wasserman, W., *Applied Linear Statistical Models*, McGraw-Hill, 1996.
[10] Pregibon, D., Logistic Regression Diagnostics. *Annals of Statistics*, 9:705-724, 1981.