Modelos Lineares Generalizados

Métodos de diagnóstico

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

17 de junho de 2021

Os diagnósticos de regressão são métodos para determinar se um modelo de regressão ajustado representa adequadamente os dados. A maioria dos diagnósticos para modelos lineares se estendem de forma relativamente diretamente aos GLMs.

Os diagnósticos aproximados são baseados diretamente na solução dos mínimos quadrados ponderados ou são derivados de estatísticas facilmente calculadas a partir desta solução. O trabalho sobre a extensão de diagnósticos de mínimos quadrados lineares para modelos lineares generalizados foi feito por Pregibon (1981), Landwehr, Pregibon e Shoemaker (1984), Wang (1985, 1987) e Williams (1987), entre outros.

Modelos lineares ajustados por mínimos quadrados fazem suposições fortes e às vezes irrealistas sobre a estrutura dos dados. Quando essas premissas são violadas, as estimativas de mínimos quadrados podem se comportar mal e podem até representar os dados de maneira completamente incorreta. Os diagnósticos de regressão podem revelar esses problemas e, muitas vezes, apontar o caminho para as soluções.

Todos os métodos discutidos estão disponíveis nas funções R padrão ou são implementados no pacote do car. Um dos objetivos do pacote car é fazer diagnósticos para modelos lineares e GLMs prontamente disponíveis em R. Nossa experiência mostra que os métodos de diagnóstico são muito mais prováveis de serem usados quando são convenientes.

Por exemplo, gráficos de variáveis adicionadas são construídos regredindo um regressor específico e a resposta em todos os outros regressores, computando os resíduos dessas regressões auxiliares e plotando um conjunto de resíduos contra o outro.

Isso não é difícil de fazer em R, embora as etapas sejam um pouco mais complicadas quando há fatores, interações ou termos polinomiais ou de spline de regressão no modelo.

A função **avPlots** no pacote **car** constrói todos os gráficos de variáveis adicionadas para um modelo linear ou um GLM e adiciona melhorias, como uma linha de mínimos quadrados e identificação de ponto.

Matriz chapéu

Os valores h_i, da matriz chapéu, para um modelo linear generalizado podem ser obtidos diretamente da iteração final do procedimento de mínimos quadrados ponderados interados para ajustar o modelo, e têm a interpretação usual - exceto que, ao contrário de um modelo linear, os valores h_i em um modelos linear generalizado dependem da variável de resposta Y, bem como na configuração dos X.

A matriz chapéu H é

$$H = W^{1/2}X(X^\top WX)^{-1}X^\top W^{1/2},$$

onde W é a matriz de peso da iteração final do procedimento de estimação.

As observações que estão relativamente longe do centro do espaço do regressor, levando em consideração o padrão correlacional entre os regressores, têm uma influência potencialmente maior nos coeficientes de regressão de mínimos quadrados; tais pontos são considerados como tendo alta alavancagem. A medida mais comum de alavancagem é o h_i ou hat-values.

O nome hat-values vem da relação entre o vetor de respostas observado e os valores ajustados. O vetor de valores ajustados é dado por $\widehat{y} = X\widehat{\beta} = Hy$, onde H, definida acima e chamada de matriz hat, projeta y, os valores observados da variável resposta Y, no subespaço estendido pelas colunas da matriz do modelo X. Como $H = H^\top H$, os valores hat h_i são simplesmente as entradas diagonais da matriz chapéu.

Os h_i são limitados entre 0 e 1; em modelos com um intercepto, eles são limitados entre 1/n e 1 e sua soma $\sum_i h_i$ é sempre igual ao número de coeficientes no modelo, incluindo o intercepto.

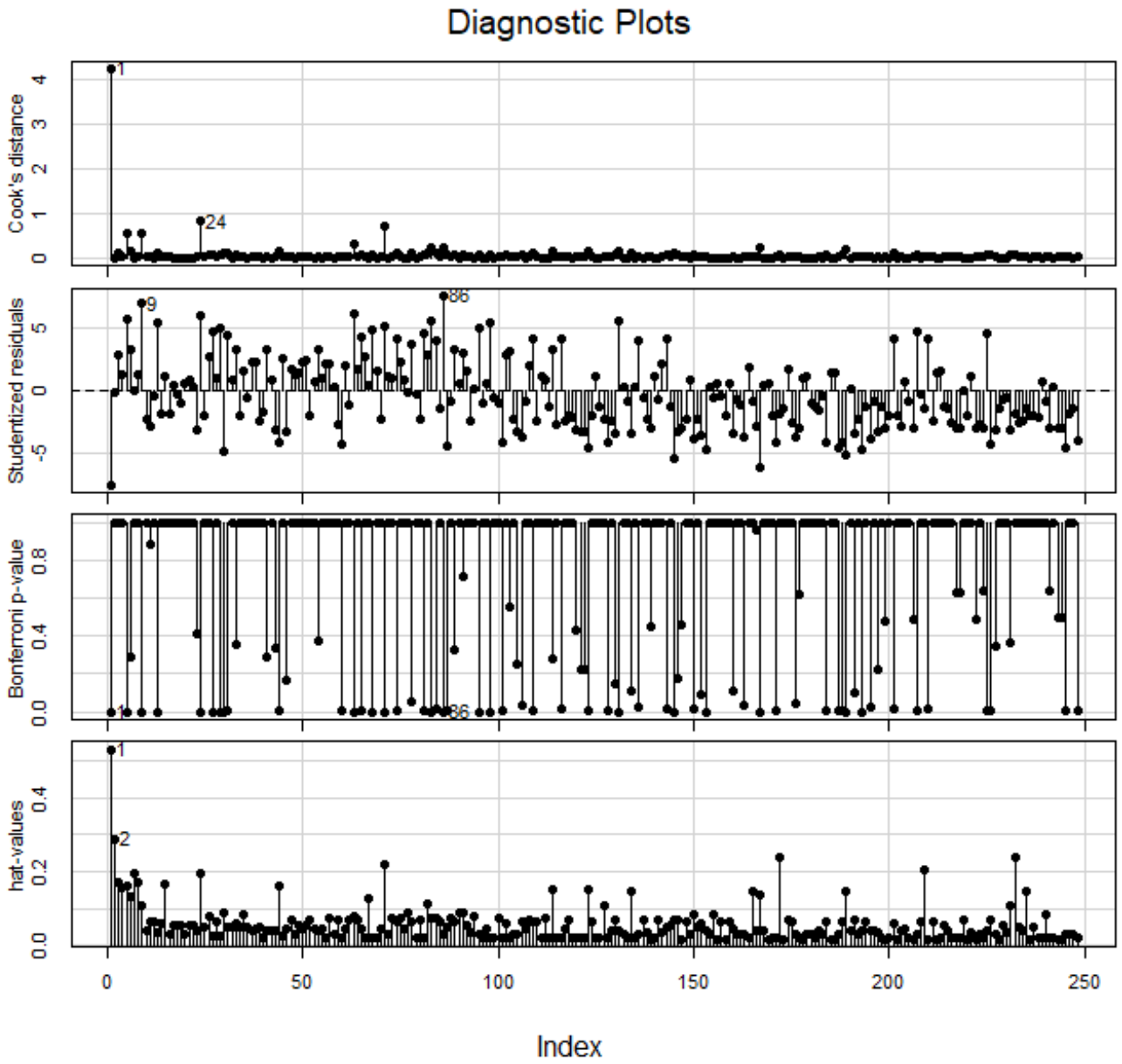
Situações nas quais há alguns h_i muito grandes podem ser problemáticas: em particular, a normalidade de grandes amostras de algumas combinações lineares dos regressores tende a falhar e as observações de alta alavancagem podem exercer influência indevida sobre os resultados.

A função **hatvalues** funciona para modelos lineares e modelos lineares generalizados. Uma maneira de examinar os valores de h_i e outras estatísticas de diagnóstico de observação individual é construir gráficos de índice, representando graficamente as estatísticas em comparação com os índices de observação correspondentes.

Por exemplo, a função **influenciaIndexPlot**, no pacote **car**, inclui gráficos de índice de resíduos estudentizados, os p-valores Bonferroni correspondentes para o teste de outlier, os hat-values e as distâncias de Cook para a regressão:

```
> library(car)
> influenceIndexPlot(ajuste)
```

Observe que as duas primeiras observações se destacam na distância de Cook e no hat-value.


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p-value
hat-values
Index
0
50
100
150
200
250
1
24
9
86
2

Vários tipos de resíduos podem ser definidos para os modelos lineares generalizados:

Mais diretamente, mas menos úteis, os **resíduos ordinários** são simplesmente as diferenças entre a resposta observada e seu valor esperado estimado: $e_i = y_i - \widehat{\mu}_i$, onde

$$\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i) = g^{-1}(\widehat{\alpha} + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_k X_{ik}).$$

Na regressão por mínimos quadrados ponderados, a soma residual dos quadrados é igual a $\sum_i e_i^2$. Caso o modelo de regressão inclua o intercepto, então $\sum_i e_i = 0$.

Os resíduos ordinários não estão correlacionados com os valores ajustados ou mesmo qualquer combinação linear dos regressores e, portanto, os padrões nos gráficos de resíduos ordinários versus combinações lineares dos regressores podem ocorrer apenas se uma ou mais suposições do modelo são inadequadas.

Se o modelo de regressão estiver correto, então os resíduos ordinários são variáveis aleatórias com média 0 e com variância dada por

$$\text{Var}(e_i) = \phi(1 - h_i)\cdot$$

A quantidade h_i é chamada de alavancagem ou hat-value. Em modelos lineares com preditores fixos, h_i é um valor não aleatório restrito a estar entre 0 e 1, dependendo da localização dos preditores para uma observação específica em relação às outras observações. Em um modelo com intercepto, o hat-value mínimo é 1/n. Valores grandes de h_i correspondem a observações com valores X_i relativamente incomuns, enquanto um pequeno valor h_i corresponde a observações próximas ao centro do espaço do regressor. Resíduos comuns para observações com grande h_i têm variâncias menores.

Para corrigir a variância não constante dos resíduos ordinários, podemos dividi-los por uma estimativa de seu desvio padrão.

Considerando que $\widehat{\phi}$ represente a estimativa de ϕ, os **resíduos padronizados** são

$$e_{i_{sd}} = \frac{e_i}{\widehat{\phi}\sqrt{1-h_i}},$$

Embora os $e_{i_{sd}}$ tenham variância constante, eles não são mais não correlacionados com os valores ajustados ou combinações lineares dos regressores, portanto, usar resíduos padronizados em gráficos não é uma melhoria óbvia.

Resíduos estudentizados são dados por

$$e_{i_T} = \frac{e_i}{\widehat{\phi}_{(-i)}\sqrt{1-h_i}},$$

onde $\widehat{\phi}_{(-i)}$ é a estimativa de ϕ calculada a partir da regressão sem a observação i. Assim como os resíduos padronizados, os resíduos estudentizados possuem variância constante. Além disso, se os erros originais são normalmente distribuídos, então e_{i_T} segue uma distribuição t com $n-k-2$ graus de liberdade e pode ser usado para testar outliers. Pode-se mostrar que

$$\widehat{\phi}_{(-i)} = \frac{\widehat{\phi}(n-k-1-e_{i_{sd}})}{n-k-2}$$

e, portanto, o cálculo dos resíduos estudentizados realmente não requer reajustar a regressão sem a observação i.

Resíduos de Pearson são componentes da estatística de qualidade de ajuste de Pearson para o modelo:

$$\frac{\widetilde{\phi}^{1/2}(y_i - \widehat{\mu}_i)}{\sqrt{\widehat{\text{Var}}(Y_i|\eta_i)}},$$

onde $\widetilde{\phi}$ é o parâmetro de dispersão estimado para o modelo e $\widehat{\text{Var}}(Y_i|\eta_i)$ é a variância condicional da resposta.

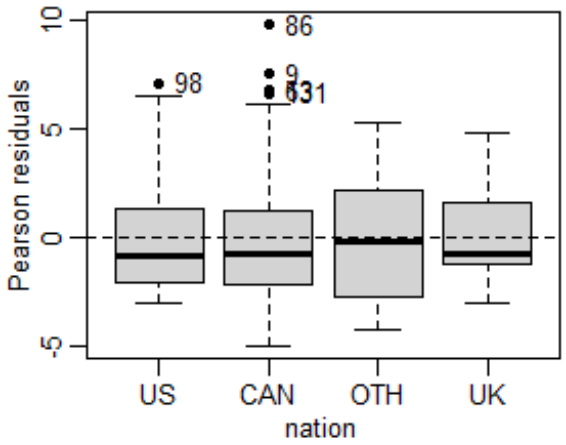
A estatística de Pearson, uma alternativa ao desvio para medir o ajuste do modelo aos dados, é a soma dos resíduos de Pearson ao quadrado.

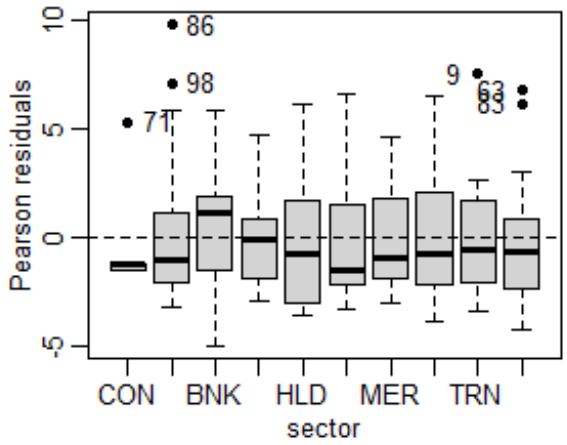
Outros resíduos são: Resíduos de trabalho, Resíduos deviance, Resíduos deviance padronizado, etc.

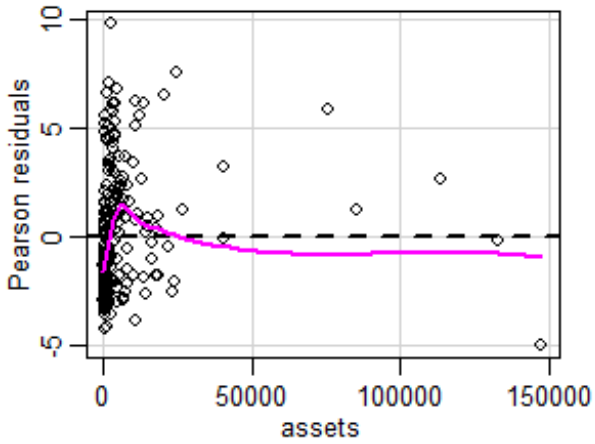
Os resíduos no podem ser encontrados utilizando a função genérica R **residuals** e podem calcular-se vários tipos de resíduos. O padrão para um modelo linear é retornar os resíduos ordinários, mesmo se houver pesos.

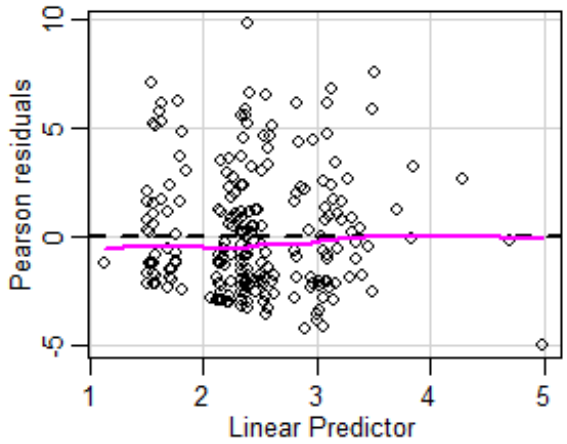
Definir o argumento **type = "pearson"**, retorna os resíduos de Pearson, que produzem resíduos corretamente ponderados se houverem pesos e resíduos ordinários se não houverem pesos. Resíduos de Pearson são o padrão quando os resíduos são usados com um modelo linear generallizado.

As funções **rstandard** e **rstudent** retornam os resíduos padronizados e estudentizados, respectivamente.


Pearson residuals
10
5
0
-5
86
98
9
63
1
US
CAN
OTH
UK
nation


Pearson residuals
10
5
0
-5
86
98
71
9
63
83
CON
BNK
HLD
MER
TRN
sector


Pearson residuals
10
5
0
-5
0
50000
100000
150000
assets


Pearson residuals
10
5
0
-5
1
2
3
4
5
Linear Predictor

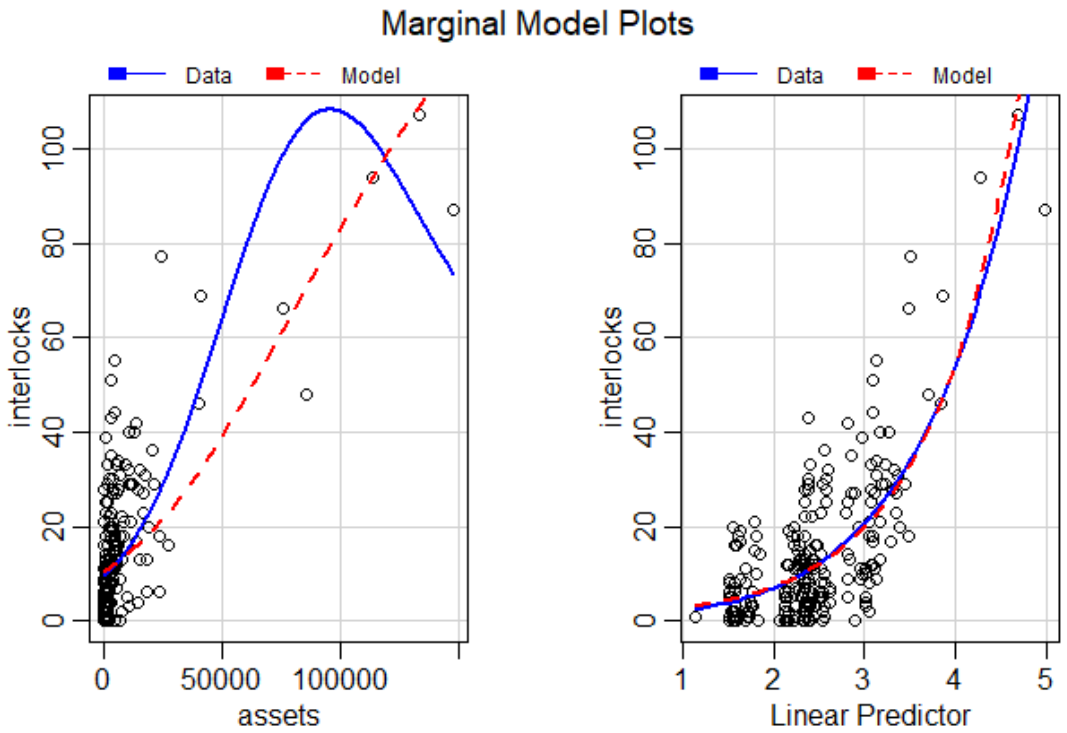
```
> dados = read.table("https://socialsciences.mcmaster.ca/jfox
      /Books/Applied-Regression-2E/datasets/Ornstein.txt",
                                                  header = T)
> dados$nation = relevel(factor(dados$nation), ref = "US")
> dados$sector = relevel(factor(dados$sector), ref = "CON")
> ajuste = glm(interlocks ~ nation + sector + assets,
      family = poisson, data = dados)
> library(car)
> par(mfrow=c(1,1), mar=c(3,2,1,0)+.5, mgp=c(1.6,.6,0), pch=19)
> residualPlots(ajuste)
       Test stat Pr(>|Test stat|)
nation
sector
assets    155.83        < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Gráficos de resíduos em relação aos valores ajustados e em relação a cada um dos preditores, por sua vez, são os gráficos de diagnóstico mais básicos. Se um modelo linear for especificado corretamente, os resíduos de Pearson são independentes dos valores ajustados e dos preditores, e esses gráficos devem ser gráficos nulos, sem características sistemáticas - no sentido de que a distribuição condicional dos resíduos, no eixo vertical do gráfico, não deve ser alterada com os valores ajustados ou com um preditor, no eixo horizontal.

A presença de características sistemáticas geralmente implica uma falha de uma ou mais suposições do modelo. De interesse nesses gráficos são as tendências não lineares, as tendências de variação no gráfico e os pontos isolados. A plotagem de resíduos em relação aos valores ajustados e preditores é útil para revelar problemas, mas menos útil para determinar a natureza exata do problema.

Uma variação no gráfico de resíduos básico é o gráfico do modelo marginal, proposto por Cook and Weisberg (1997):

```
> library(car)
> marginalModelPlots(ajuste)
```



Esses gráficos, mostrados acima, têm a variável de resposta, neste caso interlocks, no eixo vertical, enquanto o eixo horizontal é dado por sua vez por cada um dos preditores contínuos no modelo e os valores ajustados.

Os gráficos da resposta versus preditores individuais exibem a distribuição condicional da resposta dado cada preditor, ignorando os outros preditores; estes são gráficos marginais no sentido de que mostram a relação marginal entre a resposta e cada preditor contínuo. O gráfico em relação aos valores ajustados é um pouco diferente, pois exibe a distribuição condicional da resposta de acordo com o ajuste do modelo.

Podemos estimar uma função de regressão para cada um dos gráficos marginais ajustando uma suavização aos pontos do gráfico. A função **marginalModelPlots** usa uma suavização inferior, conforme mostrado pela linha sólida no gráfico.

Agora imagine um segundo gráfico que substitui o eixo vertical com os valores ajustados do modelo. Se o modelo for apropriado para os dados, então, sob condições bastante suaves, o ajuste suave para este segundo gráfico também deve estimar a esperança condicional da resposta dado o preditor no eixo horizontal.

A segunda suavização também é desenhada no gráfico do modelo marginal, como uma linha tracejada. Se o modelo se ajusta bem aos dados, então as duas suavizações devem corresponder em cada um dos gráficos do modelo marginal; se algum par de alisamentos não corresponder, então temos evidências de que o modelo não se ajusta bem aos dados.

Uma característica interessante dos gráficos do modelo marginal é que, embora o modelo que ajustamos aos dados especifique relações parciais lineares entre interlocks e assets, ele é capaz de reproduzir relações marginais não lineares para esse preditor.

Na verdade, o modelo, conforme representado pelas linhas tracejadas, faz um trabalho bastante bom em combinar as relações marginais representadas pelas linhas sólidas, embora as falhas sistemáticas descobertas nos gráficos de resíduos sejam discerníveis aqui.