

Modelos Lineares Generalizados

Métodos de diagnóstico

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

22 de junho de 2021

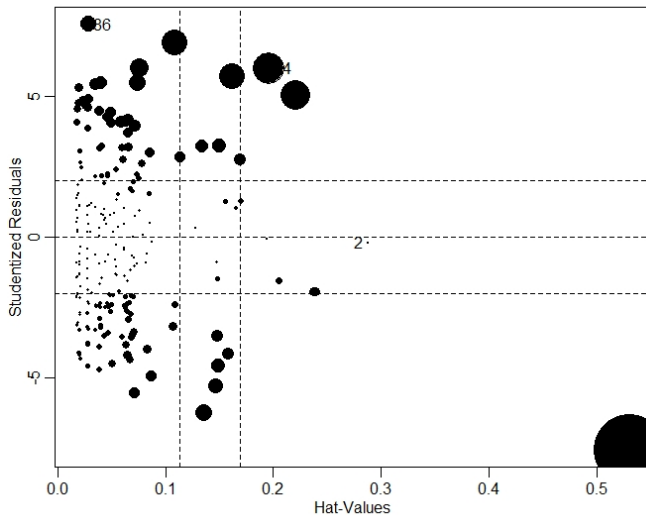
Uma aproximação da medida de influência da distância de Cook é

$$D_i = \frac{R_{P_i}^2}{\tilde{\phi}(k+1)} \times \frac{h_i}{1-h_i}.$$

Esta é essencialmente a definição de Williams, exceto que dividimos pela dispersão estimada $\tilde{\phi}$ para escalar D_i como uma estatística F em vez de uma estatística qui-quadrado.

R_{P_i} representa os **Resíduos de Pearson padronizados** corretos para a variância condicional da resposta e para a alavancagem diferencial das observações:

$$R_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}(Y_i|\eta_i)(1-h_i)}}.$$



A função **InfluencePlot** no pacote **car** fornece uma alternativa para o gráficos de índice de estatísticas de diagnóstico:

```
> influencePlot(ajuste)
```

	StudRes	Hat	CookD
1	-7.5660929	0.52990558	4.259721796
2	-0.1926656	0.28740852	0.001065426
24	5.9815914	0.19570665	0.817197475
86	7.5530132	0.02914628	0.210738087

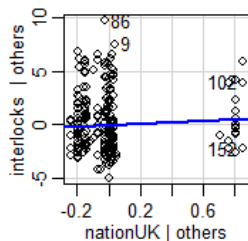
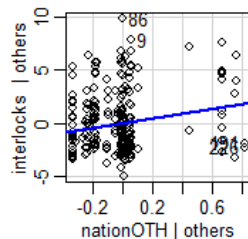
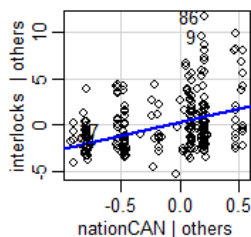
Este comando produz um gráfico de bolhas, mostrado na figura, que combina a exibição de resíduos estudentizados, valores de h_i e distâncias de Cook, com as áreas dos círculos proporcionais à distância de Cook.

Gráficos de variáveis adicionadas são um diagnóstico útil para encontrar pontos potencialmente influentes no conjunto, que corresponderão a conjuntos de pontos que estão fora de linha com o resto dos dados e estão no extremo esquerdo ou direito do eixo horizontal.

```
> avPlots(ajuste, terms = "nation")  
> avPlots(ajuste, terms = "sector")
```

Funciona da seguinte maneira: suponha que o regressor focal seja X_j . Reajuste o modelo com X_j removido, extraindo os resíduos de trabalho deste ajuste. Em seguida, regredir X_j nos outros X s por mínimos quadrados ponderados, usando os pesos da última etapa de mínimos quadrados ponderados iterados, obtendo os resíduos. Finalmente, plote os resíduos de trabalho da primeira regressão contra os resíduos de X_j da segunda regressão.

Added-Variable Plots



Em vez de resumir a influência observando todos os coeficientes simultaneamente, poderíamos criar $k + 1$ medidas de influência observando as diferenças individuais:

$$DFBETA_{ij} = \hat{\beta}_{(-i)j} - \hat{\beta}_j, \quad \text{para } j = 0, \dots, k,$$

onde $\hat{\beta}_j$ é o coeficiente calculado usando todos os dados e $\hat{\beta}_{(-i)j}$ é o mesmo coeficiente calculado com a observação i omitida.

Tal como acontece com D_i , o cálculo de $DFBETA_{ij}$ pode ser realizado de forma eficiente, sem a necessidade de reajustar o modelo. Os $DFBETA_{ij}$ são expressos na métrica, unidades de medida, do coeficiente $\hat{\beta}_j$.

Uma versão padronizada, $DFBETAS_{ij}$, divide $DFBETA_{ij}$ por uma estimativa do erro padrão de $\hat{\beta}_j$ calculado com a observação i removida.

A função **dfbeta** em R leva um modelo linear ou objeto GLM como seu argumento e retorna todos os valores de $DFBETA_{ij}$; da mesma forma, **dfbetas** calcula o $DFBETAS_{ij}$, como no exemplo a seguir para a regressão de intertravamentos:

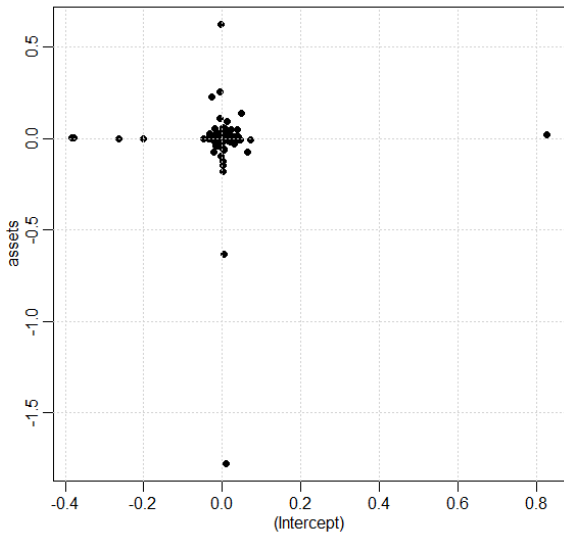
```
> dfbs.ajuste <- dfbetas(ajuste)
> head(dfbs.ajuste) # primeiras linhas
```

	(Intercept)	nationCAN	sectorBNK	sectorFIN
1	9.841257e-03	0.0784065595	0.5153892840	0.128312479
2	8.333564e-05	0.0006639458	0.0005769995	0.001086548
3	2.845862e-04	0.0022673347	0.1096039302	0.003710497

Poderíamos examinar cada coluna da matriz `dfbetas` separadamente, por exemplo, por meio de um gráfico de índices:

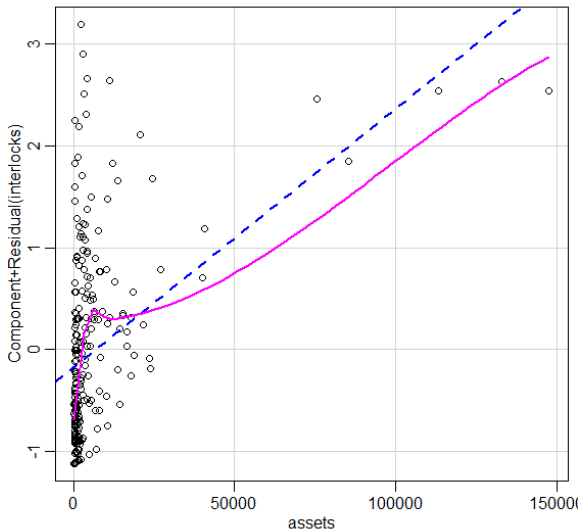
```
> par(mfrow=c(1,1), mar=c(3,2,1,0)+.5, mgp=c(1.6,.6,0), pch=19)
> plot(dfbs.ajuste[, c("(Intercept)","assets")])
> grid()
```

A falta de relação entre os valores $DFBETAS_{ij}$ para os dois regressores reflete a falta de correlação dos próprios regressores.



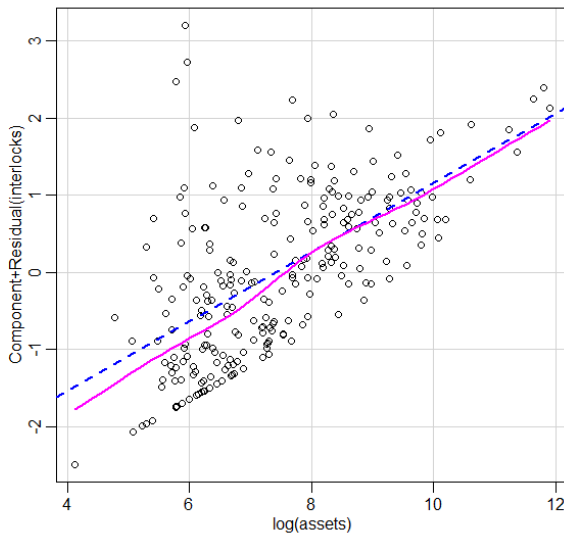
Os gráficos de componente mais residual e CERES também se estendem diretamente aos GLMs. A suavização não paramétrica dos diagramas de dispersão resultantes pode ser importante para a interpretação, especialmente em modelos para variáveis de resposta binárias, onde a descrição da resposta torna os diagramas difíceis de examinar. Efeitos semelhantes (se normalmente menos extremos) podem ocorrer para dados binomiais e de contagem.

Componente mais residual e CERES traçam um modelo linearizado da última etapa do ajuste IWLS. Por exemplo, o resíduo parcial para X_j adiciona o resíduo de trabalho $\hat{\beta}_j X_{ij}$ o componente mais residual o gráfico então representa o resíduo parcial em relação a X_j . Ao suavizar um gráfico de componente mais residual para um GLM não gaussiano, geralmente é preferível usar um suavizador não robusto.



```
> crPlots(ajuste, "assets")
```

Um gráfico de componente mais residual para ativos na regressão quase Poisson para os dados da diretoria interligados é mostrado na figura acima. Os ativos são tão inclinados positivamente que o gráfico é diferente de examinar, mas é, no entanto, aparente que a relação parcial entre o número de intertravamentos e ativos é não linear, com uma inclinação muito mais acentuada à esquerda do que à direita. Como a protuberância aponta para a esquerda, podemos tentar endireitar essa relação transformando os ativos na escada de potências e de raízes. Tentativas e erro sugerem a transformação logaritmo de ativos, após a qual um gráfico de componente mais residual para o modelo modificado (Figura abaixo) não é digno de nota.



```
> ajuste1 = update(ajuste, . ~ . - assets + log(assets),
                    family = quasipoisson)

> summary(ajuste1)

.....
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.83938      0.34563  -2.429  0.01591 *
.....
sectorWOD      0.49825      0.19124   2.605  0.00977 **
log(assets)    0.45145      0.04296  10.510  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 6.398663)

.....

Number of Fisher Scoring iterations: 5

> crPlots(ajuste1, "log(assets)")
```