# CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

11 de julho, 2018

# Aula 10 - Diagnóstico do ajuste de MLGs

# Diagnóstico do ajuste

- A análise de diagnóstico (ou diagnóstico do ajuste) configura uma etapa fundamental no ajuste de modelos de regressão.

- O objetivo principal dessa etapa da análise é a avaliação do modelo ajustado. No caso de MLGs, baseia-se, dentre outros, na verificação dos seguintes itens:
    - Avaliação da distribuição proposta;
    - Avaliação da parte sistemática do modelo;
    - Adequação da função de ligação.
    - Identificação e avaliação do efeito de observações mal ajustadas;
    - Identificação de pontos influentes.

# Diagnóstico do ajuste

- Boa parte dos métodos de diagnóstico em MLGs configuram extensões dos procedimentos utilizados em regressão linear.

- O uso de simulação no diagnóstico de MLGs (por exemplo na obtenção de qq-plots com envelopes simulados) é importante.

- O principal componente no diagnóstico de MLGs é, novamente, a análise de resíduos.

# Resíduo de Pearson

- O resíduo de Pearson é definido por:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}. \tag{1}$$

- Para um MLG Poisson, o resíduo de Pearson fica definido por:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{2}$$

- Já para um MLG binomial:

$$e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i}}. \tag{3}$$

# Resíduo de Pearson

- O resíduo de Pearson tem uma versão padronizada, com média 0 e variância aproximadamente 1, definida por:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V(\hat{\mu}_i)(1 - \hat{h}_{ii})}}, \tag{4}$$

em que $\hat{h}_{ii}$ é o $i$-ésimo elemento da diagonal da matriz

$$\hat{H} = W^{1/2} X (X' W X)^{-1} X' W^{1/2}, \tag{5}$$

que é a matriz de projeção do algoritmo de estimação dos MLGs.

# Resíduo componente da deviance

- Resgatando a definição da deviance:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} 2\omega_i \left[ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right], \tag{6}$$

o resíduo componente da deviance fica definido pela contribuição de cada observação para a deviance do modelo:

$$d_i = sinal(y_i - \hat{\mu}_i) \times \sqrt{2\omega_i \left[ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right]} \tag{7}$$

- Uma versão padronizada do resíduo componente da deviance é dada por:

$$d_i^* = \frac{d_i}{\sqrt{\hat{\phi}(1 - \hat{h}_{ii})}}. \tag{8}$$

# Análise de resíduos

- Os resíduos de Pearson e componente da deviance geralmente não tem boa aproximação com a distribuição Normal, ainda que o modelo ajustado esteja correto.

- A avaliação da qualidade do ajuste baseada em gráficos probabilísticos normais (ou meio-normais), para esses tipos de resíduos, requer simulação (envelopes simulados). Veremos adiante.

- Um tipo de resíduo que, por construção, tem distribuição normal caso o modelo ajustado esteja correto, é o **resíduo quantílico aleatorizado**.

# Resíduo quantílico aleatorizado (Dunn, 1997)

- O resíduo quantílico aleatorizado baseia-se no método método da transformação integral da probabilidade.
- Seja $y_i$ uma variável aleatória contínua com FDA $F(y_i; \mu_i, \phi)$. O método da transformação integral da probabilidade baseia-se no seguinte resultado:

$$u_i = F(y_i; \mu_i, \phi) \sim U(0, 1). \tag{9}$$

- Adicionalmente, considerando que $u_i$ tem distribuição uniforme entre 0 e 1, temos que:

$$\Phi^{-1}(F(y_i; \mu_i, \phi)) \sim N(0, 1), \tag{10}$$

resultado bastante utilizado para simular dados de uma distribuição Normal padrão.

# Resíduo quantílico aleatorizado

- Assim, o resíduo quantílico aleatorizado fica definido por:

$$q_i = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi})), \tag{11}$$

tal que, se o modelo tiver corretamente especificado, tem distribuição Normal(0,1).

# Resíduo quantílico aleatorizado

- Se a variável $y_i$ for discreta, então $F(y_i; \mu_i, \phi)$ é uma função discreta, com 'saltos' em cada valor de $y_i$ com probabilidade não nula.

- Neste caso, consideramos a seguinte adaptação:

$$F^*(y_i; \hat{\mu}_i, \hat{\phi}) = F(y_i-; \hat{\mu}_i, \hat{\phi}) + u_i f(y_i; \hat{\mu}_i, \hat{\phi}), \tag{12}$$

em que $F(y_i-; \hat{\mu}_i, \hat{\phi})$ é o limite de $F(y-; \hat{\mu}_i, \hat{\phi})$ pela esquerda, $u_i$ é um valor aleatório da distribuição $U(0,1)$ e $f(y_i; \hat{\mu}_i, \hat{\phi})$ é a massa de probabilidade em $y_i$.

# Análise gráfica de resíduos

- **Resíduos vs valores ajustados**- Para um modelo bem ajustado, deve-se observar a dispersão aleatória dos pontos, centrada em zero, com média e variância constantes e sem valores extremos.

**Nota:** É recomendável plotar os resíduos padronizados, e os valores ajustados na escala do preditor.

- **Resíduos vs variáveis incluídas no modelo:** Padrões não aleatórios indicam que a variável não está bem acomodada no modelo;
- **Resíduos vs variáveis não incluídas no modelo:** Padrões não aleatórios sinalizam a necessidade (e a forma) de inclusão da variável no modelo;
- **Gráfico de resíduos versus ordem de coleta dos dados** - Padrões não aleatórios indicam a dependência das observações gerada pela ordem de coleta (no tempo, no espaço,...).

# Análise gráfica de resíduos

- **Gráfico da variável ajustada versus preditor linear** - Plotando $z_i$ vs $\hat{\eta}_i$ podemos avaliar a adequação da função de ligação.

- Uma forma alternativa de testar a função de ligação é a seguinte:

1. Ajusta-se o modelo extrai-se $\hat{\eta}_i$;
2. Ajusta-se novamente o modelo incorporando $\hat{\eta}_i^2$ como uma nova covariável;
3. Se o efeito de $\hat{\eta}_i^2$ for significativo, então a função de ligação não é adequada.

# Gráficos meio normais com envelopes simulados

- Os gráficos meio-normais consistem na plotagem de alguma medida de diagnóstico (resíduos, distância de Cook, leverage) versus a esperança das estatísticas de ordem da distribuição meio-normal:

$$\Phi^{-1}\left(\frac{i + 1 - \frac{1}{8}}{2n + \frac{1}{2}}\right), i = 1, 2, ..., n. \tag{13}$$

- Em modelos lineares generalizados, a distribuição dos resíduos (Pearson, deviance) e das medidas de influência, dentre outros, em geral não é normal, o que pode prejudicar a avaliação dos gráficos meio-normais.
- A solução é usar simulação para poder avaliar adequadamente a disposição dos pontos em um gráfico meio-normal (envelopes simulados).

# Obtenção dos envelopes simulados para gráficos meio-normais (Moral, 2013)

1. Obter $d_{(i)}$, os valores de uma quantidade diagnóstica em valor absoluto e em ordem crescente;
2. Simular 99 amostras do modelo ajustado com os mesmos valores para as variáveis explanatórias;
3. Fazer o ajuste do modelo para as 99 amostras e, para cada ajuste, obter a quantidade diagnóstica de interesse, $d^*_{j(i)}$, $j = 1, 2, ..., 99$, em valor absoluto e em ordem crescente;
4. Para cada $i$ computar os percentis 5%, 50% e 95%;
5. Fazer o gráfico desses percentis e dos $d_{(i)}$'s observados contra as estatísticas de ordem da distribuição meio-normal.

# Gráficos meio normais com envelopes simulados

- Se a maior parte dos valores observados estiver contida no envelope simulado, há indícios de que o modelo está bem ajustado aos dados.

# Diagnóstico de influência

- Assim como no caso de modelos lineares, também para MLGs o diagnóstico de influência é útil para identificar pontos que exercem grande influência sobre o ajuste do modelo.

- A estratégia para diagnóstico de influência, novamente, é do tipo *leave-one-out*, em que se avalia o quanto resultados dos modelos (estimativas dos coeficientes, erros padrões,. . . ) mudam ao desconsiderar uma particular observação $i$, $i = 1, 2, ..., n$;

- Assim como no caso dos modelos lineares, não há a necessidade de ajustar o mesmo modelo $n$ vezes (uma para a deleção de cada observação), dispondo-se de aproximações adequadas para as medidas de influência.

# Diagnóstico de influência

- Dentre as principais medidas que fazem uso da estratégia *leave one out*, temos:
  - Resíduos studentizados;
  - DFBetas: para avaliar a mudança em coeficientes individualmente;
  - Distância de Cook: para avaliar a mudança global no ajuste do modelo.

- Para qualquer medida de influência calculada, um gráfico dos valores calculados vs índice da observação é importante para avaliação comparativa dos resultados e identificação de valores extremos.

- Gráficos meio-normais com envelopes simulados podem ser bastante apropriados para checagem de observações influentes.

# Diagnóstico de influência

- Ao detectar observações influentes ou outliers, o seguinte procedimento é recomendado:
    - Voltar à base de dados e identificar as correspondentes observações. Buscar compreender o motivo da detecção;
    - Se verificada algum erro (coleta, digitação,...) nessas observações, corrigí-las. Se não for possível, eliminá-las;
    - Se não houver erros, deve-se avaliar o impacto dessas observações no ajuste. Ajuste novos modelos eliminando-as (conjuntamente, uma a uma...) e compare os principais resultados do modelo;
    - Se alguma alteração mais significativa nos resultados for verificada ao desconsiderar tais observações, isso deverá ser reportado em seu relatório de análise.

# Multicolinearidade

- A multicolinearidade se caracteriza por uma (quase) dependência linear entre as colunas de $\boldsymbol{X}$.

- Na presença de multicolinearidade, as estimativas produzidas são bastante instáveis frente a pequenas mudanças nos dados;

- Como resultado, as conclusões produzidas pelo modelo ficam seriamente comprometidas na presença de multicolinearidade.

# Multicolinearidade

- Como o problema da multicolinearidade remete apenas à matriz do modelo ($\boldsymbol{X}$), as mesmas técnicas estudadas para os modelos lineares se aplicam diretamente aqui;

- Uma avaliação preliminar pode ser feita analisando a matriz de correlações de $\boldsymbol{X}$;

- Uma verificação mais formal baseia-se no cálculo do VIF (*variance inflation factor*), definido por:

$$VIF = \frac{1}{1 - R_j^2}, \tag{14}$$

em que $R_j^2$ é o coeficiente de determinação de $X_j$ com relação ás demais covariáveis.

- VIF acima de 5 ou 10 pode ser considerado um indicador de multicolinearidade.