Data Science and Big Data

Taconeli, C.A.

05 de outubro, 2018

Estimação de percentis e curvas centílicas

Introdução

- A metodologia GAMLSS tem se consolidado como referência para a construção de **curvas centílicas**;

- Curvas centílicas são usadas, por exemplo, para representar o padrão de diversas medidas antropométricas (peso, altura, massa muscular, massa gorda...) em humanos em função da idade;

- Nesses casos, duas variáveis contínuas são consideradas:

  * A variável resposta, usualmente uma medida antropométrica;

  * A variável explanatória, normalmente a idade.

Curvas centílicas

- O centil $100p$ de uma variável aleatória Y é o valor y_p tal que $P(Y \leq y_p) = p$;
- Assim, $y_p = F^{-1}(p)$ é a inversa da função distribuição acumulada de Y avaliada em p;
- Na construção de curvas centílicas consideramos os centis de Y condicionais ao valor de uma variável explanatória $X = x$, representados por $y_p(x) = F^{-1}_{Y|x}(p)$;
- Variando o valor de x, obtemos uma sequência de valores para $y_p(x)$ que permitirão a construção da curva centílica $100p$.

Curvas centílicas

- A Organização Mundial de Saúde utiliza as seguintes curvas centílicas como referência para padronização de medidas antropométricas:

  * $100p = (3, 15, 50, 85, 97)$, em seus gráficos;

  * $100p = (1, 3, 5, 15, 25, 50, 75, 85, 95, 97, 99)$, em suas tabelas.

- Como alternativa, em alguns casos são usados os *escores-z* (ou escores normalizados) de y, definidos por $z = \Phi^{-1}\left[F_{Y|x}(y)\right]$, onde Φ^{-1} é a inversa da distribuição normal padrão;

- No caso de GAMLSS, os escores normalizados são simplesmente os resíduos quantílicos.

Curvas centílicas

- A metodologia GAMLSS proporciona modelar a variação da medida antropométrica de interesse em função da idade acomodando relações não lineares e dispersão e forma (assimetria e curtose) da distribuição variando conforme a idade, dentre outros.

- Rigby e Stasinopoulos (2004, 2006) estabelecem a construção de curvas centílicas como:

$$
\begin{aligned}
Y &\sim D(\mu, \sigma, \nu, \tau) \\
g_1(\mu) &= s_1(u) \\
g_2(\sigma) &= s_2(u) \\
g_3(\nu) &= s_3(u) \\
g_4(\tau) &= s_4(u) \\
u &= x^{\epsilon}
\end{aligned}
\quad .
$$

Curvas centílicas

- Como distribuição para Y (D) consideramos BCCG, BCPE e BCT;
- As funções de ligação $g(\cdot)$ são escolhidas de maneira apropriada, conforme o espaço paramétrico;
- As funções $s(\cdot)$ são suavizadores não paramétricos;
- O termo ϵ é uma potência aplicada à variável explanatória (x).

Seleção de modelos na construção de curvas centílicas

- A seleção do modelo para construção de curvas centílicas requer a especificação da distribuição (D), da suavização aplicada em cada parâmetro (definida pelos respectivos graus de liberdade, $df_\mu, df_\sigma, df_\nu, df_\tau$), e o parâmetro de potência (ϵ);

- Diferentes procedimentos foram propostos para seleção de modelos, baseados, dentre outros, na minimização de GAIC ou na maximização da verossimilhança.

- A função `lms()` do pacote `gamlss` implementa a seleção de modelos e ajuste de curvas centílicas;

- Stasinopoulos *et al* (2017) apresentam com mais detalhes a metodologia de curvas centílicas.

Predição

- No contexto de curvas centílicas, as seguintes formas de predição são possíveis:

1. Para dados novos valores de x e porcentagens correspondentes aos centis (p), produzir uma matriz contendo os centis correspondentes de y;

2. Para dados novos valores de x e valores para os centis normalizados (escores z), produzir uma matriz com os centis correspondentes para y;

3. Para pares de valores de x e de y, calcular os correspondentes escores z.