Análise de resíduos nos MLG (Modelos lineares generlizados)

Prof. Caio Azevedo

(grande parte do material apresentado foi extraído do livro Modelos de regressão com apoio computacional do Prof. Gilberto A. Paula: link)

Motivação

- Podemos perceber que o resíduo studentizado (aqui), muito utilizado para verificar a qualidade de ajuste da classe de MRNLH, dificilmente apresentará normalidade assintótica (embora possa ser usado para verificar a presença de outliers e/ou em problemas na predição dos valores) para os MLG (pois corresponde à uma transformação linear dos dados).
- Portanto, a utilização de outro(s) resíduo(s) se faz necessária.
- Paula (2024) apresenta uma revisão muito boa sobre vários resíduos.
- Aqui (e referências mencionadas) também apresenta(m) alguns resíduos. .

link

Motivação

- Nos concentraremos em três tipos de resíduos:
  - Resíduo componente do desvio (RCD).
  - Resíduos quantílico (RC).
  - Resíduo quantílicos aleatorizado (RCA).
- Um fato interessante é que, na definição dos MLG, não aparece nenhum tipo de "erro", como ocorre nos MRNLH.
- Uma discussão interessante encontra-se em Cox and Snell (1968)

Introdução

- Um resíduo deve apresentar um comportamento específico quando o modelo está bem ajustado e outro(s) quando o modelo não o estiver.
- O ideal é que, dependendo de qual suposição (ou suposições, p.e., distribuição da variável resposta, independência, função de ligação e forma do preditor linear) não esteja(m) sendo satisfeita(s), alguma mudança específica ocorra em seu comportamento (conforme discutido anteriormente).
- Naturalmente, outras metodologias, para além dos resíduos, podem ser utilizadas para verificar o afastamento de suposições específicas.

RCD

- A forma geral do RCD, para a i-ésima observação, é dada por:

$$T_{D_i} = \frac{d^*(Y_i, \widehat{\mu}_i)}{\sqrt{1 - \widehat{h}_{ii}}} = \frac{\phi^{1/2} d(Y_i, \widehat{\mu}_i)}{\sqrt{1 - \widehat{h}_{ii}}}$$

em que

  - $d(Y_i, \widehat{\mu}_i) = \text{sinal}\,(Y_i - \widehat{\mu}_i)\sqrt{2}\sqrt{D(Y_i; \widehat{\mu}_i)}$.
  - $D(Y_i; \widehat{\mu}_i) = Y_i\left(\widehat{\theta}_i^{(0)} - \widehat{\theta}_i\right) + b(\widehat{\theta}_i) - b(\widehat{\theta}_i^{(0)})$ (em que $\widehat{\theta}_i^{(0)}$ representa o emv sob o modelo saturado e $\widehat{\theta}_i$ o respectivo emv sob o modelo de regressão, ver aula sobre o Desvio).
  - $\widehat{h}_{ii}$ é o i-ésimo elemento da diagonal principal da matriz $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{1/2}$, em que $\boldsymbol{X}$ e $\widehat{\boldsymbol{W}}$ são como definidas na parte de estimação.

Cont.

- Williams (1984) (veja o nome da referência completa aqui) verificou através de simulações que a distribuição de t_{D_i} tende a estar mais próxima da normalidade do que as distribuições de outros resíduos (existentes à época, veja também Paula (2024)).
- Utilizando resultados de Cox and Snell (1968), pode-se demonstrar que $\mathcal{E}(D^*(Y_i, \mu_i)) \approx 0$ e $\mathcal{V}(D^*(Y_i, \mu_i)) \approx 1 - h_{ii}$ em que os termos negligenciados são $O(n^{-1})$. Esses resultados reforçam a padronização do RCD por $\sqrt{1 - \widehat{h}_{ii}}$.

link

Cont.

- Obs: Sejam f e g duas funções definidas no mesmo subconjunto dos números reais temos que:

$$f(x) = O(g(x)), x \to \infty \leftrightarrow |f(x)| \leq \ M|g(x)| \forall x \geq x_0, \text{e algum } M > 0$$

- Na prática substituímos ϕ por um estimador consistente (emv, por exemplo).
- A estimativa do RCD é obtida substituindo-se os estimadores nele presentes por suas respectivas estimativas, bem como Y_i pelos valores observados y_i.

Exemplos

- Normal:

$$T_{D_i} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{\sigma}^2(1-\widehat{h}_{ii})}}.$$

- gama:

$$T_{D_i} = \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2\widehat{\phi}}}{\sqrt{1-\widehat{h}_{ii}}}\left[-\ln\left(\frac{Y_i}{\widehat{\mu}_i}\right) + \left(\frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right)\right]^{1/2}.$$

- Bernoulli:

$$T_{D_i} = -\frac{2|\ln(1-\widehat{\mu}_i)|^{1/2}}{\sqrt{1-\widehat{h}_{ii}}}\mathbb{1}_{\{0\}}(Y_i) + \frac{2|\ln(\widehat{\mu}_i)|^{1/2}}{\sqrt{1-\widehat{h}_{ii}}}\mathbb{1}_{\{1\}}(Y_i).$$

link

Cont.

- binomial:

$$
\begin{aligned}
T_{D_i} &= \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2}}{\sqrt{1-\widehat{h}_{ii}}}\Big\{ Y_i \ln[Y_i/(m_i\widehat{\mu}_i)] \\
&+ (m_i - Y_i)\ln\left[(1 - Y_i/m_i)/(1-\widehat{\mu}_i)\right] \times \mathbb{1}_{\{1,\ldots,(m_i-1)\}}(Y_i) \\
&- [m_i|\ln(1-\widehat{\mu}_i)|]\mathbb{1}_{\{0\}}(Y_i) - [m_i|\ln\widehat{\mu}_i|]\mathbb{1}_{\{m_i\}}(Y_i)\Big\}^{1/2}.
\end{aligned}
$$

- Poisson:

$$
\begin{aligned}
T_{D_i} &= \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2}}{\sqrt{1-\widehat{h}_{ii}}}\left\{Y_i \ln(Y_i/\widehat{\mu}_i) - (Y_i - \widehat{\mu}_i)\right\}^{1/2} I_{\{1,2,\ldots.\}}(Y_i) \\
&\quad \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2\widehat{\mu}_i}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{0\}}(Y_i).
\end{aligned}
$$

link

Comentários sobre o RCD

- Pode acontecer de que o modelo esteja bem ajustado e, mesmo assim, a distribuição do RCD não ser aproximadamente normal.
- Ainda assim podemos construir uma gráfico de quantil quantil com envelopes simulando a partir do modelo de interesse ao invés da distribuição normal.

Procedimento para se gerar o gráfico de envelopes com o RCD

1) Ajuste o modelo de regressão (estima-se os parâmetros do modelo) obtendo-se as estimativas de MV $(\widetilde{\boldsymbol{\beta}}, \widetilde{\phi})$ e calcule o RCD para cada observação, (t_{D_i}), $i = 1, 2, ..., n$.

2) De posse das estimativas de MV, repita os passos (a) e (b) m vezes.

   a) Simule n variáveis aleatórias ind. $FE(\widetilde{\theta}_i, \widetilde{\phi})$, com $\widetilde{\theta}_i = h(g^{-1}(\widetilde{\eta}_i))$, $\widetilde{\boldsymbol{\eta}}_i = \boldsymbol{X}_i' \widetilde{\boldsymbol{\beta}}$.

   b) Ajuste o modelo de regressão considerando as variáveis simuladas no item a) e obtenha o RCD para cada observação (i) em cada réplica (j).

Procedimento para se gerar o gráfico de envelopes com o RCD

3) Ao final teremos uma matriz com os RCD's, ou seja $t^*_{D_{ij}}$, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\boldsymbol{T}_1 = \begin{bmatrix} t^*_{D_{11}} & t^*_{D_{12}} & \cdots & t^*_{D_{1m}} \\ t^*_{D_{21}} & t^*_{D_{22}} & \cdots & t^*_{D_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D_{n1}} & t^*_{D_{n2}} & \cdots & t^*_{D_{nm}} \end{bmatrix}$$

Procedimento para se gerar o gráfico de envelopes com o RCD

4) Dentro de cada amostra, ordena-se, de modo crescente, os RCD's, obtendo-se $t^*_{D_{(i)j}}$ (estatísticas de ordem):

$$\boldsymbol{T}_2 = \begin{bmatrix} t^*_{D_{(1)1}} & t^*_{D_{(1)2}} & \cdots & t^*_{D_{(1)m}} \\ t^*_{D_{(2)1}} & t^*_{D_{(2)2}} & \cdots & t^*_{D_{(2)m}} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D_{(n)1}} & t^*_{D_{(n)2}} & \cdots & t^*_{D_{(n)m}} \end{bmatrix}$$

5) Obtem-se também os limites $t^*_{(i)I} = \min\limits_{1\leq j\leq m} t^*_{D_{(i)j}}$ e $t^*_{(i)S} = \max\limits_{1\leq j\leq m} t^*_{D_{(i)j}}$, $j = 1, 2, ..., m$.

Procedimento para se gerar o gráfico de envelopes com o RCD

5) Na prática considera-se $t^*_{(i)I} = \dfrac{t^*_{D_{(i)(2)}} + t^*_{D_{(i)(3)}}}{2}$ e $t^*_{(i)S} = \dfrac{t^*_{D_{(i)(m-2)}} + t^*_{D_{i(m-1)}}}{2}$ (refinamento das estimativas das bandas de confiança), em que $t^*_{D_{(i)(r)}}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ..., n$.

- Além disso, consideramos como linha de referência $t^*_{(i)} = \dfrac{1}{m}\sum_{j=1}^{m} t^*_{(i)j}, i = 1, 2, ..., n.$

link

Outros gráficos de interesse

- boxplot/histograma: simetria, outliers, multimodalidade (histograma).
- $t_{D_i} \times$ ordem da observação: pontos aberrantes, heterogeneidade (heterocedasticidade) não capturada pelo modelo.
- $t_{D_i} \times g^{-1}(\widetilde{\eta}_i)$(valor predito): pontos aberrantes.
- $\widetilde{z}_i \times \widetilde{\eta}_i$: adequabilidade da função de ligação e do preditor linear (η_i), em que $\widetilde{z}_i = \widetilde{\eta}_i + \widetilde{W}_i^{-1/2}\widetilde{V}_i^{-1/2}(y_i - \widetilde{\mu}_i)$, em que $\widetilde{(.)}$ representa uma estimativa.
- $\widehat{h}_{ii} \times \widehat{\mu}_i$ (pontos alavanca - aqueles que tem um peso desproporcional no próprio valor ajustado, devido à ter um perfil, em termos das covariáveis, diferente dos demais).

Resíduo quantílico

- Além de poderem servir para diferentes propósitos, determinados tipo de resíduos podem ter melhor desempenho do que outros, consoante a situação de interesse.
- Nas referências a seguir podem ser encontrados uma amostra de estudos, críticas, comparação e sugestões de melhoria para alguns tipos de resíduos usados em MLG: aqui, aqui, aqui, aqui, aqui.
- Veremos agora uma alternativa ao RCD chamo de Resíduo Quantílico, o qual é apropriado quanto a resposta corresponde à uma vac (variável aleatória contínua).

Cont.

- É baseado no teorema da Transformada Integral (veja aqui também).
- Teorema da transformada integral: Se X uma vac tal que sua fda (função distribuição acumulada, $F_X(.)$) seja estritamente crescente. Então $Y = F_X(X) \sim U(0,1)$.
- Dessa forma,se $Z \sim F_Z$, em que Z é uma vac com F_Z estritamente crescentes, então

$$Z = F_Z^{-1}\left(F_X(X)\right) \sim F_Z.$$

Cont.

- Se $Y_i \sim FE(\theta_i, \phi), i = 1, ..., n$ com Y_i sendo uma vac com fda (F_{Y_i}) estritamente contínua e $Z \sim N(0,1)$ com fda $F_Z \equiv \Phi$, então:

$$Z = F_Z^{-1}\left(F_{Y_i}\left(Y_i; \theta_i, \phi\right)\right) \sim N(0,1).$$

- Uma vez que $\theta_i, i = 1, 2, ..., n$ e ϕ são desconhecidos, temos que substituí-los por algum estimador consistente $(\widehat{\theta}_i, \widehat{\phi})$, assim, o resíduo quantílico para a observação RQ_i é dado por:

$$RQ_i = F_Z^{-1}\left(F_{Y_i}\left(Y_i; \widehat{\theta}_i, \widehat{\phi}\right)\right)$$

de sorte que $RQ_i \xrightarrow[n\to\infty]{D} N(0,1)$

Cont.

- No caso do RQ pode-se fazer envelopes parecido com o que fora feito par o MRNLH, uma vez que $RQ_i \approx N(0,1)$.
- Os outros gráficos podem ser utilizados de forma simular ao RCD.
- Entretanto, o resultado só vale se Y_i for uma vac com fda estritamente crescentes.
- No caso de uma vad (variável aleatória discreta) pode-se considerar, como alternativa o Resíduo Quantílico Aleatorizado.

link

Resíduo quantílico aleatorizado (RQA)

- Sejam X e Y vac's tais que F e G são, respectivamente, suas fda's. Temos então que $U = F(X) \sim U(0,1)$ (transformação integral de probabilidade) e $W = G^{-1}(U) \sim G$, ou seja W e Y possuem a mesma distribuição.
- Se $G \equiv \Phi(.)$ (fda da normal padrão) então $W \sim N(0,1)$.
- Logo, se x for um valor simulado de X então $u = F(x)$ corresponderá à uma valor simulado da distribuição U(0,1) e, consequentemente $w = G(u)$ à um valor da N(0,1).

Resíduo quantílico aleatorizado (RQA)

- Esse resultado é muito útil para construir resíduos com distribuição (ainda que aproximada) N(0,1), sob o ajuste adequado do modelo, quando a variável resposta é contínua.
- Contudo, em nosso caso, a variável resposta é discreta.
- Ocorre que se F_X for contínua $\forall q \in (0,1), \exists\, x_q, x_q = F_X^{-1}(q)$, o que não, necessariamente, ocorre quanto F_X é discreta.
- No caso discreto (sendo $y_1 < y_2 < y_3, \ldots$ o suporte de Y), temos que $P(Y = y_i) = F(y_i) - F(y_{i-1}) = P(F(y_{i-1}) \leq U \leq F(y_i))$, em que $U \sim U(0,1)$.

link

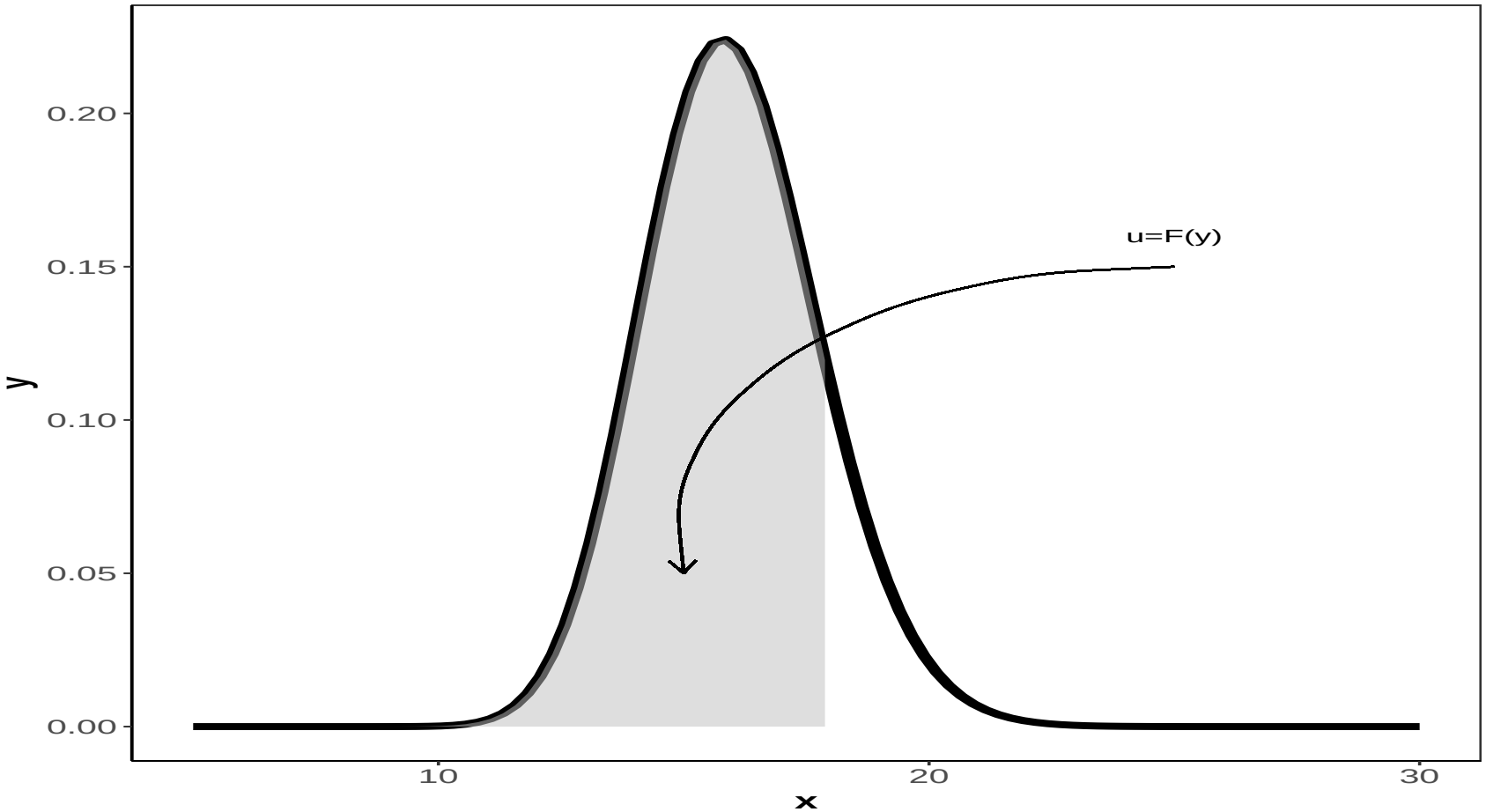
Resíduo quantílico aleatorizado (RQA)

- Ou seja, à cada valor de Y, digamos y_i, podemos associar um número uniforme no intervalo $[F(y_{i-1}), F(y_i)]$.
- Usualmente, para simular de uma vad, simula-se u, $U \sim U(0,1)$ e atribui-se $y = y_i$, se $u \in [F(y_{i-1}), F(y_i)]$.
- Por outro lado, se $y_{(1)}, ..., y_{(n)}$ forem valores (simulados) ordenados de Y, então $U_i \overset{ind.}{\sim} U([F(y_{(i-1)}), F(y_{(i)})])$ e, consequentemente $U_i \overset{ind.}{\approx} U(0,1)$. Logo $Z_i = \Phi^{-1}(U_i) \overset{i.i.d}{\sim} N(0,1)$.
- O problema é que para um conjunto de valores $y_{(1)}, ..., y_{(n)}$ temos um número virtualmente infinito de conjunto de valores $u_1, ..., u_n$.

Cálculo do RQA para o MLG

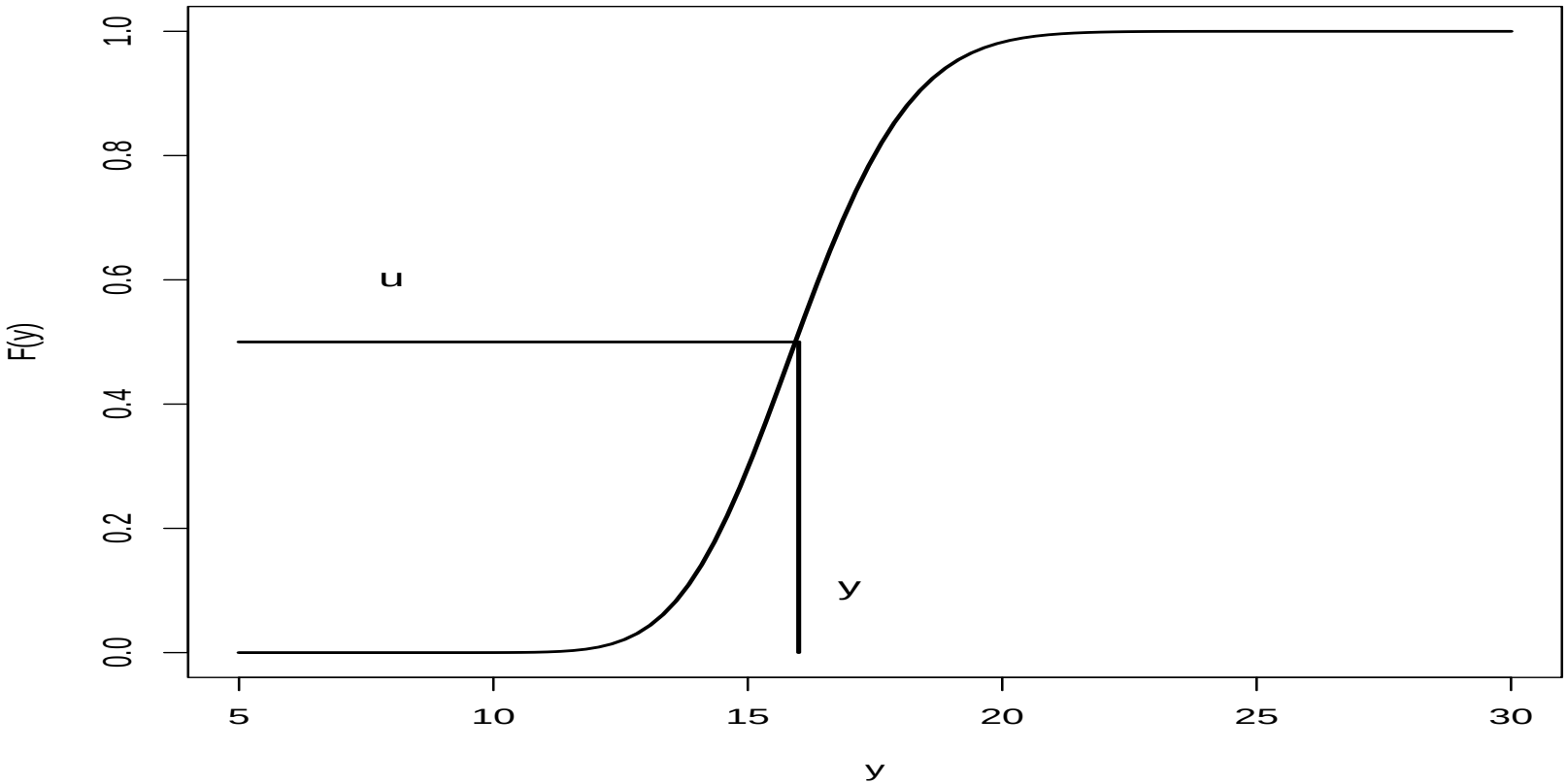
- Considera-se as observações (ordenadas) $y_{(1)}, ..., y_{(n)}$
- Com as estimativas de $\boldsymbol{\beta}$ e ϕ simula-se (m vezes) valores de $u_i \sim U([F(y_{(i-1)}), F(y_{(i)})])$, $F(.)$ é a fda da distribuição de Y_i, $Y_i \sim FE(\theta_i, \phi)$, gerando-se $u_{i1} < ... < u_{im}$, $i = 1, 2..., n$, em que $F(y_{(0)}) = 0$.
- Usualmente ordena-se os valores para cada i, ou seja, $u_{i(1)}, ..., u_{i(m)}$ e considera-se a mediana (digamos $u_{i(0,5)}$) dos valores simulados.
- Analisa-se a distribuição de $\Phi^{-1}(u_{1(0,5)}), ..., \Phi^{-1}(u_{n(0,5)})$. Sob o bom ajuste do modelo espera-se que esses resíduos tenham, aproximadamente, distribuição N(0,1).

Ilustração resíduo quantílico



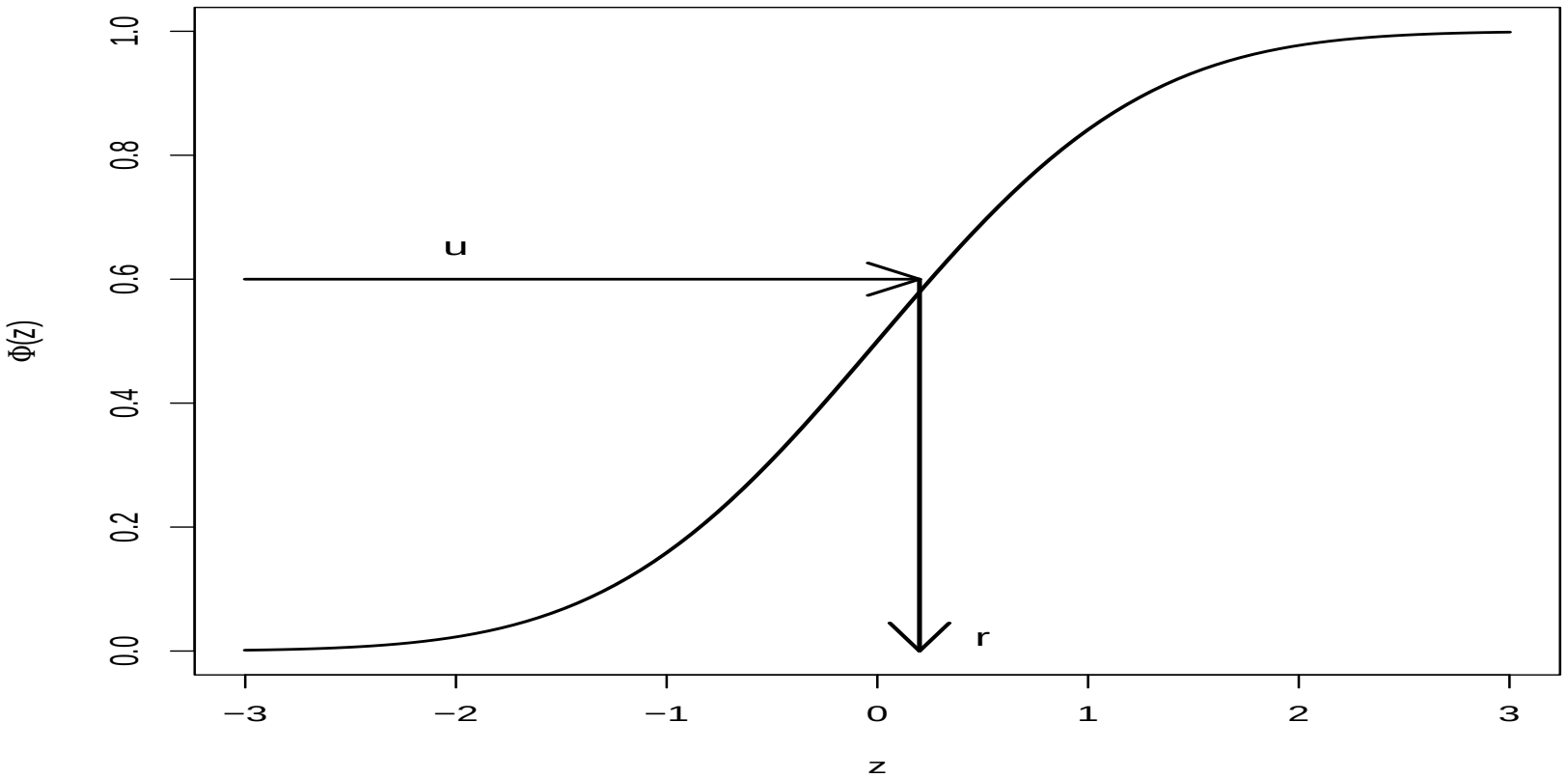
link

Ilustração resíduo quantílico



link

Ilustração resíduo quantílico



link

Ilustração resíduo quantílico aleatorizado

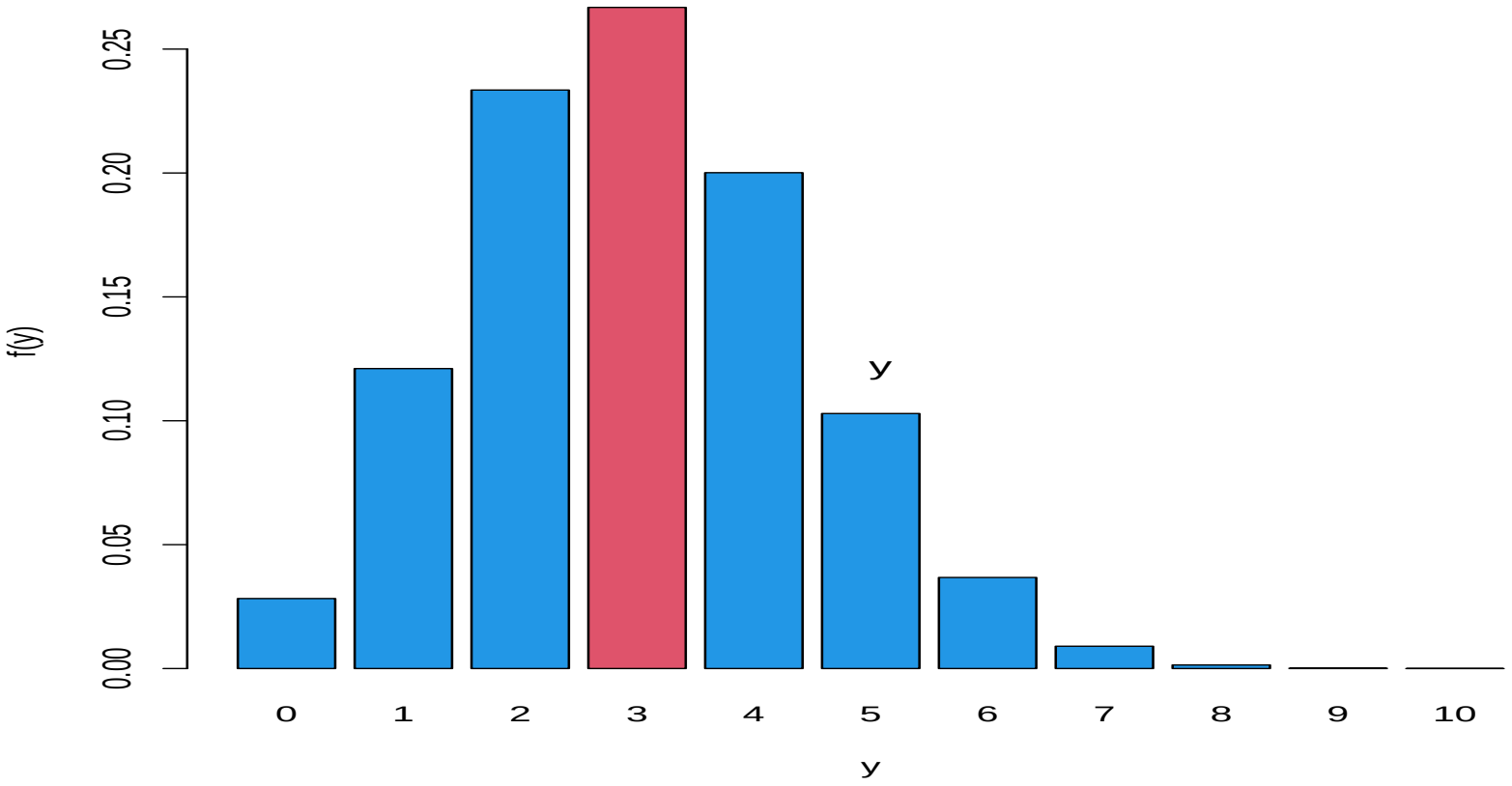
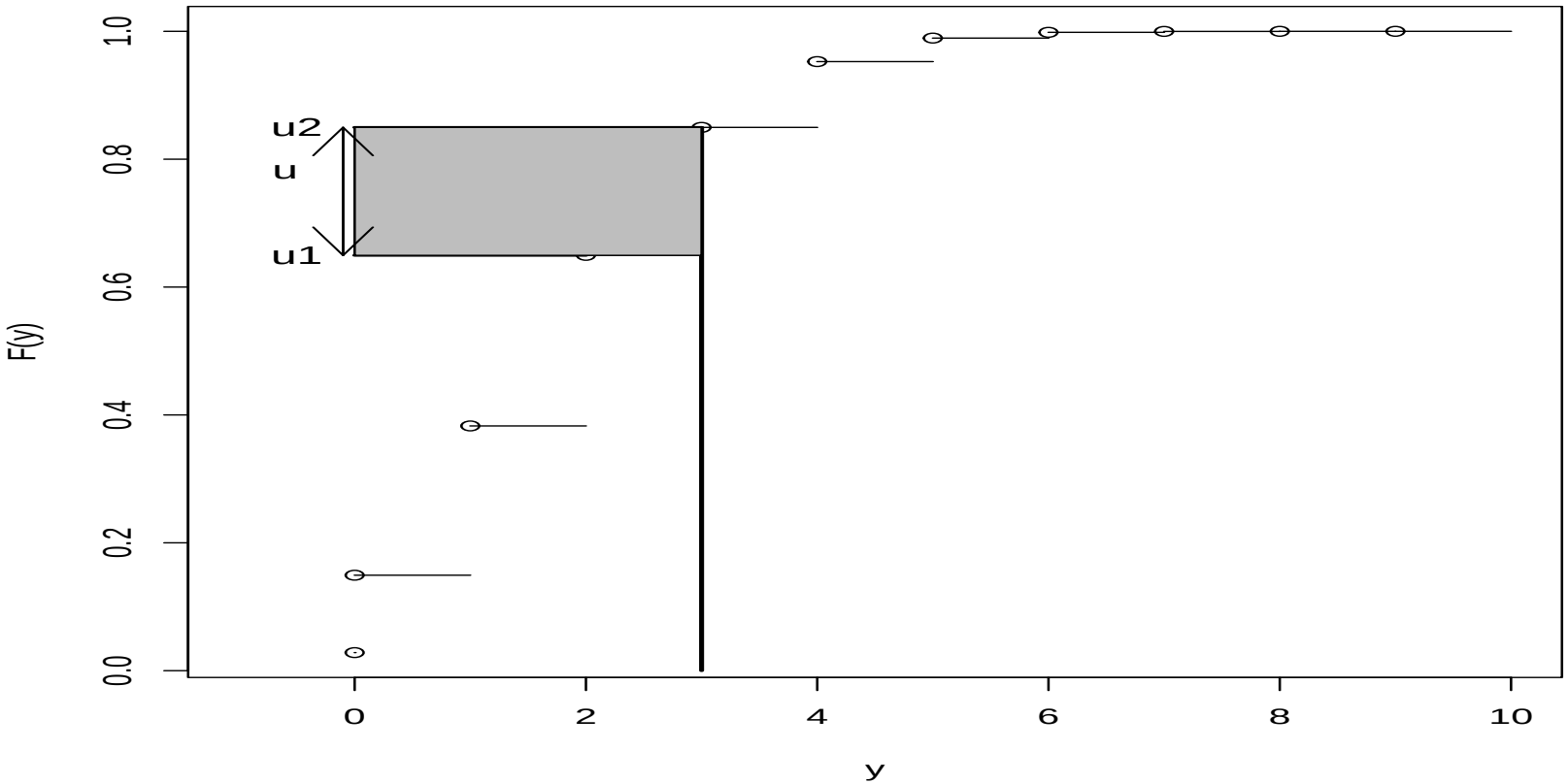
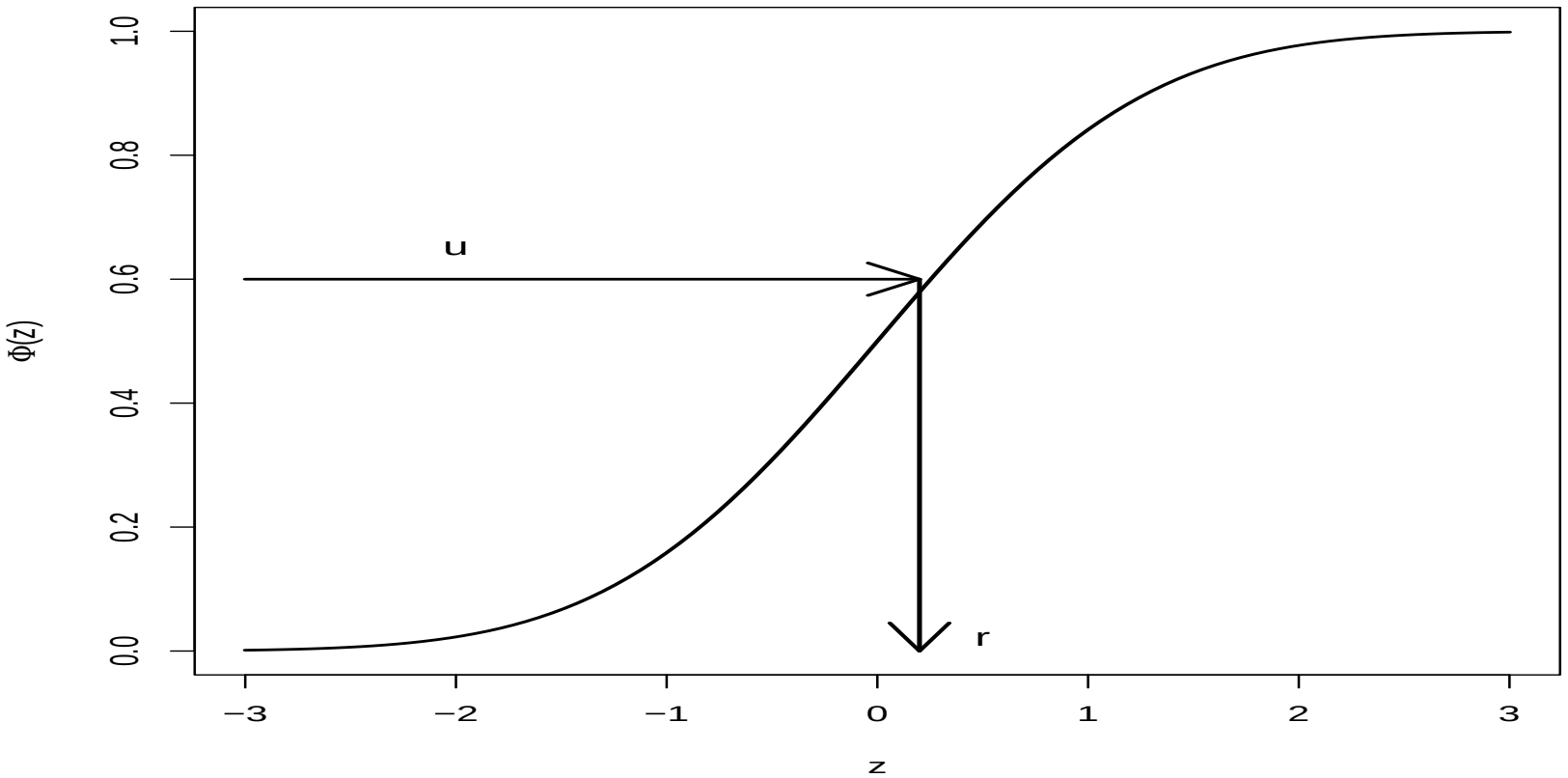


Ilustração resíduo quantílico aleatorizado



link

Ilustração resíduo quantílico aleatorizado



link

Observações

- Os gráficos de quantis-quantis em geral, tem um dos seguintes objetivos:
  - Comparam os valores ordenados (estatísticas de ordem amostrais) de duas amostras.
  - Comparam os valores ordenados de uma amostra com os quantis teóricos de uma distribuição de interesse.
  - Comparam os quantis teóricos de duas distribuições de interesse.

Observações

- Existem diferentes formas (que depende também de qual das situações acima é de interesse) de calcular os quantis teóricos de uma dada distribuição de interesse. Podemos usar resultados (exatos ou aproximados) das distribuições das estatísticas de ordem de distribuições de interesse, ou simulações. Veja, para mais detalhes, aqui, aqui, aqui, aqui.
- Vimos aqui e aqui, respectivamente, uma forma de construir gráficos de quantil-quantil com envelopes para os MRNLH e para os MLG, respectivamente. Nesses dois casos, os quantis teóricos são calculados através de simulações.

Observações

- Nos dois casos acima o procedimento usual é apresentar a dispersão entre os pontos, bem como as bandas de confiança, dentro de uma "moldura" correspondente à um gráfico (sem pontos) gerado a partir da dispersão entre os valores ordenados para os quais se quer avaliar o comportamento da respectiva distribuição e

$$\mathcal{E}(Z_{(i)}) \approx \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right),$$

(que é uma forma de calcular as estatísticas de ordem de uma N(0,1)), em que $Z_{(i)}$ é a í-ésima estatística de ordem de uma N(0,1).

Observações

- Uma alternativa ao gráfico de quantis-quantis com envelopes, é o "worm plot". A ideia é rotacionar a linha de referência de modo que ela se torne paralela ao eixo x.
- Tal gráfico é mais comumente usado para o resíduos quantílico (aleatorizado) (RQ_i), para o qual, sob o bom ajuste do modelo, espera-se normalidade (assintótica).
- Em geral, apresenta-se no worm plot $RQ_i - \mathcal{E}(Z_{(i)})$ versus $\mathcal{E}(Z_{(i)})$. Ademais, o ajuste de um polinômio cúbico é realizado e a respectiva curva é apresentada para fins de comparação entre esses duas quantidades.

Observações

- Uma forma de calcular as bandas de confiança (envelope), para se ter melhor condições de análise, para o worm plot pode ser encontradas aqui.
- Sob um bom ajuste do modelo, espera-se que os pontos se comportem de forma aleatória, de modo "plano", ao longo do valor zero (eixo horizontal, tendo a curva oriunda do plinômio cúbico como uma referência auxiliar) e dentro das bandas de confiança (aqui).