ME 720 - Modelos Lineares Generalizados
Primeiro semestre de 2016
Prova II
Data: 30/06/2016

Nome: _____ RA: _____

Leia atentamente as instruções abaixo:

- **Coloque seu nome completo e RA nesta folha, bem como em cada uma das folhas de resolução da prova.**
- **Utilize somente o espaço delimitado para cada questão/item (veja se existe limite do número de linhas que você pode escrever).**
- **Leia atentamente cada uma das questões.**
- **Enuncie, claramente, todos os resultados que você utilizar.**
- **Justifique, adequadamente, seus desenvolvimentos sem, no entanto, escrever excessivamente.**
- **Os desenvolvimentos devem ser apresentados na íntegra sendo vedado o uso da utilização de argumentos como (mas não se limitando à): "vimos em classe que a expressão geral é esta .... e, portanto, vamos particularizá-la para o caso em questão" a menos que o contrário seja mencionado na questão.**
- **Às provas cujo o código de honra não esteja devidamente preenchido e assinado, serão atribuidas nota zero.**
- **Se for detectado desvio de conduta (código de honra), à respectiva prova será atribuída nota zero.**
- **Resolva a prova, preferencialmente, à caneta (azul ou preta), e procure ser organizado(a).**
- **Contestações a respeito da nota/correção, só serão consideradas se estiverem por escrito.**
- **A nota do aluno(a) será $\frac{NP}{NT} \times 10$, em que NP é o número de pontos obtidos na prova e NT é o numero total de pontos da prova.**
- **Os resultados numéricos finais devem ser apresentados com, somente, duas casas decimais, a não ser que seja solicitado um número diferente de casas.**
- **A prova terá duração de 24 horas, das 16h do dia 30/06 às 16h do dia 01/07, improrrogáveis.**
- **A prova terá de ser entregue até as 16h00 do dia 01/07, pessoalmente. OBS: Caso o(a) aluno(a) queira entregar a prova antes do período mencionado e o professor não se encontre na sala, ele(a) pode fazê-lo, pessoalmente, da seguinte forma: deve-se redigir, de próprio punho, e assinar, um termo atestando que ele(a) está entregando pessoalmente a prova e colocar ambos (prova completa e termo) num envelope e lacrá-lo, sendo que este deve ser colocado por debaixo da porta.**

- **É vedada a consulta a terceiros e/ou a quaisquer outras fontes que não as informações constantes na própria prova, incluindo o formulário.**

Código de honra

Eu,_____, juro pela minha honra que não recebi e nem prestei auxílio durante o período de duração da prova, e tampouco consultei outras fontes, além da própria prova e formulário nela constante, e que a resolvi individualmente.

Ass:

Faça uma excelente Prova!!

1. Seja $Y_i \overset{ind.}{\sim} \text{Poisson}(\mu_i), \ln \mu_i = \beta_0 + \beta_1 x_i, \beta_i \in (-\infty, \infty), i = 1, 2$ e x_i (não aleatórias e conhecidas), $i = 1, 2..., n$. Responda os itens:

   a) Obtenha o vetor escore e a informação de Fisher associadas ao modelo e apresente o sistema de equações que deve ser resolvido para que se obtenha o estimador de máxima verossimilhança (emv) de $(\beta_0, \beta_1)'$. Além disso, apresente a distribuição assintótica desse estimador. (500 pontos)

   b) Considere o interesse em testar $H_0 : \beta_0 = \beta$ vs $H_1 : \beta_0 \neq \beta$, com $\beta \in (-\infty, \infty)$ conhecido. Proponha um teste do tipo Wald para testar essas hipóteses, utilizando o emv e sua respectiva distribuição assintótica, de sorte que a distribuição assintótica da estatística do teste, sob H_0, seja χ^2_1. Você pode propor o teste mesmo que não tenha obtido a informação de Fisher mas, nesse caso, você poderá conseguir, no máximo, a metade do valor deste item (500 pontos)

2. Sejam $Y_1|P = p, ..., Y_m|P = p \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$ em que $P \sim \text{beta}(a, b)$ e defina $Y = \sum_{i=1}^{m} Y_i$. Não se esqueça de justificar cada uma das etapas de seus desenvolvimentos. Responda aos itens:

   a) Calcule $\mathcal{E}(Y)$, $\mathcal{V}(Y)$ e $\text{Cov}(Y_i, Y_j), \forall i \neq j$ (300 pontos).

   b) Obtenha a distribuição de $Y|P = p$. Sugestão: neste caso você pode provar de modo argumentativo. (300 pontos)

   c) Obtenha a distribuição conjunta de (Y, P) e a respectiva distribuição marginal de Y. (400 pontos)

3. Seja Y uma variável aleatória discreta (vad) de sorte que sua fdp (função de probabilidade) é dada por: $g_Y(y) = [\pi + (1 - \pi) f_Z(0)] \mathbb{1}_{\{0\}}(y) + (1 - \pi) f_Z(y) \mathbb{1}_{\{1,2,...,\}}(y)$ em que $f_Z(.)$ é a fdp de uma vad (Z) com suporte nos inteiros não-negativos ($\{0,1,...,\}$). Não se esqueça de justificar cada uma das etapas de seus desenvolvimentos. Responda os itens:

   a) Prove que $g_Y(.)$ de fato é uma fdp. (200 pontos).

   b) Calcule $\mathcal{E}(Y)$ e $\mathcal{V}(Y)$. (200 pontos).

   c) Como ficam as expressões de $\mathcal{E}(Y)$ e $\mathcal{V}(Y)$ se $Z \sim$ geométrica(p), em que $p \in (0, 1)$. (200 pontos)

   d) Seja $U \sim \text{Bernoulli}(\pi)$ e $P(Y = 0|U = 1) = 1$ e $P(Y = y|U = 0) = P(Z = y)$. Prove que a fdp de Y, de fato, corresponde à $g_Y(.)$ (definida acima). (400 pontos)

4. Os dados analisados correspondem aos resultados de um estudo desenvolvido em 1990 com recrutas americanos referente a associação entre o número de infeções de ouvido diagnosticadas pelo próprio recruta (variável resposta) e alguns fatores (variáveis explicativas),

as quais são: hábito de nadar (ocasional ou frequente) e local onde costuma nadar (piscina ou praia). O interesse é analisar como os fatores afetam o número de infecções. Seja Y_{ijk} o número de infecções de ouvido do k-ésimo recruta, que possui o i-ésimo hábito de nadar, no j-ésimo local. Foram ajustados dois modelos (descritos abaixo), para analisar os dados. Considere que a aproximação do desvio pela distribuição de $\chi^2_{(n-p)}$ é adequada para os dois modelos. Alguns resultados relativos à análise descritiva e ao ajuste dos dois modelos encontram-se nas Tabelas 1, 2 e 3 e nas Figuras 1 e 2. Responda os itens abaixo:

a) Do ponto de vista descritivo, qual dos dois modelos você considera o mais apropriado? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 5 linhas. (200 pontos)

b) Do ponto de vista inferencial, qual dos dois modelos você considera o mais apropriado? Além disso, você considera o modelo sugerido por você, apropriado(com bom ajuste)? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 10 linhas.(400 pontos)

c) Com base nos resultados apresentados na Tabela 3 (suponha que eles foram obtidos a partir do modelo que você escolheu), o que é possível afirmar sobre a existência de interação e sobre as diferenças entre as médias dos quatro grupos (ou seja, como você pode classificar esses grupos com relação ao número de infecções)? Além disso, suas conclusões são compatíveis com os resultados da análise descritiva? Você continuaria com esse modelo, para realizar inferências, ou ajustaria um modelo reduzido (nesse caso descreva como seria o preditor linear)? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 15 linhas.(400 pontos)

Modelo 1

$$\begin{aligned} Y_{ijk} &\overset{ind.}{\sim} \text{Poisson}(\mu_{ij}), i = 1, 2, (1\text{: ocasional,2: frequente}), \\ &\quad j = 1, 2, (1\text{: piscina,2: praia}), k = 1, 2, ..., n_{ij} \\ \ln(\mu_{ij}) &= \alpha + \beta_i + \gamma_j + \delta_{ij}, \beta_1 = \gamma_1 = \delta_{i1} = \delta_{1j} = 0, \forall i, j \end{aligned}$$

Modelo 2

$$\begin{aligned} Y_{ijk} &\overset{ind.}{\sim} \text{BN}(\mu_{ij}, \phi), i = 1, 2, (1\text{: ocasional,2: frequente}), \\ &\quad j = 1, 2, (1\text{: piscina,2: praia}), k = 1, 2, ..., n_{ij} \\ \ln(\mu_{ij}) &= \alpha + \beta_i + \gamma_j + \delta_{ij}, \beta_1 = \gamma_1 = \delta_{i1} = \delta_{1j} = 0, \forall i, j \end{aligned}$$

Tabela 1: Medidas resumo para o número de infecções: Questão 4

Hábito	Local	Média	Var.	DP	n
frequentemente	praia	0,82	2,43	1,56	72
frequentemente	piscina	1,14	2,18	1,48	71
ocasionalmente	praia	1,28	4,85	2,20	75
ocasionalmente	piscina	2,35	11,58	3,40	69

Tabela 2: Estatísticas de comparação de modelos, desvio estimado (e respectivo p-valor): Questão 4

Modelo	AIC	BIC	desvio	p-valor
1	1143,39	1158,03	763,00	$<$ 0,0001
2	903,99	922,29	269,18	0,7132

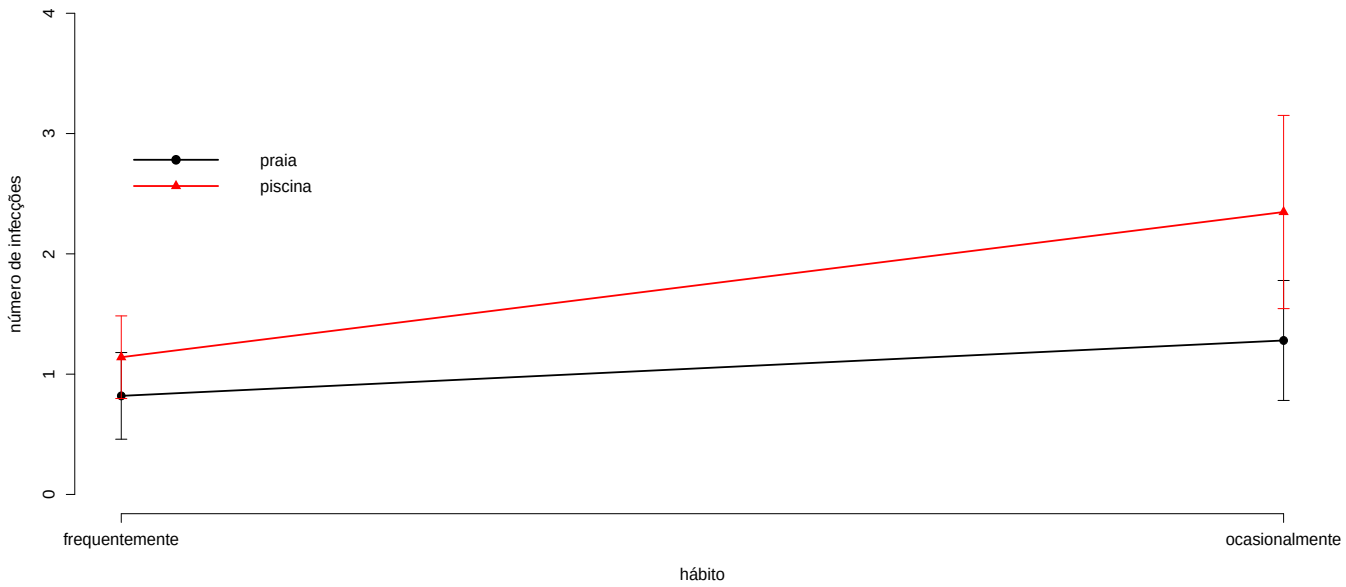

Figura 1: Gráfico de perfis médios: Questão 4

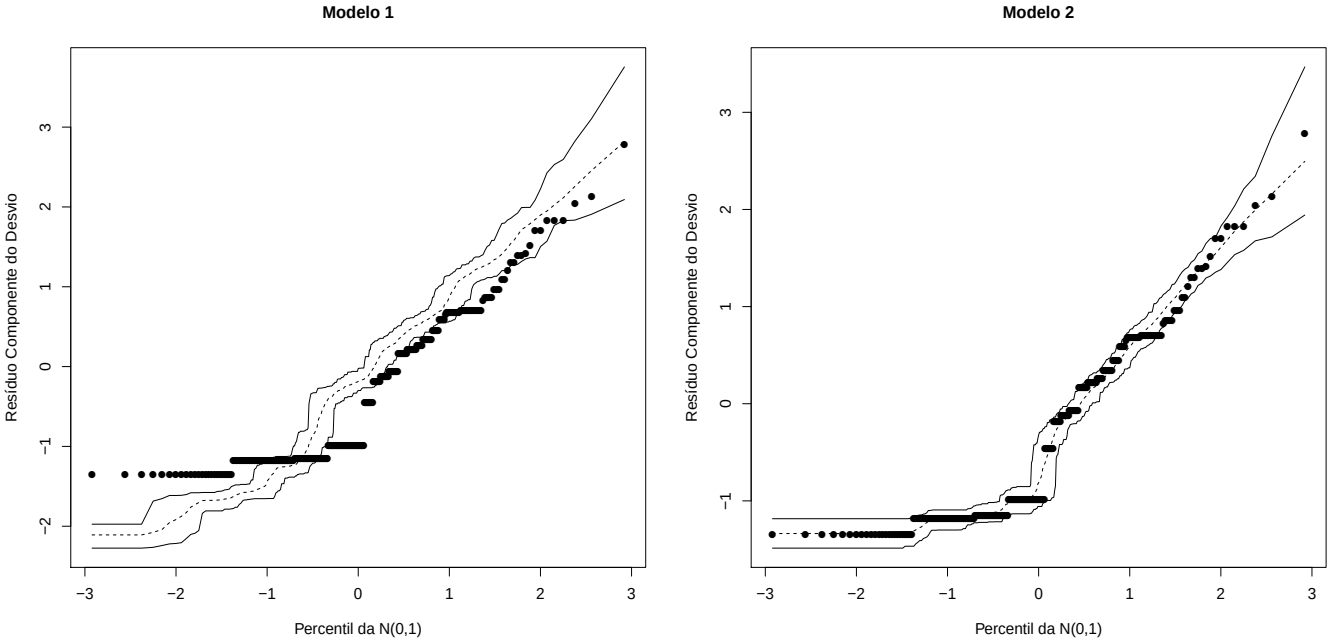

Figura 2: Gráfico de evelope para o resíduo componente do desvio dos modelos 1 e 2: Questão 4

Tabela 3: Estimativas e testes de hipótese dos parâmetros do modelo escolhido: Questão 4

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
α	-0,20	0,20	[-0,60 ; 0,20]	-0,98	0,3292
β_2	0,45	0,28	[-0,09 ; 0,99]	1,62	0,1052
γ_2	0,33	0,28	[-0,22 ; 0,88]	1,18	0,2392
δ_{22}	0,28	0,38	[-0,47 ; 1,02]	0,72	0,4691

5. Os dados analisados se referem a um experimento com duas espécies de rotifers, Polyarthra major (PM) e Keratella cochlearis (KC), um tipo microscópico de invertebrado aquático. O objetivo do experimento é medir como a densidade relativa de uma certa substância afeta a quantidade de rotifers que permanecem suspensos, ou seja, na superfície, (uma vez colocados na superfície de recipientes com essa substância, sendo 20 para cada grupo) e comparar os dois grupos quanto a esse comportamento. Os dados se encontram na Tabela 4. Seja Y_{ij} o número de animais da i-ésima espécie que permanecem suspensos num recipiente com densidade relativa d_j da solução, onde foram colocados m_{ij} rotifers. Inicialmente fora ajustado o seguinte modelo (doravante modelo 1) (alguns resultados

relativos à análise descritiva e ao ajuste dos dois modelos encontram-se nas Figuras 3, 4,5,6 e 7.):

$$\begin{aligned} Y_{ij} &\overset{ind.}{\sim} \text{binomial}(m_{ij}, \mu_{ij}), i = 1(PM), 2(KC); j = 1, 2, ..., 20 \\ \ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= (\beta_0 + \alpha_{0i}) + (\beta_1 + \alpha_{1i})(d_j - \overline{d}) \\ \overline{d} &= \frac{1}{20}\sum_{j=1}^{20} d_j; \alpha_{01} = \alpha_{11} = 0 \end{aligned}$$

a) Do ponto de vista descritivo, o que você pode dizer sobre o efeito da densidade na proporção de animais suspensos, intra e entre grupos? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 10 linhas. (100 pontos)

b) O que você pode dizer sobre a qualidade do ajuste do modelo 1? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 10 linhas. (300 pontos)

c) Se você concluiu que o modelo não está bem ajustado, qual seria a causa mais provável do mal ajuste, utilizando também o fato de que os ajustes dos outros três modelos vistos em sala (probito, extremito e cauchito) foram muito semelhantes ao ajuste do modelo em questão? Ademais, o que poderia estar causando o(s) fator(es) que estão levando, segundo sua opinião, ao mau ajuste? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 10 linhas.(300 pontos)

d) Como uma alternativa ao modelo de regressão logística, ajustou-se um outro modelo, similar ao modelo logístico, mas em que $Y_{ij} \overset{ind.}{\sim} \text{bb}(m_{ij}, \mu_{ij}, \sigma)$, doravante modelo 2. Com base em seus comentários apresentados nos itens b) e c), bem como no ajuste do modelo 2, o que você pode dizer a respeito da escolha dessa alternativa? Utilize a maior quantidade possível de informações e justifique, apropriadamente, seus comentários. Seus comentários não podem ultrapassar 10 linhas.(300 pontos)

Tabela 4: Dados relativos ao experimento com rotifers: Questão 5

		Polyarthra major		Keratella cochlearis	
Recipiente	densidade	suspensos (m_{ij})	expostos (y_{ij})	suspensos (m_{ij})	expostos (y_{ij})
1	1,019	11	58	13	161
2	1,020	7	86	14	248
3	1,021	10	76	30	234
4	1,030	19	83	10	283
5	1,030	9	56	14	129
6	1,030	21	73	35	161
7	1,031	13	29	26	167
8	1,040	34	44	32	286
9	1,040	10	31	22	117
10	1,041	36	56	23	162
11	1,048	20	27	7	42
12	1,049	54	59	22	48
13	1,050	20	22	9	49
14	1,050	9	14	34	160
15	1,060	14	17	71	74
16	1,061	10	22	25	45
17	1,063	64	66	94	101
18	1,070	68	86	63	68
19	1,070	488	492	178	190
20	1,070	88	89	154	154

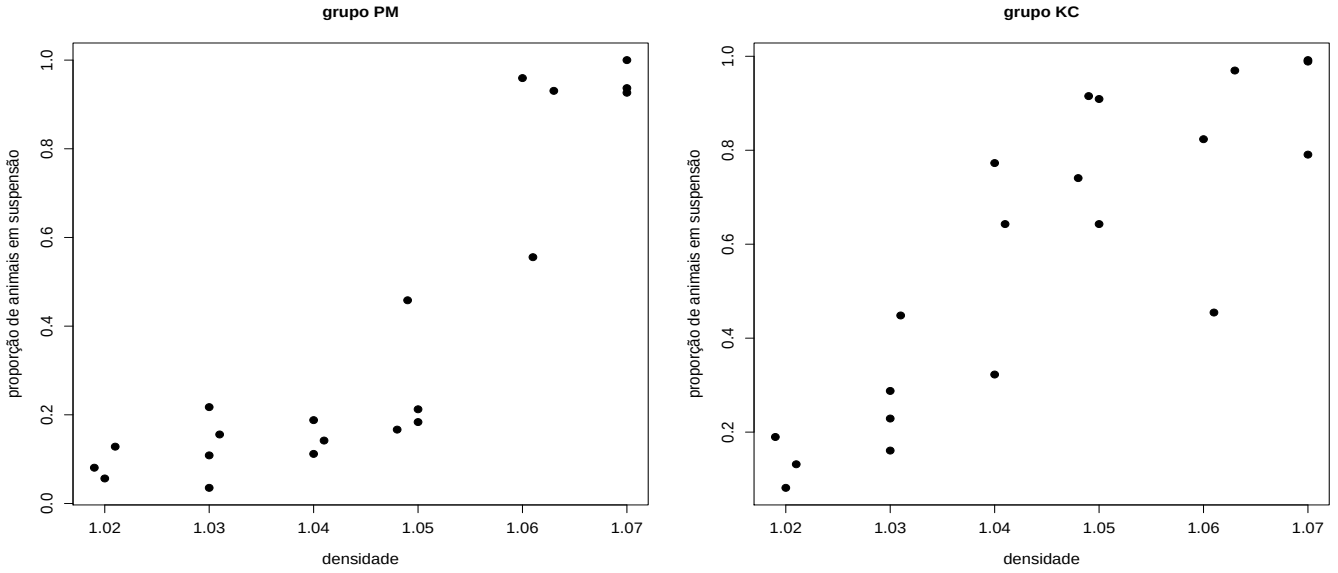

Figura 3: Gráfico de proporções observadas de animais em suspensão: Questão 5

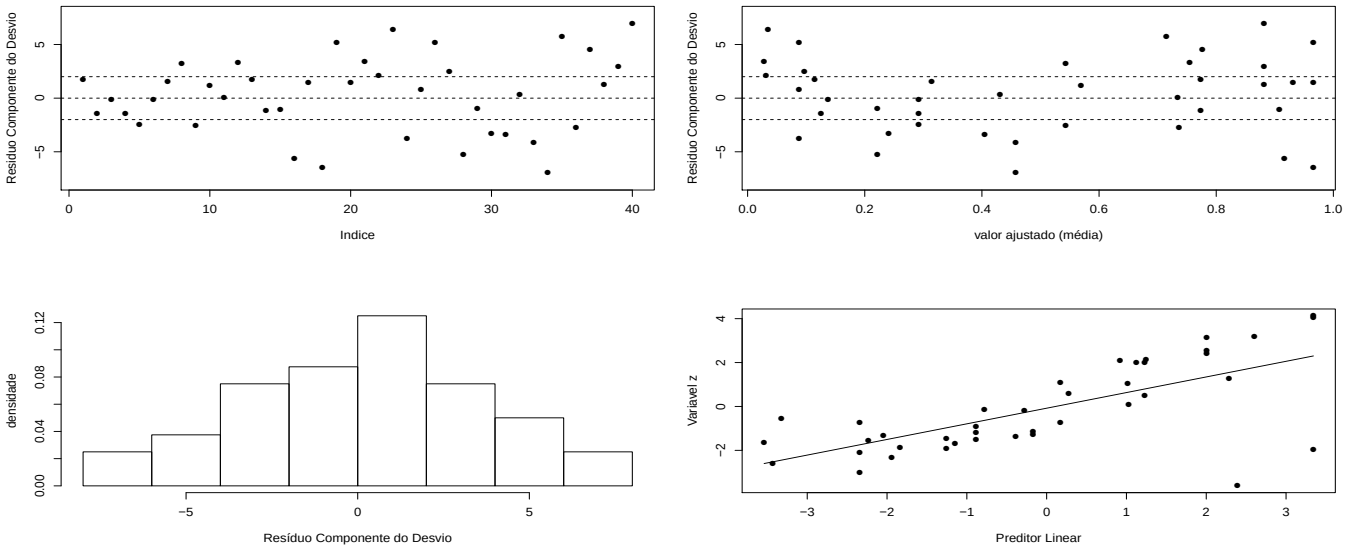

Figura 4: Gráficos de diagnóstico do resíduo componente do desvio do modelo 1: Questão 5

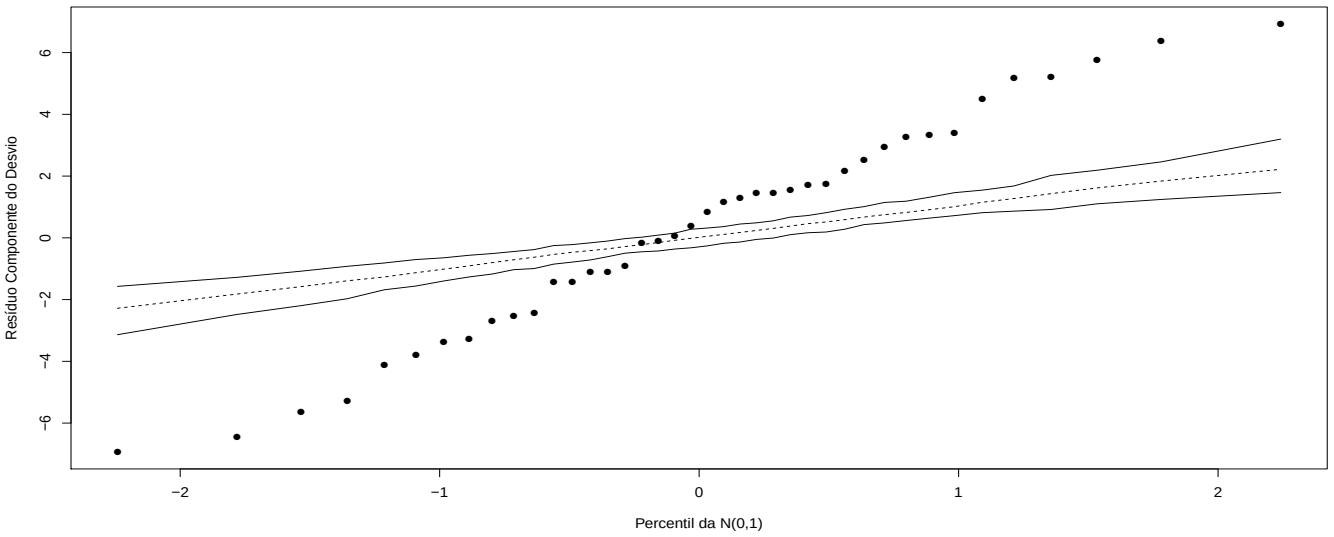

Figura 5: Gráficos de envelope para o resíduo componente do desvio do modelo 1: Questão 5

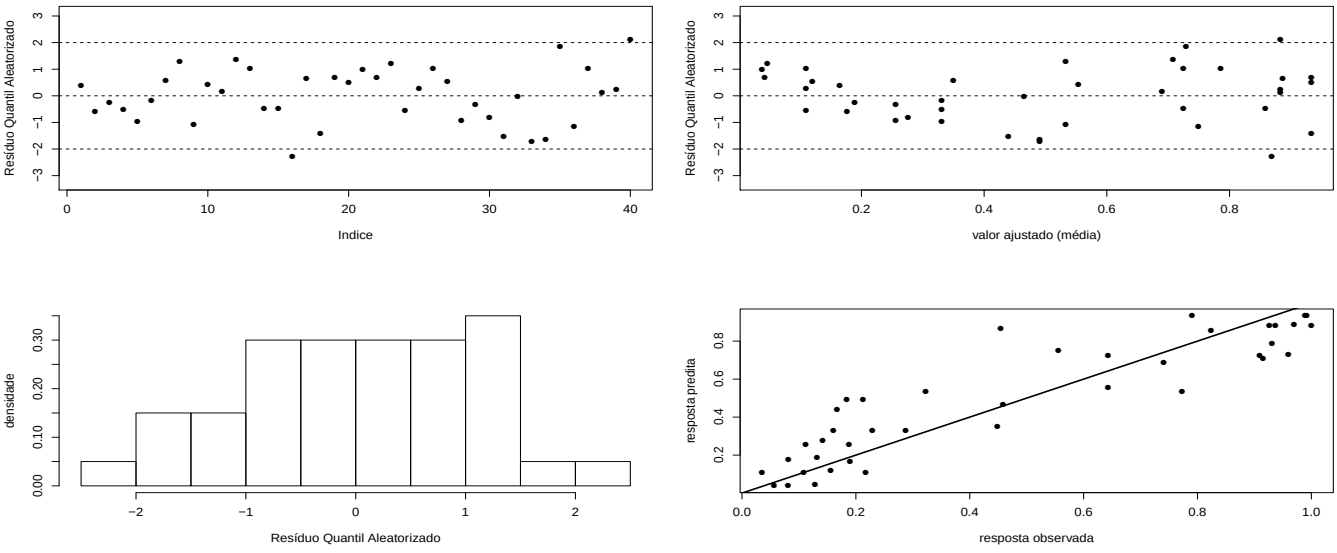

Figura 6: Gráficos de diagnóstico do resíduo quantil aleatorizado do modelo 2: Questão 5

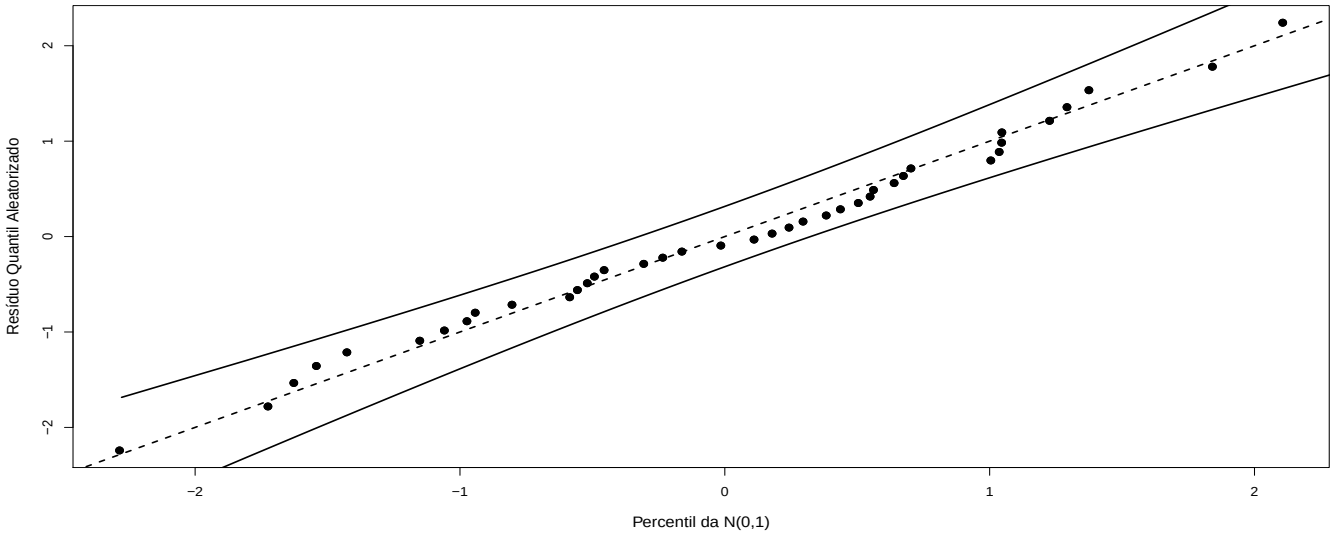

Figura 7: Gráficos de envelope para o resíduo quantil aleatorizado do modelo 2: Questão 5

Formulário

1. Se $X \sim \text{Poisson}(\mu)$ então $f(x) = \dfrac{e^{-\mu}\mu^x}{x!} I_{\{0,1,2,...\}}(x)$ e $\mathcal{E}(X) = \mathcal{V}(X) = \mu$.
2. Derivação: $f(x) = e^{g(x)}$, então $\dfrac{df(x)}{dx} = e^{g(x)} \dfrac{dg(x)}{dx}$.
3. Teste tipo Wald (uniparamétrico). Seja $\widehat{\beta}$ o estimador de máxima verossimilhança de β, tal que, para n suficientemente grande (e sob as condições de regularidade), $\widehat{\beta} \approx N(\beta, \mathcal{V}(\widehat{\beta}))$, em que $\mathcal{V}(\widehat{\boldsymbol{\beta}})$ é a variância assintótica de $\widehat{\beta}$. Então, $\dfrac{(\widehat{\beta} - \beta)^2}{\widehat{\mathcal{V}}(\widehat{\beta})} \approx \chi^2_1$, em que $\widehat{\mathcal{V}}(\widehat{\beta})$ é um estimador consistente de $\mathcal{V}(\widehat{\beta})$.
4. Se $X \sim \text{binomial}(m, \mu), m \in \{1, 2, 3, ...\}, \mu \in (0, 1)$, então $f(x) = \begin{pmatrix} m \\ x \end{pmatrix} \mu^x (1-\mu)^{m-x} \mathbb{1}_{\{0,1,...,m\}}(x)$, $\mathcal{E}(X) = m\mu, \mathcal{V}(X) = m\mu(1-\mu)$. Se $m = 1$, obtem-se a distribuição de Bernoulli(μ).
5. Se $X \sim \text{beta}(a, b)$ então $f(x) = \dfrac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1} \mathbb{1}_{(0,1)}(x)$, $\mathcal{E}(X) = \dfrac{a}{a+b}$, $\mathcal{V}(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$, em que $\beta(.,.)$ é a função beta.
6. Se $X \sim bb(m, \mu, \sigma)$ (bb representa a distribuição beta binomial), então $\mathcal{E}(X) = \mu$ e $\mathcal{V}(X) = m\mu(1-\mu)\left[1 + \dfrac{(m-1)\sigma}{1+\sigma}\right]$. OBS: Usualmente, a distribuição beta binomial é definida como $bb(m, a, b)$ e cuja fdp é dada por

$$\begin{aligned} f(x) &= \frac{\Gamma(m+1)}{\Gamma(x+1)\Gamma(m-x+1)} \frac{\Gamma(x+a)\Gamma(m-x+b)}{\Gamma(m+a+b)} \\ &\times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mathbb{1}_{\{0,1,2,3,...,m\}}(x) \end{aligned}$$

   Para obter a parametrização $bb(m, \mu, \sigma)$ basta fazer $a = \dfrac{\mu}{\sigma}$ e $b = \dfrac{1-\mu}{\sigma}$.
7. Se $X \sim$ geométrica(p), então $f(x) = (1-p)^x p \mathbb{1}_{\{0,1,2,...\}}(x)$, $\mathcal{E}(X) = \dfrac{1-p}{p}$, $\mathcal{V}(X) = \dfrac{1-p}{p^2}$.
8. $X \sim BN(\mu, \phi)$ representa uma distribuição binomial negativa com média μ e parâmetro de precisão ϕ.