ME 714 - Análise de dados discretos
Primeiro semestre de 2014
Prova II
Data: 30/06/2014

Nome: _____ RA: _____

Leia atentamente as instruções abaixo:

- **Coloque seu nome completo e RA em todas as folhas que você recebeu, inclusive nesta.**
- **Utilize somente um dos lados de cada folha de resolução.**
- **Leia atentamente cada uma das questões.**
- **Enuncie, claramente, todos os resultados que você utilizar.**
- **Justifique, adequadamente, seus desenvolvimentos, sem, no entanto, escrever excessivamente.**
- **O(a) aluno(a) só poderá sair da sala após as 16h30, mesmo que já tenha finalizado a prova. Após a saída do(a) primeiro(a) aluno(a) não será permitido a entrada de nenhum(a) outro(a) aluno(a).**
- **Não é permitdo empréstimo de material.**
- **Não serão dirimidas dúvidas de quaisquer natureza, após os 20 minutos iniciais.**
- **Resolva a prova, preferencialmente, à caneta, e procure ser organizado(a). Se fizer à lápis, destaque, à caneta, sua resposta.**
- **A resolução da prova deve seguir a ordem das questões e, dentro de cada questão, a ordem dos itens. Cada nova questão deverá ser resolvida em uma nova folha.**

- **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 120 minutos, das 16h às 18h, improrrogáveis.**

Faça uma excelente Prova!!

1. A Tabela 1 refere-se aos dados de um estudo sobre a influência do número de cigarros consumidos diariamente por gestantes (N. de cigarros) com a sobrevivência ou não dos respectivos recém-nascidos (sobrevivência). Há interesse em saber se as variáveis "N. de cigarros" e "Sobrevivência" são independentes (H_0, hipótese nula) ou dependentes (H_1, hipótese alternativa) para cada valor da variável "idade" (que representa a idade de cada gestante, em anos) de modo simultâneo. Considere que as linhas da Tabela 1 correspondem à binomiais mutuamente independentes e as quantidades observadas ao número de recem nascidos que não sobreviveram (ou sobreviveram). Denote a frequência (populacional) de cada casela por $N_{(k)ij}$ e as respectivas probabilidades de ocorrência por $\theta_{(k)ij}$, em que k corresponde ao grupo (idade), e i e j correspondem às linhas e colunas, respectivamente. Considere, ainda, que $\boldsymbol{\pi} = (\theta_{(1)11}, \theta_{(1)12}, \theta_{(1)21}, \theta_{(1)22}, \theta_{(2)11}, \theta_{(2)12}, \theta_{(2)21}, \theta_{(2)22})'$.

Tabela 1: Dados sobre o estudo

idade	N. de cigarros	Sobrevivência		
		Não	Sim	Total
<30	< 5	74	4327	4401
	5+	15	499	514
30+	< 5	55	1741	1796
	5+	5	135	140

   a) Escreva, de forma escalar e matricial ($\boldsymbol{B\pi} = \boldsymbol{D}$), as hipóteses de interesse em termos das probabilidades. (100 pontos)
   b) Escreva, de forma escalar e matricial ($\boldsymbol{A}\ln(\boldsymbol{G\pi}) = \boldsymbol{D}$), as hipóteses de interesse em termos das razões de chances. (100 pontos)
   c) Proponha um modelo de regressão logística para responder a pergunta de interesse (não é necessário apresentar as interpretações dos parâmetros). Escreva, somente de forma escalar, as hipóteses de interesse em termos dos parâmetros desse modelo. (200 pontos)

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

   a) Obtenha a função escore e a informação de Fisher associadas ao modelo e apresente a equação que deve ser resolvida para que se obtenha o estimador de máxima verossimilhança (emv) de β. Além disso, apresente a distribuição assintótica desse estimador. (200 pontos)
   b) Considere o interesse em testar $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$, com $\beta_0 \in (-\infty, \infty)$ conhecido. Proponha um teste 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 (100 pontos)

c) Obtenha o desvio do modelo (simplifique a expressão o máximo possível), escrevendo-o em função de $\widehat{\mu}_i$. Você pode obtê-lo mesmo que não tenha resolvido o item a) desta questão (nesse caso, não haverá penalização em relação à pontuação deste item). (100 pontos)

3. O conjunto de dados analisado se refere a idade de ocorrência de menarca de garotas de Varsóvia. Tem-se o interesse em saber como a idade impacta na ocorrência da menarca. Sejam (Y_i): o número de garotas que apresentaram menstruação no grupo i, (m_i): o número de garotas entrevistadas no grupo i e (x_i): a idade média no grupo i, $i = 1, 2, ..., 25$. Para analisar os dados considerou-se o seguinte modelo de regressão logística: $Y_i \overset{ind.}{\sim}$ binominal(m_i, p_i), em que logito$(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1(x_i - \overline{x}), \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, n = 25$. Um resumo dos resultados se encontram na Tabela 2 e nas Figuras 1 e 2. O desvio estimado do modelo foi $D(\boldsymbol{y}; \widetilde{\boldsymbol{\mu}}) = 26,81$. Responda os itens:

a) Prove que a razão de chances (em relação ao aumento em um ano na idade do grupo) é e^{β_1} e a estime, pontual e intervalarmente (utilizando o método delta). Use $\gamma = 0,95$ (coeficiente de confiança). OBS: Você pode estimar a razão de chances mesmo sem ter provado que ela é igual a e^{β_1}. Contudo, nesse caso, você poderá conseguir, no máximo, a metade do valor deste item. (200 pontos)

b) O que você pode afirmar sobre a qualidade de ajuste do modelo ao conjunto de dados em questão, utilizando o valor do desvio, os gráficos de diagnóstico (Figura 1) e os gráfico com as proporções observadas e preditas (Figura 2)? Comente, da forma mais completa possível, e justifique, adequadamente, seus comentários. OBS: Seus comentários não podem ultrapassar 10 linhas. (200 pontos)

Tabela 2: Estimativas e erros-padrão dos parâmetros do modelo (Questão 3)

Parâmetro	Estimativa	Erro-padrão
β_0	0,150	0,063
β_1	1,632	0,059

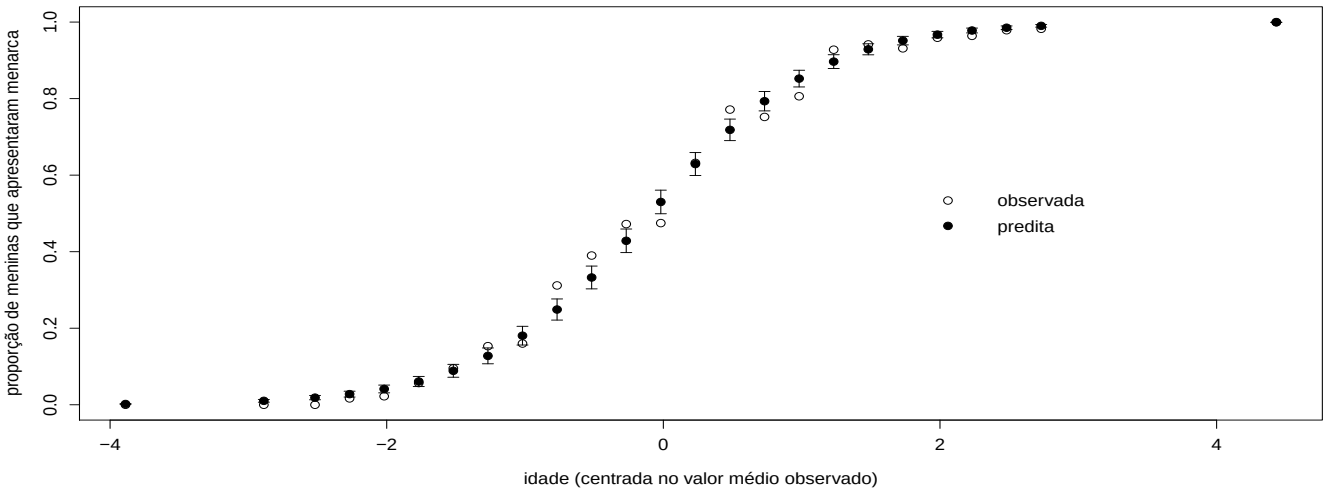

Figura 1: Proporções observadas e preditas pelo modelo (questão 3)

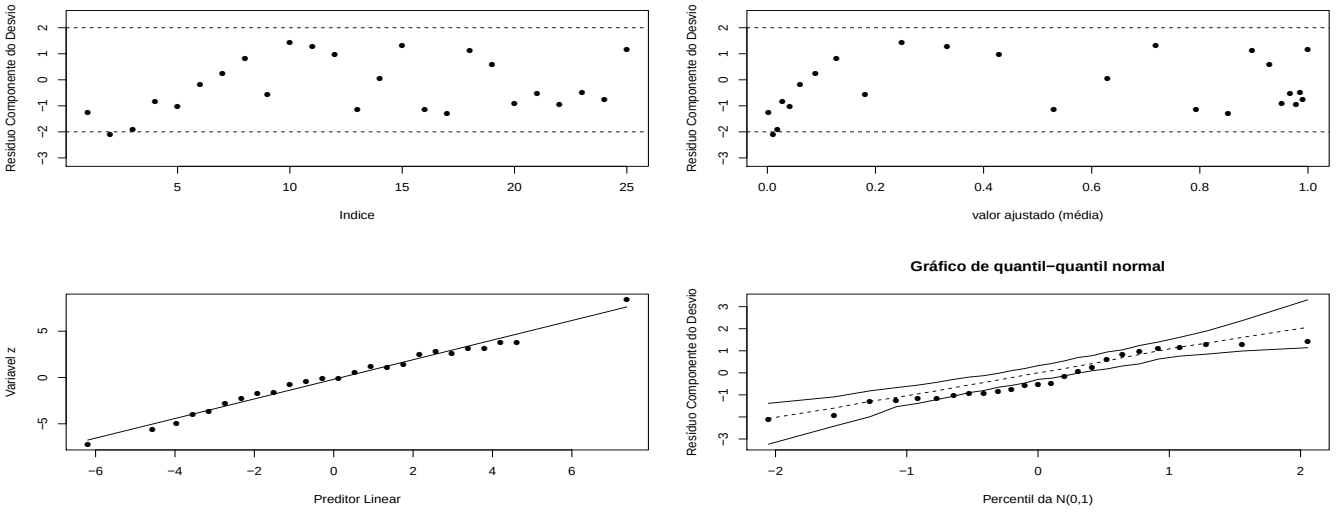

Figura 2: Gráficos de diagnóstico para o modelo (questão 3)

Formulário

1. Se $X \sim \text{binomial}(m, \theta), m \in \{1, 2, 3, ...\}, \theta \in (0, 1)$, então
   $f(x) = \left(\begin{array}{c} m \\ x \end{array} \right) \theta^x (1-\theta)^{m-x} \mathbb{1}_{\{0,1,...,m\}}(X)$, $\mathcal{E}(X) = m\theta, \mathcal{V}(X) = m\theta(1-\theta)$. Se $m = 1$, obtem-se a distribuição de Bernoulli(θ).

2. A função desvio (ou simplesmente desvio) é definida por
   $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\left\{\prod_{i=1}^n l(y_i, y_i) - \prod_{i=1}^n l(\widehat{\mu}_i, y_i)\right\}$, em que $\widehat{\mu}_i$ é o estimador de máxima verossimilhança da média da observação i e $\prod_{i=1}^n l(\mu_i, y_i)$ é a logverossimilhança do modelo. Considere que, sob as condições de regularidade, $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \approx \chi^2_{(n-p)}$, para n suficientemente grande, em que n é o tamanho da amostra e p é o número de parâmetros.

3. Razão de chances: $\eta_{(k)} = \dfrac{\lambda_{(k)1}}{\lambda_{(k)2}}$, $\lambda_{(k)i} = \dfrac{\theta_{(k)i1}}{1 - \theta_{(k)i1}}, i = 1, 2$. OBS: os índices podem variar conforme a situação em questão.

4. Método delta univariado: Seja $\widehat{\theta}$ uma variável aleatória de sorte que, para n suficientemente grande (tamanho da amostra), $\widehat{\theta} \approx N(\theta, \sigma^2)$. Defina $\widehat{\tau} = g(\widehat{\theta})$. Então, para n suficientemente grande,
   $$\widehat{\tau} \approx N(g(\theta), \sigma^2 \left[\psi(\theta)\right]^2),$$
   em que $\psi(\theta) = \dfrac{d}{d\theta} g(\theta)$.