MI 602A - Métodos computacionais em Estatística
Primeiro semestre de 2012
Prova II
Data: 27/06/2012

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.**
- **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 10h30, 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.**
- **O(a) aluno(a) deverá portar sua carteira de estudante e apresentá-la, quando for solicitada sua assinatura.**
- **Contestações a respeito da nota, 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.**
- **Use somente um lado de cada folha de resolução.**
- **Os resultados numéricos finais devem ser apresentados com duas casas decimais, apenas.**
- **A prova terá duração de 120 minutos, das 10h00 às 12h00, improrrogáveis.**

Faça uma excelente Prova!!

Parte a ser resolvida em sala

1. Seja $Y_1, ..., Y_n \sim$ Bernoulli(p_i), mutuamente independentes, $p_i = \Phi(\beta_0 + \beta_1 x_i)$ em que $\Phi(.)$ é a fda de uma normal padrão e x_i é uma covariável conhecida. Ou seja, tem-se o modelo de regressão probito usual com uma única covariável. Considere ainda a usual estrutura de aumento de dados, ou seja, $Y_i = \mathbb{1}_{(Z_i > 0)}, Z_i \sim N(\beta_0 + \beta_1 x_i, 1)$, mutuamente independentes. Responda os itens:

   a) Prove que a verossimilhança aumentada é dada por $L(\beta; \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(z_i - \beta_0 - \beta_1 x_i)^2\right\}$ a menos das funções indicadoras, em que $\boldsymbol{z} = (z_1, ..., z_n), \boldsymbol{y} = (y_1, ..., y_n), \boldsymbol{x} = (x_1, ..., x_n)$ (200 pontos).

   b) Obtenha à distribuição condicional de $Z_i | y_i, x_i, \boldsymbol{\beta}$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1)$ (200 pontos).

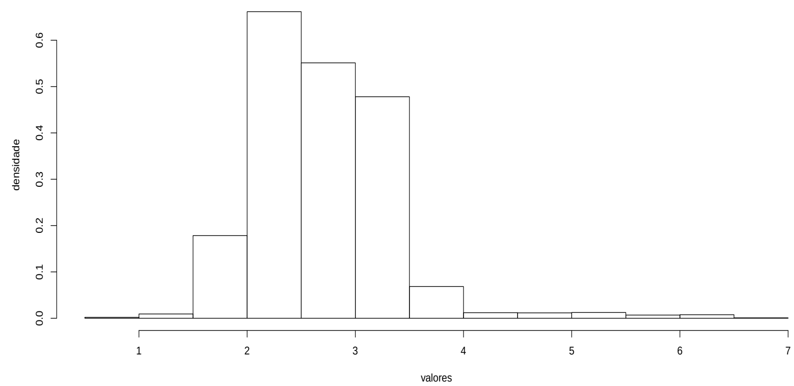
   c) Descreva a implementação do algoritmo EM, incluindo os desenvolvimentos dos Passos E e M. Você não precisa calcular as esperanças condicionais mas, precisa desenvolver as equações relativas ao Passo M (200 pontos).

2. Considere que está disponível um conjunto de valores (supostamente oriundos de uma distribuição exponencial de parâmetro λ na parametrização adotada em sala). Deseja-se estimar a mediana populacional através dos métodos de máxima verossimilhança e do bootstrap **não-paramétrico**. Você pode usar o fato de que o estimador de máxima verossimilhança de λ, neste caso é dado por $\widehat{\lambda} = \overline{X}$ e sua variância é dada por $\mathcal{V}(\widehat{\lambda}) = \frac{\lambda^2}{n}$. Responda os itens.

a) Prove que, neste caso, a mediana populacional é dada por: $\theta = Md(X) = \lambda \ln(2)$ (200 pontos).

b) Prove que, o estimador de máxima verossimilhança da mediana e o respectivo erro - padrão associados são dados por: $\widehat{\theta} = \overline{X}\ln(2)$ e $EP(\widehat{\theta}) = \frac{\lambda}{\sqrt{n}} \ln 2$. Sugestão: lembre-se da propriedade de invariância do emv (100 pontos).

c) Considere que, para uma determinado conjunto de $n = 30$ observações e selecionando-se R = 5000 amostras bootstrap (de modo não paramétrico) a estimativa bootstrap, o erro-padrão bootstrap e o intervalo de confiança bootstrap de 95% são dados por 2,69; 0,60 e [1,72;4.02] respectivamente. Considerando que $\overline{x} = 4,15$; compare, da forma mais completa possível, as estimativas de mv e bootstrap da mediana e diga qual das duas você usaria para fazer inferências com relação à esse parâmetro. Justifique, adequadamente, sua resposta. Abaixo segue o histograma das 5000 medianas calculadas para cada amostra bootstrap (300 pontos).



3. Seja $Y_1, ..., Y_n$ uma amostra de $Y_i \sim \text{Bernoulli}(p_i), i = 1, .., n$, mutuamente independentes. Considere que $p_i = F_\nu(\beta_0 + \beta_1 x_i)$, em que $F_\nu(.)$ denota a fda de uma distribuição t padrão com ν graus de liberdade (conhecido), x_i são covariáveis conhecidas e $\boldsymbol{\beta} = (\beta_0, \beta_1)$ é um vetor de parâmetros (desconhecidos). Responda os itens:

   a) Escreva a verossimilhança aumentada $L(\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, em função da representação estocástica da distribuição t padrão, através das distribuições normal e gama. Lembre-se de que a representação estocástica em questão é dada por (200 pontos):

$$\begin{aligned} Y_i &= \mathbb{1}_{(Z_i > 0)} \\ Z_i|\boldsymbol{\beta}, v_i, x_i &\sim N(\beta_0 + \beta_1 x_i, v_i^{-1}) \\ V_i|\nu &\sim \text{gama}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \end{aligned}$$

   b) Encontre as distribuições condicionais completas $Z_i|y_i, v_i, \boldsymbol{\beta}$, $\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{v}$ e $V_i|z_i, y_i, \boldsymbol{\beta}$, considerando que, à priori, $p(\boldsymbol{\beta}) \propto c$, em que c é uma constante. Você pode obter apenas os núcleos das condicionais completas e identificá-las sem calcular as constantes de normalização (300 pontos).

   c) Descreva um algoritmo MCMC geral para simular das condicionais completas (você não precisa dizer, necessariamente, qual algoritmo usaria para simular delas, apenas descrever a estrutura de simulação) (100 pontos).

Formulário

(a) Se $X \sim \text{binomial}(m, \theta), m \in I\!N, \theta \in (0,1)$ então, $f_X(x) = \dfrac{m!}{x!(m-x)!}\theta^x (1-\theta)^{m-x} \mathbb{1}_{\{0,1,2,\ldots,m\}}(x)$, $\mathcal{E}(X) = m\theta$, $\mathcal{V}(X) = m\theta(1-\theta)$. Fazendo-se m=1, tem-se a distribuição de Bernoulli.

(b) Se $X \sim exp(\theta), \theta > 0$, então $f_X(x) = \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{(0,\infty)}(x), \mathcal{E}(X) = \theta, \mathcal{V}(X) = \theta^2$.

(c) Se $X \sim t(\nu)$ então $f_X(x) = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathbb{1}_{(-\infty,\infty)}(x)$

(d) Se $X \sim N(\mu, \sigma^2)$, então $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}$ $\mu \in (-\infty, \infty), \sigma^2 \in (0, \infty), \mathcal{E}(X) = \mu, \mathcal{V}(X) = \sigma^2$.

(e) Se $X \sim \text{gama}(r, \lambda), r > 0, \lambda > 0$, então $f_X(x) = \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1} \mathbb{1}_{(0,\infty)}(x), \mathcal{E}(X) = r/\lambda, \mathcal{V}(X) = r/\lambda^2$.

Parte a ser resolvida em casa: entrega até dia 04/07/2012 às 18h00, junto com a resolução das questões acima. As resoluções das questão acima poderão ser entregues manuscritas. A parte a ser resolvida em casa deverá ser digitada e entregue impressa e por e-mail em PDF.

4. Com relação à Questão 1 acima (200 pontos):

   Implemente o algoritmo EM desenvolvido e o utilize para ajustar o modelo probito para os dados descritos na Tabela 3.12, página 222 do livro do Prof. Gilberto, Modelo de regressão e aplicações. Considere que Y_i é a resposta e X_i é a variável "razão". Não se esqueça de fornecer os erros-padrão associados às estimativas.

5. Com relação à Questão 3 acima (200 pontos): Implemente o algoritmo MCMC desenvolvido, no WinBUGS, e o utilize para analisar os dados mencionados na Questão 4, com o modelo de regressão descrito na Questão 3. Considere que $\nu = 13$. Na implementação no WinBUGS utilize prioris normais independentes para β_0, β_1, com médias iguais 0 e variâncias à sua escolha.

OBS: Será atribuida uma nota entre 0 e 10 para cada parte. As partes feitas em classe e em casa terão o mesmo peso na composição final da nota da prova.