Segunda Prova
Fundamentos Estatísticos para Ciência dos Dados

01/06/2017

1. O vetor aleatório $\mathbf{X} = (X_1, X_2, X_3)$ segue uma normal multivariada com valor esperado $\mu = (\mu_1, \mu_2, \mu_3) = (10, 20, -50)$ e com matriz de covariância

$$\mathbf{\Sigma} = \begin{bmatrix} 4 & 9 & -14 \\ 9 & 30 & -44 \\ -14 & -44 & 94 \end{bmatrix}$$

   - Obtenha a distribuição de $Y_1 = 0.3X_1 + 0.2X_2 + 0.5X_3$
   - Obtenha a distribuição conjunta de Y_1 e $Y_2 = (X_1 + X_2)/2$.
   - Obtenha a distribuição condicional da v.a. $(X_2|X_1 = 14)$

2. Uma análise de regressão linear de y, o comprimento do pé (em cm), em x, a altura (em cm) envolvendo indivíduos adultos do sexo masculino produziu a saída abaixo:

```
call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96172 -0.44332  0.02362  0.47609  1.34005

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.419020   0.956217   10.90   <2e-16 ***
x            0.093410   0.005652   16.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.654 on 98 degrees of freedom
Multiple R-squared:  0.736,Adjusted R-squared:  0.7333
F-statistic: 273.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

   - Um adulto tem altura igual a 1.80 *METROS*. Qual o tamanho predito para seu pé?
   - O adulto acima tem pé de tamanho igual a 28.50. Qual o valor do resíduo para este indivíduo?
   - Qual o valor da estatística R^2?

3. Mostre que a matriz de covariância

$$\mathbf{\Sigma} = \begin{bmatrix} 1.24 & 0.48 & 0.16 \\ 0.48 & 0.86 & 0.12 \\ 0.16 & 0.12 & 0.14 \end{bmatrix}$$

do vetor aleatório $\mathbf{X} = (X_1, X_2, X_3)^t$ pode ser gerado pelo modelo de análise fatorial com um único fator dado por

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ -1.0 \end{bmatrix} + \begin{bmatrix} 0.8 \\ 0.6 \\ 0.2 \end{bmatrix} F + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

Obtenha a matrix Ψ com as variâncias específicas dos termos ε_k implicadas pelo modelo.

4. O vento é uma força vetorial que está sempre mudando de direção e de intensidade (magnitude). Por causa, disso, sua medição num instante de tempo qualquer é tratada como uma variável aleatória. O modelo usualmente adotado para a intensidade X do vento (uma v.a. contínua) é a densidade

$$f(x) = \begin{cases} \frac{x}{\phi} e^{-x^2/(2\phi)}, & \text{se } x > 0 \\ 0, & \text{se } x \leq 0 \end{cases}$$

Estamos considerando apenas a magnitude escalar da intensidade do vento, ignorando a sua direção. Encontre o MLE de ϕ se uma amostra i.i.d. de medições $X_1, X_2, \ldots, X_n$ do vento é feita.

5. Considere o problema de regressão ou mínimos quadrados ponderados no qual você recebe um conjunto de dados $(\mathbf{x}_i, y_i, w_i)$ com $i = 1, \ldots, n$, onde w_i é um peso associado com o i-ésimo dado. O objetivo é minimizar a soma de quadrados dos resíduos ponderada (inversamente) por w_i:

$$\sum_i \frac{1}{w_i} (y_i - \mathbf{x}' \boldsymbol{\beta})^2$$

Forneça uma expressão para calcular os coeficientes $\boldsymbol{\beta}$ em forma fechada (não precisa deduzir, basta apresentar a expressão).

6. Responda T ou F com uma curta justificativa: Num problema de regressão linear com a notação usual:
   - A matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é a matriz de projeção no espaço das combinações lineares das colunas da matriz $\mathbf{X}$.
   - O vetor de resíduos $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ é ortogonal ao vetor dos coeficientes estimados $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
   - O vetor $\hat{\mathbf{Y}}$ é ortogonal ao vetor dos dados $\mathbf{Y}$.
   - $\hat{\mathbf{Y}}$ pertence ao espaço das combinações lineares das colunas de $\mathbf{X}$.
   - O vetor de resíduos $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$ pertence ao espaço das combinações lineares das colunas de $\mathbf{X}$.
   - Se $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, então $\mathbf{H}\mathbf{X} = \mathbf{X}$.
   - O R^2 sempre aumenta quando acrescentamos preditores ‘linearmente independentes na matriz de desenho $\mathbf{X}$.

7. Seja a matriz de covariância

$$\boldsymbol{\Sigma} = \begin{bmatrix} 25 & -12 \\ -12 & 9 \end{bmatrix}$$

do vetor aleatório $\mathbf{X} = (X_1, X_2)$ com média $\boldsymbol{\mu} = (0, 0)$. Os seus dois autovetores são $\mathbf{v}_1 = (-0.88, 0.47)$ e $\mathbf{v}_2 = (-0.47, -0.88)$ com respectivos autovalores $\lambda_1 = 31.4$ e $\lambda_2 = 2.6$. Usa-se o primeiro componente principal para reduzir a dimensionalidade do vetor $\mathbf{X}$ para uma dimensão $Y = \mathbf{v}_1'\mathbf{X}$. Se $\mathbf{x} = (3, 1)$ qual o valor do componente principal Y para este ponto?

8. QUESTÃO ADICIONAL, OPTATIVA, ALÉM DOS 20 PONTOS: VOCÊ GANHA 1 PONTO EXTRA SE FIZER ESTA QUESTÃO. A Desigualdade de Tchebyshev diz que, para qualquer variável aleatória X com $\mathbb{E}(X) = \mu$ e $\mathbb{V}(X) = \sigma^2$, temos $\mathbb{P}(|X - \mu| > k\sigma) \leq 1/k^2$. Use esta desigualdade para encontrar um limite superior para a probabilidade de que, ao jogar uma moeda honesta 100 vezes, o número de sucessos não difira de 50 por mais que 30 unidades. DICA: Seja X o número de sucessos. Obtenha μ e σ e use Tchebyshev.