1o. Trabalho Computacional
(TIP7077 – Inteligência Computacional Aplicada)

Programa de Pós-Graduação em Engenharia de Teleinformática (PPGETI)
Departamento de Engenharia de Teleinformática (DETI)
Universidade Federal do Ceará (UFC)

Responsável: Prof. Guilherme de Alencar Barreto
Data: 15/04/2015 - Data de entrega: 30/04/2015

Questão 1 (Regressão Linear Simples e validação de modelo):

Por volta da metade do século 19, o físico escocês James D. Forbes estudou a relação entre pressão e pontos de ebulição. O objetivo do seu trabalho era (entre outras coisas) estimar a altitude acima do nível do mar ao medir o ponto de ebulição da água (barômetros era frágeis e difíceis de transportar naquele tempo). Suas próprias medições a partir dos Alpes e da Escócia foram as seguintes:

Location	1	2	3	4	5	6	7	8
Boiling point (°F)	194.5	194.3	197.9	198.4	199.4	199.9	200.9	201.1
Pressure (inches Hg)	20.79	20.79	22.40	22.67	23.15	23.35	23.89	23.99

9	10	11	12	13	14	15	16	17
201.4	201.3	203.6	204.6	209.5	208.6	210.7	211.9	212.2
24.02	24.01	25.14	26.57	28.49	27.76	29.04	29.88	30.06

Fonte: Weisberg (1985): Applied Linear Regression; data from Forbes, J. D. (1857): Further experiments and remarks on the measurement of heights by the boiling point of water, Trans. R. Soc. Edinburgh 21, 135–143.

Pede-se para analizar estes dados e responder aos seguintes itens:

a) Formular um modelo estatístico (escolha das variáveis dependentes e independentes).

b) Validar o modelo verificando a distribuição dos resíduos via coeficiente de determinação (R2), histograma e boxplot. Existe alguma medida discrepante (i.e. outlier)? A relação linear é apropriada? Se não, verifique a necessidade de se fazer ajustes para obter um bom modelo.

c) Estimar os parâmetros do modelo escolhido no Item a) e estimar a variância do ruído a partir dos resíduos.

d) A partir de fundamentos teóricos, Forbes escolheu analisar a dependência do logaritmo da pressão em relação ao ponto de ebulição. Conduzir esta análise também, e comparar os resultados com a análise anterior.

Questão 2 (Regressão Linear Múltipla e validação de modelo):

O conjunto de dados **bodyfat** do repositório Statlib fornece estimativas da porcentagem de gordura corporal determinada a partir de pesagem debaixo d'água e várias medidas da circunferência de várias partes do corpo de 252 homens. É interessante desenvolver uma equação para estimar a gordura corporal em função da medida da circunferência abdominal (vide a descrição do conjunto de dados e o contexto biológico em http://lib.stat.cmu.edu/datasets/bodyfat).

A variável de saída é a porcentagem de gordura corporal estimada a partir da equação de Siri (%), e as variáveis preditoras de interesse são idade (anos), peso (libras), altura (polegadas) e circunferências (todas em cm) do pescoço, abdômen, cintura, coxa, joelho, calcanhar, bíceps, antebraço e punho.

1) Usar os dados para determinar um bom modelo para estimar a porcentagem de gordura corporal a partir das outras variáveis (exceto densidade corporal – *body density*).

2) Avaliar as suposições do modelo a partir de uma análise cuidadosa dos resíduos. Avaliar se qualquer uma das observações deve ser eliminada da análise (i.e. se há outliers e se estes devem ser eliminados do conjunto de dados usado para construir o modelo).

3) Executar uma análise de sensibilidade do modelo a fim de determinar quais das variáveis preditoras são mais relevantes (tem mais importância) na estimação da gordura corporal.

Boa Sorte!