ESCOLA DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE MATEMÁTICA

Segundo Trabalho de Análise Categórica de Dados (2017/18)
Mestrado em Modelação Estatística e Análise de Dados

Entrega até dia 22 de Junho
Parte II: Apresentação e discussão no dia 22 de Junho às 11h na sala 133 CLV

Parte I

1. O ficheiro **loja.txt** contém os dados sobre o perfil dos clientes de uma determinada loja oriundos de 110 zonas de uma cidade. O objetivo do estudo é relacionar o número de clientes em cada zona com as seguintes variáveis explicativas em cada zona: número de habitações (em milhares), rendimento médio anual (em €), idade média das habitações (em anos), distância entre a área e a loja concorrente mais próxima (em Km) e a distância entre a zona e a loja (em Km).
   a) Constate a não adequabilidade do modelo normal.
   b) Ajuste um modelo de regressão de Poisson e interprete os seus coeficientes. Quais as variáveis que melhor explicam o número esperado de clientes em cada zona?
   c) Verifique a bondade do ajustamento obtido e realize uma análise de resíduos.
   d) Estime o número esperado de clientes numa área em que o número de domicílios é de 500 milhares, o rendimento médio anual é de 38000€, a idade média das habitações é de 45 anos, a distância da zona à loja concorrente mais próxima é de 5 Km e a distância entre a zona e a loja é de 7 Km.

2. A Tabela 1 apresenta um conjunto de dados em que pacientes com leucemia foram classificados segundo a ausência ou presença de uma característica morfológica nos glóbulos brancos. Pacientes classificados de AG positivo foram aqueles com a presença da característica e pacientes classificados de AG negativo não apresentaram a característica. É apresentado também o tempo de sobrevivência do paciente (em semanas) após o diagnóstico da doença e o número de glóbulos brancos (WBC) no momento do diagnóstico. Supondo que o tempo de sobrevivência após o diagnóstico segue uma distribuição Gama, proponha um modelo para explicar o tempo médio de sobrevivência em função de log(WBC) e AG(=1 positivo, =0 negativo). Emita um relatório para a entidade interessada com as principais conclusões que o modelo lhe permitiu retirar. (Não esquecer de fazer uma análise de diagnóstico com o modelo ajustado e interprete as estimativas).

Tabela 1

AG Positivo		AG Negativo	
WBC	Tempo	WBC	Tempo
2300	65	4400	56
750	156	3000	65
4300	100	4000	17
2600	134	1500	7
6000	16	9000	16
10500	108	5300	22
10000	121	10000	3
17000	4	19000	4
5400	39	27000	2
7000	143	28000	3
9400	56	31000	8
32000	26	26000	4
35000	22	21000	3
100000	1	79000	30
10000	1	100000	4
52000	5	100000	43
100000	65		

Parte II: Apresentação e discussão no dia 22 de Junho às 11h na sala 133 CLV

Deverão preparar uma apresentação para 20/25 minutos sobre os seguintes temas:

Francisco – Regressão logística em amostras emparelhadas

João – Modelos lineares mistos generalizados

Jéssica – Regressão não paramétrica (*splines* e *wavelets*)

Fábio – Modelos aditivos generalizados (GAM)

Ana – Modelos inflacionados de Zeros

Dário – Modelo de regressão binomial negativa

Bom trabalho!