ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

INTRODUÇÃO AO APRENDIZADO ESTATÍSTICO

TRABALHO 1

Assuma que a variável aleatória *Y* denote custos de cancelamento de contratos (em milhares de Reais) registrados por uma empresa. O arquivo **custos.txt** contém 1000 registros de custos de cancelamento, selecionados aleatoriamente da base de dados da empresa. Na primeira coluna, estão os custos na escala original e, na segunda coluna do arquivo, tem-se o logaritmo natural de cada custo, que será usado como veremos adiante.

Obs: Você pode ler o arquivo de dados e conferir se a leitura foi correta usando as linhas de comando R:

```
dados<-matrix(scan(file="custos.txt"),ncol=2,byrow=T) #Leitura do arquivo
head(dados) #Visualização das primeiras linhas do arquivo lido
dim(dados) #Conferindo a dimensão dos do arquivo
```

ETAPA 1 (20%): Especificação de um modelo observacional

(a) Construa um histograma dos custos de cancelamento na sua escala original, ou seja, os dados contidos na primeira coluna do arquivo.

(b) Admita que se deseje modelar os custos de cancelamento usando um modelo Lognormal(μ, σ^2). Justifique brevemente a escolha desse modelo, com base na visualização do histograma do item (a). Caso não conheça o modelo Lognormal, faça uma pesquisa sobre o aspecto de sua função densidade de probabilidade. Enuncie pelo menos um outro modelo probabilístico que você julga que poderia ser usado para descrição desses dados, justificando.

A partir de agora, nos fixaremos no estudo do modelo Lognormal(μ, σ^2) para os custos de cancelamento.

Obs: Dizer que Y (custos) tem distribuição Lognormal(μ, σ^2) é equivalente a dizer que Z=ln(Y) tem distribuição Normal(μ, σ^2). Passamos a estudar a informação contida nos dados com respeito a esses dois parâmetros, μ *e* σ^2, que são idênticos no modelo Lognormal ou Normal.

ETAPA 2 (40 %): Obtenção numérica de estimas de máxima verossimilhança e seu uso para alimentar o modelo probabilístico:

(c) Obtenha estimativas de máxima verossimilhança para μ *e* σ^2.

Obs1: Utilize, a título de exemplo, o código R apresentado em aula para obtenção de estimativas de máxima verossimilhança para o caso Gama e adapte-o para o problema que estamos tratando. Considere verossimilhança Normal, alimentada pelos dados transformados Z=ln(Y) (segunda coluna da base de dados).

Obs2: No R, as distribuições Normal e Lognormal são parametrizadas por σ e não por σ^2, então a estimativa que você vai obter é para σ (desvio-padrão da Normal). Esteja atento para esse detalhe.

Obs3: Você precisará informar valores iniciais para a rotina de otimização da verossimilhança. Inicialize o algoritmo de otimização com par de valores (1,1). Repita o procedimento algumas vezes, fazendo uma análise de sensibilidade dos seus resultados ao valor inicial do algoritmo de otimização.

(d) Utilizando o modelo observacional com valores dos parâmetros substituídos por suas estimativas de máxima verossimilhança, determine a probabilidade de que um custo de cancelamento de contrato futuro, nessa empresa, seja superior a 9 (mil Reais).

ETAPA 3 (40%): Estudo visual do comportamento da função de verossimilhança:

Nesta etapa, desejamos visualizar o comportamento do gráfico da função de verossimilhança perfilada para μ. Note que se trabalharmos com os dados transformados Z=ln(Y), teremos verossimilhança obtida a partir da densidade normal, cuja forma analítica é bem conhecida. Expresse analiticamente a função de verossimilhança, notando que esta pode ser escrita em função das estatísticas $t1 = \sum_{i=1}^{n} z_i$ e $t2 = \sum_{i=1}^{n} z_i^2$. Então, a exemplo do que foi feito em aula, escreva a expressão da função de verossimilhança em função de $t1$ e $t2$. Note que essas duas estatísticas contêm toda a informação necessária, vinda dos dados, para avaliarmos a função de verossimilhança (são chamadas estatísticas suficientes – para estimar os parâmetros de interesse, não precisamos guardar toda a amostra observada – é suficiente guardar os valores dessas estatísticas). Agora, tome os dados transformados *Z*, disponíveis na segunda coluna do arquivo custos.txt e, com esses dados, no R:

(e) Faça uma grade para o parâmetro μ , com valores variando de 0 a 5.
(f) Faça o gráfico da função de verossimilhança normal (alimentada pelos dados transformados, Z), contra a grade de valores para μ, fixando o valor de σ^2 na estimativa de máxima verossimilhança obtida na etapa 2.
(g) Comente o gráficos e relacione seus comentários à estimativa de μ obtida na etapa 2.

Obs: Fizemos, em aula, gráficos exploratórios do comportamento da função de verossimilhança para os modelos Bernoulli e Poisson. Os códigos R usados naqueles exemplos podem ser facilmente adaptados para a execução da etapa 3.