CEFET/RJ
Bacharelado em Ciência da Computação
Inferência Estatística - Trabalho 04

Prof. Eduardo Bezerra (ebezerra@cefet-rj.br)

9/6/2018

Conteúdo

1 Diagramas de Dispersão

Nessa parte, você irá realizar testes com o conjunto de dados `trees`. Para visualizar as primeiras linhas desse conjunto, assim como para obter documentação associada, utilize os comandos abaixo no `R`:

```
head(trees)
help(trees)
```

Agora, você deve gerar diagramas de dispersão (também com o `R`) para visualizar eventuais dependências entre as variáveis deste conjunto. Para isso, utilize os comandos abaixo.

```
plot(trees$Volume~trees$Height, main = 'Black Cherry Tree
    Volume Relationship', xlab = 'Height', ylab = 'Volume',
    pch = 16, col = 'blue')
plot(trees$Volume~trees$Girth, main = 'Black Cherry Tree
    Volume Relationship', xlab = 'Girth', ylab = 'Volume',
    pch = 16, col = 'blue')
```

Ao analisar apenas os diagramas de dispersão produzidos, quais os tipos de dependência existentes em cada um dos dois pares de variáveis? Justifique sua resposta.

2 Correlação

Agora, você deve implementar uma função no `R` denominada `coefPearson`. Essa função deve receber dois vetores coluna com dados de duas variáveis X e Y e retornar o valor do coeficiente de correlação de Pearson correspondente.

Na implementação desta função, você deve necessariamente utilizar a expressão matemática apresentada a seguir. Considere que os valores da primeira variável X são $\{x_1, \ldots, x_n\}$ e que os valores da outra variável Y são $\{y_1, \ldots, y_n\}$. Considere também que $\bar{x}$ e $\bar{y}$ são as médias das duas variáveis X e Y, respectivamente. Então sua função deve computar a seguinte fórmula para cálculo do coeficiente de correlação de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Você pode verificar a corretude de sua implementação comparando os resultados que sua função produz com os resultados produzidos pelo comando `cor` do `R`.

Os valores produzidos pela sua função confirmam a sua análise prévia, feita apenas por meio dos diagramas de dispersão? Justifique sua resposta.

3 Regressão Linear

Nessa parte do trabalho, você irá implementar a regressão linear para predizer o lucro para uma cadeia de *food truck*. Essa cadeia já possui diversas filiais em diferentes cidades. Você possui dados do lucro e população para cada uma dessas cidades.

O arquivo `ex1data1.txt` contém os dados a serem usados nessa parte do trabalho. A primeira coluna corresponde à população de cada cidade, enquanto que a segunda coluna corresponde ao lucro da filial daquela cidade. Um valor negativo para o lucro indica que a filial correspondente está dando prejuízo.

3.1 Visualização dos Dados

Para a maioria dos conjuntos de dados do mundo real, não é possível criar um gráfico para visualizar seus pontos. Mas, para o conjunto de dados fornecido, isso é possível. Gere um *gráfico de dispersão* (*scatter plot*) dos dados fornecidos. Aqui, podem ser úteis os comandos já fornecidos na parte 1 deste trabalho.

3.2 Ajuste do Modelo Linear

Nessa parte, sua tarefa é determinar os parâmetros do modelo de regressão linear por meio do comando `lm` do `R`. Apresente os valores dos coeficientes de regressão produzidos. Em seguida apresente outro diagrama de dispersão, dessa vez apresentando também a linha de regressão produzida.

Como última tarefa nessa parte do trabalho, você deve usar o modelo de regressão linear produzido pelo seu código para predizer o lucro em regiões com populações de 35.000 e 70.000 habitantes. Forneça no seu relatório o código (em `R`) para isso, assim como os valores correspondentes do lucro para cada uma daquelas duas cidades.