Trilha 04

Indicações de Soluções das Atividades de Aprofundamento

Atividade de Aprofundamento

A atividade de fixação desta trilha será a resolução de alguns problemas de Análise de Regressão Simples utilizando o **R**.

Problema 1

Para uma amostra de oito operadores de máquina, foram coletados o número de horas de treinamento (x) e o tempo necessário para completar o trabalho (y). Os dados coletados encontram-se na tabela abaixo:

x	y
5.2	13
5.1	15
4.9	16
4.6	20
4.7	19
4.8	17
4.6	21
4.9	16

Com estes dados, faça:

a) Um gráfico de dispersão para os dados

<< Você pode usar a função de PLOT do R para fazer o gráfico de dispersão >>

<< Pagina 26 e 27 ajuda a resolver >>

b) Determine o modelo de regressão linear simples entre as variáveis x e y, sendo y a variável resposta.

<< Você pode atribuir o resultado da função da regressão à uma variável, e então, apresentar os dados desta regressão (a partir da análise da variável) >>

<< Pagina 36 ajuda a resolver >>

c) Faça uma análise do modelo de regressão utilizando a função **summary**:

i) resíduos, significância estatística dos coeficientes, percentual de variância explicada pelo modelo.

<< Explicar o comportamento do resultado que encontrou a partir do summary, em forma textual. Também é esperado que a plotagem dos gráficos gerados a partir da regressão seja apresentado. Faça a explicação dos resultados visíveis nos gráficos. >>

<< Pagina 40 ajuda a resolver >>

d) Trace, no gráfico anterior, a reta de regressão.

<< imprimir o gráfico de dispersão, porém, colocando a reta >>

<< Pagina 26 e 27 ajuda a resolver >>

Problema 2

O site Gapminder compilou uma base com dados sobre população, expectativa de vida e PIB per capta de 142 países, em 12 anos diferentes.
Com estes dados, disponibilizados no arquivo `pib_gapminder.csv`, faça:

a) Faça a importação dos dados, verifique a estrutura e faça um sumário estatístico.
<< Já fizemos isso algumas vezes, o sumário deve ser algo automático já >>

b) Verifique a estrutura dos dados (`str`)
<< entender a estrutura dos dados com o STR já é algo corriqueiro neste momento >>

c) Classifique cada variável de acordo com seu tipo (qualitativa ordinal, nominal, quantitativa discreta, contínua, etc).
<< Já fizemos isso no trabalho desenvolvido na trilha 1 >>

d) Faça um sumário estatístico dos dados
<< Já foi feito no item A >>

e) Faça uma tabela de frequência absoluta e uma tabela de frequência relativa para verificar o número de observações por continente.
<< Com a função Table você consegue pegar a frequência das ocorrências na variável específica. Você precisa entender a diferença das frequências absolutas e relativas para calcular. É mais fácil armazenar cada uma em um novo Data Frame e fazer as operações. >>

f) Faça um gráfico de barras da tabela de frequência absoluta dos continentes.
<< Imprimir um gráfico já não é mais novidade. Você consegue fazer com o BarPlot ou com o GGplot2. O que lhe for mais conveniente >>

g) Faça um gráfico apropriado para relacionar o PIB per capta à expectativa de vida.
<< Aqui será um gráfico de dispersão, onde terão 2 eixos, um sendo o PIB e o outro a Expectativa. Você pode usar cores diferentes para os continentes analisados >>

h) Crie duas novas colunas nesta base de dados com o logaritmo de PIB per capta, e o logaritmo da expectativa de vida. Estas colunas devem ter os nomes: lpibPercap e lexpVida, respectivamente.

<< O cálculo do LOG já foi visto antes também, e acredito que é fácil de aplicar simplesmente colocando essas novas colunas com o resultado do LOG >>

i) Faça um gráfico apropriado para relacionar estas duas novas variáveis.

<< Da mesma forma que fez o gráfico de dispersão na atividade G, você pode trocar as variáveis dos eixos X e Y pelas novas variáveis e resolver este item >>

j) Ajuste um modelo linear aos dados, utilizando as duas novas variáveis criadas, sendo lexpVida a variável resposta.

<< Crie uma nova regressão linear utilizando a variável lpibPercap (independente) como explicação da variável lexpVida (dependente). >>

<< Na página 30 pode acompanhar a sintaxe na qual a variável independente é a tannin e a dependente é a growth (ou seja, tannin explica growth) >>

k) Faça todas as análises da regressão, julgando:
   i) Os gráficos diagnósticos
   ii) Os parâmetros obtidos (avaliar o summary do modelo)
   iii) O poder de explicação do modelo.

<< Você já fez algumas explicações do modelo no item C do Problema 1. Aqui é similar, mas com o novo modelo. >>

Problema 3

Neste exercício vamos fazer uma análise de regressão com a base de dados `autos.csv` para tentar prever o preço de carro a partir de sua potência. Na nossa base de dados, estas variáveis são horsepower e price.

Utilizando então a base disponibilizada, você deve:

a) Criar um dicionário de dados, para entender o significado o tipo de cada variável; veja no link fornecido se há documentação disponível.

<< A URL é esta: http://archive.ics.uci.edu/ml/datasets/Automobile >>

<< Dicionário de dados já foi trabalhado na trilha 1 >>

b) Carregar a base para o R, certificando-se de que os dados estão corretos, de acordo com o dicionário de dados.
<< Já fizemos isso algumas vezes >>

c) Explore a base de dados:
   i) Sumários estatísticos dos dados
   ii) Tabelas quando apropriado
   iii) Gráficos exploratórios apropriados.

<< Também já fizemos atividades solicitadas no item C algumas vezes >>

d) Considerando então apenas uma variável preditora (explicativa) no modelo (horsepower), tente ajustar um modelo para explicar o preço (price) dos carros:
   i) Qual sua intuição sobre o relacionamento da "potência" de um carro com o seu preço?

<< Crie um modelo de Regressão Linear com as variáveis HorsePower e Price, sendo que a HorsePower é a variável explicativa e a Price é a resposta. Com isso, você responderá a pergunta sobre o relacionamento da potência (que é o HorsePower) com o preço >>

e) Faça todas as análises da regressão (modelo), julgando:
   i) Os gráficos diagnósticos
   ii) ii) Os parâmetros obtidos (avaliar o summary do modelo)

<< Você já fez algumas explicações do modelo no item C do Problema 1 e no K do Problema 2. Aqui é similar, mas com o novo modelo. >>

f) Interprete os resultados do ajuste:

i) Em que posição a reta corta o eixo Y? Isso faz sentido?

<< Você deve observar os coeficientes do modelo gerado. O que é esperado é o Intercept >>

ii) ii) Como corrigir um modelo que apresenta este comportamento?

<< Qual é a sua abordagem para corrigir o problema?? >>

g) Analise: Será que apenas a potência de um carro é suficiente para termos uma boa previsão do preço deste carro?

i) O que indica isso no seu ajuste?

<< Exponha sua análise em forma de texto, do que acredita sobre esta frase. Após sua resposta, explique como esta resposta interage com o seu ajuste >>