COE241 - Estatística e Modelos Probabilísticos
Segundo Semestre de 2018 - Professora: Rosa Maria Meri Leão

Projeto do Curso

1 Objetivo

O objetivo deste trabalho é estudar um conjunto de dados aplicando a teoria aprendida em classe. É muito importante que seja realizada uma análise crítica dos resultados encontrados.

Iremos utilizar dados reais fornecidos gentilmente pelo Professor Claudio Gil Soares de Araujo (até recentemente professor do Instituto do Coração Edson Saad da UFRJ) da CLINIMEX, através da aluna de doutorado da UFRJ Christina G. de Souza e Silva. Os dados foram obtidos a partir de uma extensa base de dados do Prof. Claudio Gil, coletada durante muitos anos e usada em suas pesquisas. Os dados mostram uma medida da condição aeróbica do paciente (o VO_2 max) (por quilo de peso do indivíduo) e ainda as variáveis idade, peso e a carga máxima atingida durante um teste ao qual o paciente foi submetido. Todos os pacientes são masculinos. O VO_2 max é a taxa máxima de consumo de oxigênio medida durante um teste de esforço, e reflete a capacidade aeróbica do paciente, expressa em volume de oxigênio por massa corporal por minuto (ml/(Kg.min). É uma importante métrica usada na avaliação cardiovascular de indivíduos. As características dos dados fornecidos são: idade do paciente, peso (kg), carga final (watts) e VO_2 máximo (mg/Kg/min). Fornecemos um (1) arquivo com dados de 1.172 pacientes coletados pelo Professor e sua equipe: Dados-medicos.csv.

2 Análises a serem realizadas

Inicialmente devem ser obtidas estatísticas preliminares de cada uma das variáveis. Em uma segunda etapa deve-se estudar se as variáveis podem ser representadas por distribuições da literatura. Posteriomente, deve-se analisar possíveis dependências entre as variáveis e obter um modelo que permita prever o VO_2 máximo a partir das outras variáveis coletadas.

2.1 Histograma e Função Distribuição Empírica

Calcular o histograma e a função distribuição empírica para as seguintes variáveis: **idade**, **peso**, **carga final**, VO_2 **máximo**.

Lembre-se que o tamanho do *bin* deve ser estimado de forma que o histograma represente de forma adequada os dados da população. Comente o que você observou a partir dos resultados obtidos.

2.2 Média, Variância e Boxplot

Calcular a média, variância e construir o gráfico BoxPlot para as seguintes variáveis: **idade**, **peso**, **carga final**, VO_2 **máximo**.

Comente o que você observou a partir dos gráficos. É importante interpretar os resultados obtidos.

2.3 Parametrizando distribuições

Neste item o objetivo é parametrizar um conjunto de distribuições da literatura para as variáveis coletadas dos pacientes. Após a parametrização você irá verificar se alguma das variáveis pode ser representada por uma distribuição da literatura. Utilize o método da máxima verossimilhança para estimar os paramêtros das seguintes distribuições: exponencial, gaussiana, lognormal, weibull.

Você deve descrever no seu relatório como foram calculados os parâmetros das distribuições usando o método da máxima verossimilhança. Ou seja, as expressões analíticas usadas para a estimativa dos parâmetros.

As variáveis que devem ser consideradas são: **idade**, **peso**, **carga final**, VO_2 **máximo**. Você deve obter para cada dessas variáveis o valor dos parâmetros das distribuições citadas acima. Após a obtenção dos valores dos parâmetros, você deve fazer um gráfico para cada uma das variáveis aleatórias com a função distribuição empírica e as quatro distribuições que você parametrizou. Observando o gráfico você deve identificar se existe ou não uma das distribuições da literatura que pode ser usada para representar a variável aleatória. O gráfico ProbabilityPlot e o teste de hipótese solicitados nos itens abaixo vão ajudar você a escolher a distribuição que melhor representa os dados coletados.

Se você achar que nenhuma das distribuições que você parametrizou representa de forma adequada os dados dos pacientes, você pode escolher uma outra distribuição da literatura que seja mais adequada e parametrizá-la com os dados dos pacientes.

2.4 Gráfico QQplot ou ProbabilityPlot

Os gráficos QQplot ou ProbabilityPlot servem para analisar se duas variáveis aleatórias possuem a mesma distribuição. Você deve traçar os gráficos abaixo para verificar se as distribuições que você parametrizou se adequam as variáveis dos pacientes.

1. **idade** x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).
2. **peso** x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).
3. **carga final** x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).
4. VO_2 x cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull).

Comente o que você pôde observar a partir dos resultados.

2.5 Teste de Hipótese

Uma outra forma de você verificar se uma determinada variável pode ser representada por uma distribuição de probabilidade é formulando um teste de hipótese. Formule um teste para cada das variáveis dos pacientes **idade**, **peso**, **carga final**, VO_2 **máximo**, em relação a cada uma das distribuições parametrizadas (exponencial, gaussiana, lognormal, weibull). Você deve usar o teste Komolgorov-Smirnov.

Comente o que você pôde observar a partir dos resultados dos testes.

2.6 Análise de dependência entre as variáveis, modelo de regressão

.

Você deve calcular o coeficiente de correlação amostral e fazer o gráfico scatter plot para as seguintes variáveis: **idade e** VO_2 **máximo**, **peso e** VO_2 **máximo**, **carga final e** VO_2 **máximo**. Você deve analisar os resultados e indicar se existe alguma dependência entre as variáveis acima. É possível usar o modelo de regressão para representar a relação entre alguma das variáveis e o VO_2 **máximo** ? Caso a resposta seja afirmativa, calcule os parâmetros do modelo de regressão.

2.7 Inferência Bayesiana

No item acima você calculou o coeficiente de correlação amostral entre as variáveis analisadas. Escolha o par de variáveis que possuem o maior coeficiente de correlação amostral e construa a tabela de inferência bayesiana conforme apresentado em aula. A tabela deve conter cinco colunas: as hipóteses, a pmf da prior, a likelihood, o numerador da fórmula de Bayes e a pmf da posterior.

As hipóteses são os valores da variável escolhida: **idade** ou **peso** ou **carga final**. Você pode dividir esses valores em faixas. A pmf da prior é a distribuição empírica da variável escolhida (você já calculou no primeiro item do trabalho). A likelihood pode ser obtida dos dados dos pacientes, é a probabilidade de obter um certo valor de VO_2 **máximo** para uma dada faixa de valores da variável escolhida. Você deve considerar duas possíveis faixas de valores para a variável VO_2 **máximo**. A primeira faixa é VO_2 **máximo** $<$ **35** indicando que o paciente está abaixo da média de um valor considerado normal para homens. A segunda faixa é VO_2 **máximo** $\geq$ **35** indicando que o paciente está na média ou acima da média, de um valor considerado normal para homens. O numerador da fórmula da Bayes e a pmf da posterior você calcula com os dados das outras colunas da tabela. A distribuição da posterior vai indicar qual a probabilidade de uma dada hipótese (valores da variável escolhida: **idade** ou **peso** ou **carga final**) dada uma certa faixa de VO_2 **máximo**.

ATENÇÃO: você deve construir duas tabelas, uma para cada faixa de VO_2 **máximo**. Uma tabela para VO_2 **máximo** $<$ **35** e outra para VO_2 **máximo** $\geq$ **35**.

Agora você deve fazer uma previsão usando as tabelas que você construiu. Considere a tabela que você construiu para VO_2 **máximo** $<$ **35**. Faça uma previsão de uma melhora no VO_2 **máximo** de uma pessoa que possui um VO_2 **máximo** abaixo da média. Ou seja, você deve calcular P [VO_2 **máximo** $\geq 35|$ VO_2 **máximo** < 35] usando os dados das tabelas que você construiu.

3 Relatório

Você deve fazer um relatório contendo todos os resultados que você obteve e explicando como você os obteve. É importante comentar cada um dos resultados e explicar como o resultado que você obteve poderá auxiliar os médicos no estudo dos pacientes. A avaliação do projeto será feita com base na qualidade do relatório.

Você deve entregar o seu relatório impresso em pdf. No relatório deve estar indicado um link para o código que você usou para obter os resultados do trabalho.