ACH2036 - Métodos Quantitativos para Análise Multivariada

Atividade 1: Aplicação de Modelos de Regressão

O objetivo desta atividade é apresentar um relatório da aplicação de Regressão Linear Múltipla sobre uma base de dados. Você pode utilizar dados secundários do IBGE, IPEA ou outros órgãos, ou ainda datasets públicos disponíveis em repositórios públicos. Seguem abaixo alguns exemplos de repositórios (mas você pode pesquisar em outros que conhecer):

- **Ipeadata**: http://ipeadata.gov.br/Default.aspx
- **UCI Machine Learning Repository**: https://archive.ics.uci.edu/ml/index.php
  Este repositório é um dos mais conhecidos e possui um conjunto enorme de bases de dados.
- **KDnuggets**:
  http://www.kdnuggets.com/datasets
  https://www.kdnuggets.com/2015/04/awesome-public-datasets-github.html
- **Ferramenta Google Dataset Search**: https://toolbox.google.com/datasetsearch
- **Kaggle**: https://www.kaggle.com/datasets
- **Microsoft Research Open Data**: https://msropendata.com

Se tiver interesse em utilizar dados relacionados ao seu estágio/trabalho/iniciação científica, é perfeitamente possível – mas fica sob sua responsabilidade solicitar seu(sua) superior(a) ou orientador(a) a autorização do uso desses dados. Isso é especialmente crítico se envolver dados pessoais de terceiros ou dados críticos da empresa.

Recomendação para a escolha da base:

- Naturalmente, a base deve possuir uma variável resposta contínua.

  Se a variável resposta for limitada dentro de um intervalo, poderá ser necessária a aplicação de alguma transformação. Por exemplo, para variáveis limitadas ao intervalo $(0, 1)$, uma transformação possível (mas não a única) é a transformação logito (já apresentada em aulas anteriores):

$$y_{trans} = \log\left(\frac{y}{1-y}\right) = log(y) - log(1-y)$$

  Atenção: Esta transformação só vale dentro do intervalo aberto $(0, 1)$. Se houver valores $y = 0$ ou $y = 1$, você deve "truncar" esses valores para ficarem dentro do intervalo aberto: $y = 0 \rightarrow y = \epsilon$ e $y = 1 \rightarrow y = 1 - \epsilon$, onde ϵ é uma constante pequena (p.ex. 10^{-3} ou 10^{-4}). Naturalmente, variáveis descritas em percentuais (entre 0% e 100%), a adaptação é imediata, bastando dividi-las por 100 e aplicando a transformação acima.

- O ideal é que a base contenha pelo menos cinco variáveis preditoras. Se o número for muito maior do que isso (por exemplo, mais do que 10), algumas variáveis poderão ser eliminadas com a estratégia comentada mais adiante.

  Se sua base de escolha é referente a um estudo prévio de seu grupo de pesquisa e não contém este número mínimo de variáveis, procure-me para conversar no final da aula, para que possamos discutir como tentar enriquecer as análises.

- Algumas bases contêm identificadores dos registros (código de paciente, código da amostra, etc). Eles até podem ser mantidos por você para facilitar eventual identificação de registros anômalos (com outliers, com dados faltantes etc), mas raramente são utilizados em análises estatísticas e não deverão ser utilizados como variáveis preditoras. Leve isso em consideração no item acima.

- Idealmente, ao menos uma das variáveis preditoras deve seja categórica, para permitir a você explorar esse tipo de variável em um modelo de regressão. Se não for possível utilizar uma base atendendo esta recomendação, você

deverá explicar isso no relatório (ou seja, por que razão a base escolhida era importante, mesmo não contendo variáveis categóricas).

Relatório:

Você deverá apresentar um pequeno relatório, cuja estrutura de seções é a seguinte:

1. *Título e nomes dos integrantes do grupo*
   (2 ou 3 integrantes, que podem ser de turmas distintas)

2. *Problema de regressão a ser tratado*

   Inicie com um ou dois parágrafos gerais sobre a base utilizada. Se tiver sido obtida de um repositório, explicitar o nome do repositório e o nome da base dentro deste repositório.

   Em seguida, apresente um breve enunciado do problema de regressão a ser resolvido.

   Um exemplo (bastante simplório) seria: "O objetivo original nesta base de dados é avaliar a associação entre o consumo de um veículo (em Km/L) e algumas de suas características, como potência do motor, peso do veículo, entre outras."

   (Você verá que, mesmo nos repositórios que mencionamos, alguns datasets disponibilizados contêm uma descrição implícita ou explícita do problema de regressão.)

   Opcionalmente, você pode complementar o texto explicando a importância do problema, ou por que ele motivou sua escolha.

3. *Descrição da base*

   3.1 Apresente um dicionário com descrições de cada uma das principais variáveis. Idealmente, a descrição de cada variável deve conter:

      a) Abreviação da variável (tente usar nomes curtos ou abreviações que permitam uma fácil identificação; evite nomes codificados (V01, V02, etc).

b) Nome original da variável, se achar conveniente, e sua descrição sucinta

c) Tipo da variável – contínua/inteira/categórica ordinal/categórica não ordinal.

Nota: Por conveniência, em algumas bases é comum que as variáveis categóricas sejam codificadas em números (1,2, etc). Mas a natureza dessas variáveis permanece categórica e elas não devem ser tratadas nas análises estatísticas como variáveis numéricas. Utilize os nomes das categorias (mesmo que abreviados) em vez dos códigos numéricos!

d) Se a variável necessitou de algum tratamento de ajuste de escala ou reagrupamento de níveis, comente. (Ver o item 4 abaixo para maiores esclarecimentos.)

3.2 Se você identificou a necessidade de criação de uma ou mais variáveis derivadas a partir das variáveis originais, inclua uma descrição dessas variáveis também.

4. Análise exploratória das principais variáveis (obs: Tópico em construção)

Nesta seção, você apresentará gráficos ou tabelas para visualização das principais variáveis. Esta etapa é importante para você identificar, por exemplo:

a) Assimetrias nas distribuições das variáveis numéricas (o que pode demandar a transformação dessas variáveis, p.ex. logaritmo base 10).

b) Necessidade de agrupar alguns níveis de uma variável categórica, seja pela presença de categorias com frequências muito baixas, ou pela presença de categorias equivalentes. Por exemplo, para a variável "Estado Civil", é comum que algumas bases apresentem as categorias "Casado" e "União estável" separadamente; nesses casos, se esta distinção não for objeto de interesse direto da pesquisa, pode ser mais conveniente agrupar as duas categorias em uma só ("Casado/União estável").

Observação: Se você identificar variáveis que precisem dos tratamentos acima, não é necessário apresentar os gráficos/tabelas antes desses tratamentos - bastam os gráficos/tabelas depois do tratamento. Mas é importante mencionar esses tratamentos na Seção 3-d.

c) Associações entre as principais variáveis: gráficos de dispersão dois a dois e/ou matrizes de correlação podem ser bastante úteis.

d) *Aplicação do modelo e procedimento de seleção de variáveis* (obs: Tópico em construção)

e) *Resultados e Discussões* (obs: Tópico em construção)

Nesta seção, você apresentará os resultados da análise de regressão.

- Tabela dos coeficientes das variáveis selecionadas - valores estimados, erros padrão, intervalos de confiança etc.
- Tabela da análise de variância.
- Valor do coeficiente de determinação R^2 (original e ajustado).
- Scatter Plot Matrix - Matriz com os diagramas de dispersão entre as variáveis preditoras numéricas e a variável resposta.
- Discussão final sobre as variáveis mais relevantes.

Prazo para entrega: 12/10