ACH2036 - Métodos Quantitativos para Análise Multivariada

Atividade 1: Aplicação de Modelos de Regressão

1 Introdução

O objetivo desta atividade é a aplicação de Regressão Linear Múltipla sobre uma base de dados. Você pode utilizar:

- Dados secundários do IBGE, IPEA ou outros órgãos – embora nesses casos o seja desafio é maior, visto que usualmente há necessidade de uma curadoria de dados mais cuidadosa e a integração entre *datasets* distintos;
- *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 ao (à) seu(sua) superior(a) ou orientador(a) a autorização para uso desses dados. Isso é especialmente crítico se envolver dados pessoais de terceiros ou dados críticos da empresa.

2 Critérios para a escolha da base:

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

2. A base deve conter 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 não contém este número mínimo de variáveis mas tem uma motivação clara (p.ex. está sendo usada em algum projeto de pesquisa em que você participa ou tem relação com algum tema de seu interesse), isso deverá ser justificado (ver adiante).

3. 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.

4. 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, isso deverá ser justificado (ou seja, por que razão a base escolhida era importante, mesmo não contendo variáveis categóricas).

3 Comentários sobre a análise exploratória e pré-processamento dos dados:

A análise exploratória dos dados (AED) é uma etapa importante e obrigatória, para a verificação da necessidade de eliminação de campos identificadores de registros, pré-processamento de variáveis categóricas e transformações de escala em variáveis numéricas.

Campos identificadores ou descritivos

Campos como código, nome, endereço etc. devem ser suprimidos na análise. Somente devem ser mantidas variáveis com as características potencialmente relevantes para a identificação de padrões.

Variáveis categóricas

A inspeção de variáveis categóricas também é importante, com especial atenção à eventual necessidade de tratamento de categorias equivalentes ou com frequências relativas muito baixas.

- No caso de categorias equivalentes para o propósito da análise estatística, pode ser conveniente sua fusão em uma só. Por exemplo, em bases de dados com descrições pessoais, uma variável usual é "Estado Civil"; nessas bases, é comum apresentar "Casado" e "União estável" como duas categorias distintas. Nesses casos, se esta distinção não for considerada relevante , pode ser mais conveniente agrupá-las em uma só ("Casado/União estável"). O mesmo vale para as categorias "Divorciado" e "Separado judicialmente".

- No caso de categorias com frequências muito baixas (o que pode ser analisado através de tabelas de frequências ou gráficos de barras), há duas situações frequentes:

  – Categorias do tipo "Não informado": nesses casos, pode-se considerar esta categoria como valor faltante. Nesta Atividade 1, não consideraremos o tratamento de valores faltantes, e os registros nesta situação poderão ser eliminados (a menos que sua proporção seja muito alta).

  – Categorias informadas, porém com prevalência muito baixa na população. Se essas categorias não forem consideradas como equivalentes a outras mais frequentes, tipicamente seu tratamento deve ser discutido com os especialistas na área (manutenção da categoria ou eliminação dos respectivos registros). Para os propósitos desta Atividade 1, você terá liberdade de escolher uma dessas duas opções (mas justifique).

Variáveis numéricas

Entre os diversos tipos de análise para variáveis numéricas, consideramos aqui o aspecto de suas distribuições individuais e associações entre pares de variáveis.

A análise das distribuições individuais das variáveis pode ser feita de forma visual, através de histogramas ou *boxplots*, ou ainda com indicadores numéricos de assimetria (*skewness*) ou curtose. Para os propósitos desta Atividade 1, nossa principal preocupação é a assimetria – pois é uma das características com maior impacto negativo na análise de regressão.

Para análise gráfica, o notebook `Secao_7_ResidualDiagnostics.ipynb`, disponibilizado no diretório do Google Colab desta disciplina, contém a implementação de uma função para gerar uma *scatter plor matrix* adaptada, e que pode ser utilizada em seu trabalho.

Para variáveis com alta assimetria, pode ser necessário aplicar algum tipo de transformação. A Seção `9:Data Transformations` do material de aula utilizado (https://online.stat.psu.edu/stat501/lesson/9) descreve várias situações e as respectivas transformações. Uma transformação interessante é a *transformação potência*, cuja formulação é a seguinte:

$$y^* = \begin{cases} y^\lambda, & \text{se } \lambda \neq 0, \\ \ln(y), & \text{se } \lambda = 0, \end{cases} \tag{1}$$

em que y é a variável original a ser transformada, y^* é a versão transformada e λ é o expoente utilizado como parâmetro. Um aspecto interessante dessa família é a generalização para vários

casos específicos:

$$\begin{aligned}
\lambda = -2 &\rightarrow y^* = \frac{1}{y^2} \\
\lambda = -1 &\rightarrow y^* = \frac{1}{y} \\
\lambda = -0.5 &\rightarrow y^* = \frac{1}{\sqrt{y}} \\
\lambda = 0 &\rightarrow y^* = \ln y \\
\lambda = 0.5 &\rightarrow y^* = \sqrt{y} \\
\lambda = 1 &\rightarrow y^* = y \text{ (identidade)} \\
\lambda = 2 &\rightarrow y^* = y^2
\end{aligned}$$

Para variáveis limitadas dentro do intervalo $(0, 1)$, uma possibilidade possível (mas não a única) é a transformação *logito* (já apresentada em aulas anteriores):

$$y^* = \log\left(\frac{y}{1-y}\right) = log(y) - log(1-y)$$

Atenção: Esta transformação só vale dentro do intervalo aberto $(0, 1)$. Se houver valores $y = 0$ ou $y = 1$, você deve "truncá-los" 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, para variáveis descritas em percentuais (entre 0% e 100%), a adaptação é imediata, bastando dividi-las por 100 e aplicando a transformação acima.

4 Relatório:

Você deverá apresentar um pequeno relatório, com título e nomes dos integrantes do grupo (2 ou 3 integrantes, que podem ser de turmas distintas).

A estrutura de seções é a seguinte:

1. *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, nome da base e sua respectiva URL. Se for de outra origem, informar também.

   Em seguida, apresente um breve enunciado do problema de regressão a ser resolvido. P.ex: *O objetivo original nesta base de dados é avaliar a associação entre o consumo de combustível de veículos (em Km/L) e algumas de suas características, como potência do motor, peso do veículo, entre outras.*

   Alguns datasets públicos 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.

2. *Descrição da base*

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

   a) Abreviação da variável: Se os nomes originais dos campos forem muito longos, crie nomes curtos ou abreviações que permitam uma fácil identificação. Isso é importante para facilitar a leitura dos gráficos e interpretação dos resultados. Evite nomes codificados (V01, V02, etc). Se a base original já continha nomes curtos, você pode suprimir este campo, mantendo apenas seu nome original (campo abaixo).

   b) Nome original da variável.

   c) Breve descrição da variável.

   d) 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!

   Se algumas variáveis necessitaram de algum tratamento de ajuste de escala, reagrupamentos de categorias e outras situações comentadas na Seção 3, detalhe esses casos.

   Se a base não atende algum dos requisitos indicados na Seção 2 (quantidade mínima de variáveis preditoras ou inexistência de uma variável categórica), justifique a razão pela escolha dessa base.

3. *Caracterização das principais variáveis <u>após eventuais tratamentos</u>*

   Nesta seção, você apresentará gráficos ou tabelas para visualização das distribuições das principais variáveis e das associações entre elas. Para evitar que o relatório fique muito extenso, essas visualizações podem ser apresentadas somente sobre os dados pré-processados (conforme descrito na Seção 3).

   Observação: Se sua base original continha um número muito grande de variáveis, você pode apresentar os gráficos e tabelas somente das variáveis selecionadas em seu modelo (ver item abaixo).

4. *Seleção de variáveis*

   Descreva o procedimento de seleção de variáveis utilizado. Minha recomendação é utilizar um procedimento automatizado simples, como o *stepwise*, descrito em https://online.stat.psu.edu/stat501/lesson/10/10.2, mas você tem liberdade para utilizar outra abordagem de seleção.

5. *Resultados e Discussões*

   Nesta seção, você apresentará os resultados da análise de regressão, *seguidos de seus comentários e interpretações em cada etapa*:

   - Tabela dos coeficientes das variáveis selecionadas: valor estimado do coeficiente, erro padrão, intervalo de confiança etc.
   - Valor do coeficiente de determinação R^2 (original e ajustado).
   - Resultado do teste F de significância do modelo.
   - Gráficos dos resíduos:
     - Histograma dos resíduos (com a respectiva densidade sob a distribuição normal, se possível)
     - QQ-plot dos resíduos usando a distribuição normal como referência
     - Gráfico dos resíduo vs. valores preditos
   - Resultados dos testes estatísticos sobre os resíduos:
     - Testes de normalidade dos resíduos
     - Testes de homogeneidade das variâncias

     Vários gráficos e testes estão implementados no notebook `Secao_7_ResidualDiagnostics.ipynb`, e você poderá utilizá-los
   - Apresente uma síntese final sobre as *qualidades e limitações do modelo*:
     - Ele fornece um bom ajuste? Consegue explicar uma fração importante da variabilidade da variável resposta?
     - O modelo atende às quatro premissas **LINE**?
       – **L**inear function,
       – **I**ndependent errors,
       – **N**ormal distribution,
       – **E**qual variance.
       Ver https://online.stat.psu.edu/stat501/lesson/1/1.3

5 Formato e prazo para entrega

- A entrega deverá ser realizada em um documento PDF.
- Você pode entregar apenas o relatório ou, se preferir, a versão impressa em PDF de seu Notebook Python com código e texto integrados.

  Atualização em 28/10 para quem for entregar o Notebook em PDF:
  Na versão anterior deste enunciado, eu havia sugerido "imprimir" o notebook em formato PDF, mas alguns alunos relataram que a versão PDF gerada perde a fidelidade (algumas

figuras ficam "cortadas" e o texto se sobrepõe às figuras). Para resolver este problema, disponibilizei no drive compartilhado da disciplina (https://drive.google.com/drive/folders/19H5V3mnNc9KMnXke3m5YKGMbVNedGu4l?usp=drive_link) um script chamado `Conversao_ipynb_2_pdf.ipynb`, que faz esta conversão de forma apropriada. Os testes iniciais indicaram que a versão PDF gerada tem boa fidelidade visual em relação ao Notebook original. Mas recomendo que *testem este script com antecedência e me avisem se encontrarem algum problema.*

- O texto deve seguir a estrutura indicada na Seção 4.
- A entrega deverá ser feita exclusivamente através do e-disciplinas.
- **Prazo para entrega: 10/11**