Análise de Variância Multivariada para Dados Não Gaussianos via Teste Wald

Uma visão geral

Lineu Alberto Cavazani de Freitas
lineuacf@gmail.com

PPG Informática
Data Science & Big Data
Universidade Federal do Paraná

https://lineu96.github.io/st/

1

Etapas do processo de análise

O processo de análise consiste em:

- Definição do problema.
- Planejamento do estudo.
- Coleta de dados.
- Análise dos dados:
  - Análise exploratória.
  - Aplicação de métodos mais sofisticados que permitam generalizar os resultados para a população.
- Interpretação dos resultados.

2

Conjunto de dados

Variáveis

- Denominam-se variáveis as características observadas em cada um dos elementos que pertencem à amostra.
- Podemos coletar variáveis de diferentes tipos e naturezas.

Organização dos dados

- Uma forma conveniente de se organizar os dados é da seguinte forma:
  - Cada coluna representa uma variável.
  - Cada linha representa uma observação.
  - Cada célula representa o valor observado no elemento i na variável j.

3

Modelos de Regressão

- Nos casos univariados mais gerais, estes modelos associam uma única variável resposta, a uma ou mais variáveis explicativas.
- De forma geral, um modelo de regressão é uma expressão matemática que relaciona a média da variável resposta às variáveis preditoras (covariáveis).
- A variável resposta segue uma distribuição de probabilidade condicional às covariáveis e a média é descrita por um preditor linear.
- Há casos em que são coletadas mais de uma resposta por unidade experimental e há o interesse de modelá-las em função de um conjunto de variáveis explicativas.
- Saímos então do cenário univariado (apenas uma resposta) e passamos para o multivariado (mais de uma resposta).

São uma das principais e mais difundidas ferramentas utilizadas em diversas áreas do conhecimento, sendo comum o interesse em:

- Explicar a associação entre uma variável resposta e um conjunto de variáveis explicativas.
- Utilizar o modelo para realizar predições para uma população.

Existem inúmeras classes de modelos de regressão:

- Modelo Linear Normal.
- Modelos Lineares Generalizados (GLM).
- Modelos de regressão local.
- Modelos de regressão de splines.
- Modelos aditivos generalizados (GAM).
- Modelos de efeitos aleatórios.
- Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS).
- Modelos de regressão multivariados.
- Modelos Multivariados de Covariância Linear Generalizada (McGLM)

4

McGLM

- Os GLMs são uma forma de modelagem univariada para dados de diferentes naturezas, tais como dados contínuos simétricos e assimétricos, contagens, dentre outras.
- Tais características tornam essa classe de modelos uma flexível ferramenta de modelagem aplicável a diversos tipos de problemas.

Contudo, por mais flexível e discutida na literatura, essa classe apresenta duas principais restrições:

1. A incapacidade de lidar com observações dependentes.
2. A incapacidade de lidar com múltiplas respostas simultaneamente.

Com o objetivo de solucionar esses problemas, foi proposta uma estrutura geral para análise de dados denominada Modelos Multivariados de Covariância Linear Generalizada (MCGLM).
Esta classe de modelagem comporta:

- Múltiplas respostas.
- Respostas de diferentes naturezas.
- Respostas correlacionadas.
- Observações não independentes.
- Extensões multivariadas para modelos de:
  - Séries temporais.
  - Dados longitudinais.
  - Dados espaciais.

5

MANOVA

A MANOVA clássica é um assunto com vasta discussão na literatura e possui diversas propostas com o objetivo de verificar a nulidade conjunta dos parâmetros de um modelo de regressão multivariado, como:

- Lambda de Wilk's.
- Traço de Hotelling-Lawley.
- Traço de Pillai.
- Maior raiz de Roy.

Para o McGLM (cenário com múltiplas respostas não gaussianas) não existia teste similar. No trabalho:

- Propomos e implementamos o teste de Wald para análise de variância multivariada para dados não gaussianos.
- Discutindo as propriedades e comportamento do teste proposto com base em estudos de simulação e aplicação a conjuntos de dados reais.

6

Proposta para o mestrado

A proposta incial era:

- Construção de outros testes multivariados para dados não gaussianos no contexto dos MCGLMs.
- Ampliar o estudo de simulação, de forma a considerar cenários não abrangidos.
- Encontrar formas de melhorar o tempo computacional dos estudos de simulação.

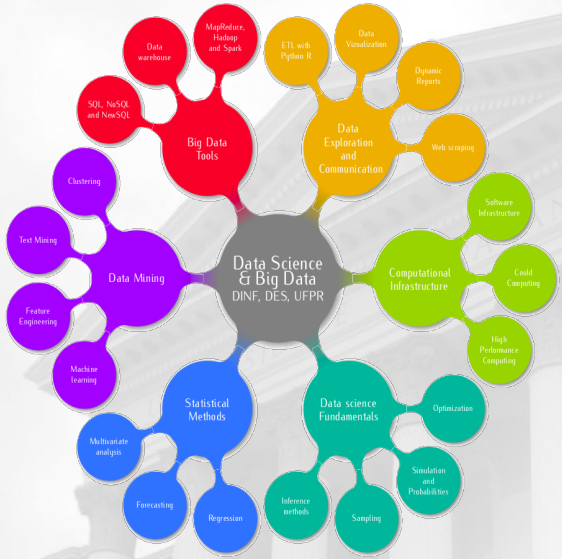
Proposta atual:

- Artigo 1: reimplementar o pacote car com foco em objetos mcglm.
  - Diferentes tipos de output.
  - Gráficos.
  - Testes de comparações múltiplas.
  - Etc.
- Artigo 2: estudo de simulação.
  - Simular o maior número de casos possíveis de conjuntos de dados.
  - Ajustar o modelo para cada um deles, aplicar o teste e reportar em que casos o teste funcionou ou não.
  - Verificar como as revistas fazem para simular coisas do tipo.
- Artigo 3: Analisar conjuntos de dados reais.
  - Ajustar o modelo.
  - Aplicar o teste proposto.
  - Reportar os resultados.

Fica na lista de afazeres propor e implementar os outros testes porque é um caminho muito nebuloso.

Considerando o grupo de pesquisa (Data Science & Big Data) o trabalho teria as seguintes contribuições:

1. Propor um teste novo para uma classe de modelos não usual mas com alto potencial de aplicação.
2. Realizar um estudo pesado de simulação para verificar o funcionamento do teste proposto.
3. Análise de dados.


Data Science & Big Data
DINF, DES, UFPR
Big Data Tools
Data warehouse
MapReduce, Hadoop and Spark
SQL, NoSQL and NewSQL
Data Exploration and Communication
ETL with Python R
Data Visualization
Dynamic Reports
Web scraping
Computational Infrastructure
Software Infrastructure
Could Computing
High Performance Computing
Data Mining
Clustering
Text Mining
Feature Engineering
Machine learning
Statistical Methods
Multivariate analysis
Forecasting
Regression
Data science Fundamentals
Optimization
Simulation and Probabilities
Sampling
Inference methods