**CEBRAP.LAB**

**Tema central:** Introdução à análise de regressão para pesquisas sociais

**Ministrante:** Thiago Cordeiro Almeida

**E-mail de contato:** thiagocordalmeida@gmail.com

**Datas de realização:** 15 a 18 de setembro

**Horário das aulas:** 19 às 22 horas

**Objetivo(s) do curso:**

Este curso tem por objetivo introduzir pesquisadoras/es na utilização de modelos de análise de regressão à investigação de fenômenos sociais. Serão abordadas técnicas para análise de dados contínuos (regressão linear) e categóricos (regressão logística), com a exposição de suas propriedades, procedimentos de implementação usando o software R e a realização de exercícios práticos a partir de bases de dados disponíveis gratuitamente online. Ao fim do curso, os/as participantes estarão familiarizados com os procedimentos necessários para a implementação de modelos de regressão, de forma que possam adaptar para cada problema de pesquisa que trabalhem. Para participar do curso é necessário ter noções básicas dos softwares Excel, R e ter instalado o RStudio em seu computador. É recomendado ter noções de estatística básica (frequências, tabelas cruzadas, médias, medianas, etc.).

**Requisitos:**

Noções básicas de Excel, R e ter instalado em sua máquina o RStudio.

**Estrutura do curso (temas e subtemas previstos e materiais complementares):**

**DIA 1** – Fundamentos para a modelagem de dados quantitativos (conceitos, definições e pressupostos)

1. **Aula expositiva e prática**

Introdução ao curso e apresentações

* Apresentação minha e do pessoal, bem rapidamente, para dar uma quebrada de gelo
  + Quem é você?
  + Qual sua afinidade com dados quantitativos?
  + Qual a afinidade com modelos de regressão?
  + Motivação em fazer o curso
  + Qual sua afinidade com o R?

Algumas definições acerca de dados quantitativos

* Natureza dos dados
  + Qualitativa
  + Quantitativa
* Tipos de dados quantitativos
  + Dados primários
  + Dados secundários
    - Vantagens e desvantagens
  + Dados públicos
    - Exemplos

Conceitos estatísticos

* População e amostra
* Variáveis
  + Definição
  + Natureza das variáveis
* Tipos de medidas
* Tipos de análises
  + Univariada
  + Bivariada
  + Multivariada
* **Exercício**: Tipos de variáveis e formas de análise uni, bi e multivariada

Relação entre fenômenos sociais: associação ou efeito?

* Intuição: quando e como podemos dizer que um fenômeno social interage com outro?
  + Formas de se captar a interação entre fenômenos sociais
  + Formas de se avaliar se há (e a magnitude de) a interação
* Correlação
  + Definição
  + Formas de mensurar estatisticamente
* Causalidade
  + Definição
  + Diferenças para a correlação
  + Formas de mensurar estatisticamente: modelos de regressão!
* **Exercício:** Site: correlação espúria.
  + Apresentarei um exemplo criado por mim com base em diferentes exemplos dados no site
  + Apresentarei a forma de analisar este exemplo (estatísticas descritivas e como mensurar a correlação)
  + Darei um tempo para eles para fazer isso (em R ou Excel)
  + [Para a aula seguinte, disponibilizarei a resolução em ambos os formatos]

Modelagem de fenômenos sociais

* Intuição da modelagem de fenômenos sociais
* Motivação para se modelar fenômenos de diferentes naturezas
  + Momento de convencer eles sobre a potencialidade disso!
  + Por que se modela?
    - Predição de um fenômeno complexo
      * Envolve a relação com mais de uma variável
    - Compreensão da relação de um (ou mais de um) fator específico para o fenômeno estudado
      * Explicação/associação
* Formalização geral de uma modelagem
  + Estabelecer a lei/modelo no âmbito populacional
  + Ajustes a serem feitos para se implementar em uma dada amostra
* Qual(is) o(s) limite(s) de uma modelagem?
  + Brainstorm!
    - Dados
    - Conhecimento teórico sobre a relação entre fenômenos
    - Conhecimento empírico sobre a relação entre fenômenos
    - Conhecimento técnico sobre a relação entre fenômenos
* Fundamentos de uma modelagem
  + Propriedades dos fenômenos a serem analisados
    - Tipos de variáveis e suas distribuições
  + Forma como os fenômenos se relacionam
    - Qual é o fenômeno de interesse?
    - Qual é o fenômeno que interage/relaciona/associa/explica o fenômeno de interesse?

1. **Materiais de referência**

Criar uma pasta com os materiais que serão usados como referência para esta aula

1. **Parte prática**

Criar uma pasta com os materiais que serão usados na parte prática para acesso/download.

* Adicionar a resolução do exercício sobre “correlação espúria” em R e Excel

1. **Vídeo(s)**

* Vídeo de como instalar o R e o Rstudio

**DIA 2** – Modelagem de dados contínuos – regressão linear

* + - 1. **Aula expositiva e prática**

Aquecimento: dúvidas, comentários, críticas sobre aula anterior

* Dúvidas sobre exercício ou houve alguém que fez e queira compartilhar?
* Comentários gerais

Modelagem de fenômenos sociais contínuos

* Fluxo de análise de dados
* Conceitos e definições
  + Variável aleatória
  + Variável a ser explicada
    - Variável resposta, dependente, ...
  + Variável a explicar
    - Variável explicadora, independente, ...
  + Outras variáveis
    - Covariáveis, controle, confundidoras, ...
  + Função de distribuição
    - Importância (cada função de distribuição faz parte de um conjunto mais amplo, carregando pressupostos e propriedades)
    - Tipos de funções de distribuição
    - Links do modelo
* Exemplo prático inserindo no fluxo de análise
  + Criar um exemplo com uma base de dados fictícia
    - Pergunta de pesquisa
    - Explorar a variável dependente e independente
    - Estabelecer o modelo
    - Analisar a relação entre elas
    - Avaliar a validade do modelo
  + Relacionar os conceitos apresentados anteriormente às etapas do exemplo prático

Regressão linear simples

* Intuição
* Formalização na população
* Função de distribuição
* Propriedades
* Pressupostos
* Exemplo da aula – como fazer isso na prática?
  + Voltar à etapa de estabelecimento do modelo
  + Análise da relação entre variáveis
    - Interpretação 1 – coeficientes
    - Interpretação 2 – *Average Marginl Effect (AME)*
  + Avaliar a validade do modelo à luz das propriedades/pressupostos

[break!]

Regressão linear multivariada

* Intuição
* Formalização na população
* Função de distribuição
* O que muda na prática entre a regressão linear simples e multivariada?
* **Exercício:**
  + Replicar o que foi feito na regressão linear simples em R
  + Dar orientações para que façam o mesmo adicionando algumas variáveis.
* Interpretação em um cenário de regressão multivariada
  + Vamos focar na interpretação
    - Exercício somente com uma variável adicional
    - Exercício com 3 variáveis adicionais
    - Adicionando variável binária ou categórica
      1. **Materiais de referência**

Criar uma pasta com os materiais que serão usados como referência para esta aula.

* + - 1. **Parte prática**

Criar uma pasta com os materiais que serão usados como referência para a aula anterior

* + - 1. **Vídeo(s)**
* PDF de como instalar o R e RStudio
* Vídeos de introdução ao R e RStudio para quem não teve este momento ainda.
* Vídeo de alguma parte que não tenhamos coberto em aula

**DIA 3** – Modelagem de dados contínuos – regressão linear multivariada com variáveis categóricas binárias/nominais/ordinais

1. **Aula e prática**

...

* Exercício (para fazer em casa)

1. **Materiais de referência**

Criar uma pasta com os materiais que serão usados como referência para esta aula

* *Materiais do pessoal do Zen do R (curso-R) pode ser uma boa*
* *Material dos autores da UFRN*
* *Ver outros materiais complementares (tanto do próprio IBGE quanto outros)*
* *Posso inserir um txt com links de aulas do Gabriel Assunção (IBGE) sobre manipulação de dados da PNADC em R via o pacote.*

1. **Parte prática**

Criar uma pasta com os materiais que serão usados na parte prática para acesso/download.

**DIA 4** – Modelagem de dados categóricos binários – regressão logística

1. **Aula expositiva e prática**

Aquecimento: dúvidas, comentários, críticas sobre aula anterior

* Dúvidas sobre exercício ou houve alguém que fez e queira compartilhar?
* Como incorporar variáveis categóricas e como interpretar, dúvidas?
* Comentários gerais

Modelagem de fenômenos sociais categóricos

* Fluxo de análise de dados
* Conceitos e definições
  + Retomar isso só apresentando as diferenças ou semelhanças aqui
  + Variável aleatória
  + Variável a ser explicada
    - Variável resposta, dependente, ...
  + Variável a explicar
    - Variável explicadora, independente, ...
  + Outras variáveis
    - Covariáveis, controle, confundidoras, ...
  + Função de distribuição
    - Importância (cada função de distribuição faz parte de um conjunto mais amplo, carregando pressupostos e propriedades)
    - Tipos de funções de distribuição
    - Links do modelo
* Exemplo prático inserindo no fluxo de análise
  + Criar um exemplo com uma base de dados fictícia
    - Pergunta de pesquisa
    - Explorar a variável dependente e independente
    - Estabelecer o modelo
    - Analisar a relação entre elas
    - Avaliar a validade do modelo
  + Relacionar os conceitos apresentados anteriormente às etapas do exemplo prático

Regressão logística

* Família de modelos de regressão linear generalizados
  + Apresentar a família desses modelos
  + Tipos de distribuições que eles assumem e a relação com a ideia de “generalizados”
* Regressão logística
  + Intuição
  + Formalização na população
  + Função de distribuição
  + Propriedades
  + Pressupostos
    - Quais são flexibilizados e como eles são flexibilizados
* Exemplo da aula – como fazer isso na prática?
  + Voltar à etapa de estabelecimento do modelo
  + Análise da relação entre variáveis
    - Interpretação 1 - coeficientes
    - Interpretação 2 – *odds ratio* (explicar as potencialidades e limitações
    - Interpretação 3 – probabilidades preditas
  + Avaliar a validade do modelo à luz das propriedades/pressupostos

[break!]

Regressão logística multivariada – exemplo empírico

* Intuição, semelhanças e diferenças com os modelos anteriores
  + Não há diferenças além da distribuição assumida para a relação linear entre os parâmetros
* Formalização do modelo populacional
* Função de distribuição
* **Exemplo:**
  + Replicar o que foi feito na regressão linear simples em R
  + Dar orientações para que façam o mesmo adicionando algumas variáveis.
* Interpretação em um cenário de regressão multivariada
  + Vamos focar na interpretação
    - Inserir algumas variáveis de diferentes naturezas
    - Apresentar a interpretação em diferentes modos
      * Coeficientes
      * Odds Ratio
      * Probabilidades preditas

Finalização do curso

* Retomar outline e ferramentas que eles adquiriram
* Próximas etapas
  + Regressões lineares
    - O que fazer quando pressupostos são violados?
    - Para cada pressuposto há uma forma de lidar, dar exemplos
      * Modelos robustos a heterocedasticidade
  + Regressões não lineares (GLM)
    - O que fazer quando pressupostos são violados?
    - Outros tipos de variáveis?
      * Categórica ordinal e não ordinal
      * Contínuas que seguem distribuições não normais
      * Contínuas discretas (modelos de contagem)

1. **Materiais de referência**

Criar uma pasta com os materiais que serão usados como referência para esta aula.

1. **Parte prática**

Criar uma pasta com os materiais que serão usados como referência para a aula anterior

1. **Vídeo(s)**

* PDF de como instalar o R e RStudio
* Vídeos de introdução ao R e RStudio para quem não teve este momento ainda.
* Vídeo de alguma parte que não tenhamos coberto em aula.