

Lego III

Métodos estatísticos para análise de dados categóricos: abordagem clássica e paradigma bayesiano

Profs: Carlos A. Costa Ribeiro e Vinicius Pinheiro Israel

Horário: Quarta-feira, das 13h às 16h

Consultas: a combinar com os coordenadores

OBJETIVOS:

Os métodos para análises de dados categóricos ocupam lugar central nas metodologias estatísticas em sociologia e ciências políticas. Há uma variedade enorme de técnicas e métodos usados nestas disciplinas que são baseados nas análises de categorias, como, por exemplo, classe social, gênero, partidos políticos, transições educacionais, número de filhos nascidos vivos, etc.

Adicionalmente, pretende-se apresentar o principal debate inferencial da estatística que consiste nas vertentes de análise: clássica ou frequentista, por um lado, e na abordagem da estatística bayesiana, por outro.

Este curso pretende apresentar aos alunos as principais técnicas de análises estatísticas de dados categóricos. Espera-se que ao concluir com sucesso o curso eles sejam capazes de entender métodos básicos e alguns relativamente sofisticados. Embora o curso não seja exaustivo sobre o tema, pretende encorajar os alunos a continuar a estudar métodos estatísticos mais avançados para análise de dados categóricos.

PRÉ-REQUISITOS:

Conhecimento de conceitos básicos de amostragem e inferência; distribuições binomial, normal, e qui-quadrado; técnicas básicas para a análise de tabelas de contingência 2x2. É recomendável o conhecimento de princípios básicos de análise de regressão.

PRESENÇA NAS AULAS:

As aulas e discussões em sala de aula são indispensável para o entendimento da matéria do curso. A falta em até mesmo uma única aula pode ser prejudicial para o aprendizado dos tópicos seguintes do curso.

EXERCÍCIOS:

O entendimento dos tópicos do curso depende em grande medida do trabalho dos exercícios que os alunos devem fazer fora da sala de aula. Durante o decorrer do curso serão distribuídos um conjunto de exercícios (lista de exercícios). Eles devem ser entregues em aulas estipuladas (em média de duas em duas semanas).

Os exercícios da primeira metade do curso envolvem parte teórica e prática. Na segunda parte do curso os exercícios serão feitos usando os seguintes programas de computador: R, GLIM (Generalized Linear Models) e STATA. Os professores providenciarão cópias do programa.

TRABALHO FINAL:

Um trabalho final, para ser feito fora de sala de aula, envolvendo uso do programa de computador será proposto ao final do curso. A nota final será estabelecida a partir deste trabalho (50% da nota) e dos exercícios (50% da nota) feitos ao longo do curso.

CONSULTAS:

Os alunos são encorajados a discutir os exercícios entre si. Os professores estarão disponível para consultas em determinados horários de atendimento.

O curso não seguirá os textos detalhadamente. Os livros propostos são necessários na medida em que apresentam perspectivas alternativas e adicionais aos tópicos discutidos nas aulas. Outras fontes de leitura poderão ser sugeridas ao longo do curso.

PROGRAMA DO CURSO:

Aula 01: Apresentação do curso:

- Tipos de dados, análise exploratória de dados, noções de probabilidade, Teorema Central do Limite (TCL), inferência estatística (população, amostra, estimadores) e teste de hipóteses (abordagem clássica e paradigma bayesiano).
- Principais distribuições de probabilidade para dados categóricos.
- Revisão dos modelos de regressão linear simples.

Leituras:

Power; Xie (2008) Capítulo 1: *Introduction*; e Capítulo 2: *Review of Linear Regression Models*. pp. 11-28;
Gill (2015) Capítulo 1: *Introduction*;
Bussab; Morettin (2010) Capítulo 10: *Introdução à Inferência Estatística*.

Primeira lista de exercícios.

Aula 02: Modelos de regressão linear simples e múltipla e modelos lineares generalizados MLG:

- Caracterização dos modelos de regressão linear simples e múltipla.
- Ajuste e diagnóstico dos modelos.
- Enfraquecendo as condições dos modelos anteriores.
- Caracterização dos modelos lineares generalizados. Família exponencial.

Leituras:

Gill (2015) Capítulo 1: *Introduction*; e Capítulo 3: *The Normal and Student's t-Models*;
Dobson (2002) Capítulo 3: *Exponential Family and Generalized Linear Models*.

Aula 03: Estatística clássica *versus* paradigma bayesiano

- Inferência estatística: parâmetros fixos ou aleatórios?
- Estimadores clássicos: estimação por máxima verossimilhança (EMV), métodos de mínimos quadrados, métodos dos momentos.
- Teorema de Bayes e estimadores bayesianos. Verossimilhança, distribuição *a priori*, preditiva e distribuição *a posteriori*. Elicitação de distribuições *a priori*.

Leituras:

O'Hagan, *Bayesian statistics: principles and benefits*;
O'Hagan (1998);
Lock e Gelman (2010).

Entrega da primeira lista de exercícios e início da segunda lista de exercícios.

Aula 04. Análise de Tabelas de Contingência 2x2

- Modelo de independência
- Medidas de associação
- Estatísticas de ajuste dos modelos
- Conceitos básicos para modelos log-lineares

Leituras:

Fienberg (1994), Capítulos 1 e 2;
Knoke; Burke (1980), pp. 7-22, 30-33;
Agresti (1996), Capítulo 2
Silva (1988), Capítulos 2 e 3;
Giolo (2017), Capítulos 1, 2 e 3.

Aula 05. Análises de Tabelas de Contingência com 3 variáveis

- Independência, associação, e modelos de não-interação
- Algoritmos iterativos de ajuste proporcional (estimativa iterativa dos valores esperados)
- Comparações hierárquicas

Leituras:

Fienberg (1994), Capítulo 3;

Agresti (1996), Capítulo 3.

Entrega da segunda lista de exercícios e início da terceira lista de exercícios.

Aula 06: Dividindo Tabelas de Contingência

- Modelos de independência em partes das tabelas
- Dividindo o Chi-quadrado

Leitura: Fienberg (1994), Capítulo 4.

Aula 07: Análises de Tabelas de Contingência Multidimensionais

- Modelos de interação (*higher order*)
- Procedimentos para seleção de modelos
- Comparação hierárquica de modelos interligados

Leituras:

Fienberg (1994), Capítulos 5;

Knoke; Burke (1980), pp. 33-47.

Entrega da terceira lista de exercícios e início da quarta lista de exercícios.

Aula 08: Análise de Tabelas de Mobilidade

- Mobilidade Total, fluxos de saída e de entrada
- Índice de Dissimilaridade
- Modelo de Mobilidade Quasi-perfeita
- Modelos de distância social
- Modelos Topológicos
- Modelos de associação escalar

Leituras:

Hout (1983);

Powers; Xie (2000) Capítulo 4;

Ribeiro; Israel (2016).

Aula 09: Modelos Logit e Probit para Variáveis Binárias

- Relação entre o modelo log-lineares e o modelo logit

Leituras:

⑩ Fienberg, capítulos 6 e 7;

⑩ Knoke; Burke, pp. 24-20;

⑩ Powers; Xie, Capítulo 3;

⑩ Agresti, pp. 162-167;

⑩ Long; Freese, Capítulo 4.

Entrega da quarta lista de exercício e início da quinta lista de exercício.

Aula 10: Modelos para Variáveis Dependentes Ordinais

- Modelos Logit e Probit dados ordinais

Leituras: Long e Freese, capítulo 5; Powers and Xie, capítulo 6.

Aula 11: Modelos para Variáveis Dependentes Nominais

- Modelo Logit Multinomial
- O Modelo Logit Multinomial Básico

Leituras:

Long; Freese, Capítulo 6;

Powers; Xie, Capítulo 7.

Entrega da quinta lista de exercício. Lista extra.

Aula 12: Modelos Hierárquicos

- Modelos hierárquicos básicos
- Programas e pacotes: Winbugs, JAGS e STAN.

Leituras:

Gill (2015) Capítulo 12;

Israel (2016).

Aula 13: Aula de dúvidas, revisão e discussão sobre o trabalho final do curso.

Entrega da lista extra.

TEXTOS DE REFERÊNCIA:

AGRESTI, Alan. *An Introduction to Categorical Data Analysis*. New York: Wiley Series in Probability and Statistics, 1996.

BUSSAB, Wilton O.; MORETTIN, Pedro A. *Estatística Básica*. 6ª ed. São Paulo: Saraiva, 2010.

DOBSON, Annette. *Generalized Linear Models*. 2ª ed. Chapman and Hall/CRC, 2002.

FIENBERG, Stephen E. *The Analysis of Cross-Classified Categorical Data*. 3rd Edition. Cambridge, Mass.: MIT Press, 1994.

GILL, Jeff. *Bayesian Methods: A Social and Behavioral Sciences Approach*. 3ª ed. S. Louis, Chapman & Hall/CRC, 2015.

GIOLO, Suely Ruiz. *Introdução à Análise de Dados Categóricos com Aplicações*. São Paulo, Bluncher, 2017.

HOUT, Michael. *Mobility Tables*. California: Sage Publications, 1983.

ISRAEL, V. P. Macrossociologia sobre relações causais de encarceramento: comparações entre países e estados brasileiros. Atas do 10º Encontro da Associação Brasileira de Ciência Política, 2016.

JACKMAN, SIMON. *Bayesian Analysis for the Social Science*. Stanford, John Wiley and Sons, 2009.

KNOKE, David, BURKE, Peter J. *Log-linear Models*. California: Sage Publications, 1980.

LOCK, Kari; GELMAN, Andrew *Bayesian combination of state polls and election forecasts*. Political analysis. Vol. 18, pp. 337-348, 2010.

LONG; FREESE. *Regression Models for Categorical Dependent Variables using Stata*. Texas: Stata Press Corporation, 2001.

O'HAGAN, Anthony. Bayesian statistics: principles and benefits.

O'HAGAN, Antony. *Eliciting expert beliefs in substantial practical application*. The Royal Statistical Society, series D. Vol. 47, n. 1, pp. 21-35, 1998.

POWERS, Daniel A., XIE, Yu. *Statistical Methods for Categorical Data Analysis*. New York: Academic Press, 2000.

RIBEIRO, C. A. C; ISRAEL, V. P. *Voto assimétrico, classes e mobilidade social no Brasil*. Tempo Social, revista de sociologia da USP, Vol. 28, n. 2, pp. 105-129, 2016.

SILVA, Nelson do Valle. *Análise de Dados Categóricos*. Vértice, 1988.