Instituto de Estudos Sociais e Políticos
Universidade do Estado do Rio de Janeiro

Análise de Dados Categóricos

Lego III
1º Semestre de 2014
Prof. Dr. Ricardo Ceneviva
ceneviva@iesp.uerj.br

Objetivos do Curso

Este é um curso de introdução à análise de dados categóricos para alunos de pós-graduação em Ciência Política e Sociologia. O objetivo principal do curso é fornecer uma introdução básica de métodos estatísticos para a análise de variáveis dependentes limitadas; isto é, variáveis dependentes que assumem valores binários ou categóricos (que podem ser ordenados ou não), variáveis que se expressam na forma de eventos ou contagem, ou variáveis onde uma parte dos dados estão censurados ou truncados. Perguntas muito antigas da Ciências Sociais como, por exemplo: existe alguma relação entre o nível de desenvolvimento econômico dos países e seu regime político? Candidatos que concorrem à reeleição no exercício do cargo têm mais chances de sucesso eleitoral? Sexo, renda ou educação afetam a opinião das pessoas sobre a pena de morte? Tais perguntas podem ser respondidas com o auxílio de métodos e técnicas estatísticas apropriados para situações nas quais a variável de interesse não é contínua.

Há uma grande variedade de métodos estatísticos usados nas ciências sociais, que são baseados na análise de variáveis categóricas. Este curso apresenta uma visão geral desses modelos (em sua maioria, estimados via máxima verossimilhança) em que os pressupostos fundamentais do modelo de regressão linear (via mínimos quadrados ordinários) são violados porque a variável dependente é discreta. Alguns modelos específicos que serão abordados incluem, modelos para variáveis binárias (logit e probit), modelos para variáveis multinomiais (ordenadas e não-ordenadas), modelos para variáveis censuradas e truncadas e modelos de contagem de eventos. Além disso, os alunos deverão aplicar esses métodos em uma série de listas de exercícios e em um projeto de final.

Ao final do semestre espera-se que os alunos sejam capazes de produzir e analisar seus dados, além de ler e criticar textos que apliquem as técnicas discutidas nas aulas. Para tanto, o curso combina aulas expositivas e sessões práticas no laboratório, onde os alunos serão familiarizados com o manuseio de *softwares* de análise quantitativa de dados. O conteúdo do curso não é exaustivo, contudo, pretende-se encorajar os alunos a continuar estudando métodos mais sofisticados para a análise de dados categóricos.

A quem esse curso se destina?

O foco do curso são os alunos de pós-graduação de Ciência Política e Sociologia e os exemplos e casos analisados nas sessões teóricas e práticas serão retirados, principalmente, de artigos e trabalhos dessas duas disciplinas. Embora o curso tenha um conteúdo bastante técnico, sua abordagem será conceitual e aplicada, ao invés de técnica e teórica. Isto é, as aulas expositivas e sessões de laboratório buscarão se concentrar na intuição básica por traz de cada um dos modelos estudados e na sua aplicação prática nas ciências sociais. Os cursos de Introdução a Estatística (Lego I) e Análise de Regressão (Lego II) são considerados pré-requisitos para este curso. Portanto, conhecimento de conceitos básicos de amostragem e inferência estatística; distribuição de probabilidade (normal, binomial, qui-quadrado, poisson, etc.) e os fundamentos do modelo de regressão linear (simples e múltipla) são fundamentais para a compreensão do conteúdo do curso.

Avaliação

Como se trata de um curso aplicado, a participação em sala de aula terá grande importância. Além das atividades em laboratório e das listas de exercícios, os alunos serão encorajados a coletar, analisar e apresentar seus próprios dados ou a replicar algum trabalho ou artigo já publicado. Para tanto, será proposto um projeto final que consistirá em uma apresentação de uma análise de algum banco de dados escolhido pelo aluno.

1. Os alunos deverão entregar 5 (cinco) listas de exercícios ao longo do curso. Cada lista de exercício vale 10% da nota final do aluno. Não serão aceitas listas de exercícios entregues fora do prazo.
2. Projeto final valendo 50% da nota final do aluno.

Software para Análise de Dados

Neste curso, tanto nas sessões de laboratório como para a realização das listas de exercícios, serão usados *softwares* para a análise de dados. Hoje, há vários programas computacionais concebidos especialmente para a análise estatística de dados sociais e políticos, além de outros de uso mais geral. Parte importante desse curso destina-se à introdução ao uso de *softwares* para a análise estatística de dados. Espera-se que ao final do semestre o aluno seja capaz de realizar operações básicas para a implementação, análise, disgnóstico e apresentação gráfica dos modelos estatísticos estudados ao longo do curso. Dentre as várias opções de *softwares* para a análise de dados, as sessões práticas em laboratório fornecerão treinamento básico para o manuseio de dois programas computacionais: R e Stata. Caberá ao aluno escolher a opção que mais lhe convém.

R

O R é uma linguagem para computação estatística e gráficos. Uma das grandes vantagens do R é que se trata de um *Software* Livre, que pode ser baixado gratuitamente na página oficial do Projeto R na internet: http://cran.r-project.org/. Recentemente um artigo do *New York Times* ("Data Analysts Captivated by R's Power", 6 de janeiro de 2009) caracterizou o R como:

> *a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca [...] whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it. [...] "The great beauty of R is that you can modify it to do all sorts of things," said Hal Varian, chief economist at Google. "And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."*

Como mencionado, o R não é apenas um pacote estatístico; trata-se de uma linguagem para computação estatística e gráfica. Como tal, é muito mais flexível, rápido e poderoso do que os demais *softwares* de análise de dados comerciais disponíveis no mercado, como o SPSS, Stata ou SAS, além de ser gratuito. O R é também mais difícil de ser usado, já que exige alguma noção de programação. Há uma imensa variedade de materiais de apoio, apostilas, e recursos (em português e inglês) destinados à aprendizagem do R disponíveis gratuitamente na Internet. Alguns desses recursos serão comentados e usados nas sessões de laboratório.

Stata

O Stata é um software comercial muito popular para análise estatística de dados. Este *software* foi criado em 1985 pela StataCorp. O Stata é largamente empregado, tanto por instituições do setor privado como na academia, particularmente, por pesquisadores que lidam com gerenciamento e manipulação de bancos de dados, o que engloba áreas tão diversas como: biologia, epidemiologia , economia, sociologia, demografia e ciência política, entre outras.

Os recursos do Stata incluem não apenas o gerenciamento e manipulação de bancos de dados, mas também, análises estatísticas, gráficos, simulações, além de oferecer ferramentas de programação personalizada para execução de tarefas específicas. O Stata oferece uma interface gráfica o que permite ao usuário a análise via programação ou utilizando os menus (*point-and-click*). O aprendizado no Stata é relativamente fácil, sendo superado apenas pelo SPSS, que não será usado no curso devido a seu elevado custo comercial.

Programa

***Aula 1**: "Revisão do Modelo de Regressão Linear"*

Apresentação do curso e revisão do modelo de regressão linear.

Leitura obrigatória:

Powers & Xie (2008) capítulo 2: *"Review of Linear Regression Models"*

Long (1997) capítulo 2: *"Continuous Outcomes: The Linear Regression Model"*

Leitura complementar:

Wooldridge (2010) capítulo 6: *"Análise de Regressão Múltipla: Problemas Adicionais."*

Angrist & Pischke (2008) capítulo 3: *"Making Sense of Regression"*.

***Aula 2**: "Revisão do Modelo de Regressão Linear II"*

O Modelo de regressão múltipla com informações qualitativas, A descrição de informações qualitativas, o uso de variáveis *dummy* para categorias binárias e múltiplas, interações e a interpretação dos coeficientes.

Leitura obrigatória:

Brambor, Thomas, William Roberts Clark, and Matt Golder. (2006) "Understanding interaction models: Improving empirical analyses."Political analysis 14.1: pp. 63-82.

Leitura complementar:

Wooldridge (2010) capítulo 7: *"O Modelo de Regressão Múltipla com Informações Qualitativas: Variáveis Binárias (ou Dummy)"*.

Gelman e Hill (2007) capítulo 4: *"Linear Regression: before and after fitting the model"*.

***Aula 3**: "Estimação por Máxima Verossimilhança"*

Breve introdução à estimação por máxima verossimilhança (MLE).

Leitura obrigatória:

Powers & Xie (2008) capítulo 2: *"Review of Linear Regression Models"*.

Long (1997) capítulo 2: *"Continuous Outcomes: The Linear Regression Model"*

Leitura complementar:

Wooldridge (2010), capítulo 17. *"Modelos com Variáveis Dependentes Limitadas e Correções de Seleção Amostral"*, Apêncide 17.A e Apêndice17.B.

OBSERVAÇÃO: **Devolução da "Lista de Exercícios 1"**

***Aula 4**: "Modelos de Variáveis Dependentes Binárias"*

Os Modelos de escolha binária: *"Logit"* e *"Probit"*

Leitura obrigatória:

Long (1997) capítulo 3: *"Binary Outcomes: The Linear Probabilty, Probit and Logit Models"*.

Powers & Xie (2008) capítulo 3: *"Models for Binary Data"*.

Leitura complementar:

King (1989) capítulo 5: *"Discrete regression models"*, pp. 97-115.

Wooldridge (2010), capítulo 17. *"Modelos com Variáveis Dependentes Limitadas e Correções de Seleção Amostral"*.

***Aula 5**: "Modelos de Variáveis Dependentes Binárias II"*

Algumas aplicações dos modelos de escolha binária: *"Logit"* e *"Probit"*

Leitura obrigatória:

Bartels, Larry M. 2000. *"Partisanship and Voting Behavior, 1952-1996."*, American Journal of Political Science, 44 (1):35-50.

Long (1997), capítulo 4: *"Hypothesis Testing and Goodness of Fit"*, pp. 85-113.

Leitura complementar:

Herron, Michael C. 1999. *"Postestimation Uncertainty in Limited Dependent Variable Models"*, Political Analysis. 8(1):83-98.

Kerrissey, J., & Schofer, E. (2013). *"Union membership and political participation in the United States"*. Social forces, 91(3), 895-928.

***Aula 6**: "Modelos de Variáveis Dependentes Ordenadas"*

Os modelos de escolha ordenada: *"Ordered Logit"* e *"Ordered Probit"*

Leitura obrigatória:

Long (1997) capítulo 5: *"Ordinal Outcomes: Ordered Logit and Ordered Probit Analysis"*.

Powers & Xie (2008) capítulo 7: *"Models for Ordinal Dependent Variables"*.

Leitura complementar:

King (1989) capítulo 5: *"Discrete regression models"*, especialmente seção 5.4. *"Ordered Categorical Models"*, pp. 115-117.

OBSERVAÇÃO: **Devolução da "Lista de Exercícios 2"**

***Aula 7**: "Modelos de Escolha Multinomial"*

Os Modelos de escolha multinomial: *"Multinomial Logit"* e *"Multinomial Probit"*

Leitura obrigatória:

Long (1997) capítulo 6: *"Nominal Outcomes: Multinomial Logit and Related Models"*.

Powers & Xie (2008) capítulo 8: *"Models for Nominal Dependent Variables"*.

Leitura complementar:

Alvarez, R.M. and Jonathan Nagler. 1995. "When Politics and Models Collide: Estimating Models of Multiparty Elections. American Journal of Political Science 42(1):55-96

OBSERVAÇÃO: **Devolução da "Lista de Exercícios 3"**

***Aula 8**: "Modelos de Escolha Multinomial II"*

Algumas aplicações dos modelos de escolha multinomial*"Multinomial Logit"* e *"Multinomial Probit"*.

Leitura obrigatória:

Whitten, Guy D. & Harvey D. Palmer. (1996). *"Heightening Comparativists' Concern for Model Choice: Voting Behavior in Great Britain and the Netherlands."* American Journal of Political Science, 40:231-260.

Dow, Jay. K., & Endersby, James. W. (2004). *"Multinomial probit and multinomial logit: A comparison of choice models for voting research"*. Electoral Studies, 23, 107-122.

Leitura complementar:

Glasgow, Garrett. 2001. *"Mixed Logit Models for Multiparty Elections."* Political Analysis, 9(2): 116-136.

McVeigh, R. & Christian S. (1999). *"Who Protests in America: An Analysis of Three Political Alternatives – Inaction, Institutionalized Politics, or Protest."* Sociological Forum, 14, 4:685-702.

Mullen, Ann L., Kimberly A. Goyette, and Joseph A. Soares. 2003. *"Who Goes to Graduate School? Social and Academic Correlates of Educational Continuation After College."* Sociology of Education, 76,2:143-169.

Gerber, Theodore P. 2000. "Market, State, or Don't Know? Education, Economic Ideology, and Voting in Contemporary Russia." Social Forces, 79, 2:477-521.

***Aula 9**: "Modelos de Variáveis Dependentes Censuradas ou Truncadas"*

Os modelos *"Tobit"* censurado e *"Tobit"* truncado suas aplicações

Leitura obrigatória:

Long (1997) capítulo 7: *"Limited Outcomes: The Tobit Model"*.

King (1998) capítulo 9: *"Models with nonrandom selection"*.

Leitura complementar:

Wooldridge (2010), capítulo 17. *"Modelos com Variáveis Dependentes Limitadas e Correções de Seleção Amostral"*.

Observação: **Devolução da "Lista de Exercícios 4"**

***Aula 10**: "Modelos para Dados de Contagem'*

O modelo de poisson e os modelos com distribuição binomial negativa.

Leitura obrigatória:

Long (1997) capítulo 8: *"Count Outcomes: Regression Models for Counts"*.

Powers & Xie (2008) capítulo 6: *"Models for Event Ocurrence"*.

Leitura complementar:

King, Gary. (1988). *"Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model."* American Journal of Political Science 32(3):838-63

Wooldridge (2010), capítulo 17. *"Modelos com Variáveis Dependentes Limitadas e Correções de Seleção Amostral"*.

***Aula 11**: "Modelos para Dados de Contagem II"*

Algumas aplicações do modelo de poisson e dos modelos com distribuição binomial negativa.

Leitura obrigatória:

Land, K., McCall, P., and Nagin, D. (1996). *"A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models: With Empirical Applications to Criminal Careers Data."* Sociological Methods & Research, May (24): 387-442.

Krain, M. (1998). *"Contemporary Democracies Revisited: Democracy, Political Violence, and Event Count Models."* Comparative Political Studies April (31): 139-164.

Leitura complementar:

King (1989) capítulo 5: *"Discrete regression models"*;

Cole, Wade. 2006. "Accrediting Culture: An Analysis of Tribal and Historically Black College Curricula." Sociology of Education, 79:355-388.

Haynie, Dana L. 2001. "Delinquent Peers Revisited: Does Network Structure Matter?" American Journal of Sociology, 106, 4:1013-1057.

Isaac, Larry and Lars Christiansen. 2002. "How the Civil Rights Movement Revitalized Labor Militancy." American Sociological Review, 67:722-746.

Observação: **Devolução da "Lista de Exercícios 5"**

PROJETO FINAL

Bibliografia

As leituras exigidas para as aulas dividem-se em obrigatórias e complementares. Espera-se que os alunos já tenham feito a leitura dos textos indicados como obrigatórios antes das respectivas aulas. Cabe observar que o conteúdo dessa disciplina tem uma caráter fortemente cumulativo. Ou seja, a compreensão dos tópicos abordados numa determinada aula dependem do domínio dos conteúdos abordados nas aulas passadas. Uma bibliografia complementar será apresentada ao final de cada aula como sugestão de leitura para aqueles que desejem se aprofundar no tema.

Livros textos

1. Gelman, Andrew, and Jennifer Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
2. King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: Michigan University Press.
3. Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Vol. 7. Thousand Oaks: Sage Publications.
4. Long, J. Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*, 2nd ed. College Station, Tex.: StataCorp LP.
5. Powers, Daniel A., and Yu Xie. 2008. *Statistical methods for categorical data analysis*. Emerald Group Publishing.

6. Wooldridge, Jeffrey M. 2010. *Introdução à Econometria: Uma Abordagem Moderna.* Tradução da 4 ed. norte-americana. São Paulo: Cengage Learning.

Leitura Complementar

1. Alvarez, R. Michael, and Jonathan Nagler. 1998. *"When Politics and Models Collide: Estimating Models of Multiparty Elections."* American Journal of Political Science 42 (1):55-96.

2. Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

3. Bartels, Larry M. 2000. *"Partisanship and Voting Behavior, 1952-1996."* American Journal of Political Science 44 (1):35-50.

4. Brambor, Thomas, William Clark and Matt Golder. 1996. *"Understanding Interaction Models: Improving Empirical Analyses."* Political Analysis 14.1: 63-82.

5. Cole, Wade. 2006. *"Accrediting Culture: An Analysis of Tribal and Historically Black College Curricula."* Sociology of Education, 79:355-388.

6. Dow, Jay. K., & Endersby, James. W. 2004. *"Multinomial probit and multinomial logit: A comparison of choice models for voting research".* Electoral Studies, 23, 107-122.

7. Glasgow, Garrett. 2001. *"Mixed Logit Models for Multiparty Elections."* Political Analysis, 9(2): 116-136

8. Gerber, Theodore P. 2000. *"Market, State, or Don't Know? Education, Economic Ideology, and Voting in Contemporary Russia."* Social Forces, 79, 2:477-521.

9. Haynie, Dana L. 2001. *"Delinquent Peers Revisited: Does Network Structure Matter?"* American Journal of Sociology, 106, 4:1013-1057.

10. Herron, Michael C. 1999. *"Postestimation Uncertainty in Limited Dependent Variable Models"*, Political Analysis. 8(1):83-98.

11. Isaac, Larry and Lars Christiansen. 2002. *"How the Civil Rights Movement Revitalized Labor Militancy."* American Sociological Review, 67:722-746.

12. Land, Kenneth C., McCall, Patricia L., and Nagin, Daniel S. 1996. *"A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models: With Empirical Applications to Criminal Careers Data."* Sociological Methods & Research, May (24): 387-442.

13. Kellstedt, Paul M. & Whitten, Guy D. *The Fundamentals of Political Research.* Cambridge: Cambridge University Press, 2009.

14. Kerrissey, J., & Schofer, E. (2013). Union membership and political participation in the United States. Social forces, 91(3), 895-928.

15. King, Gary. 1988. *"Statistical Models for Political Science Event Counts:Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model."* American Journal of Political Science 32(3):838-63.

16. King, Gary, Michael Tomz, and Jason Wittenberg. 2000. *"Making the Most of Statistical Analyses: Improving Interpretation and Presentation."* American Journal of Political Science 44: 341- 355.

17. Krain, Matthew. 1998. *"Contemporary Democracies Revisited: Democracy, Political Violence, and Event Count Models."* Comparative Political Studies April (31): 139-164.

18. Miller, Jane E. 2005. *The Chicago Guide to Writing about Multivariate Analysis.* Chicago, University of Chicago Press.

19. McVeigh, Rory. & Christian Smith. 1999. *"Who Protests in America: An Analysis of Three Political Alternatives - Inaction, Institutionalized Politics, or Protest."* Sociological Forum, 14, 4:685-702.

20. Mullen, Ann L., Kimberly A. Goyette, and Joseph A. Soares. 2003. *"Who Goes to Graduate School? Social and Academic Correlates of Educational Continuation After College."* Sociology of Education, 76,2:143-169.

21. Whitten, Guy D. and Harvey D. Palmer. 1996. *Heightening Comparativists' Concern for Model Choice: Voting Behavior in Great Britain and the Netherlands.* American Journal of Political Science 40:231-260.

Versão preliminar e sujeita a (pequenas) alterações, atualizada em: 16 de fevereiro de 2014.
`http://ricardoceneviva.com/Lego3/`