Modelos *logit* para Escolhas Discretas com Enfoque em Excel

Guilherme Lopes de Oliveira[1]

[1]Departamento de Estatística
Universidade Federal de Minas Gerais

Stat4Good, 20 de maio de 2016

Outline

Motivação

Princípio da Utilidade Máxima

Derivação dos Modelos de Escolhas Discretas

Modelo logit para Escolhas Binárias
- Logit Multinomial Nominal (LMN)
- Logit Multinomial Ordinal (LMO)
- Logit Condicional Nominal (LMO)

Motivação

- Modelos de escolha discreta são usados na modelagem das escolhas feitas por *indivíduos* dentre um conjunto **finito** de alternativas **exaustivas** e **mutuamente exclusivas**.

- As respostas são discretas. Então, em vez de examinar **quanto** como em problemas com variáveis de escolha contínuas, a análise da escolha discreta examina **qual**.

Motivação

- Modelos são úteis para prever como as escolhas das pessoas vão mudar sob mudanças na demografia e/ou atributos das alternativas. Exemplo: demanda ou propensão a compra de produtos.

- As probabilidades de escolha podem ser relacionadas aos atributos do indivíduo e/ou aos atributos das alternativas disponíveis.

- Daniel McFadden ganhou o prêmio Nobel de Economia em 2000 por seu trabalho pioneiro no desenvolvimento da base teórica neste tipo de problema (Modelo Logit Condicional).

Princípio da Utilidade Máxima

- Assume-se que o indivíduo escolhe a alternativa que lhe fornece a maior utilidade, ou seja, alternativa k é escolhida pelo indivíduo i se $U_{ik} > U_{ij}$, $\forall j \neq k$, sendo

$$U_{ij} = \beta' x_{ij} + \epsilon_{ij}.$$

  Diferentes modelos de escolha são obtidos através de diferentes especificações para a densidade conjunta do vetor $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{iJ})$.

Derivação das Probabilidades de Escolha

Only differences matter: "A rising tide raises all boats."

- A probabilidade de que o indivíduo i escolha a alternativa k é dada por

$$\begin{aligned} P_{ij} &= \text{Prob}(U_{ik} > U_{ij},\ \forall k \neq j) \\ &= \text{Prob}(\beta' x_{ik} + \epsilon_{ik} > \beta' x_{ij} + \epsilon_{ij},\ \forall k \neq j) \\ &= \text{Prob}(\epsilon_{ij} - \epsilon_{ik} < \beta' x_{ik} + \beta' x_{ij},\ \forall k \neq j) \\ &= \int_{\boldsymbol{\epsilon}} I\{\epsilon_{ij} - \epsilon_{ik} < \beta' x_{ik} + \beta' x_{ij},\ \forall k \neq j\} f(\boldsymbol{\epsilon}_i) d\boldsymbol{\epsilon}_i. \end{aligned}$$

Modelo logit para Escolhas Binárias

- Se $\epsilon_{ij} \overset{iid}{\sim}$ Valor Extremo, $j = 0, 1$, pensando em $P(U_{i1} > U_{i0})$ temos

$$P_{i1} = \frac{e^{\beta' x_{i1}}}{e^{\beta' x_{i1}} + e^{\beta' x_{i0}}}.$$

Modelos para Escolhas Múltiplas

Example

- Suponha que você é um diretor de escola de ensino fundamental interessado em que as crianças se identifiquem com sua matéria favorita e quer saber como isso está associado com a sua idade ou o método de ensino aplicado [debate (1) ou palestra (0)]. Trinta crianças, de 10 a 13 anos de idade, tiveram aulas teóricas em artes da linguagem (1), ciências (2) e matemática (3) que empregaram técnicas de debate ou palestra. No final do ano, elas foram solicitadas a identificar suas matérias favoritas. Queremos modelar as probabilidades de escolha das disciplinas.

Modelos para Escolhas Múltiplas

Example

- Suponha que você é um diretor de escola de ensino fundamental interessado em que as crianças se identifiquem com sua matéria favorita e quer saber como isso está associado com a sua idade ou o método de ensino aplicado [debate (1) ou palestra (0)]. Trinta crianças, de 10 a 13 anos de idade, tiveram aulas teóricas em artes da linguagem (1), ciências (2) e matemática (3) que empregaram técnicas de debate ou palestra. No final do ano, elas foram solicitadas a identificar suas matérias favoritas. Queremos modelar as probabilidades de escolha das disciplinas. Usa-se a modelo de regressão logística nominal porque a resposta é categórica e não possui nenhuma ordenação implícita nas categorias.

Modelos para Escolhas Múltiplas

Logit Multinomial Nominal (LMN):

- Indivíduo i escolhe uma dentre as alternativas $j = 1, 2, ..., J$ (sem ordenação intrínseca). O modelo LMN é construído supondo $J - 1$ modelos Logísticos Binários usando umas das J alternativas, digamos a k-ésima, como referência e pensando em $U_{ij} > U_{ik}$, $\forall j$. Daí, assume-se

$$\log\left(\frac{P_{ij}}{P_{ik}}\right) = e^{\beta_j' x_{ij}}, \ \forall j \leq k$$

tal que

$$P_{ik} = \frac{1}{1 + \sum\limits_{\forall l \neq k} e^{\beta_j' x_{il}}} \quad e \quad P_{ij} = \frac{e^{\beta_j' x_{ij}}}{1 + \sum\limits_{\forall l \neq k} e^{\beta_j' x_{il}}}, \ \forall j \leq k.$$

Modelos para Escolhas Múltiplas

Example

- Suponha que você seja um biólogo e acredita que a população adulta de salamandras no Nordeste diminuiu de tamanho nos últimos anos. Você gostaria de determinar se existe alguma relação entre o tempo de sobrevivência de uma salamandra e o nível de toxicidade da água e se existe algum efeito regional. O tempo de sobrevivência é codificado como 1 se < 10 dias, 2 se 10 a 30 dias e 3 se 31 a 60 dias. Queremos modelar as probabilidades associadas a cada categoria da resposta.

Modelos para Escolhas Múltiplas

Example

- Suponha que você seja um biólogo e acredita que a população adulta de salamandras no Nordeste diminuiu de tamanho nos últimos anos. Você gostaria de determinar se existe alguma relação entre o tempo de sobrevivência de uma salamandra e o nível de toxicidade da água e se existe algum efeito regional. O tempo de sobrevivência é codificado como 1 se < 10 dias, 2 se 10 a 30 dias e 3 se 31 a 60 dias. Queremos modelar as probabilidades associadas a cada categoria da resposta. Usa-se a modelo de regressão logística ordinal porque a resposta é categórica e possui uma ordenação implícita nas categorias.

Modelos para Escolhas Múltiplas

Logit Multinomial Ordinal (LMO) - *Odds Proporcionais*:

- Suponha agora que existe uma ordenação intrínseca entre as alternativas (como pensamos U_{ij} neste caso?!). O modelo LMO é baseado nas probabilidades acumuladas assumindo que

$$\log\left(\frac{\mathrm{P}(y_i \leq j)}{\mathrm{P}(y_i > j)}\right) = \alpha_j - \beta' x_{ij},$$

sendo $\alpha_1 < \alpha_2 < ... < \alpha_{J-1}$. Daí, para $j = 1, ..., J-1$

$$\mathrm{P}(y_i \leq j) = \frac{e^{\alpha_j - \beta' x_{ij}}}{1 + e^{\alpha_j - \beta' x_{ij}}} \quad e \quad \mathrm{P}(y_i = J) = 1.$$

Com isso,

$$\mathrm{P}(y_i = j) = \mathrm{P}(y_i \leq j) - \mathrm{P}(y_i \leq j-1).$$

Modelos para Escolhas Múltiplas

Example

- Uma determinada revendedora de *tablets* está interessada em estudar a demanda destes produtos com base em seus atributos como marca (Saamsung, Apple e LG), tamanho da tela (7' e 10'), memória (8Gb e 16Gb) e preço (R$400, R$600 e R$800). São confeccionadas fichas contendo combinações aleatórias destes atributos e então é coletada uma amostra da preferência de clientes em potencial. Para tal, simula-se um total de doze cenários, cada um destes com três fichas escolhidas aleatoriamente, e então cada cliente escolhe uma das cartas (*tablets*) em cada cenário. Queremos estudar as probabilidades de escolha (e potencial compra!) de determinado produto com base em suas características. Note que os clientes não precisam necessariamente opinar sobre todas as possíveis configurações dos produtos.

Modelos para Escolhas Múltiplas

Example

- Uma determinada revendedora de *tablets* está interessada em estudar a demanda destes produtos com base em seus atributos como marca (Saamsung, Apple e LG), tamanho da tela (7' e 10'), memória (8Gb e 16Gb) e preço (R$400, R$600 e R$800). São confeccionadas fichas contendo combinações aleatórias destes atributos e então é coletada uma amostra da preferência de clientes em potencial. Para tal, simula-se um total de doze cenários, cada um destes com três fichas escolhidas aleatoriamente, e então cada cliente escolhe uma das cartas (*tablets*) em cada cenário. Queremos estudar as probabilidades de escolha (e potencial compra!) de determinado produto com base em suas características. Note que os clientes não precisam necessariamente opinar sobre todas as possíveis configurações dos produtos. Neste contexto, surge o modelo de regressão logística condicional nominal.

Modelos para Escolhas Múltiplas

Logit Condicional Nominal (LCN) [McFadden]:

- Considere agora que a j-ésima alternativa é escolhida se ocorrer $U_{ij} > U_{ik}$, $\forall j \neq k$. Daí surge o modelo LCN no qual

$$\log\left(\frac{P_{ij}}{P_{ik}}\right) = e^{\beta_j' x_{ij}},$$

tal que

$$P_{ij} = \frac{e^{\beta' x_{ij}}}{\sum_{k=1}^{J} e^{\beta' x_{ik}}}.$$

Propriedade destes Modelos Logit

Independência sob Alternativas Irrelevantes (IAI):

- A razão de chances para quaisquer duas alternativas é independente das demais alternativas e de seus atributos. Isto implica *substituição proporcional* entre as alternativas. **Isto pode não der verdade em muitos casos!**

- Porém, se IAI de fato acontece, as estimativas dos parâmetros obtidas para um subconjunto de parâmetros não terão uma diferença significativa com relação às estimativas que seriam obtidas sob o conjunto de todas as alternativas. **Isto é interessante em termos práticos!**

Outros Modelos

- Modelos *probit* flexibilizam a restrição de que não há correlação entre os fatores não-observados para as alternativas. Modelos Valor Extremo Generalizados seriam outra opção.

- Modelos probit podem acomodar qualquer padrão de correlação ou heterocedastividade, mas somente sob normalidade.

- Modelos mais gerais: Nested Logit e Logit Misto.

Referências I

T. Kenneth.
Discrete Choice Methods with Simulation, University of California, Berkeley. National Economic Research Associates.
Cambridge University Press, 2002.

D. Schroeder.
Seminar on econometric analysis of accounting choice: Discrete choice models.
Papers and background materials for summer '06 seminar, chapter5:77–95, 2006.