Nearest neighbour prediction method in mixed logit discrete choice model.

Awo Sitsofe Tsagbey

Thesis presented
to
Institute of Mathematics and Statistics
of
University of São Paulo
to
obtain the title
of
Doctorate in Science

Program: Statistics

Advisor: Profa. Dra. Viviana Giampaoli

During the development of this work the author received
financial support from CNPq/CAPES

São Paulo, August 2021

Método de predição usando vizinho mais próximo em modelo misto logito de escolha discreta.

Awo Sitsofe Tsagbey

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
Obtenção do Título
de
doutorado em Ciências

Programa: Estatística

Orientador: Profa. Dr. Viviana Giampaoli

Durante o desenvolvimento deste trabalho o autor recebeu auxílo financeiro do CNPq/CAPES

São Paulo, agosto de 2021

Método de predição usando vizinho mais próximo em modelo misto logito de escolha discreta.

Versão Corrigida Simplificada

Comissão Julgadora:

- Profa. Dra. Viviana Giampaoli (orientadora) - IME-USP
- Prof. Dr. Gilberto Alvarenga Paula - IME-USP
- Prof. Dr. Juvêncio Santos Nobre - UFC
- Prof. Dr. Francisco José de Azavedo Cysneiros - UFPE
- Profa. Dra. Alejandra Andrea Tapia Silva - Univ. Cat. del Maule

Acknowledgements

I will praise the Lord at all times. I will constantly speak his praises. I will boast only in the Lord.

Psalm 34:1-2a

Thanks to God Almighty for how far He has brought me.

My heartfelt gratitude to my advisor Dr. Viviana Giampaoli for her guidance, motivation, support and patience during this work. I couldn't have asked for a better advisor. Thank you. I also want to appreciate my thesis committee for their comments and suggestions.

To all the professors and staff of the Department of Statistics, Institute of Mathematics and Statistics, University of Sao Paulo, I say thank you for your guidance, training and support in various ways.

I am forever grateful to my husband, my children, my siblings and my parents for their unwavering love and support. Also, my appreciation goes to all my friends both home and abroad and my colleagues in IME-USP who helped me to adapt to the Brazilian culture and also in various ways made my time in the department less stressful and meaningful.

God bless you all.

Abstract

TSAGBEY, A. S. **Nearest neighbour prediction method in mixed logit discrete choice model.** 2021. Thesis (Doctorate in Science - Statistics) - Institute of Mathematics and Statistics, São Paulo, 2021.

Discrete choice models (DCMs) are a group of models that are used to analyze choice data basically because they accommodate the nature of the process that generates the data. The most common types of discrete models include the logit, probit, multinomial logit, nested logit, mixed logit, and most recently the generalized multinomial logit. Discrete choice models have been mostly used in the area of economics, transportation, energy, psychology, etc. Prediction in these models isn't uncommon, in contexts such as engineering, marketing, and production, discrete choice models are mostly used to forecast demand. Making out-of-sample predictions at the individual level is an aspect of DCMs that has not been exploited much unlike in linear models. We found no work in literature on out-of-sample prediction at the individual level for complex models such as mixed logit, which involves predicting the random effects also known as the individual-specific parameters. Thus, in this is work we propose a method for this scenario in mixed logit discrete models using the nearest neighbour concept. This involves comparing both the characteristics of the individuals and also the choice set faced by them and using the nearest neighbour technique to predict the individual-specific parameters for a new individual based on the sample set. The predicted individual-specific parameters are then used to calculate the utility from each alternative which leads to estimating the probability of each alternative to be chosen and finally deducing the most probable choice for the individual. To evaluate the performance of our proposed method, we carry out simulations under various situations considering cross-sectional data

and panel data, quality of model fitted and sample sizes. We also compare our method to the rudimentary method where the population parameters are used. In summary, we see that of the quality of the model fitted slightly affects the performance of our method. Our proposed method performs as well as the rudimentary method when the better model was used whereas, with a good-fit model, our method overshadows the rudimentary methods in terms of performance/accuracy. This is true for both cross-sectional and panel data and likewise for all sample sizes. Applying our proposed method in two real-life datasets; cross-sectional data and panel data. We find that the prediction accuracy of our method is better than the rudimentary method. Even though our method is a bit more computational than the rudimentary method, it is a minor price to pay for a higher prediction accuracy while maintaining the explanatory power of the model.

Keywords: Discrete choice model, Mixed logit, Nearest Neighbour Prediction, Random effects.

Resumo

TSAGBEY, A. S. **Método de predição usando vizinho mais próximo em modelo misto logito de escolha discreta.** 2021. Tese (Doutorado em Ciências - Estatística) - Institute of Mathematics and Statistics, São Paulo, 2021.

Modelos de escolha discreta (DCMs) são um grupo de modelos usados para analisar dados de escolha basicamente porque eles acomodam a natureza do processo que gera os dados. Os tipos mais comuns de modelos discretos incluem o logito, probito, logito multinomial, nested logito, misto logito e, mais recentemente, o logito multinomial generalizado. Modelos de escolha discreta têm sido usados principalmente na área de Economia, Transporte, Energia, Psicologia, etc. Usando modelos de escolha discreta para predição é comum; em áreas como Engenharia, marketing e produção, são usados principalmente para prever a demanda. Fazendo out-of-sample predições no nível individual é um aspecto da DCMs que não tenham sido exploradas muito ao contrário de modelos lineares. Não encontramos nenhum trabalho na literatura sobre out-of-sample predição no nível individual para modelos complexos, como misto logito, que envolve a predição dos efeitos aleatórios também conhecidos como parâmetros específicos do indivíduo. Assim, neste trabalho propomos um método para este cenário em modelos discretos misto logito usando o conceito de vizinho mais próximo. Isso envolve comparar as características do indivíduo e também o conjunto de escolhas enfrentado por eles e usar a técnica do vizinho mais próximo para prever os parâmetros específicos do indivíduo para um novo indivíduo baseado da amostra. Os parâmetros específicos individuais preditos são usados para calcular a utilidade de cada alternativa, o que leva a estimar a probabilidade de cada alternativa ser escolhida e, finalmente, deduzir a escolha mais provável para o indivíduo.Para avaliar o desempenho do método proposto, realizamos

simulações em várias situações, considerando o tipo de conjunto de dados (cross-sectional e panel), qualidade do modelo ajustado e tamanhos de amostra. Também comparamos nosso método com o método rudimentar, onde os parâmetros populacionais são usados. Em resumo, vemos que a qualidade do modelo ajustado afeta ligeiramente o desempenho do nosso método. Nosso método proposto tem um desempenho tão bom quanto o método rudimentar quando o melhor modelo foi usado, ao passo que, com um modelo de bom ajuste, nosso método ofusca os métodos rudimentares em termos de desempenho / acurácia. Isso é verdadeiro para ambos tipos de conjunto de dados e da mesma forma para todos os tamanhos de amostra. Aplicando nosso método proposto em dois conjuntos de dados reais, descobrimos que a acurácia da predição de nosso método é melhor do que o método rudimentar. Mesmo que nosso método seja um pouco mais computacional do que o método rudimentar, é um preço menor a pagar por uma maior acurácia de predição, mantendo o poder explicativo do modelo.

Palavras-chave: Modelo de escolha discreta, misto logito, predição, vizinho mais próximo, efeitos aleatórios.