Robust beta regression through the logit transformation

Yuri Sampaio Maluf

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Doutorado em Estatística
Orientadora: Prof. Dra. Silvia Lopes de Paula Ferrari

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES e do CNPq

Março, 2023

Robust beta regression through the logit transformation

Versão original simplificada

Tese apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Março, 2023

Ficha catalográfica elaborada com dados inseridos pelo autor
Biblioteca Carlos Benjamin de Lyra
Instituto de Matemática e Estatística
Universidade de São Paulo

Maluf, Sampaio
Robust beta regression through the logit transformation / Sampaio Maluf; orientadora, Silvia Lopes de Paula Ferrari. - São Paulo, 2023.
121 p.: il.

Tese (Doutorado) - Programa de Pós-Graduação em Estatística / Instituto de Matemática e Estatística / Universidade de São Paulo.
Bibliografia
Versão corrigida

1. Estimadores robusto. 2. Dados de proporção. 3. Inferência robusta. 4. L_q-verossimilhança. 5. Regressão Beta. I. Lopes de Paula Ferrari, Silvia. II. Título.

Bibliotecárias do Serviço de Informação e Biblioteca
Carlos Benjamin de Lyra do IME-USP, responsáveis pela
estrutura de catalogação da publicação de acordo com a AACR2:
Maria Lúcia Ribeiro CRB-8/2766; Stela do Nascimento Madruga CRB 8/7534.

Agradecimentos

Primeiramente aos meus pais, Ana Helena, minha mentora na Terra, e Fares, por todo o apoio irrestrito, carinho, suporte e amor. Esses personagens são centrais em minha vida por sempre acreditarem em minha capacidade. Aos outros membros da minha família Ana Júlia, Ana Cristhina e Lindinalva meu todo agradecimento. Rendo aqui também meus agradecimentos póstumos a outros integrantes da família que infelizmente não puderam me assistir ao fim dessa trajetória, minha tia Ana Maria e aos meus tios Timos e Augusto.

Agradeço a minha professora e orientadora Silvia Ferrari pela atenção e a ajuda fundamental na construção deste trabalho. Suas criteriosas avaliações e cobranças me permitiram evoluir como futuro pesquisador. Cabe destacar ainda sua imensa paciência, compreensão e trato ao longo de minha trajetória no doutorado que em especial foi recheada de problemas pessoais e de saúde. Portanto, sua dedicação foi de vital importância pra mim a qual reconheço e pela qual sou grato imensamente.

Toda gratidão ao meu amigo Felipe Queiroz, por suas enormes contribuições, ajudas, dicas, conversas, encorajamentos em um dos momentos mais difíceis na construção desse trabalho.

Agradeço a todos os membros da banca pelas contribuições postas de maneira tão generosa. Aos meus professores do doutorado por sua entrega e doação no exercício mais nobre que a civilização humana tem, a transmissão do conhecimento.

À Regiane pela imensa e reiterada ajuda prestada com toda eficiência e celeridade nos assuntos da pós-graduação, uma profissional incrível e sempre com um sorriso e carregada de leveza.

À Paula, minha grande companheira, por todo amor, amizade, e força. Não foram poucos os momentos difíceis e de obstáculos aos quais não faltaram seu apoio, ombro e carinho. Sem sua presença tudo seria cinza e nublado.

Aos diversos amigos e companheiros que conquistei ao longo do doutorado, Calado, César, Felipe, Daniel, Jadevilson, Eurípedes, Elivaldo, Oscar, Jaime, Jairo, Terezinha e entre outros, com os quais compartilhei muitos momentos bons e com grande trocas de conhecimentos e experiências. Incluo aqui também a participação fundamental do meu psicólogo Paulo e psiquiatra Amanda por todo cuidado e suporte.

Um agradecimento especial a minha ex-companheira Denise, por todo apoio, compreensão e solidariedade. Gratidão por estar sempre presente ao longo de toda minha vida acadêmica, desde as graduações, passando pelo mestrado e doutorado.

Todo meu respeito e gratidão à cena do Psytrance, que me resgatou e me convidou a vivenciar integralmente o meu ser. Foi nela que encontrei uma nova casa e cura. O meu muito obrigado!

A todos que não foram citados aqui mas, que de alguma maneira, contribuíram nas mais diversas formas como uma simples conversa, gestos e incentivos. À Universidade de Brasília, meu grande templo intelectual, e a Universidade de São Paulo por terem proporcionado a minha formação. À Capes e ao CNPq, pelo apoio financeiro.

Abstract

Beta regression models are employed for modeling continuous response variables in the unit interval, like rates, percentages, or proportions. Their applications appear in several areas, such as medicine, environmental research, and finance. Under beta regression models, the usual inference procedure is essentially based on the classical maximum likelihood approach. Nevertheless, it is well known that the maximum likelihood-based inference is easily affected by the presence of outliers. The lack of robustness may bring a severe bias and misleading conclusions. Recently, robust estimators for beta regression models have been proposed. These estimators require non-trivial restrictions in the parameter space, which limit their application. This thesis proposes two new robust estimators, namely the logit minimum power divergence estimator (LMDPDE) and the logit surrogate maximum likelihood estimator (LSMLE), that overcome this drawback. The LMDPDE and the LSMLE belong to the general class of M-estimators, which allows us to derive some important properties, such as the asymptotic distribution and the influence function. We also develop robust versions of the Wald-type test. All robust estimators, statistical tests, and other tools treated in this work are implemented in the R package robustbetareg. The package is available on the official repository of R (CRAN). Through Monte Carlo simulation studies, we examine the performance of the proposed estimators and the robust Wald tests and compare them with the corresponding inference procedures presented in the literature. Real data applications in the context of sports of Australian athletes and access to health insurance coverage in a Brazilian state using the proposed methodologies are presented. The thesis closes with concluding remarks about its main contributions and points out some important aspects that deserve special attention for future research.

Keywords: Beta regression, L_q-likelihood, Proportional data, Robust estimators, Robust inference.

Resumo

Os modelos de regressão beta são empregados para modelar variáveis de resposta contínua no intervalo da unidade, como taxas, porcentagens ou proporções. Suas aplicações surgem em diversas áreas como medicina, pesquisa ambiental e finanças. Em modelos de regressão beta, o procedimento de inferência usual é essencialmente baseado na abordagem clássica de máxima verossimilhança. No entanto, é sabido que a inferência baseada em máxima verossimilhança é facilmente afetada pela presença de *outliers*. A falta de robustez pode trazer um viés severo e conclusões enganosas. Recentemente, estimadores robustos para modelos de regressão beta foram propostos. Esses estimadores requerem restrições não triviais no espaço paramétrico, o que limita sua aplicação. Esta tese propõe dois novos estimadores robustos, a saber, o estimador logit de divergência de potência mínima (LMDPDE) e o estimador logit de máxima verossimilhança substituta (LSMLE), que superam essa desvantagem. O LMDPDE e o LSMLE pertencem à classe geral dos M-estimadores, o que nos permite derivar algumas propriedades importantes, como a distribuição assintótica e a função de influência. Também desenvolvemos versões robustas do teste do tipo Wald. Todos os estimadores robustos, testes estatísticos e outras ferramentas tratadas neste trabalho são implementados no pacote robustbetareg em linguagem R. O pacote está disponível no repositório oficial do R (CRAN). Através de estudos de simulações de Monte Carlo, examinamos o desempenho dos estimadores propostos, do teste robusto de Wald e os comparamos com os procedimentos de inferência correspondentes apresentados na literatura. São apresentadas aplicações de dados reais no contexto esportivo de atletas australianos e acesso à cobertura de plano de saúde em um estado brasileiro utilizando as metodologias propostas. A tese encerra com considerações finais sobre suas principais contribuições e aponta alguns aspectos importantes que merecem atenção especial para pesquisas futuras.

Palavras-chave: Estimadores robusto, Dados de proporção, Inferência robusta, L_q-verossimilhança, Regressão Beta.