Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

Rio de Janeiro, Brazil

14 de junho de 2021

Lucas Machado Moschen

Bayesian analysis of respondent-driven surveys with outcome uncertainty

Monograph Project presented to the School of Mathematics Applied (FGV) as a partial requirement for continuing the monograph work.

Getulio Vargas Foundation – FGV

School of Applied Mathematics

Undergraduate Course in Applied Mathematics

Rio de Janeiro, Brazil

14 de junho de 2021

Contents

1 Introduction

> O texto deve ser constituído de uma parte introdutória, na qual devem ser expostos o tema do projeto, o problema a ser abordado, a(s) hipótese(s), quando couber(em), bem como o(s) objetivo(s) a ser(em) atingido(s) e a(s) justificativa(s). É necessário que sejam indicados o referencial teórico que o embasa, a metodologia a ser utilizada, assim como os recursos e o cronograma necessários à sua consecução.

The proposal of this work is to study the survey method Respondent-Driven Sampling, a chain-referral method with the objetive of sample from hard-to-reach populations, when it is necessary to estimate the prevalence of some characteristic from this population. It is also regarded the imperfection of the tests of this characteristic, so sensibility and sensitivity are accounted in the model.

Hidden or hard-to-reach populations have two main features: no sampling frame exists, so size and boundaries of the population are unknown; and there are heavy privacy concerns because the subjects are stigmatized or have illegal behavior (HECKATHORN, 1997). Fear of exposition of prosecution complicates the enumeration of the populations and the learning about them. Moreover, if the occurrence frequency of the characteristic is low, there is high logistic cost involved. Some examples are heavy drug users, sex workers, homeless people, and men who have sex with men.

Some methods were developed to reach these populations, such as, for example, snowball method (GOODMAN, 1961), key important method (DEAUX; CALLAGHAN, 1985), and targeted method (WATTERS; BIERNACKI, 1989). HECKATHORN introduced the Respondent-Driven Sampling (RDS) to fill some gaps he depicted in his work. In his proposed method, the researches select a handful of individuals from the target population and give them coupons to recruit others from the population. The individuals receive a reward for being recruited and for each recruitment, which creates a dual incentive system. Several papers after 1997 were written.

After sampling individuals from the target population, a questionnaire is conducted, and in this work takes questions with binary outcome.

Falar um pouco de como esse estudo vai ser feito: estatística bayesiana

Falar um pouco do objetivo deste trabalho.

1.1 Respondent-driven sampling

1.2 Regression with binary outcome with imperfect tests

1.3 Bayesian statistics

Respondant-Driven Sampling (RDS) é um procedimento utilizado para amostrar populações de difícil acesso, como exemplo a população de usuários de drogas pesadas e profissionais do sexo. Ele funciona de forma similar a um processo de ramificação em formato de rede, em que os participantes em cada estágio recrutam, em sua própria sub-rede, os próximos participantes e o primeiro estágio é chamado de semente.

Esse método pode ser utilizado em forma de pesquisa a fim de estimar a prevalência de alguma característica, isto é, o número total de indivíduos que possuem determinada característica. Nessa pesquisa, cada participante responde uma série de perguntas relacionadas ao objeto de estudo e outras covariáveis. Vamos considerar neste trabalho que o desfecho de interesse é uma variável binária e sujeita a erro de medição, isto é, não é possível ter certeza sobre a veracidade da resposta dada. Usamos os conceitos de sensitividade e especificidade para lidar com isso.

Todavia, em vista de nosso desconhecimento sobre a natureza em si, se faz necessário modelar a incerteza dessas variáveis e, para tanto, a Estatística bayesiana é a área de estudo indicada. A ideia, portanto, é propagar a incerteza sobre a resposta dos participantes pela rede de contatos.

Por fim, pretende-se aplicar esse framework de forma eficiente, em particular, comparando os algoritmos de Markov chain Monte Carlo e Aproximação de Laplace aninhada (INLA) e programando-os com ajuda de alguma linguagem de programação como R, Stan ou Python.

2 Justification

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3 Objectives

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4 Methodology

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

5 Preliminary results

a) Descrição do problema em termos matemáticos e revisão bibliográfica: material sobre RDS (formalização matemática em forma de cadeia ou processo de ramificação), regressão logística em que a resposta tem incerteza e aplicações em usuários de drogas, infecções transmissíveis, entre outros.

b) Incerteza sobre especificidade e sensitividade do teste e como propagar a classificação errada na rede. Comparação de prioris e, por isso, estudo de métodos Bayesianos. Justificar utilização desses métodos com argumento da incerteza.

c) Estudo do MCMC e Aproximação de Laplace, comparação dos algoritmos em alguns artigos e, quem sabe, codificação em Python e R.

d) Implementação de inferência eficiente em INLA, com possibilidades abertas em Python (talvez Julia?)

6 Schedule

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Final considerations

Sed consequat tellus et tortor. Ut tempor laoreet quam. Nullam id wisi a libero tristique semper. Nullam nisl massa, rutrum ut, egestas semper, mollis id, leo. Nulla ac massa eu risus blandit mattis. Mauris ut nunc. In hac habitasse platea dictumst. Aliquam eget tortor. Quisque dapibus pede in erat. Nunc enim. In dui nulla, commodo at, consectetuer nec, malesuada nec, elit. Aliquam ornare tellus eu urna. Sed nec metus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Phasellus id magna. Duis malesuada interdum arcu. Integer metus. Morbi pulvinar pellentesque mi. Suspendisse sed est eu magna molestie egestas. Quisque mi lorem, pulvinar eget, egestas quis, luctus at, ante. Proin auctor vehicula purus. Fusce ac nisl aliquam ante hendrerit pellentesque. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi wisi. Etiam arcu mauris, facilisis sed, eleifend non, nonummy ut, pede. Cras ut lacus tempor metus mollis placerat. Vivamus eu tortor vel metus interdum malesuada.

Sed eleifend, eros sit amet faucibus elementum, urna sapien consectetuer mauris, quis egestas leo justo non risus. Morbi non felis ac libero vulputate fringilla. Mauris libero eros, lacinia non, sodales quis, dapibus porttitor, pede. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi dapibus mauris condimentum nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam sit amet erat. Nulla varius. Etiam tincidunt dui vitae turpis. Donec leo. Morbi vulputate convallis est. Integer aliquet. Pellentesque aliquet sodales urna.

Bibliography

DEAUX, E.; CALLAGHAN, J. W. Key informant versus self-report estimates of health-risk behavior. *Evaluation Review*, v. 9, n. 3, p. 365–368, 1985. Disponível em: <https://doi.org/10.1177/0193841X8500900308>. Citado na página 3.

GOODMAN, L. A. Snowball Sampling. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 32, n. 1, p. 148–170, 1961. Disponível em: <https://doi.org/10.1214/aoms/1177705148>. Citado na página 3.

HECKATHORN, D. D. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, [Oxford University Press, Society for the Study of Social Problems], v. 44, n. 2, p. 174–199, 1997. Disponível em: <http://www.jstor.org/stable/3096941>. Citado na página 3.

WATTERS, J. K.; BIERNACKI, P. Targeted sampling: Options for the study of hidden populations. *Social Problems*, Oxford University Press, Society for the Study of Social Problems, v. 36, n. 4, p. 416–430, 1989. Disponível em: <http://www.jstor.org/stable/800824>. Citado na página 3.

Introdução