Subprojeto de Iniciação Científica

Edital:	Edital PIBIC 2020 / 2021
Título do Projeto:	Ciência de Dados e Aprendizado Estatístico Aplicados à Saúde
Título do Subprojeto:	Modelos de regressão para desfecho escalar e preditores funcionais - soluções para problema da medicina obstétrica
Candidato a Orientador:	Agatha Sacramento Rodrigues
Candidato a Bolsista:	A definir a (o) estudante
Membros Equipe do Projeto:	Maria de Lourdes Brizot (Departamento de Obstetrícia da USP)

1 Resumo

É comum encontrar situações na área da medicina obstétrica nas quais o interesse consiste em estudar um desfecho escalar (isto é, avaliada em um único momento), em geral alguma informação do parto, com relação a variáveis explicativas (covariáveis) medidas em diferentes momentos ao longo do pré-natal, nomeadas de covariáveis funcionais. Na prática, pesquisadores consideram no modelo de regressão uma única medida da covariável funcional (uma medida resumo, como a média das medições, ou a última avaliação antes do desfecho de interesse), o que leva à perda de informação por impossibilitar a análise da relação da variação dessa covariável com o desfecho escalar e por não representar todo o comportamento da covariável ao longo das avaliações. A utilização dessas soluções se deve, muitas vezes, por desconhecimento de métodos estatísticos adequados para lidar com dados funcionais e/ou devido à ausência de ferramentas mais acessíveis aos pesquisadores aplicados. Com esse estudo se objetiva estudar modelos que corretamente posicionam e consideram covariáveis funcionais com desfecho escalar sob o paradigma Bayesiano, considerando também componentes de interação entre duas covariáveis funcionais e interação entre uma covariável funcional e uma covariável fixa. Além do desenvolvimento teórico, esse projeto também tem como objetivo tornar os métodos propostos aplicáveis ao disponibilizar ferramentas para seus ajustes, além de discuti-los sob a perspectiva de sua interpretabilidade. Os métodos propostos serão aplicados aos dados reais do Departamento de Obstetrícia da Faculdade de Medicina da Universidade de São Paulo, que motivam esse estudo. Com esse projeto, espera-se fortalecer ainda mais a pesquisa teórica em dados funcionais, assim como fazê-la acessível aos pesquisadores práticos para que analisem seus dados de maneira correta, sem a perda de informações.

Palavras chaves: Covariáveis funcionais. Desfecho escalar. Ferramentas aplicadas. Modelos paramétricos. Paradigma Bayesiano.

2 Introdução

Em pesquisas na área da medicina obstétrica, é muito comum encontrar situações nas quais o interesse consiste em estudar um desfecho escalar (isto é, avaliado em um único momento), em geral alguma informação do parto, com relação a variáveis explicativas (covariáveis) medidas em diferentes momentos ao longo do pré-natal, nomeadas de covariáveis funcionais.

No problema que soluções serão propostas e que motiva esse projeto, o objetivo consiste em estudar a relação entre variáveis clínicas, histórico familiar, histórico obstétrico e medidas gestacionais com a idade gestacional do parto (desfecho escalar) em gestações de gemelares (de gêmeos). Além de variáveis fixas (seus status ou valores não se alteram ao longo do tempo), também foram registrados os valores das variáveis avaliadas em diferentes momentos do pré-natal: medida do colo do útero e número de contrações durante o exame de cardiotocografia. Esses dados fazem parte do estudo realizado no Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (FMUSP) e liderado pela médica obstetra e parceira nesse projeto de pesquisa, a Profa. Dra. Maria de Lourdes Brizot.

O estudo da idade gestacional do parto em gestações gemelares é importante devido ao elevado risco de prematuridade em gestações múltiplas. Entre as mulheres com gestação gemelar, o parto prematuro que ocorre antes das 37 semanas é observado em mais de 50% dos casos e quase 12% antes de 32 semanas completas de gestação (Silva, 1995). Devido a este fato, observa-se uma taxa de mortalidade neonatal nas gestações gemelares de 6,4 vezes maior que nas gestações de único feto (into Maternal, 2009). Por esse motivo se faz necessário estudar os fatores de riscos relacionados ao parto prematuro em gestações gemelares e dentre esses fatores, há também variáveis funcionais.

O fato de ter uma variável resposta escalar e querer relacioná-la a covariáveis funcionais é um desafio. Na prática, pesquisadores consideram no modelo de regressão uma única medida da covariável funcional (uma medida resumo, como a média das medições, ou a última avaliação antes do desfecho de interesse), o que leva à perda de informação por impossibilitar a análise da relação da variação dessa covariável com o desfecho escalar e por não representar todo o comportamento da covariável ao longo das avaliações. A utilização dessas medidas que perdem informação dos dados se deve, muitas vezes, por desconhecimento de métodos estatísticos adequados para lidar com dados funcionais e/ou devido à ausência de ferramentas mais acessíveis aos pesquisadores aplicados.

Apesar de serem utilizadas as soluções citadas no parágrafo anterior, alguns trabalhos teóricos propõem modelos para desfecho escalar e covariáveis funcionais são encontrados na literatura e que auxiliarão no desenvolvimento desse projeto. Goldsmith et al. (2011a) expandem a função coeficiente associada à covariável funcional por splines e Malloy et al. (2010) realizam essa expansão por ondaletas. Li et al. (2013) consideram a análise de componentes principais

funcional ao adotar uma abordagem orientada por dados para a expansão da função coeficiente. Sob uma abordagem não paramétrica, Zhang et al. (2014) utilizam o estimador de Nadaraya–Watson para a esperança da variável resposta escalar condicional a covariável funcional. Lian (2013) propõe um método de seleção de variáveis quando há mais de uma covariável funcional. Todos trabalhos citados até aqui consideram que a variável resposta é contínua e Müller et al. (2005) generalizam para o cenário que a distribuição da variável resposta que pertence à família exponencial (modelo linear generalizado), permitindo soluções para variável resposta de contagem ou binária, por exemplo.

O interesse desse projeto consiste em estudar modelos que corretamente posicionam e consideram covariáveis funcionais com desfecho escalar, motivados pelas necessidades metodológicas dos dados da motivação. Este trabalho não se resume apenas em desenvolver esses modelos do ponto de vista teórico, mas também torná-los aplicáveis ao disponibilizar ferramentas para seus ajustes, além de discuti-los sob a perspectiva de interpretabilidade. Esta etapa será bastante discutida com o grupo de estudos de gemelares do Departamento de Obstetrícia da FMUSP.

Esse subprojeto é ligado ao projeto de pesquisa “Ciência de Dados e Aprendizado Estatístico Aplicados à Saúde” (número de registro 10225/2020 e certificado no CNPq - dgp.cnpq.br/dgp/espelhogrupo/2755579833615592), coordenado pela professora orientadora e que objetiva aplicar e desenvolver metodologias na área de ciência de dados para resolver problemas e inovação na área da saúde.

Esta proposta está dividida como segue: na Seção 3 serão descritos os objetivos geral e específicos e na Seção 4 serão discutidas as abordagens propostas para alcançar os objetivos desse projeto. Por fim, na Seção 5 será apresentado o plano de desenvolvimento das atividades com o respectivo cronograma.

3 Objetivos

O objetivo geral desse projeto é avaliar os efeitos das covariáveis fixas de históricos familiar, obstétrico e clínico e covariáveis avaliadas durante o pré-natal, incluindo covariáveis funcionais, na idade gestacional do parto, ao modelar conjuntamente as covariáveis funcionais (expansão por splines, ondaletas ou orientada por dados) e a variável resposta, descrita por um modelo linear generalizado, sob o paradigma inferencial Bayesiano.

São os objetivos específicos:

1. Comparar as abordagens que serão utilizadas para expansão das funções coeficientes associadas as covariáveis funcionais por meio de estudos de simulação em cenários baseados nos dados da motivação do projeto;

2. Discutir a interpretabilidade das abordagens consideradas - ponto importante quando fala-

se da aplicabilidade dos modelos propostos e de grande relevância prática;

3. Desenvolver códigos computacionais bem documentados para permitir a reprodutibilidade e também a aplicabilidade dos métodos propostos e desenvolvidos.
4. Desenvolver uma plataforma web por meio de um aplicativo Shiny (https://www.shinyapps.io/) para melhor acessibilidade dos modelos aos usuários.
5. Apresentar e discutir os resultados com o grupo de pesquisa de gemelares do Departamento de Obstetrícia da FMUSP, liderado pela Profa. Dra. Maria de Lourdes Brizot.

4 Metodologia

Considere uma amostra de n indivíduos e sejam $(y_i, \mathbf{z}_i, \mathscr{X}_{1i}, \mathscr{X}_{2i}, \ldots, \mathscr{X}_{qi})$ os dados para o i-ésimo indivíduo, com $i = 1, \ldots, n$, em que y_i denota a observação da variável resposta, $\mathbf{z_i}$ é o vetor de covariáveis fixas (não funcionais) e $\mathscr{X}_{ji}$ denota a j-ésima covariável funcional, com $j = 1, \ldots, q$, que pertence ao espaço $\mathscr{F}$ de funções reais em um intervalo finito $\mathscr{T} \subset \mathscr{R}$, ou seja, $\{\mathscr{X}_{ji}(t),\ t \in \mathscr{T}\}$.

É assumido aqui que a distribuição da variável resposta Y pertence à família exponencial (Müller et al., 2005) e sejam $\mu = \mathrm{E}(Y \mid \mathbf{z}, \mathscr{X}_1(t), \mathscr{X}_2(t), t \in \mathscr{T})$ e $\sigma^2(\mu) = \mathrm{Var}(Y \mid \mathbf{z}, \mathscr{X}_1(t), \mathscr{X}_2(t), t \in \mathscr{T})$ a esperança de Y condicional às covariáveis e a função variância, respectivamente. Seja $g(\cdot)$ uma função monótona e diferenciável, que liga a média condicional de Y ao preditor linear η, ou seja, $\mu = g(\eta)$. Assim, seja

$$\eta = \mathbf{z}^\top \alpha + \sum_{j=1}^{q} \int_{\mathscr{T}} \beta_j(t) \mathscr{X}_j(t) dt,$$

em que α denota o vetor de coeficientes associados às covariáveis fixas $\mathbf{z}$ e $\beta_j(\cdot)$ é a função coeficiente associada à j-ésima covariável funcional $\mathscr{X}_j(\cdot)$, com $j = 1, \ldots, q$.

Sob o paradigma Bayesiano, três abordagens para a expansão dos parâmetros associados as covariáveis funcionais serão consideradas: a abordagem por splines de Goldsmith et al. (2011a), por ondaletas de Malloy et al. (2010) e a abordagem orientada por dados de Li et al. (2013). Uma das contribuições teóricas desse projeto é estender esses métodos para cenários com mais de uma covariável funcional. As abordagens serão comparadas em estudos de simulação em cenários baseados nos dados que motivam esse projeto, e também serão discutidas as abordagens consideradas na questão de interpretabilidade.

Outra contribuição teórica é a inclusão de componentes de interação entre uma covariável fixa e uma covariável funcional (McKeague & Qian, 2014) e interação entre duas covariáveis funcionais, questão pouco discutida na literatura e com grande demanda prática.

Algumas vantagens podem ser citadas ao considerar o processo inferencial sobre o paradigma Bayesiano, como a utilização de informações a priori (bastante relevante em aplicações

da área da saúde, em que se pode elicitar a distribuição a priori com base em estudos anteriores e/ou experiência dos pesquisadores) e a possibilidade de estimação das medidas da distribuição a posteriori, seja via métodos Monte Carlo Markov-Chain (Bigelow & Dunson, 2009) ou por métodos variacionais de Bayes (Goldsmith et al., 2011b), ao invés de maximizar a função de verossimilhança, o que não é algo tão trivial no cenário posto.

Nos dados que motivam esse projeto, são as covariáveis fixas (denotadas por $\mathbf{z}$): corionicidade (se monocoriônica, dicoriônica ou massa placentária única), idade da gestante, escolaridade da gestante, cor da pele, IMC pré gestacional, paridade (número de partos anteriores da gestante), aborto (número de abortos anteriores), fumante (sim ou não), se bebe alcool (sim ou não), se usa alguma droga ilícita (sim ou não) e comorbidade materna. As covariáveis funcionais (denotadas por $\mathscr{X}.$) medida do colo do útero e número de contrações avaliadas durante o exame de cardiotocografia. A variável resposta escalar é a idade gestacional do parto (IGP, denotada por Y) que será analisada como variável contínua (em semanas) e como variável binária, sendo a indicadora de prematuridade (sim, se IGP $<$37 semanas e não, caso contrário).

Todo o desenvolvimento computacional será realizado no software livre R (www.r-project.org/) e tudo será documentado e disponível no GitHub (https://github.com), conforme orientam as boas práticas na pesquisa e possibilidade de reprodutibilidade. A prática de disponibilidade computacional no GitHub já é rotina da orientadora proponente, como pode ser visto em seu repositório: github.com/agathasr.

Um aplicativo Shiny (https://www.shinyapps.io/) será desenvolvido para a acessibilidade dos modelos aos usuários. Um exemplo de aplicativo Shiny desenvolvido pela orientadora proponente pode ser visto em https://obstetriciafmusp.shinyapps.io/growthcurves/, em que são disponibilizadas as curvas de crescimento fetal construídas em um projeto coordenado pela candidata a orientadora.

Esse projeto faz parte da linha de pesquisa certificada no CNPq “Ciência de Dados e Aprendizado Estatístico Aplicados à Saúde” (dgp.cnpq.br/dgp/espelhogrupo/2755579833615592), liderado pela orientadora proponente, que objetiva aplicar e desenvolver metodologias na área de ciência de dados para resolver problemas e inovação na área da saúde. Por entender que é necessário estreitar a distância que muitas vezes separam a pesquisa em Estatística com as demais pesquisas na área da saúde, a proponente lidera outros dois projetos com aplicação na área da medicina obstétrica: 1) Métodos de aprendizado de máquinas aplicados ao desenvolvimento de curvas de crescimento fetal personalizadas para a população brasileira (mais informações em https://daslab-ufes.github.io/projetos/); e 2) Observatório Obstétrico (mais informações em https://daslab-ufes.github.io/projetos/projeto-observatório-obstétrico). Como consequências da aproximação entre pesquisadores estatísticos e pesquisadores aplicados, são destacadas as análises estatísticas mais acuradas nas pesquisas baseadas em evidências e oportunidades para desenvolvimento de novos métodos estatísticos baseados em problemas reais.

Todos esses projetos fazem parte do laboratório de *Data Science* (DaSLab), coordenado pela orientadora candidata. O DaSLab visa promover a discussão, apresentar o desenvolvimento e as potencialidades e oxigenar a área de Ciência de Dados, enfatizando o elemento multidisciplinar e fundamental da pesquisa aplicada e teórica. Os projetos desenvolvidos no DaSLab visam trazer soluções nas áreas de saúde, segurança pública, confiabilidade de componentes e sistema, ao combinar metodologias estatísticas, big data e *machine learning*. Mais informações sobre os projetos do DaSLab podem ser encontradas em https://daslab-ufes.github.io/.

5 Plano de trabalho/Cronograma

A seguir serão apresentadas as atividades que serão desenvolvidas com os respectivos prazos para realização.

ATIVIDADES

Lista de atividades
1- Revisão sistemática - Nessa atividade será aprofundado o levantamento bibliográfico realizado para a confecção desse projeto.
2- Análise inicial dos dados da motivação - Uma análise descritiva e exploratória dos dados será realizada de maneira profunda para entender com detalhes os dados que motivam esse projeto. Nessa etapa, serão realizadas reuniões com os pesquisadores do grupo de pesquisa de gemelares do Departamento de Obstetrícia da FMUSP para discussão e entendimento dos dados.
3- Desenvolvimento dos métodos - Uma vez realizada a revisão sistemática e entendidos todos os detalhes dos dados, os métodos serão desenvolvidos de maneira teórica, estudando suas possíveis postulações teóricas e possíveis restrições. Também nessa etapa, os métodos serão discutidos na questão de sua interpretabilidade.
4- Desenvolvimento computacional - Os métodos propostos serão implementados computacionalmente no software R. Tudo será documentado e disponível no GitHub (https://github.com), conforme orienta as boas práticas na pesquisa e reprodutibilidade na Ciência.
5- Realização do relatório parcial da pesquisa.
6- Desenvolvimento de plataforma web – Os métodos propostos serão disponíveis em um aplicativo Shiny (https://www.shinyapps.io/) para melhor acessibilidade dos modelos aos usuários.
7- Estudos de simulação e comparação dos métodos - Os métodos propostos serão comparados em estudos de simulação em cenários baseados nos dados que motivam esse projeto.
8- Aplicação dos métodos aos dados de motivação - Os métodos propostos e implementados serão aplicados aos dados da medicina obstétrica que motivam esse estudo.
9- Apresentação dos resultados - Os resultados desse projeto serão divulgados e apresentados aos pesquisadores do Departamento de Obstetrícia da FMUSP.
10- Realização do relatório final da pesquisa.

CRONOGRAMA (Ago/2020 a Jul/2021)

Atividade	ago	set	out	nov	dez	jan	fev	mar	abr	mai	jun	jul
1	X	X										
2		X	X									
3			X	X	X							
4					X	X	X					
5						X						
6							X	X				
7								X	X	X		
8										X	X	
9											X	X
10												X

Referências

Bigelow, J. L. & Dunson, D. B. (2009). Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association*, 104(485):26–36.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2011a). Penalized functional regression. *Journal of computational and graphical statistics*, 20(4):830–851.

Goldsmith, J., Wand, M. P., & Crainiceanu, C. (2011b). Functional regression via variational bayes. *Electronic journal of statistics*, 5:572.

into Maternal, C. E. (2009). Child health (cemach) perinatal mortality 2007. *United Kingdom. London: CEMACH*.

Li, Y., Wang, N., & Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108(504):1284–1294.

Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression. *Statistica Sinica*, pages 51–74.

Malloy, E. J., Morris, J. S., Adar, S. D., Suh, H., Gold, D. R., & Coull, B. A. (2010). Wavelet-based functional linear mixed models: an application to measurement error–corrected distributed lag models. *Biostatistics*, 11(3):432–452.

McKeague, I. W. & Qian, M. (2014). Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24(3):1461.

Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *the Annals of Statistics*, 33(2):774–805.

Silva, J. (1995). Prematuridade: aspectos obstétricos. *Neme B. Obtetricia básica*, 2:372–379.

Zhang, X., King, M. L., & Shang, H. L. (2014). A sampling algorithm for bandwidth estimation in a nonparametric regression model with a flexible error density. *Computational Statistics & Data Analysis*, 78:218–234.