Universidade Federal do Espírito Santo
Programa Institucional de Iniciação Científica
Relatório Final de Pesquisa
Ciências Exatas e da Terra

Modelos de Regressão Para Desfecho Escalar e Preditores Funcionais - Soluções Para Problema da Medicina Obstétrica

Edital:	Edital PIIC 2020 /2021
Grande Área do Conhecimento (CNPq):	Ciências Exatas e da Terra
Área do Conhecimento (CNPq):	Probabilidade e Estatística
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
Professora Orientadora:	Agatha Sacramento Rodrigues
Estudante:	José Carlos Soares Junior

Resumo

Este trabalho consiste no estudo de um desfecho escalar (variável resposta), cuja a informação é a idade gestacional do parto de gestações de gemelares, relacionando-o à covariáveis funcionais (medida do colo uterino e número de contrações), às quais as medidas foram obtidas ao longo do pré-natal. Na prática, quando se trata de covariáveis com essa característica temporal, profissionais da área utilizam uma medida resumo da covariável (média das avaliações ou a última avaliação antes do desfecho) para a realização de análises estatísticas, desconsiderando sua característica temporal e gerando perda de informações. O objetivo desta pesquisa é estudar modelos que corretamente consideram que a(s) covariável(is) é(são) curva(s). Foi utilizado o método de regressão funcional penalizada (PFR) para a estimação dos parâmetros e diferentes bases splines foram testadas. Estudos de simulação foram realizados para comparar a abordagem de ADF com as abordagens feitas na prática (uso de uma medida resumo), em que a abordagem de ADF se mostrou melhor que os demais em todos os cenários considerados. Os modelos PFR foram ajustados aos dados reais de gestações gemelares do Departamento de Obstetrícia da Universidade de São Paulo, obtendo resultados interessantes do ponto de vista prático.

Palavras-chave: Análise de dados funcionais. Covariáveis funcionais. Desfecho escalar. Gestação gemelar. Modelos paramétricos. Obstetrícia.

1 Introdução

É comum na área da Obstetrícia que o interesse consista em estudar um desfecho escalar (avaliado em um único momento), em geral alguma informação do parto, relacionando-o à variáveis medidas ao longo do pré-natal, estas chamadas de covariáveis funcionais. Nos dados que motivam este trabalho (dados do Departamento de Obstetrícia da USP), o desfecho (variável resposta) é a idade gestacional do parto de gestações gemelares, e as covariáveis funcionais avaliadas em diferentes momentos do pré-natal são a medida do colo do útero e o número de contrações.

O estudo da idade gestacional do parto em gestações gemelares é importante devido ao elevado risco de prematuridade em gestações múltiplas (com mais de um feto). 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 do que nas gestações de um único feto (Into Maternal, 2009).

Ao lidar com covariável funcional, pesquisadores na prática costumam fazer uso de uma única medida dessa covariável, geralmente a média ou a última avaliação antes do desfecho. Essa abordagem pode gerar perda de informação, pois a relação da variação dessa covariável com o desfecho e seu comportamento ao longo das avaliações não poderão ser analisados. Sendo assim, é de nosso interesse uma abordagem que considere essa característica temporal.

Quando o assunto é covariáveis funcionais, a análise estatística que tem se desenvolvido nas últimas décadas é comumente chamada de análise de dados funcionais (ADF). No âmbito deste estudo, as medidas da covariável funcional registradas ao longo do pré-natal são discretas e a ideia da ADF é de considerar que existe uma curva gerando esses valores, então, para n gestantes teríamos uma amostra de n curvas. Logo, a análise estatística é feita em uma amostra de curvas, onde podemos considerar a variação da covariável ao longo do pré-natal. Essa abordagem tem sido objeto de estudo nos últimos anos e Ramsay & Silverman (2005) é uma contribuição de destaque.

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.

Os objetivos desse estudo são apresentados na Seção 2. Na Seção 3 é apresentado o embasamento teórico e na Seção 4 é discutida a metodologia adotada. Os resultados do estudo de simulação e da aplicação aos dados gemelares são considerados na Seção 5 e, por fim, na Seção 6 são descritas as conclusões e considerações finais.

2 Objetivos

O objetivo geral deste projeto é estudar modelos que corretamente posicionam e consideram covariáveis funcionais com desfecho escalar, mantendo as características funcionais dessas covariáveis.

São os objetivos específicos:

1. Verificar se os modelos estudados são de fato melhores do que modelos que consideram covariáveis funcionais como escalares (média e a última avaliação) por meio de estudos de simulação.
2. Estudar a inclusão de covariáveis fixas em modelos que consideram covariáveis funcionais.
3. Desenvolver códigos computacionais bem documentados para permitir a reprodutibilidade e também a aplicabilidade dos métodos propostos e desenvolvidos.
4. Discutir a acessibilidade de ADF aos usuários.
5. Ajustar os modelos de ADF aos dados de gemelares que motivam esse estudo e discutir os resultados com o grupo de pesquisa de gemelares do Departamento de Obstetrícia da USP.

3 Embasamento Teórico

A abordagem com a análise de dados funcionais (ADF) tem sido objeto de estudo nos últimos anos e, estando entre os trabalhos pioneiros sobre ADF, temos Ramsay & Silverman (2005) como uma das principais fontes teóricas utilizadas neste projeto. Outro destaque é o trabalho de Kokoszka & Reimherr (2017), que foi importante tanto para a compreensão de vários conceitos como também para o desenvolvimento do estudo de simulação desta pesquisa.

Considere um número finito de medidas no qual a i-ésima observação é uma função real $X_i(t)$, com $i = 1,\ldots,n$, e $t \in \mathscr{T}$, em que $\mathscr{T}$ é um intervalo finito nos reais. Assim, dizemos que cada observação $X(t)$ que pertence ao espaço de funções reais $\mathscr{F}$ é um dado funcional (Ramsay & Silverman, 2005).

Para a análise de regressão da variável resposta Y_i, o modelo linear postulado é

$$Y_i = \alpha + \int_{\mathscr{T}} \beta(t) X_i(t) dt + \varepsilon_i, \tag{1}$$

em que o parâmetro $\beta(t)$ a ser estimado é uma função. No que segue, desenvolvemos a abordagem de regressão funcional penalizada (PFR - *Penalized Functional Regression*) para estimar $\beta(t)$, proposta por Goldsmith et al. (2011). O método consiste resumidamente de dois passos:

1. Decompor as funções (observações) da covariável funcional na forma

$$X_i(t) = \sum_{j=1}^{K_z} c_{ij} \psi_j(t),$$

   sendo $\psi_j(t)$ o termo de K_z autofunções obtidas da matriz de covariâncias estimada.
2. Utilizar K_b bases splines para expandir o termo $\beta(t)$.

Goldsmith et al. (2011) propuseram uma adaptação que após esses passos resulta no modelo em questão em um modelo de efeitos mistos e, sendo assim, estimam os parâmetros do modelo pelo método de máxima verossimilhança restrita.

É necessário escolher o número de autofunções em que $X_i(t)$ será decomposto, e o número de bases para a funcão coeficiente $\beta(t)$, ou seja, dependerá de K_z e K_b. Na prática, o mais comum é considerar $K = K_z = K_b$, sendo a dimensão da base usada para a expansão de $\beta(t)$ o maior interesse, como é mostrado em Kokoszka & Reimherr (2017).

Em um cenário com duas covariáveis funcionais e com covariáveis fixas, o modelo apresentado em (2) é dado por:

$$Y_i = \alpha + \int_{\mathscr{T}} \beta_1(t) X_{1i}(t) dt + \int_{\mathscr{T}} \beta_2(t) X_{2i}(t) dt + z_i^\top \gamma + \varepsilon_i, \tag{2}$$

com $z_i = (z_{1i},\ldots,z_{pi})^\top$ o vetor de p covariáveis fixas para a i-ésima observação e $\gamma = (\gamma_1,\ldots,\gamma_p)^\top$ o vetor dos p coeficientes associados às covariáveis fixas. O desenvolvimento é análogo ao que apresentado anteriormente e pode ser visto com maiores detalhes em Goldsmith et al. (2011).

4 Metodologia

O método PFR foi inicialmente utilizado por meio do pacote *fda* (Ramsay et al., 2020) no software R (R Core Team, 2020) - linguagem utilizada em todo desenvolvimento computacional deste trabalho. Os ajustes foram instáveis, como já anteriormente relatado na literatura. Por outro lado, ao utilizar o pacote *refund* (Goldsmith et al., 2020), também disponível no R, os resultados possuem mais coerência com o resultado obtido na análise exploratória dos dados funcionais.

Além do método PFR, inicialmente estava sendo considerado a abordagem por ondaletas de Malloy et al. (2010), mas após a revisão sistemática realizada, a abordagem foi descartada ao considerar as características dos dados da aplicação. Além disso, o método orientado por dados FPC (*functional principal component regression*) desenvolvido por Reiss & Ogden (2007), embora se assemelhe ao PFR por fazer uso de componentes principais, ele possui uma abordagem não tão direta onde depende da dimensão do problema. Como na aplicação desse estudo a dimensão é baixa (apenas 5 pontos que melhor será explicado nos resultados), o método FPC não foi adequado. Desta forma, apenas o método PFR é considerado.

Ajustamos modelos pelo método PFR ao considerar as combinações de $k \in \{3,4,\ldots,K\}$, em que K é o número de pontos observados, $fx \in \{TRUE, FALSE\}$ que indica se será uma regressão por splines com ou sem penalização, e spline $\in \{ps, tp, cr\}$ indica o tipo de base (*P-spline*, *thin-plate spline*, *cubic spline*) que será utilizada para expansão de $\beta(t)$.

Na aplicação (Seção 5.2), o modelo escolhido será aquele que apresentar o menor valor do critério de Akaike (AIC). Ainda, a análise exploratória dos dados funcionais será realizada de acordo com o apresentado por Frozza (2010).

No estudo de simulação, comparamos o modelo que corretamente posiciona a covariável funcional com o modelo linear escalar com a última avaliação (UA) e o modelo linear escalar com a média das avaliações (MA). Para isso, foram realizados dois estudos de simulação. O primeiro é baseado nos dados reais deste estudo (Subseção 5.2.1), e o outro baseado nos dados de simulação utilizados por Kokoszka & Reimherr (2017). Em ambos os casos, realizamos $B = 1000$ simulações para cada tamanho amostral $n \in \{50, 100, 300, 500, 1000\}$. Para cada amostra, separamos 70% dos dados como amostra treinamento e 30% como amostra teste. Os modelos são ajustados na amostra treinamento e a amostra teste é usada para calcular o erro quadrático médio (EQM), dado por: $EQM = \sum_{i=1}^{n_{tt}} (\hat{y}_i - y_i)^2 / n_{tt}$, em que $\hat{y}_i$ é o valor predito da i-ésima observação da amostra teste, com $i = 1,2,\ldots,n_{tt}$, e n_{tt} é o tamanho da amostra teste. Por fim, teremos $B = 1000$ valores de EQM de cada método e apresentamos os gráficos boxplots do método UA, MA e do método PFR (utilizando o modelo com menor EQM). Os dados são gerados da seguinte maneira:

1. Definir o tamanho do grid de pontos observados de t.
2. Gerar os coeficientes $\beta(t)$ por meio da função considerada ou definir os valores para esses coeficientes.
3. Gerar a covariável X para todo i, $i = 1,\ldots,n$.
4. Gerar $Y = X\beta + \varepsilon$, para todo i, $i = 1,\ldots,n$.

Todos os procedimentos utilizados na realização do estudo de simulação encontram-se em mais detalhes na Seção 4 do material de apoio em rpubs.com/Soju-JC/805876.

5 Resultados e Discussão

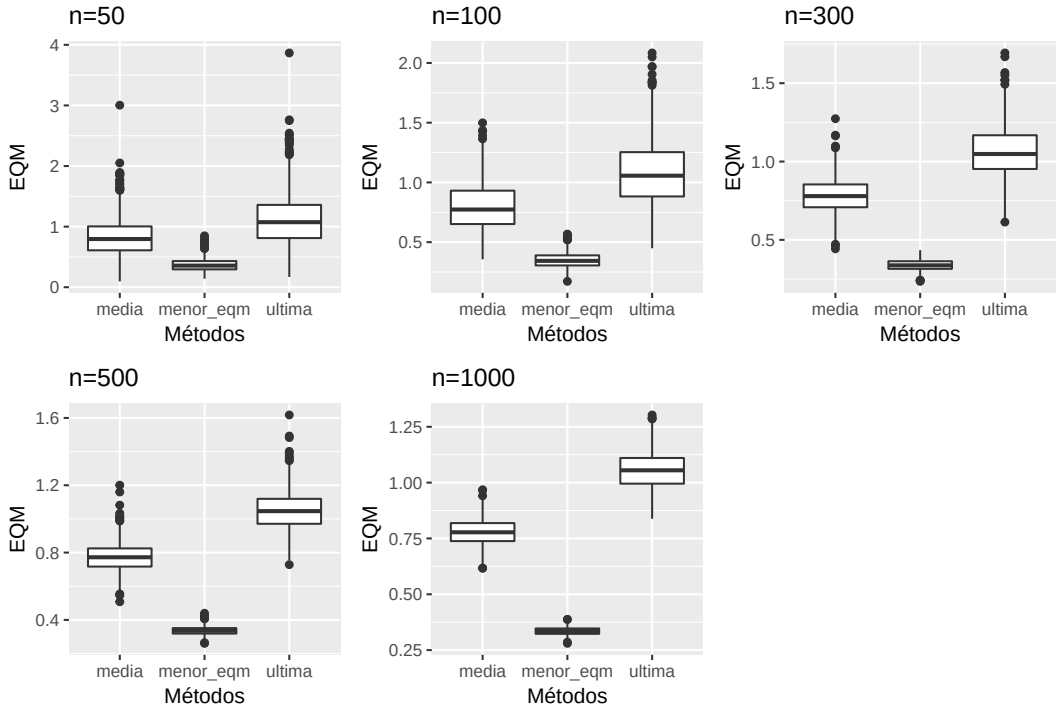
Esta seção está dividida em duas partes. A primeira (Subseção 5.1) é dedicada para os resultados dos estudos de simulação e na segunda parte (Subseção 5.2) serão apresentados os resultados da aplicação dos dados gemelares.

5.1 Estudo de simulação

Como dito anteriormente, foram realizados dois estudos de simulação. O primeiro baseado nos dados da aplicação e o outro inspirado nos dados de simulação utilizados por Kokoszka & Reimherr (2017).

No estudo de simulação baseado nos dados da aplicação, os valores de β considerados foram os valores que se encontram na Tabela 2. Foram ajustados 18 modelos PFR (combinações de $k = 3,4,5$, $fx \in \{TRUE, FALSE\}$ e spline $\in \{ps, tp, cr\}$) e, para todos os possíveis valores de $n \in \{50, 100, 300, 500, 1000\}$, o modelo que apresentou o menor EQM com maior frequência foi o que utiliza bases *cubic splines* de dimensão 3 sem penalização. Pela Figura 1 podemos observar que o modelo PFR apresenta os melhores resultados para todos os tamanhos amostrais, ou seja, apresenta os menores valores de EQM quando comparado com os métodos que utilizam média e última avaliação. Além disso, o melhor desempenho do método PFR fica mais evidente conforme o tamanho da amostra aumenta.

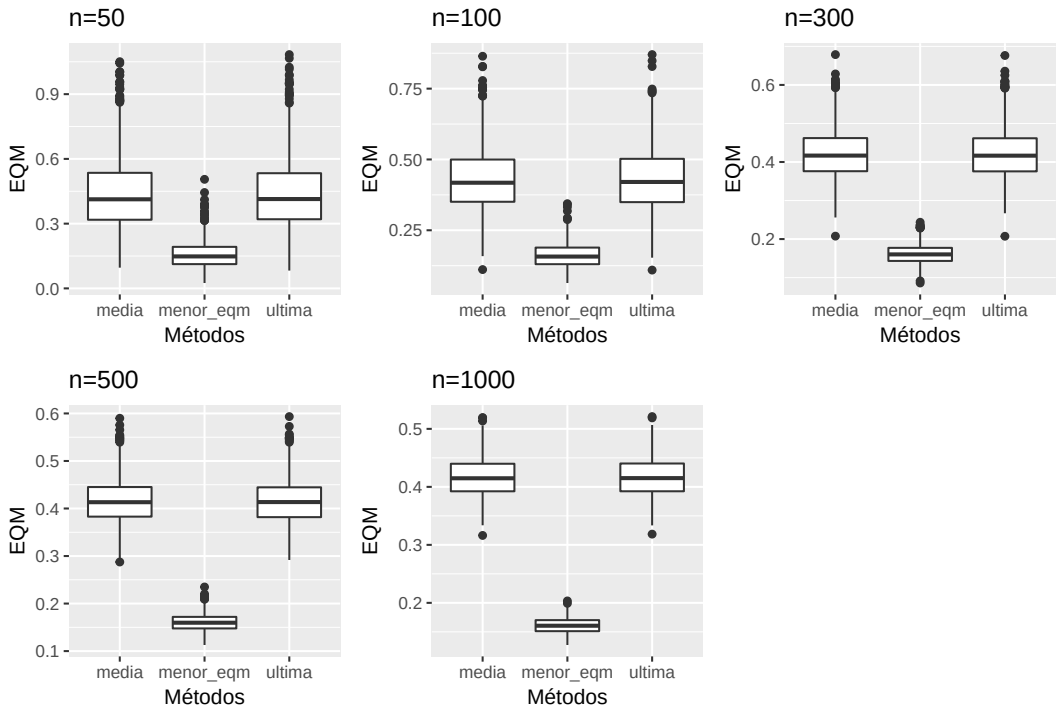
Figura 1: Gráficos boxplot do EQM dos três métodos considerados (media = MA, menor_eqm = PFR, ultima = UA), em que a simulação foi baseada nos dados da aplicação (Subseção 5.2.1) para diferentes tamanhos de amostra (n), sendo 1000 repetições para cada n. Observação: escalas não são as mesmas.



Fonte: Produção do próprio autor.

Para o estudo de simulação baseado nos dados utilizados por Kokoszka & Reimherr (2017), os valores de β considerados foram gerados pela função $\beta(t) = \sin(2\pi t)$ no passo 2 do processo de geração dos dados da simulação, descrito na Seção 4. O total de modelos PFR considerados foi de 594, isto devido ao número de pontos t, 101 neste caso. Para todos os possíveis valores de $n \in \{50, 100, 300, 500, 1000\}$, o modelo que apresentou o menor EQM com maior frequência foi o que utiliza *P-splines* de dimensão 4 sem penalização. Pela Figura 2 podemos observar que o modelo PFR apresenta os menores valores de EQM quando comparado com os métodos que utilizam média e última avaliação. Ainda, o melhor desempenho do método PFR fica mais evidente conforme o tamanho da amostra aumenta.

Figura 2: Gráficos boxplot do EQM dos três métodos considerados (media = MA, menor_eqm = PFR, ultima = UA), em que a simulação foi baseada nos dados utilizados por Kokoszka & Reimherr (2017) avaliando para diferentes tamanhos de amostra (n), sendo 1000 repetições para cada n. Observação: escalas não são as mesmas.



Fonte: Produção do próprio autor.

Vale ressaltar que o eixo do EQM tanto da Figura 1 quanto da Figura 2 não estão na mesma escala, isso porque o intuito é comparar os métodos para cada tamanho amostral e, além disso, padronizar as escalas faria com que alguns boxplots ficassem difíceis de interpretar.

5.2 Aplicação aos dados reais

Foram acompanhadas 336 gestações múltiplas que realizaram o pré-natal no Ambulatório de Gestações Múltiplas e concordaram em participar do estudo. As gestantes foram separadas em dois grupos, um ao qual durante a gestação receberam pílulas com progesterona ($n = 166$), e o outro ao qual receberam placebo ($n = 170$). O objetivo do estudo

é avaliar se fazer o uso da progesterona aumenta a idade gestacional do parto. Para isso, além da variável de grupo (chamada de óvulos), há interesse em avaliar o efeito na idade gestacional do parto de outras variáveis. Dentre elas, são as covariáveis fixas: corionicidade (se monocoriônica ou dicoriônica - 78% são dicoriônicas), idade materna (em anos - média: 28 anos e desvio padrão: 6,08), tempo de escolaridade (em anos - média: 10,46 anos e desvio padrão: 2,34), cor (branca ou não branca - 54% são brancas), indicador de doença prévia (sim ou não - 26% tem doença prévia), tabagismo (sim ou não - 9% são tagabistas), consumo de álcool (sim ou não - 1% consomem álcool), uso de drogas ilícitas (sim ou não - 1% usam drogas) e índice de massa corporal (IMC - kg/m^2 - média: 25,12 e desvio padrão: 4,94). As covariáveis funcionais são 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.

A média da idade gestacional do parto (desfecho de interesse) das gestantes é de 36,1 semanas de gestação (desvio padrão de 1,73), valor similar à mediana (36,30). Como os partos são considerados prematuros se ocorrem antes de 37 semanas de gestação, essas informações nos mostram que pelo menos 50% das gestantes da amostra tiveram parto prematuro. De fato, a porcentagem de gestantes que tiveram parto prematuro foi de 58,9%. A análise descritiva de todas as variáveis da aplicação pode ser vista na Seção 2 do material de apoio em rpubs.com/Soju-JC/805876.

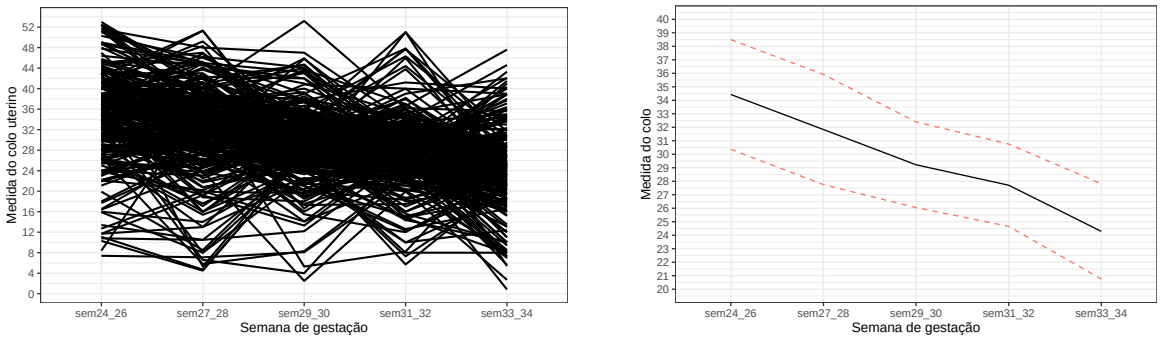
Inicialmente, das 336 gestantes da base de dados, era esperado que cada uma tivesse 5 medidas registradas de cada covariável funcional de acordo com o planejamento amostal (1ª avaliação da semana gestacional 24 a 26, 2ª avaliação da semana gestacional 27 a 28, 3ª avaliação da semana gestacional 29 a 30, 4ª avaliação da semana gestacional 31 a 32, 5ª avaliação da semana gestacional 33 a 34). Porém, ao verificar os dados, observamos que a maioria das gestantes não possuíam todas as 5 medidas, o que dificultava a continuidade do estudo nos obrigando a estudar alguns métodos de imputação de dados. Selecionamos as observações com pelo menos duas avaliações no pré-natal, resultando em $n = 263$ gestações.

Os métodos de imputação considerados para cada variável funcional foram: imputação por regressão utilizando modelo linear misto ao considerar algumas covariáveis fixas de importância e imputação por trajetória spline, que é utilizada em dados longitudinais (pacote *longitudinalData*). Dentre estes, a imputação por regressão considerando covariáveis fixas foi a mais coerente e foi a imputação aplicada. Todo o processo de aplicação e escolha do método de imputação foi devidamente descrito na Seção 3 do material de apoio (rpubs.com/Soju-JC/805876).

Na Figura 3(a) apresentamos o gráfico de perfis individuais da medida do colo ao longo do pré-natal, em que observamos um comportamento decrescente do comprimento do colo uterino de acordo com que a gravidez se aproximava do momento do parto. É de conhecimento dos profissionais da área que esse comportamento de fato ocorra, pois o declínio do comprimento cervical no decorrer da evolução da gestação é uma forma do corpo da gestante se preparar para o parto. Além disso, considerando que a medida cervical do colo do útero menor que 25mm é um critério de risco para prematuridade (critério já consagrado na literatura obstétrica), é possível observar pela Figura 3(a) que muitas gestantes chegaram a ter o comprimento cervical abaixo do valor de risco, precisamente 64,6% delas.

Na Figura 3(b) podemos observar o gráfico da média das curvas de todas as 263 gestantes ao considerar a medida do colo uterino registrada ao longo do pré-natal, o qual reforça o comportamento decrescente apresentado anteriormente. Também observamos que, em média, o comprimento cervical das gestantes se manteve acima de 25mm até um pouco antes da semana 33 de gestação. Após esse momento, em média, o comprimento cervical delas entrou em zona de risco.

Figura 3: Análise descritiva da medida do colo (em milímetros), onde o eixo *x* é o momento em que as avaliações do pré-natal ocorreram (ex: em sem27_28, a avaliação ocorreu em um momento durante as semanas 27 e 28 de gestação)



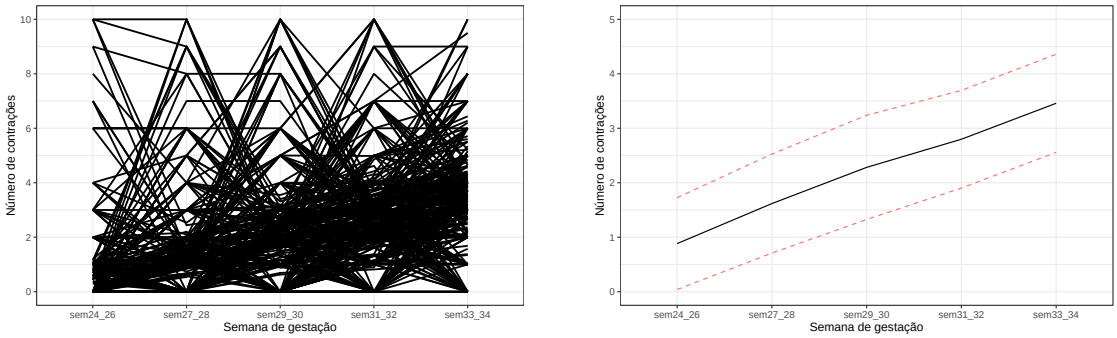
(a) Gráfico de perfis individuais

(b) Gráfico da média das curvas para medida do colo (curva preta) e desvio padrão de $\pm$ 0,5 (curva vermelha)

Fonte: Produção do próprio autor.

Na Figura 4(a) apresentamos o gráfico de perfis individuais do número de contrações, em que observamos um comportamento crescente de acordo com que a gravidez se aproxima do momento do parto. Este comportamento é o esperado pelo mesmo motivo do comprimento do colo do útero decrescer no decorrer da gestação. Observe que neste caso não é tão evidente esse comportamento esperado como ocorreu para a medida do colo, pois há uma boa quantidade de curvas de perfis com comportamento irregular. A princípio, essas irregularidades deveriam ser estudadas, pois podem ter ocorrido por vários motivos, mas isso demandaria um novo estudo já que temos que tomar um maior cuidado por estarmos lidando com uma variável funcional, o que não faz parte do objetivo desta pesquisa. Na Figura 4(b) podemos observar o gráfico da média das curvas de todas as 263 gestantes ao considerar o número de contrações registrado ao longo do pré-natal, o qual reforça o comportamento crescente esperado dessa variável.

Figura 4: Análise descritiva do número de contrações (em milímetros), onde o eixo x é o momento em que as avaliações do pré-natal ocorreram (ex: em sem27_28, a avaliação ocorreu em um momento durante as semanas 27 e 28 de gestação).



(a) Gráfico de perfis individuais.

(b) Gráfico da média das curvas para número de contrações (curva preta) e desvio padrão de $\pm$ 0,5 (curva vermelha).

Fonte: Produção do próprio autor.

Nas subseções que seguem, os modelos de ADF são ajustados aos dados da aplicação. Na Subseção 5.2.1 será considerada apenas a medida do colo como covariável, na Subseção 5.2.2 a única covariável é o número de contrações e na Subseção 5.2.3 são consideradas no ajuste todas as covariáveis funcionais e as covariáveis fixas.

5.2.1 Ajuste do modelo considerando apenas medida do colo

Considerando a covariável referente à medida do colo registrada durante o pré-natal das gestantes como a única covariável, a Tabela 1 apresenta 9 modelos com diferentes combinações de expansões de base para o processo de estimação da função coeficiente $\beta(t)$. Podemos observar que o melhor modelo funcional apresenta o critério de Akaike (AIC) de 717,10 e este modelo faz uso da base *cubic spline* de dimensão 3 na expansão da função coeficiente $\beta(t)$.

Tabela 1: AIC dos modelos funcionais candidatos que consideram apenas a medida do colo como covariável.

Dimensão da base	**ps**	**tp**	**cr**
k = 3	NA	717,14	717,10
k = 4	719,23	719,13	719,13
k = 5	720,97	720,97	720,97

[1] ps = P-spline [2] tp = thin-plate spline [3] cr = cubic spline
[4] NA = não funciona para o k especificado

Fonte: Produção do próprio autor.

Na Tabela 2 estão os valores dos coeficientes da curva estimada pelo modelo vencedor e na Figura 5(a) podemos visualizar o comportamento da função coeficiente do ajuste realizado.Os cinco momentos gestacionais estão padronizados no intervalo [0,1], em que $t = 0$ representa a avaliação nas semanas gestacionais 24 a 26, $t = 0{,}25$

representa a avaliação nas semanas gestacionais 27 a 28, $t = 0{,}50$ representa a avaliação nas semanas gestacionais 29 a 30, $t = 0{,}75$ representa a avaliação nas semanas gestacionais 31 a 32 e $t = 1$ representa a avaliação nas semanas gestacionais 33 a 34. Como esperado e também corroborando o resultado visto na análise descritiva da covariável funcional, os coeficientes diminuem para idades gestacionais do pré-natal mais avançadas.

Tabela 2: Coeficientes do ajuste do modelo que considera a covariável referente à medida do colo, modelo ao qual foi utilizado bases *cubic splines* de dimensão 3 para expansão de sua função coeficiente $\beta(t)$.

Momentos gestacionais padronizados no intervalo [0,1]	**Coeficientes**
0,00	0,084
0,25	0,256
0,50	0,271
0,75	0,027
1,00	-0,372

Fonte: Produção do próprio autor.

Ao efetuar um teste de hipóteses em que nosso interesse é saber se os coeficientes são estatisticamente nulos, obtivemos uma estatística F de 9,716 e um valor-p menor do que 0,001, levando à conclusão de que os coeficientes não são nulos, e portanto, o modelo está conseguindo identificar alguma variabilidade em relação à idade gestacional do parto das gestantes.

5.2.2 Ajuste do modelo considerando apenas número de contrações

Considerando o número de contrações registrado durante o pré-natal das gestantes como a única covariável no modelo, a Tabela 3 nos mostra que o melhor modelo funcional apresenta o critério de Akaike (AIC) de 737,06. Este modelo faz uso da base *thin-plate spline* de dimensão 3 na expansão da função coeficiente $\beta(t)$.

Tabela 3: AIC dos modelos funcionais candidatos que consideram apenas o número de contrações como covariável.

Dimensão da base	**ps**	**tp**	**cr**
k = 3	NA	737,06	737,07
k = 4	738,80	738,81	738,81
k = 5	740,66	740,66	740,66

[1] ps = P-spline [2] tp = thin-plate spline [3] cr = cubic spline
[4] NA = não funciona para o k especificado

Fonte: Produção do próprio autor.

Na Tabela 4 encontra-se os valores dos coeficientes da curva estimada pelo modelo vencedor e na Figura 5(b) podemos visualizar o comportamento da função coeficiente do ajuste realizado. Podemos notar um comportamento diferente do observado na análise descritiva e da opinião das especialistas, uma vez que os coeficientes são negativos e com comportamento decrescente a partir de $t = 0{,}5$. No entanto, ao efetuar o teste de hipóteses com hipótese nula de que coeficientes são todos iguais a zero, obtivemos uma estatística F de 2,547 e um valor-p de 0,057, o

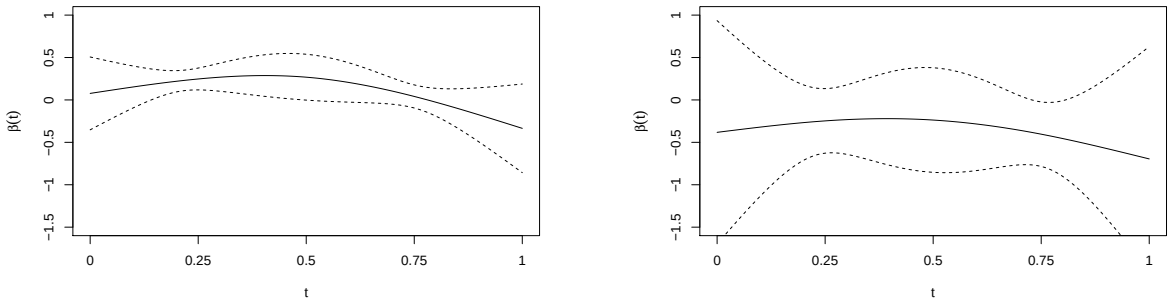
que leva à conclusão de que não podemos rejeitar a hipótese de nulidade dos coeficientes ao considerar um nível de significância de 5%, e a covariável referente ao número de contrações se mostrou não significativa para idade gestacional do parto.

Tabela 4: Coeficientes do ajuste do modelo que considera a covariável referente ao número de contrações registrado das gestantes, modelo ao qual foi utilizado bases *thin-plate splines* de dimensão 3 para expansão de sua função coeficiente $\beta(t)$.

Momentos gestacionais padronizados no intervalo [0,1]	**Coeficientes**
0,00	-0,351
0,25	-0,279
0,50	-0.279
0,75	-0.380
1,00	-0.553

Fonte: Produção do próprio autor.

Figura 5: Ajuste dos modelos individuais.



(a) Função $\beta(t)$ estimada do modelo vencedor referente à covariável medida do colo, ao qual foi utilizado bases *cubic splines* de dimensão 3 para sua expansão.

(b) Função $\beta(t)$ estimada do modelo vencedor referente à covariável número de contrações, ao qual foi utilizado bases *thin-plate splines* de dimensão 3 para sua expansão.

Fonte: Produção do próprio autor.

5.2.3 Ajuste do modelo considerando duas covariáveis funcionais e as covariáveis fixas

Uma das questões que surgiu durante o estudo, é se seria possível a construção de um modelo que considere tanto variáveis funcionais (ambas variáveis funcionais) quanto variáveis fixas. A Tabela 5 nos mostra as informações das covariáveis fixas no ajuste do modelo que considera ambas covariáveis funcionais e 10 covariáveis fixas diferentes. Foi considerada a base *cubic splines* de dimensão 3 para a medida do colo e a base *thin-plate splines* de dimensão 3 para o número de contrações, bases vencedoras para medida do colo e número de contrações (Seções 5.2.1 e 5.2.2, respectivamente). Ao considerar todas essas covariáveis fixas, vemos que apenas a covariável referente à corionicidade foi significante ao modelo.

Tabela 5: Coeficientes das variáveis fixas do modelo que considera ambas covariáveis funcionais (medida do colo e número de contrações) e as covariáveis ovulos (ovulosprogesterona), corionicidade (corionmonocoriônica), idade materna (im), tempo de escolaridade (tescola), cor (cor_brancosim), doenças prévias (ind_apsim), tabagismo (hv_tabagismosim), consumo de álcool (hv_alcoolsim), uso de drogas (hv_drogassim) e índice de massa corporal (imc).

	Coeficientes	**Erro padrão**	**Estatística t**	**Valor-p**
(Intercepto)	34,176	1,437	23,788	$< 0{,}001$
ovulosprogesterona	-0,048	0,253	-0,190	0,850
corionmonocoriônica	-0,932	0,315	-2,959	0,004
im	-0,008	0,025	-0,318	0,751
tescola	0,022	0,059	0,380	0,704
cor_brancosim	0,408	0,285	1,431	0,154
ind_apsim	-0,301	0,292	-1,029	0,305
hv_tabagismosim	-0,639	0,471	-1,357	0,176
hv_alcoolsim	-1,098	1,402	-0,783	0,435
hv_drogassim	1,798	1,452	1,238	0,217
imc	-0,015	0,032	-0,458	0,648

Fonte: Produção do próprio autor.

A Tabela 6 mostra os valores dos coeficientes para as covariáveis funcionais cujo modelo considera apenas corionicidade como covariável fixa. Para o teste de hipótese da nulidade dos coeficientes referentes à medida do colo, obtivemos uma estatística F de 8,181 e valor-p menor que 0,001, em que rejeitamos a hipótese de nulidade dos coeficientes. O mesmo teste aplicado aos coeficientes referentes ao número de contrações gerou uma estatística F de 0,574 e um valor p de 0,633, indicando a não rejeição da hipótese de nulidade.

Tabela 6: Coeficientes (para as variáveis funcionais) do ajuste do modelo que considera ambas covariáveis funcionais e a covariável referente à corionicidade.

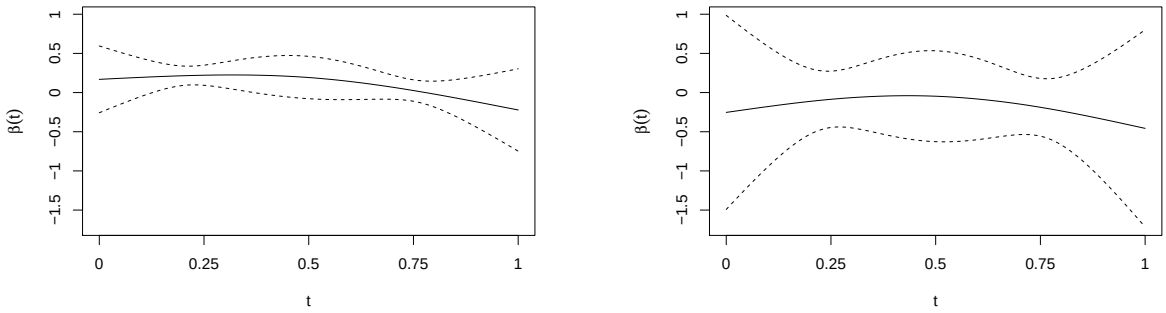
Momentos gestacionais padronizados no intervalo [0,1]	**Coeficientes (medida do colo)**	**Coeficientes (número de contrações)**
0,00	0,169	-0,253
0,25	0,221	-0,086
0,50	0,192	-0,047
0,75	0,026	-0,188
1,00	-0,222	-0,457

Fonte: Produção do próprio autor.

Podemos observar que na Figura 6(a) o comportamento de decréscimo da curva esperado para a medida do colo, resultado consiste com a análise exploratória na Figura 3 e com o ajuste na Subseção 5.2.1. Na Figura 6(b) vemos que a curva se comporta abaixo do valor zero, enfatizando os coeficientes negativos encontrados na Tabela 6, além de

a covariável número de contrações não ser significante, resultado coerente também com a Subseção 5.2.2. Além disso, para este ajuste obtivemos um coeficiente estimado de $-0{,}842$ (erro padrão: 0,302; T: $-2{,}785$; valor-p: 0,006) para a covariável corionicidade, indicando que gestações monocoriônicas apresentam menor idade gestacional do parto esperada.

Figura 6: Ajuste do modelo que considera ambas covariáveis funcionais e a covariável referente à corionicidade.



(a) Função $\beta(t)$ estimada referente à medida do colo, ao qual foi utilizado bases *cubic splines* de dimensão 3 para sua expansão.

(b) Função $\beta(t)$ referente ao número de contrações, ao qual foi utilizado bases *thin-plate splines* de dimensão 3 para sua expansão.

Fonte: Produção do próprio autor.

6 Conclusões

Nesse trabalho, estudamos modelos de regressão para um desfecho escalar na presença de covariáveis funcionais. Os modelos estudados foram aplicados aos dados reais de gestações gemelares do Departamento de Obstetrícia da USP, em que, além de covariáveis fixas, há duas covariáveis funcionais e o desfecho de interesse é a idade gestacional do parto. Na prática, não são utilizados métodos de análise de dados funcionais (ADF) e, no lugar da covariável funcional, a média ou a última avaliação é utilizada.

Nos estudos de simulações realizados, foi possível observar em termos de erro quadrático médio que, os modelos que corretamente posicionam a covariável funcional (construídos por regressão funcional penalizada - PFR) são melhores do que modelos que consideram um vetor escalar (média ou última avaliação) dessa covariável. Ao efetuar dois estudos de simulações em cenários diferentes, vemos que essa conclusão é consistente.

Na aplicação aos dados dos gemelares, inicialmente, foram ajustados os modelos PFR com covariável funcional individuais. No ajuste considerando a medida do colo como a única covariável, foram considerados 9 modelos e o melhor foi o que usa base do tipo *cubic splines* de dimensão 3 para a expansão da função coeficiente. Ao interpretar os resultados do modelo vencedor, podemos observar que a curva dos coeficientes decresce para elevadas idades gestacionais do pré-natal, como esperado, e rejeitamos a hipótese de nulidade dos coeficientes.

Já no ajuste considerando o número de contrações apenas, dos 9 modelos ajustados, o melhor foi o que usa bases do tipo *thin-plate splines* de dimensão 3 para a expansão da função coeficiente. Nesse ajuste, a hipótese de nulidade dos coeficientes não poder ser rejeitada, e essa covariável não se apresentou significativa para a idade gestacional

do parto.

Um desafio desse trabalho foi avaliar um modelo de regressão que considere mais de uma covariável funcional, além das covariáveis fixas. O ajuste do modelo ao utilizar o pacote *refund* ocorreu sem erros ou problemas. A presença de covariáveis fixas e de outra covariável funcional não mudou o comportamento das covariáveis funcionais anteriormente ajustadas de maneira individual. Além disso, dentre todas as covariáveis fixas utilizadas, apenas a variável relacionada à corionicidade se mostrou significativa, em que o grupo monocoriônico apresenta menor idade gestacional do parto esperada.

Dessa forma, não há indícios que o uso de progesterona aumenta a idade gestacional do parto esperada e, além da corionicidade, a covariável funcional medida do colo é importante para o desfecho de interesse.

Com base nas análises feitas e nos resultados obtidos neste trabalho, podemos observar que 58,8% das gestantes de gemelares acabam tendo parto prematuro (idade gestacional do parto menor que 37 semanas). Além disso, essa informação se mostrou condizente com o número de gestantes que tiveram a medida do colo na zona de risco durante a gestação, sendo 64,6% delas, o que mostra que foram porcentagens próximas.

Inicialmente foi planejado a construção de uma plataforma web por meio de um aplicativo Shiny (shinyapps.io), porém, após algumas conversas por email com Jeff Goldsmith, professor de Bioestatística da Columbia University School of Public Health, um dos criadores do pacote refund (Goldsmith et al., 2020), foi concluído que, devido a falta de opções gráficas para ADF no caso particular em que o desfecho é escalar, não há muito o que fazer com uma plataforma para este caso. Esse fato não ocorre para o caso da ADF em que o desfecho também é funcional e, para este caso, o próprio Jeff Goldsmith possui projetos que utilizam o *refund* com o aplicativo Shiny. Mais informações podem ser encontradas em sua página web no endereço https://jeffgoldsmith.com/software.html.

Embora o pacote refund (Goldsmith et al., 2020) seja extremamente útil, ele é novo no mundo da ADF e novas versões com novas implementações estão continuamente sendo feitas. Na data da finalização desta pesquisa, está disponível uma nova versão do *refund* que permite a ADF considerando dados faltantes, não necessitando, obrigatoriamente, de imputação de dados antes do ajuste dos modelos. O fato de ter que recorrer à imputação nesse trabalho levou à observações com medida do colo e/ou número de contrações imputados após o parto já ter ocorrido. Embora o número de observações com esse problema tenha sido de apenas 12 (4,56%) e o impacto não tenha sido grande, é algo que vale ressaltar.

Outra recomendação para análises futuras é realizar uma transformação na variável resposta que leve a uma distribuição mais simétrica. Para esta pesquisa foi feito uma transformação, porém, seria necessário mais tempo para estudar como seria feito a interpretação para o nosso problema, pois o desfecho transformado fica com uma direção diferente do desfecho original. Devido ao prazo de conclusão do estudo, foi decidido realizar esta pesquisa sem essa transformação, priorizando a interpretação dos modelos e considerando também que a média (36,10) e a mediana (36,30) do desfecho são bastante próximas. O processo de transformação do desfecho de interesse se encontra na Seção 5 do material de apoio em rpubs.com/Soju-JC/805876.

Esta pesquisa foi parcialmente apresentada em formato pôster no 2nd Workshop on Data Science and Statistical Learning (WDSSL) que aconteceu em 16 a 18 de junho de 2021, cujos os responsáveis pela coordenação e organização foram o Data Science Lab (DaSLab) e o Laboratório de Estatística e Computação Natural (LECON), ambos os laboratórios estão vinculados ao Departamento de Estatística da UFES.

Todos os códigos criados e utilizados durante esta pesquisa estão disponíveis no endereço https://github.com/Soju-

JC/Undergraduate-research, onde qualquer pessoa tem acesso e pode reproduzir os resultados aqui apresentados.

Agradecimentos

Agradecemos à FAPES (Fundação de Amparo à Pesquisa e Inovação do Espírito Santo) pelo financiamento desta pesquisa.

Referências Bibliográficas

Frozza, M. (2010). Introdução à análise descritiva de dados funcionais.

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

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., & Reiss, P. T. (2020). *refund: Regression with Functional Data*.

Into Maternal, C. E. (2009). *Perinatal Mortality 2007: United Kingdom*. CEMACH.

Kokoszka, P. & Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.

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.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. & Silverman, B. (2005). *Functional Data Analysis (2nd ed)*. Springer.

Ramsay, J. O., Graves, S., & Hooker, G. (2020). *fda: Functional Data Analysis*. R package version 5.1.9.

Reiss, P. T. & Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

Silva, J. (1995). Prematuridade: aspectos obstétricos. In Neme, B., editor, *Obstetrícia Básica*. Sarvier.