Victor Hugo Soares Ney

Análise de Curvas ROC na Presença de Medidas Repetidas Irregulares

Niterói - RJ, Brasil

19 de dezembro de 2023

Universidade Federal Fluminense

Victor Hugo Soares Ney

Análise de Curvas ROC na Presença de Medidas Repetidas Irregulares

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dr. Jony Arrais Pinto Junior

Niterói - RJ, Brasil

19 de dezembro de 2023

Universidade Federal Fluminense

Victor Hugo Soares Ney

Análise de Curvas ROC na Presença de Medidas Repetidas Irregulares

Monografia de Projeto Final de Graduação sob o título *"Análise de Curvas ROC na Presença de Medidas Repetidas Irregulares"*, defendida por Victor Hugo Soares Ney e aprovada em 19 de dezembro de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Jony Arrais Pinto Junior
Departamento de Estatística – UFF

Profa. Dra. Ana Beatriz Monteiro Fonseca
Universidade Federal Fluminense

Prof. Dra. Ludmilla da Silva Viana Jacobson
Universidade Federal Fluminense

Niterói, 19 de dezembro de 2023

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

N571a Ney, Victor Hugo Soares
Análise de Curvas ROC na Presença de Medidas Repetidas Irregulares / Victor Hugo Soares Ney. - 2023.
53 f.: il.

Orientador: Jony Arrais Pinto Junior.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Medidas repetidas. 2. Curvas ROC. 3. Modelos mistos. 4. Produção intelectual. I. Pinto Junior, Jony Arrais, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

Um dos principais pontos para o exercício da saúde pública é o diagnóstico de doenças de forma confiável, acessível e que possa ser disponibilizada à população. Nesse sentido, a análise de curvas ROC desempenha um papel crucial no desenvolvimento de testes de diagnóstico com alto desempenho. Um cenário muito comum na saúde é o acompanhamento de pacientes ao longo do tempo, em que diversas observações são coletadas sob os mesmos pacientes durante um certo período de tempo – caracterizando, assim, um estudo com a presença de medidas repetidas. Entretanto, é comum que alguns pacientes inicialmente envolvidos no estudo abandonem logo após a primeira coleta de dados e entre os que continuam, muitas vezes não conseguem comparecer em todas as datas pré-estabelecidas, constituindo, assim, uma base de dados irregular: indivíduos com diferentes quantidades de observações e diferentes tempos entre as observações. Em estudos de medidas repetidas, cada paciente observado constitui o que se chama de *cluster*. Devido às irregularidades anteriormente citadas, é comum a ocorrência de *clusters* com apenas uma observação, o qual é denominado *singleton* – e estes são identificados como a principal fonte de problemas nas análises. De forma a realizar a análise de curvas ROC no cenário descrito, supondo que se tenha o interesse de investigar diversos fatores mais facilmente coletados que possam estar associados com o diagnóstico – podendo constituir uma alternativa de diagnóstico ao método de referência, padrão-ouro –, é proposto na literatura um modelo misto de efeitos aleatórios, em que é incluído um intercepto para cada paciente na modelagem. Essa abordagem, no cenário descrito, pode ser um problema por diversos motivos. O principal deles é o fato de incluir um intercepto aleatório por paciente, o que causa *overfitting* do modelo quando há grande presença de *singletons*. O trabalho busca realizar um estudo de simulação em diversos cenários, avaliando como a presença de *singletons* afetam a análise de curvas ROC. Além disso, é proposto uma composição da verossimilhança de forma a minimizar o problema observado. Nos cenários simulados, realizar a análise de curvas ROC, com a metodologia proposta de modelos mistos com efeitos aleatórios, resultou em áreas abaixo da curva (AUC) viesadas e pontos de cortes sem interpretação. A modificação proposta trouxe uma melhor interpretação das curvas ROC e dos possíveis fatores associados com os diagnósticos das doenças.

Palavras-chave: Medidas repetidas. Curvas ROC. Modelos mistos.

Agradecimentos

Primeiramente, gostaria de agradecer imensamente aos meus pais, Cássia e Blaud, pela incansável dedicação, disponibilidade e amor de ambos desde sempre. Mesmo estando fisicamente distante de ambos, vocês são minha base todos os dias e nada disso teria sido possível sem vocês. Sou também extremamente grato pela minha avó, Cleonice, e a minha falecida tia Andrea por terem sempre me incitado a curiosidade, a vontade de aprender e o carinho pelo mundo acadêmico. Obrigado também ao meu padrasto Wilton pelo cuidado com minha família e também pelos meus queridos irmãos Davi, Daniel e Giovanna. Sinto saudades de bagunçar com vocês e obrigado por me lembrarem de ser criança e brincar, mesmo sendo adulto.

Como importantes partes da minha vida acadêmica, sou grato aos professores Ana Beatriz Monteiro Fonseca, Rafael dos Santos Erbisti e ao Jony Arrais Pinto Junior. Para mim, vocês foram muito mais que apenas professores: quando necessário, eram amigos e conselheiros e, sempre que deviam, ótimos professores. Estendo meus agradecimentos às professoras Thaís Cristina Oliveira de Fonseca, Mariane Branco Alves e Viviana das Graças Ribeiro Lobo, que me acompanham e orientam no Laboratório de Matemática Aplicada da UFRJ. Em especial, agradeço muitíssimo ao Jony, meu orientador, pelos mais de dois anos seguidos de orientações, pela paciência, amizade, confiança e pelos ensinamentos. Espero tê-los na minha vida enquanto possível for.

Aos meus amigos da vida, Vítor Sperandio, Pedro Lucas e Victoria Maura, peço desculpas pela minha ausência em todos importantes momentos das vidas de vocês. Agradeço por sempre me receberem de braços abertos quando era necessário, pelos infinitos momentos de risos que me proporcionaram – mesmo à distância. Aos principais amigos que fiz na UFF, Lucas, Daniel e Nathan, espero levá-los para toda a vida. A companhia de vocês é que tornou minha rotina agradável e todos dias sem pelo menos um de vocês foi um pouco mais difícil que o normal.

À minha melhor amiga e namorada, Pamela. Obrigado por toda sua compreensão nos momentos que não pude estar totalmente presente e por fazer parte da minha vida, e por me permitir fazer parte da sua. Você tem sido essencial e espero que juntos possamos superar todas as dificuldades que venham a aparecer no caminho.

Agradeço também a todos que de alguma forma colaboraram com minha formação, seja

no dia a dia na UFF ou à distância. Obrigado ao CNPq pelo financiamento de ambas minhas pesquisas e ao Laboratório de Estatística da UFF pelo financiamento em eventos acadêmicos e pela estrutura de apoio fornecida na universidade.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Entre os principais pontos para o exercício da saúde pública de forma eficaz está a capacidade de diagnosticar doenças de forma confiável e abrangente, principalmente as transmissivas e infecciosas. Por abrangente se entende a realização de testes de diagnóstico de formas simples e acessíveis – tanto à população como a clínicas de saúde –, capaz de providenciar resultados confiáveis em tempo suficiente para intervenção clínica efetiva no combate à doença. (BANOO et al., 2007)

Testes utilizados para diagnósticos de doenças apresentam algumas características intrínsecas aos mesmos como, por exemplo, a sensibilidade e a especificidade. Por si, a sensibilidade representa a proporção da população doente que é identificada como doente, enquanto que a especificidade representa a proporção populacional saudável que apresenta resultado negativo para o teste (CARDOSO et al., 2014).

Os **testes padrão-ouro** são aqueles utilizados como de referência para avaliar a presença ou ausência de uma doença, apresenta a máxima confiabilidade possível. Apesar disso, dificilmente são os que apresentam as características desejadas de um teste, principalmente em questões de acessibilidade e abrangência. Muitas vezes podem apresentar elevados custos, recursos de aplicação que não podem ser facilmente atendidos – como os equipamentos necessários para diagnóstico –, tempo de obtenção dos resultados, procedimentos invasivos – algumas vezes, antiéticos –, além de outros problemas relacionados com acessibilidade pela população (MORSE et al., 2009). Torna-se necessário, do ponto de vista da saúde pública, o estudo, avaliação e desenvolvimento de outros métodos de diagnóstico em alternativa aos testes padrão-ouro quando estes não apresentam características esperadas – principalmente quando as doenças diagnosticadas pelo mesmo possuem alta incidência na população, geralmente as que apresentam comportamento infeccioso ou transmissivo: estas necessitam de intervenção clínica o quanto antes para a contenção da transmissão e a redução de contaminação.

No cenário clínico, a metodologia análise de curvas ROC (*Receiver Operating Characteristic* – Característica de Operação do Receptor) é amplamente utilizada no desenvolvimento

e estudo de testes de diagnósticos alternativos ao padrão-ouro no qual o principal interesse é utilizar uma **variável quantitativa** para, a partir de um valor desta, **diagnosticar** uma condição. Essa metodologia é amplamente utilizada na literatura como um dos principais métodos para avaliação de performance de testes no caso de um desfecho binário – por exemplo, doente ou não-doente, aderente ou não-aderente – avaliando uma métrica quantitativa (SWETS, 1988; PEPE, 2003; ZHOU; MCCLISH; OBUCHOWSKI, 2009). Inicialmente proposta durante a Segunda Guerra Mundial para analisar a precisão na classificação dos radares ao diferenciar sinais – ou seja, detecções verdadeiras de alvos – de ruídos (LUSTED, 1971; EGAN, 1975) –, rapidamente encontrou aplicações em diversas áreas: em *machine learning*, foi inicialmente utilizada por Spackman (1989) para avaliação e comparação de algoritmos; na área de saúde, sua aplicação é majoritariamente percebida pela comunidade médica nos cenários de tomadas de decisões, a qual inclui também a avaliação e criação de testes de diagnósticos (ZOU, 2002).

Nesta metodologia é suposto a existência de um teste padrão-ouro que permita classificar a população em doentes e não-doentes e também a existência de uma variável quantitativa – denominada usualmente como **variável classificadora** – a qual se acredita estar associada com o desfecho de interesse, ou seja, acredita-se que esta pode ser utilizada como alternativa ao teste padrão-ouro. Em geral, a análise de curvas ROC consiste em avaliar quão boa é a performance dessa variável classificadora – em termos de sensibilidade e especificidade – em alternativa ao teste padrão-ouro por meio da definição de pontos de corte da variável classificadora. Esses pontos de cortes são valores a partir dos quais se classifica como "doente" ou "não-doente" qualquer paciente, ficando a critério do pesquisador definir a categoria de interesse do desfecho. Por exemplo, a subnutrição é uma condição que, para ser diagnosticada com extrema precisão, necessita de equipamentos complexos e avaliações demoradas, como o DXA – dupla emissão de raios X –, o que é inviável de se disponibilizar em clínicas de saúde para acesso à população. Como alternativa ao DXA, costuma-se utilizar o ângulo de fase da corrente elétrica obtida por um equipamento de bioimpedância (KYLE; GENTON; PICHARD, 2013), onde baixos valores dessa estão associados com a subnutrição. Para construir um teste de diagnóstico nesse cenário seriam avaliados diversos pontos de corte no ângulo de fase da corrente elétrica e seria determinado como melhor ponto de corte aquele que possuísse maiores valores de sensibilidade e especificidade.

A análise de curvas ROC usual assume independência entre as observações da variável classificadora, mas alguns tipos de estudos possuem dependência entre observações como, por exemplo, dados coletados em diferentes momentos do tempo sobre uma mesma unidade observacional. Esse tipo de estudo é conhecido como estudos de medidas repetidas, ou estudo longitudinal, e é comum que uma estrutura de dependência esteja presente nesses tipos de dados: os

pacientes são independentes entre si, mas as repetidas observações feitas no mesmo paciente são dependentes entre si. A abordagem usual de curvas ROC poderia ser aplicada nesse contexto se fosse adotada a estratégia de manter a primeira observação de cada paciente, onde as observações são independentes entre si, visto que cada observação é provinda de um paciente e não há mais observações repetidas sobre um mesmo paciente.

Tratando-se de dados longitudinais, a abordagem usual da análise de curvas ROC não permite avaliar a eficácia de um teste de diagnóstico para um determinado desfecho devido à falta de independência entre observações. A literatura apresenta algumas alternativas para avaliar um teste de diagnóstico na presença de múltiplas observações de um mesmo indivíduo. O trabalho de Obuchowski (1997) expande a metodologia de análise de curvas ROC para cenários longitudinais, tornando assim não mais necessário o requisito de independência entre observações. Ainda assim, se é de interesse avaliar como diversos fatores podem afetar a classificação de um paciente em doente ou não-doente, é possível utilizar a modelagem estatística para avaliação conjunta desses fatores, construindo uma variável classificadora que será utilizada de forma usual na análise de curvas ROC.

Como bem avaliado em Evans e Johnson (2001), diferentes indivíduos apresentam taxas de absorção de fármacos impactadas por características individuais e fisiológicas, além de comportamentos individuais que possam estar associados com a ocorrência de um determinado desfecho. Logo, a abordagem de considerar a primeira observação de cada paciente não parece ser eficiente, visto que é descartada a estrutura de dependência presente. Além disso, se o interesse for avaliar como diferentes fatores podem influenciar a presença de um determinado desfecho, a abordagem sugerida em Obuchowski (1997) torna-se inviável, visto que essa estende-se apenas ao uso da variável classificadora. Sendo assim, uma das abordagens mais utilizadas na literatura, que busca acomodar a estrutura de dependência dos dados – também, a inclusão de diversos fatores –, consiste em utilizar um modelo estatístico que seja adequado e, por meio do ajuste deste modelo, usar as estimativas dos parâmetros para construir a variável classificadora, por exemplo, estimando o valor esperado de cada observação. Com isso, pode-se encontrar um ponto de corte ótimo nos valores estimados, utilizando um modelo que acomode a estrutura de dependência e utilizando todas as observações.

No cenário de modelagem estatística para dados longitudinais em análises de curvas ROC, os trabalhos de Liu e Wu (2003) e Liu et al. (2005) propuseram soluções para avaliar como diversas variáveis podem ser usadas para diagnosticar, na presença de medidas repetidas, e propuseram também o ajuste de um Modelo Linear Generalizado Misto – *Generalized Linear Mixed Model* - GLMM – para lidar com a estrutura de dependência. Em específico, é ajustado um mo-

delo de regressão logística misto – contando tanto com efeitos fixos como efeitos aleatórios, que variam segundo cada unidade observacional – em que a variável resposta, aquela que se tenta explicar, é o desfecho do indivíduo (doente ou não-doente) e como variáveis explicativas são utilizadas todas as covariáveis de interesse do pesquisador na análise. Ajustando um modelo de regressão logística, é possível calcular a probabilidade estimada de um paciente apresentar o desfecho de interesse e, por meio desses valores, assim avaliar pontos de corte na probabilidade ajustada para a realização da análise de curvas ROC.

A abordagem mais simples para modelos GLMM é considerar um intercepto aleatório por *cluster* – paciente –, representando assim o deslocamento que o indivíduo possui quanto ao desfecho de interesse – por exemplo, se considerar o desfecho de interesse como aderência – uma variável binária –, indivíduos que seguem o tratamento de forma mais apropriada que a média terão um intercepto positivo (efeito individual positivo), ou seja, esse indivíduo em específico tem uma característica pessoal de ser aderente ao tratamento. Ao contrário, se um indivíduo não segue o tratamento de forma apropriada, encontra-se um efeito individual negativo para esse paciente, isto é, característica pessoais do mesmo reduzem a classificação deste como aderente ao tratamento.

Por conta de sua importância na aplicação clínica, considerável desenvolvimento tem sido feito na literatura para lidar com cenários de medidas repetidas e análise de curvas ROC: Michael, Tian e Ghebremichael (2019) propõe extensões da curva ROC para lidar com estudos longitudinais em que os biomarcadores são regularmente observados – isto é, os indivíduos comparecem para avaliação clínica em tempos igualmente espaçados; Foulkes et al. (2010) e Liu e Albert (2014) desenvolvem regras de predição em desfechos binários para biomarcadores longitudinalmente avaliados. Apesar do considerável desenvolvimento recente nesta área, muitos consideram cenários restritos em suas aplicações como, por exemplo, o caso em que todos os pacientes possuem mais de uma observação e todas essas observações são feitas em tempos igualmente espaçados – cenário de medidas repetidas regularmente observadas. No caso com menor restrição, é apenas necessário que todos pacientes possuam medidas repetidas, isto é, mais de uma observação. No entanto, o que muitas vezes acontece, na realidade, é que os pacientes não conseguem comparecer no tempo pré-determinado: comparecem antes por algum motivo de saúde emergencial ou abandonam o estudo. Encontra-se então cenários onde há diversos pacientes com apenas uma observação e, para aqueles que possuem mais de uma observação, os intervalos entre observações são irregulares.

Tratando-se de medidas repetidas, cada grupo é denominado por *cluster* e os grupos formados por uma única observação – ou seja, cada paciente que só compareceu uma única vez no

estudo – são frequentemente denominados na literatura como *singletons*. O problema em considerar a abordagem proposta em Liu e Wu (2003) e Liu et al. (2005) de ajustar um GLMM nestes cenários não está nos vieses nas estimativas pontuais dos efeitos das covariáveis, como suspeitado: os estudos de Maas e Hox (2005), Clarke e Wheaton (2007), Bell, Ferron e Kromrey (2008), Clarke (2008), Austin e Leckie (2018) sugerem que não há um tamanho mínimo de observações por *cluster*. Além disso, concluem que aumentar o número de observações intra-*cluster* reduz apenas o viés do intercepto do *cluster* e sua incerteza associada, mas que o número de *clusters* considerados no estudo é mais importante do que o tamanho dos mesmos. Dos estudos anteriores, especialmente o estudo de Bell, Ferron e Kromrey (2008) avalia como diferentes proporções de *singletons* afetam, em um estudo de simulação, o viés dos efeitos estimados das covariáveis sob diferentes números de grupos, encontrando que o viés existente em relação às estimativas dos efeitos fixos é praticamente ignorável. A abordagem proposta por Liu e Wu (2003) e Liu et al. (2005) possui principal desvantagem na análise de curvas ROC quando há grande presença de *singletons*, o que resultaS em resultados ilusórios da análise de curva ROC. Como será visto nas simulações realizadas, considerar interceptos aleatórios para os *singletons* ocasiona o que se chama de sobreajuste/ ou *overfitting* do modelo estatístico que, emborap areça descrever perfeitamente os dados, não possui capacidade preditiva.

O atual trabalho tem como objetivo o desenvolvimento de métodos estatísticos para lidar com cenários desbalanceados – especialmente com grande presença de *singletons*, intervalos irregulares entre observações repetidas e *clusters* com pouco número de observações. Inicialmente, são realizados estudos de simulação para avaliar como as metodologias desenvolvidas se comportam em diferentes cenários de análise de curvas ROC, além de explorar algumas abordagens que podem ser utilizadas de forma imediata para resolver o problema da dependência dos dados – como considerar a primeira observação de cada indivíduo. Além disso, é proposta uma modificação na função de verossimilhança do modelo linear generalizado misto que se apresentou capaz de minimizar o impacto da presença de *singletons* na análise de curvas ROC, recuperando a capacidade preditiva dos modelos ajustados e permitindo uma melhor interpretação dos fatores associados, auxiliando assim no desenvolvimento de testes de diagnósticos por meio desta metodologia.

A organização do trabalho se dá da seguinte forma: Capítulo 2, Metodologia, em que serão discutidas as metodologias estatísticas utilizadas ao longo do desenvolvimento do trabalho; Capítulo 3, Resultados e Discussões, em que são apresentados os produtos das análises estatísticas e feita e investigação dos resultados obtidos; Capítulo 4, Conclusão, no qual são apontados os principais resultados do trabalho e apontados futuros direcionamentos para a pesquisa.

2 Metodologia

Neste capítulo são apresentadas as metodologias estatísticas voltadas a testes de diagnósticos, modelos de regressão logística e estruturas de dados longitudinais. Além disso, será introduzido o modelo logístico Bayesiano de verossimilhança composta proposto para medidas repetidas irregulares.

2.1 Performance de Testes e Estudos de Classificação

No âmbito de estudos de classificação, é de grande ajuda construir uma tabela com duas linhas e duas colunas conhecida como **matriz de confusão** que auxilia na visualização da performance do teste e das estatísticas que compõem esse conceito. Nas células desta tabela são relacionadas as concordâncias e discordâncias dos resultados entre dois testes de diagnósticos – um padrão-ouro e um teste alternativo – em que se tem o interesse de se avaliar a performance de classificação do teste alternativo em relação ao de referência. A matriz é construída é da seguinte forma:

Tabela 1: Matriz de confusão relacionando a classificação do teste de diagnóstico alternativo e o do teste padrão-ouro.

		Real status do paciente	
		Doente	Não-doente
Diagnóstico	Doente	VP	FP
	Não-doente	FN	VN

A quantidade VP é referente aos Verdadeiros Positivos (indivíduos com a doença e diagnosticados como positivos pelo teste de diagnóstico), FP aos Falsos Positivos (indivíduos sem a doença e diagnosticados como positivos), FN aos Falsos Negativos (indivíduos com a doença e diagnosticados como negativos) e VN aos Verdadeiros Negativos (indivíduos sem a doença e diagnosticados como negativo). Com essas quantidades, é possível calcular algumas estatísticas

de particular interesse ao avaliar performance de testes:

Sensibilidade: a fração de verdadeiros positivos diagnosticados pelo teste,

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} .$$

Especificidade: a fração de verdadeiros negativos diagnosticados pelo teste,

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} .$$

1 - Especificidade: a fração de falsos positivos diagnosticados pelo teste,

$$1 - \text{Especificidade} = \frac{\text{FP}}{\text{VN} + \text{FP}} .$$

Essas estatísticas são, por fim, utilizadas para resumir a performance de discriminação de um teste de diagnóstico e também utilizadas na aplicação de análises de curvas ROC, especificamente ao avaliar os classificadores construídos.

2.2 Análise de Curvas ROC

A análise de curvas ROC é uma técnica estatística para visualizar, organizar e selecionar classificadores – como testes de diagnósticos – baseado em suas performances. Inicialmente, é necessário que cada paciente pertença à classe de positivos ou negativos quando avaliado em relação ao desfecho de interesse. Este último, pode ser qualquer desfecho binário: aderência ou não-aderência ao tratamento, presença ou ausência de uma doença/condição, entre outros.

A principal motivação desta metodologia é avaliar como uma variável quantitativa – denominada de **variável classificadora** – pode ser utilizada para, a partir da avaliação de **pontos de corte** no seu domínio, diagnosticar quanto ao desfecho de interesse. Uma das opções para satisfazer a condição de pré-existência de dois grupos pode ser a realização de um teste padrão-ouro antes do estudo, diferenciando assim os pacientes entre doentes e não-doentes. Para avaliar a performance de um determinado ponto de corte, é necessário que se defina a categoria de interesse do desfecho – a ausência ou a presença deste. Então, para cada ponto de corte definido no domínio da variável classificadora, todas observações que possuam **valores acima deste** são classificadas como **positivas** – em relação à categoria de interesse – e todas **abaixo deste** são classificadas como **negativas**. São calculadas, então, a sensibilidade e a especificidade para este ponto de corte em particular.

Por fim, a curva ROC é construída utilizando um gráfico bidimensional em que se tem no eixo das ordenadas a **sensibilidade** – fração de verdadeiros positivos para cada ponto de corte – e, no eixo das abscissas, a quantidade **1 - especificidade** – fração de falsos positivos para cada ponto – apesar de serem possíveis outras configurações, como a especificidade no eixo das ordenadas e a sensibilidade no eixo das abscissas. Tomando como exemplo a primeira orientação especificada, busca-se selecionar como melhor ponto de corte aquele mais próximo do ponto $(0,\ 1)$, isto é, o ponto que apresenta 100% de especificidade e 100% de sensibilidade. Um exemplo de gráfico de curva ROC pode ser visualizado na figura abaixo:

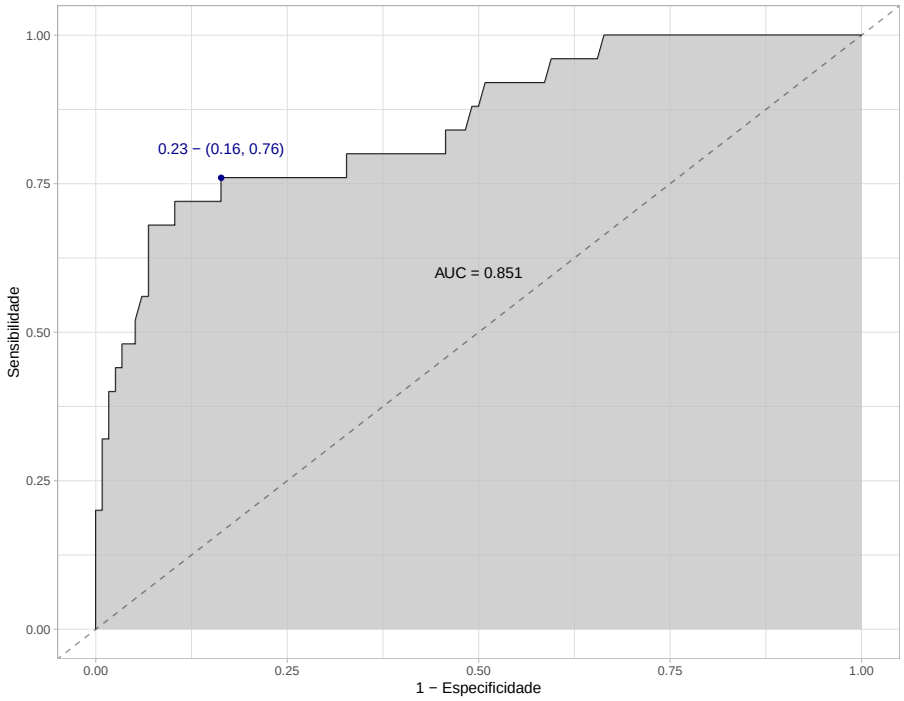

Figura 1: Exemplo de uma curva ROC.

O ponto marcado em preto no Gráfico 1 é referente ao ponto de corte que apresenta maiores valores de sensibilidade e especificidade – melhor ponto de corte entre os avaliados – e a quantidade no centro do gráfico se refere à estatística AUC – *Area Under the Curve*, área abaixo da curva. Acima do melhor ponto de corte estão representados, respectivamente, o valor do melhor ponto de corte, a quantidade $1-$ especificidade e a sensibilidade que este ponto de corte fornece ao ser utilizado como teste de diagnóstico. A linha diagonal conectando os pontos $(0,\ 0)$ e $(1,\ 1)$ representa uma reta tal que sua área abaixo da curva seja $0{,}5$ e esta é útil para verificar a capacidade discriminatória do teste em questão. Caso a AUC esteja muito próxima ou abaixo de $0{,}5$, a performance do teste é tão efetiva quanto classificar o indivíduo aleatoriamente entre doente e não-doente, ou seja, o teste não apresenta boa – ou nenhuma – capacidade discriminatória.

Um dos índices mais recorrentes na literatura é o **Índice de Youden**, que busca maximizar a

sensibilidade e especificidade conjuntamente. Para um ponto de corte c do domínio, esse índice é obtido como:

$$\text{Youden}(c) = \text{Sensibilidade}(c) + \text{Especificidade})(c) - 1 \ ,$$

em que Sensibilidade(c) e $(1 - \text{Especificidade})(c)$ são quantidades referentes ao ponto de corte c. O melhor ponto de corte é aquele que apresenta maior Índice de Youden.

A AUC é uma das principais estatísticas resultantes da análise de curvas ROC e assume valores entre 0 e 1. É possível utilizá-la para avaliar a capacidade discriminatória de um teste e, apesar de seus valores não possuírem uma interpretação direta, existem na literatura sugestões para a interpretação desta. Jr, Lemeshow e Sturdivant (2013) sugere a seguinte classificação:

Tabela 2: Valores de AUC para classificação discriminatória de um teste.

Intervalo AUC	**Capacidade Discriminatória**
AUC $\leq 0{,}5$	Nenhuma
$0{,}5 <$ AUC $\leq 0{,}6$	Fraca
$0{,}6 <$ AUC $\leq 0{,}7$	Aceitável
$0{,}7 <$ AUC $\leq 0{,}8$	Boa
AUC $> 0{,}8$	Excelente

Como a AUC representa uma área, esta é calculada, idealmente, como a integral da curva. Em geral, essa curva não possui uma fórmula de construção e, portanto, não é possível calcular, de formula analítica, a área abaixo desta. São utilizados métodos de aproximação numérica ou estimadores estatísticos para a obtenção da AUC. De forma a apresentar um dos estimadores mais utilizados na literatura para o cálculo da AUC, é necessário que aqui seja introduzida uma necessária notação estatística. Essa notação será introduzida para o caso de apenas uma observação por paciente, mas nas próximas seções será estendida e adaptada para o caso em que pacientes podem possuir mais de uma observação.

Sejam$X$ uma variável quantitativa, que será utilizada como variável classificadora na análise de curvas ROC, e N_c o número total de pacientes. Cria-se o vetor $\underset{\sim}{\boldsymbol{X}}^+$ para representar os valores dessa variável para os pacientes que possuem o desfecho – por exemplo, para os pacientes que estão doentes – e $\underset{\sim}{\boldsymbol{X}}^-$ para representar os valores da variável classificadora para os pacientes que não possuem o desfecho – por exemplo, que estão saudáveis. Para cada $i = 1, \ \ldots, \ N_c$, se o i-ésimo paciente possui o desfecho, então o valor da variável classificadora X desse paciente é adicionado como entrada no vetor $\underset{\sim}{\boldsymbol{X}}^+$ e, caso contrário, é adicionado como entrada no vetor

$\underset{\sim}{\boldsymbol{X}}^-$. Então, o número total de pacientes que possuem o desfecho é denotado como $M = \#\underset{\sim}{\boldsymbol{X}}^+$ enquanto que daqueles que não possuem o desfecho é $N = \#\underset{\sim}{\boldsymbol{X}}^-$, em que $\#\underset{\sim}{\boldsymbol{V}}$ é a dimensão de um vetor $\underset{\sim}{\boldsymbol{V}}$.

Bamber (1975) fornece um estimador não-paramétrico para a área abaixo da curva para dados independentes, dado por:

$$\widehat{AUC} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{i'=1}^{N} \psi \left(X_i^+,\ X_{i'}^- \right) \ , \tag{2.1}$$

em que a função ψ é dada por:

$$\psi(X, Y) = \begin{cases} 1 \text{ se } Y < X, \\ 0{,}5 \text{ se } Y = X, \\ 0 \text{ se } Y > X. \end{cases} \tag{2.2}$$

Na seção a seguir, serão feitas mudanças de notação e revisados alguns conceitos sobre a perspectiva de medidas repetidas para a análise de curvas ROC.

2.3 Análise de Curvas ROC – Medidas Repetidas

A notação anteriormente utilizada, assim como o estimador da AUC, devem ser adaptados para o caso de medidas repetidas. A principal motivação para tal é o cálculo da variância da AUC que quando está na presença de medidas repetidas, deve incluir a correlação entre observações de um mesmo *cluster*. A variância calculada pode ser utilizada em testes de hipóteses, em especial, avaliando se a medida de diagnóstico utilizada possui alguma capacidade discriminatória, isto é, se a área apresenta valor significativamente acima de $0{,}5$. A motivação para tal teste vem do fato de que áreas abaixo de $0{,}5$ indicam testes com capacidade discriminatória pior que a classificação aleatória da ausência/presença do desfecho para cada observação.

Obuchowski (1997) propõe um estimador não-paramétrico para a variância da AUC quando há a presença de medidas repetidas, permitindo assim o cálculo do intervalo de confiança. Para apresentar tais estimadores, é necessário que as notação utilizadas até então sejam readequadas para o cenário de medidas repetidas.

Como antes, X segue sendo a variável classificadora – aquela utilizada para classificar indivíduos entre possuem e não possuem o desfecho – e N_c o número total de pacientes. Cada paciente $i = 1, \ \ldots, \ N_c$ possui m_i observações com o desfecho e n_i observações sem o

desfecho. O valor X_{ij}^{+} refere-se ao valor da variável classificadora para a j-ésima observação ($j = 1,\ \ldots,\ m_i$) do paciente i ($i = 1,\ \ldots,\ N_c$) quando esta possui o desfecho de interesse e, em contra-partida, X_{ij}^{-} refere-se ao valor da variável classificadora para a j-ésima observação do paciente i quando este não possui o desfecho. O número total de observações que possuem o desfecho pode ser calculado como $M = \sum_i m_i$ e o de observações que não possuem o desfecho é calculado como $N = \sum_i n_i$. Obuchowski (1997) fornece uma releitura do estimador da AUC, de Bamber (1975), para quando há presença de medidas repetidas, definindo-o por

$$\widehat{\text{AUC}}_c = \frac{1}{MN} \sum_{i=1}^{N_c} \sum_{i'=1}^{N_c} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{i'}} \psi(X_{ij}^{+},\ X_{i'k}^{-}) \ , \tag{2.3}$$

em que ψ é a mesma utilizada na Equação 2.2.

Destaca-se que anteriormente foi utilizada a palavra "releitura" pois, em termos práticos, a AUC que estaria sendo estimada é a mesma para ambos estimadores aqui citados. A verdadeira contribuição de Obuchowski (1997) é o estimador não-paramétrico da variância da AUC que não é apresentado no presente trabalho, mas encontra-se detalhadamente descrito no próprio artigo.

Toda metodologia citada e mostrada até então é limitada pelo fato de classificar observações entre *possuem o desfecho* e *não possuem o desfecho* com base nas medições diretas de apenas uma variável classificadora. Caso exista apenas uma única variável classificadora e a presença de medidas repetidas, o trabalho de Obuchowski (1997) é um dos possíveis caminhos a serem seguidos. Entretanto, a pergunta que fica – também, principal motivação para este trabalho – é a seguinte: e se o interesse estiver em avaliar a capacidade discriminatória de diversos fatores conjuntamente? Isto é, se mais de uma variável tiver sido observada para cada paciente, qual metodologia seria apropriada para realizar uma análise de curvas ROC utilizando a maior quantidade de informação disponível?

Uma das abordagens mais utilizadas nesse caso é a "redução de dimensionalidade" por meio de um modelo estatístico, isto é, a construção de uma única variável classificadora utilizando toda informação disponível. Uma das abordagens mais imediatas é utilizar o modelo logístico em que as diversas variáveis são utilizadas como variáveis explicativas e a variável resposta – aquela que se deseja estimar – é a ausência/presença do desfecho. Assim, estima-se a probabilidade de uma observação possuir o desfecho a partir de um conjunto de covariáveis e, por fim, a probabilidade estimada é utilizada como variável classificadora na análise de curvas ROC.

De forma a lidar com a presença de medidas repetidas, Liu e Wu (2003) utilizam a abordagem descrita acima junto de um modelo logístico de efeitos mistos que será visto, detalhada-

mente, nas próximas seções. Uma das principais desvantagens dessa abordagem são as condições às quais está sujeita um modelo estatístico, como o sobreajuste e a sua especificação incorreta.

2.4 Modelos Lineares Generalizados

A classe de modelos denominada de Modelos Lineares Generalizados – mais conhecidos por *Generalized Linear Models* - GLM – foi inicialmente proposta em Nelder e Wedderburn (1972) e engloba distribuições de probabilidade que pertençam à família exponencial e compartilham diversas características.

Denomina-se por Y a variável resposta – aquela sobre a qual se quer fazer inferência – e $\boldsymbol{X}$ uma matriz de covariáveis, fatores, em que cada coluna da matriz $\boldsymbol{X}$ representa os valores para uma variável explicativa – aquelas que serão utilizadas para tentar explicar Y. Seja $i = 1,\ 2,\ \ldots,\ N$, o índice denotando a i-ésima observação de um total de N observações, tal que y_i representa a i-ésima observação da variável resposta Y e o vetor $\underset{\sim}{\boldsymbol{x}}_i^T$ representa a i-ésima linha da matriz de covariáveis $\boldsymbol{X}$, isto é, os valores das covariáveis associados com a resposta y_i. Seja, também, o vetor $\underset{\sim}{\boldsymbol{\beta}}$ o vetor de efeitos fixos associados à cada covariável presente em $\boldsymbol{X}$. Destaca-se que utiliza-se o sinal de transposto $(^T)$ no vetor de variáveis explicativas pois este representa um vetor linha. A notação sem o sinal de transposto é utilizada em todo o texto para representar vetores coluna.

No *framework* de GLM é assumida uma função g, denominada função de ligação – ou função *link* – que permite conectar a esperança da variável resposta Y_i, denominada por μ_i, ao componente linear $\underset{\sim}{\boldsymbol{x}}_i^T\underset{\sim}{\boldsymbol{\beta}}$, ou seja,

$$g(\mathrm{E}[Y_i]) = g(\mu_i) = \underset{\sim}{\boldsymbol{x}}_i^T\underset{\sim}{\boldsymbol{\beta}}$$

Os modelos que pertencem à classe GLM são definidos em termos de variáveis aleatórias independentes $Y_1,\ Y_2,\ \ldots,\ Y_N$, cada uma com distribuição pertencente à família exponencial (BARNDORFF-NIELSEN, 2014) de distribuições com um parâmetro θ_i associado. O principal interesse está em um menor conjunto de parâmetros $\underset{\sim}{\boldsymbol{\beta}} = [\beta_0,\ \beta_1,\ \ldots,\ \beta_{p-1}]^T$, em que $p < N$ representa o número de covariáveis. Supondo que $\mathrm{E}[Y_i] = \mu_i$, em que μ_i é alguma função de θ_i, existe uma transformação de μ_i tal que

$$\mu_i = g^{-1}(\underset{\sim}{\boldsymbol{x}}_i^T\underset{\sim}{\boldsymbol{\beta}})\ ,$$

em que devem ser verificadas as seguintes propriedades:

1. g, denominada função *link*, é monótona e diferenciável;
2. O vetor $\underset{\sim}{\boldsymbol{x}}_i^T$ é um vetor $p \times 1$ de variáveis explicativas;
3. $\underset{\sim}{\boldsymbol{\beta}}$ é um vetor $p \times 1$ de parâmetros, denominados como efeitos fixos das covariáveis.

Um GLM está definido um preditor linear $\eta_i = \underset{\sim}{\boldsymbol{x}}_i^T \underset{\sim}{\boldsymbol{\beta}}$ que está relacionado a μ_i, a esperança de Y_i, através da conexão $\eta_i = g(\mu_i)$.

2.5 Modelo de Regressão Logística

O modelo de regressão logística é um tipo de modelo que pertence à classe GLM. Neste, a variável resposta Y assume valores 0 ou 1, em que se define 0 como falha e 1 como sucesso e, sendo assim, $Y_i \sim \text{Bernoulli}(\pi_i)$, sendo π_i a probabilidade de sucesso para o i-ésimo indivíduo. O particular interesse neste tipo de modelo é avaliar como as covariáveis estudadas afetam a probabilidade de sucesso $\Pr(y_i = 1)$, ou seja, estimar π_i e o efeito – impacto que possui na variável resposta – de cada covariável.

A distribuição de *Bernoulli* possui sua função de probabilidade dada por

$$Y_i \sim \text{Bernoulli}(\pi_i) \implies \Pr(y_i) = \pi_i^{y_i}(1-\pi_i)^{(1-y_i)} \tag{2.4}$$

e, como a distribuição de *Bernoulli* pertence à família exponencial de distribuições, é possível escrever 2.4 da seguinte forma:

$$\Pr(y_i) = \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right] .$$

Se for definido $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, então é possível escrever a equação anterior como

$$\Pr(y_i) = \exp\left\{y_i\theta_i - \ln\left[1+\exp(\theta_i)\right]\right\} . \tag{2.5}$$

Portanto, a função *link* canônica obtida é dada por

$$g(\pi_i) = \theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) ,$$

muito conhecida na literatura como função logística. Utilizando a relação de conectar o parâmetro natural θ_i ao preditor linear $\underset{\sim}{\boldsymbol{x}}_i^T \underset{\sim}{\boldsymbol{\beta}}$,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \underset{\sim}{\boldsymbol{x}}_i^T \underset{\sim}{\boldsymbol{\beta}} ,$$

de onde se encontra

$$\pi_i = \frac{\exp(\underset{\sim}{\boldsymbol{x}}_i^T \underset{\sim}{\boldsymbol{\beta}})}{1 + \exp(\underset{\sim}{\boldsymbol{x}}_i^T \underset{\sim}{\boldsymbol{\beta}})} = \text{logit}^{-1}(\underset{\sim}{\boldsymbol{x}}_i^T \underset{\sim}{\boldsymbol{\beta}}) \,. \tag{2.6}$$

A função logit, vista na Equação 2.6, é uma função aplicada sobre um número p entre 0 e 1:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

2.6 Modelos Lineares Generalizados de Efeitos Mistos – Intercepto Aleatório

O termo **medidas repetidas** é utilizado para descrever dados agrupados, coletados em diferentes momentos, como, por exemplo, informações sobre o mesmo paciente coletadas a cada determinado intervalo de tempo. Em cenários de medidas repetidas, a utilização de modelos da classe GLM não são adequados da forma original em que são propostos, visto que uma das suposições mais importantes é a independência entre observações. Como já discutido anteriormente, dados coletados sobre os mesmos indivíduos possuem uma estrutura de dependência individual e, portanto, torna-se necessário lidar com essa estrutura de formas apropriada.

Para lidar com esse problema, a classe de Modelos Lineares Generalizados Mistos uma extensão do *framework* de GLM, foi desenvolvida. Esta permite lidar com a estrutura de dependência ao realizar uma modelagem hierárquica multinível nos dados. Por exemplo, em um estudo com medidas repetidas por pacientes cada paciente é independente de qualquer outro, porém há uma dependência intra-paciente, ou seja, as observações realizadas no mesmo paciente são dependentes entre si.

Uma das abordagens mais simples de lidar com o problema da dependência utilizando GLMM é considerar efeitos aleatórios, que são efeitos que variam por *clusters*.

Em um GLM considerou-se um vetor de efeitos fixos das covariáveis $\underset{\sim}{\boldsymbol{\beta}} = [\beta_0,\ \beta_1,\ \ldots,\ \beta_{p-1}]^T$, que era único entre todas as observações da variável aleatória $Y_i,\ i = 1,\ 2,\ \ldots,\ N$. Aqui, para incluir efeitos aleatórios no modelo, é necessário mudar um pouco a notação e incluir um índice adicional.

Seja y_{ij} a j-ésima observação do i-ésimo *cluster*, $i = 1,\ \ldots,\ N_c$ e $j = 1,\ \ldots,\ n_i$, em que N_c denomina o número total de *clusters* presentes nos dados e n_i representa o número total de observações para o i-ésimo *cluster*.

Seja, também, $\underset{\sim}{\boldsymbol{x}}_{ij}^T$ referente ao vetor de covariáveis para a j-ésima observação do i-ésimo

cluster, $\underset{\sim}{\boldsymbol{\beta}} = [\beta_0, \beta_1, \ \ldots, \ \beta_{p-1}]^T$ um vetor de efeitos fixos – presente em todos os *clusters* – e $\underset{\sim}{\boldsymbol{\alpha}} = [\alpha_1, \ \ldots, \ \alpha_i]$ representa o vetor de efeitos individuais, correspondendo a interceptos aleatórios a cada *cluster*.

A formulação acima é a mais simples para um cenário de GLMM, em que se é considerado um intercepto aleatório por *cluster*. A tradução dessa implementação é que todos os *clusters* possuem efeitos que variam em nível populacional – ou seja, que não depende especificamente do *cluster* (as covariáveis fixas) mas cada um deles possui um deslocamento que é independente de outros *clusters* – os efeitos aleatórios.

Utilizando a formulação descrita anteriormente GLM, também é assumido aqui existir uma função de ligação g que conecta a esperança da variável resposta Y_{ij}, representada aqui por μ_{ij}, ao componente linear $\underset{\sim}{\boldsymbol{x}}_{ij}^T\underset{\sim}{\boldsymbol{\beta}} + \alpha_i$ para a j-ésima observação do i-ésimo *cluster*. Em si, o termo α_i representa o deslocamento do intercepto β_0 e esse deslocamento pode ser compreendido como comportamentos individuais ao *cluster* i que afetam – tanto positiva como negativamente – a variável resposta.

2.7 Modelo de Regressão Logística Misto

Aplicando o raciocínio anterior ao modelo de regressão logística, descrito em 2.5, é possível lidar com a estrutura de dependência incluindo, por exemplo, o intercepto aleatório em adição ao modelo. Neste caso, Y_{ij} é uma variável binária – assume valores 0 ou 1 – e, portanto, $Y_{ij} \sim$ Bernoulli(π_{ij}), em que π_{ij} é a probabilidade de sucesso para a j-ésima observação do i-ésimo *cluster*. Nesse caso,

$$\Pr(y_{ij} \mid i) = \pi_{ij}^{y_{ij}} \left(1 - \pi_{ij}\right)^{(1-y_{ij})}, \ i = 1, ; \ \ldots, \ N_c, \ \ j = 1, \ \ldots, \ n_i \, .$$

Assim como feito anteriormente, a equação acima pode ser colocada na forma conhecida da família exponencial de distribuições:

$$\Pr(y_{ij} \mid i) = \exp\left[y_{ij} \ln\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) + \ln(1 - \pi_{ij})\right] \, .$$

Ao definir $\theta_{ij} = \ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$, reduz-se a equação acima para:

$$\Pr(y_{ij} \mid i) = \exp\left\{y_{ij}\theta_{ij} - \ln\left[1 + \exp(\theta_i j)\right]\right\} \, . \tag{2.7}$$

Desta forma, utiliza-se a função de ligação canônica neste caso e, portanto,

$$\pi_{ij} = \frac{\exp(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i)}{1 + \exp(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i)} \, . \tag{2.8}$$

Logo, o modelo logístico misto é dado por

$$\text{logit}\left[\Pr(y_{ij} = 1 \mid i)\right] = \ln\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i \, , \tag{2.9}$$

em que $\underset{\sim}{\boldsymbol{x}}_{ij}^T$ é o vetor de covariáveis, $\underset{\sim}{\boldsymbol{\beta}}$ é o vetor de coeficientes de regressão fixos e α_i é o intercepto aleatório do i-ésimo grupo. Nesse modelo, se assume que $\alpha_1, \ \alpha_2, \ \ldots, \ \alpha_{Nc}$ são independentes e identicamente distribuídos e, além disso, $\sum_{i=1}^{N_c} \alpha_i = 0$ para garantir a identificabilidade desses parâmetros (GUILLAUME et al., 2019).

2.8 Análise de Curvas ROC por Meio de Modelo Estatístico

A análise de curva ROC consiste em avaliar diferentes pontos de corte em uma variável classificadora e determinar o ponto de corte que possui a melhor performance para classificar um desfecho de interesse. Como visto anteriormente, a principal limitação na abordagem de Obuchowski (1997) está no fato de apenas uma variável quantitativa poder ser utilizada para a construção da curva ROC, sendo necessário considerar outra abordagem se o pesquisador possuir interesse em avaliar como um conjunto de covariáveis está associado com o desfecho de interesse.

Para acomodar diferentes covariáveis na construção de uma curva ROC em um cenário de medidas repetidas, é proposto em Liu e Wu (2003), Liu et al. (2005) o ajuste de um modelo de regressão logística misto com a inclusão de um intercepto aleatório por paciente. Trata-se como variável resposta Y_{ij}, o desfecho do paciente: por exemplo, $Y_{ij} = 0$ se a j-ésima observação do i-ésimo *cluster* não possui o desfecho (não está doente, não é aderente, etc.) ou $Y_{ij} = 1$ se a j-ésima observação do i-ésimo *cluster* possui o desfecho de interesse (está doente, é aderente, etc.).

Dessa forma, e também utilizando da Equação 2.8, é possível calcular uma estimativa **probabilidade de um indivíduo possuir o desfecho de interesse** ao avaliar uma combinação das variáveis explicativas. A probabilidade ajustada é utilizada como **variável classificadora**, ou seja, são definidos **pontos de corte na probabilidade ajustada** dos indivíduos serem positivos para o desfecho de interesse e se seleciona o ponto de corte com a melhor performance nesse

intervalo.

Aqui, a principal questão está na construção dessa variável classificadora. Como dito anteriormente, quando a construção da variável classificadora é o valor estimado oriundo de um modelo estatístico, ela está sujeita às mesmas condições e pressupostos de um modelo. Em especial, essa abordagem se torna problemática quando se está na presença de medidas irregulares e há a presença de grandes proporções de *singletons*, visto que será incluso um parâmetro adicional para cada paciente, mesmo que este possua apenas uma única observação. Será visto mais detalhadamente na seção de Resultados e Discussões como a massiva presença de *singletons* torna o modelo estatístico extremamente suscetível a sobreajustes e como isso se reflete na análise de curvas ROC.

2.9 Composição de Verossimilhança – Modelo de Regressão Logística Misto

Utilizando a notação anteriormente vista na seção de modelos mistos, aqui é denotado por $i = 1, \ldots, N_c$ o i-ésimo *cluster* e $j = 1, \ldots, n_i$ a j-ésima observação deste *cluster*. Assumindo que a variável aleatória Y_{ij} seja distribuída segundo uma distribuição de *Bernoulli*, com probabilidade de sucesso π_{ij}, então a função de verossimilhança é, neste caso,

$$\mathcal{L}(\underset{\sim}{\boldsymbol{\pi}}, \underset{\sim}{\boldsymbol{y}}) = \prod_{i=1}^{N_c} \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \ , \tag{2.10}$$

visto que os *clusters* são independentes entre si. Além disso, de 2.8,

$$\pi_{ij} = \frac{\exp(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i)}{1 + \exp(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i)} \ . \tag{2.11}$$

A composição de verossimilhança (VARIN; REID; FIRTH, 2011) é uma técnica estatística empregada quando partições dos dados seguem diferentes distribuições. Aqui, a proposta de modificação de verossimilhança é, ao contrário do modelo de efeitos mistos, definir interceptos aleatórios apenas para *clusters* com mais de uma observação. Assim, espera-se que todos os pacientes possuam uma probabilidade de sucesso modelada pelas covariáveis – efeitos fixos a todos pacientes nas análises – mas que a estrutura de dependência decorrente das medidas repetidas seja representada por meio de interceptos aleatórios. A principal motivação para não considerar interceptos aleatórios para *singletons*, assim como é feito no modelo de efeitos mistos, está na redução de parâmetros no modelo. É feita, inicialmente, a suposição de que os *clusters*

são independentes entre si, *singletons* são independentes de outras observações na base de dados e não há razões para ser acomodada uma estrutura de dependência para estes. Em contrapartida, observações de um mesmo *cluster* possuem dependência entre si e então é incluso o intercepto aleatório de forma a acomodara a dependência.

Se inclusos interceptos aleatórios para *singletons*, então os mesmos possuirão amplitudes extremamente grandes para o intervalo de confiança/credibilidade e, também, poderão ter um efeito pontual destoante do real valor, visto que só há uma observação para ser utilizada na estimação deste parâmetro. Aqueles *clusters* que possuem mais observações possuem mais informações válidas para a estimação do intercepto e, consequentemente, existirá uma menor incerteza sobre seu valor.

A função de verossimilhança composta se utiliza da mesma verossimilhança representada em 2.10, ou seja, segue sendo dada por

$$\mathcal{L}(\underset{\sim}{\boldsymbol{\pi}}, \underset{\sim}{\boldsymbol{y}}) = \prod_{i=1}^{N_c} \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}}$$

mas, agora, a probabilidade π_{ij} de sucesso para a j-ésima observação do i-ésimo *cluster* é dada por

$$\pi_{ij} = \begin{cases} \dfrac{\exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i\right)}{1 + \exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i\right)}, \text{ se } n_i > 1 \\ \\ \dfrac{\exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}}\right)}{1 + \exp\left(\underset{\sim}{\boldsymbol{x}} ij^T \underset{\sim}{\boldsymbol{\beta}}\right)}, \text{ c.c.} \end{cases} . \tag{2.12}$$

Substituindo esse valor de π_{ij} na função de verossimilhança, tem-se que a equação de verossimilhança para o modelo logístico de verossimilhança composta é, então, dada por:

$$\mathcal{L}(\underset{\sim}{\boldsymbol{\beta}}, \underset{\sim}{\boldsymbol{\alpha}};\ \underset{\sim}{\boldsymbol{y}}, X) = \prod_{i=1}^{N_c} \prod_{j=1}^{n_i} \left\{ \left[\frac{\exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i\right)}{1 + \exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i\right)} \right]^{y_{ij}} \left[1 - \frac{\exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i\right)}{1 + \exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i\right)} \right]^{1-y_{ij}} \right\}^{\mathcal{I}(n_i > 1)}$$
$$\times \left\{ \left[\frac{\exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}}\right)}{1 + \exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}}\right)} \right]^{y_{ij}} \left[1 - \frac{\exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}}\right)}{1 + \exp\left(\underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}}\right)} \right]^{1-y_{ij}} \right\}^{1 - \mathcal{I}(n_i > 1)}$$

Aqui, $\mathcal{I}(n_i > 1)$ é a função indicadora, em que $\mathcal{I}(n_i > 1) = 1$ se $n_i > 1$ e 0 caso contrário.

2.10 Inferência no Modelo Proposto

O modelo de verossimilhança composta foi implementado sob uma perspectiva Bayesiana, implementada no *software R* (R Core Team, 2023) com uso do pacote *RStan* (Stan Development Team, 2023). Em particular, foi utilizado o *RStan* por conta do algoritmo Monte-Carlo Hamiltoniano – um método MCMC –, que oferece melhor convergência das cadeias geradas e maior eficiência computacional (BETANCOURT, 2017).

Supondo independência *a priori*, foram especificadas as seguintes distribuições:

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\
\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \begin{cases} \underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}} + \alpha_i, & \text{se } n_i > 1, \\ \underset{\sim}{\boldsymbol{x}}_{ij}^T \underset{\sim}{\boldsymbol{\beta}}, & \text{se } n_i = 1, \end{cases} \\
\underset{\sim}{\boldsymbol{\alpha}} &\sim \mathcal{N}(\underset{\sim}{\mathbf{0}}, \tau^{-1} I_{N_c}) \\
\sum_{i:\, n_i > 1} \alpha_i &\sim \text{Normal}(0,\ 0{,}001 N_c) \\
\underset{\sim}{\boldsymbol{\beta}} &\sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ 100 I_p) \\
\tau &\sim \text{Gama}(0{,}1,\ 0{,}1)
\end{aligned}
\tag{2.13}
$$

em que π_{ij} é a probabilidade de sucesso calculada a partir da equação 2.12, $\mathcal{N}$ denota a distribuição normal multivariada, I_p representa a matriz identidade de ordem p (tamanho do vetor $\underset{\sim}{\boldsymbol{\beta}}$) N_c é o número de *clusters* com mais de uma observação e I_{N_c} é uma matriz identidade de ordem N_c.

3 Resultados e Discussões

De forma a motivar o desenvolvimento de uma metodologia capaz de lidar com a presença de medidas irregulares na análise de curvas ROC, são feitos aqui estudos de simulação em que é avaliada a sensibilidade do modelo Bayesiano proposto à escolha da distribuição *a priori* e também será avaliada a capacidade de recuperação dos parâmetros para os modelos discutidos aqui no texto.

Após isso, serão utilizados dados reais em que se tem o principal interesse de realizar uma análise de curvas ROC em dados altamente desbalanceados, com muita presença de medidas irregulares. Para fins de comparação, será apresentada uma abordagem que mantém a primeira observação de cada paciente quando há a presença de medidas repetidas – uma abordagem que pode imediadamente ser tomada e que se beneficia de utilizar metodologias usuais de análise estatística e modelagem, mas que acaba realizando o descarte de informações e ignorando a estrutura de dependência entre observações de um mesmo *cluster*.

3.1 Estudos de Sensibilidade

De forma a verificar como os parâmetros das distribuições *a priori* utilizadas podem afetar o modelo, foram conduzidos estudos de sensibilidade em alguns cenários simulados, com 50 conjuntos de dados gerados em cada cenário. Os cenários simulados foram pensados de forma tal que os dados gerados possuíssem características semelhantes ao conjunto de dados reais ao qual será aplicada a metodologia, visando também validar e justificar a abordagem tomada para aplicação. Nestes estudos de sensibilidade busca-se, principalmente, verificar se o modelo logístico Bayesiano de verossimilhança composta possui capacidade de recuperação dos parâmetros utilizados em ambientes controlados e, também, qual foi o impacto das variâncias das distribuições *a priori* utilizadas.

Para todos elementos do vetor $\underset{\sim}{\beta}$ – efeitos associados às covariáveis – são consideradas as mesmas distribuições *a priori*, baseadas em uma distribuição normal. Cada efeito aleatório,

elementos do vetor $\underset{\sim}{\boldsymbol{\alpha}}$, possui também uma distribuição normal mas com variância dada por τ^{-1}. O hiperparâmetro τ, no contexto da abordagem Bayesiana, possuirá uma distribuição Gama e será verificado como a variância atribuída desta distribuição pode afetar o modelo utilizado.

Para todos os dados simulados é definido que $\underset{\sim}{\boldsymbol{\beta}} = (\beta_0,\ \beta_1,\ \beta_2)^T = (1{,}4,\ -0{,}7,\ 0{,}93)^T$ e $\tau = 0{,}6$. São gerados de 160 a 220 *clusters*, dos quais sorteia-se uma porcentagem entre 55% a 70% para serem *singletons* – isto é, que terão somente uma observação. Para todos os outros *clusters* que não são previamente definidos como *singletons*, é sorteado um tamanho máximo de *cluster* entre 5 e 9, definido como m, e para cada *cluster*, gera-se um número de observações de uma distribuição Binomial com um total de m tentativas e probabilidade de sucesso $0{,}6$. Os valores da covariável X_1 são gerados a partir de uma distribuição *Poisson* com esperança 2 e de X_2, de uma distribuição normal com esperança 1 e variância $1{,}7$. Como dito anteriormente, serão gerados 50 bases de dados com essas características e busca-se, dessa forma, criar uma variedade de cenários simulados nos quais serão avaliados como o modelo proposto performa na recuperação dos parâmetros.

A primeira configuração de distribuições *a priori* a ser testada nos dados gerados foi a seguinte:

$$\underset{\sim}{\boldsymbol{\beta}} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ 100 I_p)$$

$$\underset{\sim}{\boldsymbol{\alpha}} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ \tau^{-1} I_{N_c})$$

$$\tau \sim Gama(0{,}1,\ 0{,}1)$$

A princípio, preferiu-se que fosse deixada a variância dos efeitos aleatórios como hiperparamêtros pois a restrição de lhe atribuir um valor fixo para esta pode ser considerada uma imposição muito forte. Logo, para obter uma distribuição menos informativa, a variância foi considerada como hiperparâmetro e utilizou-se uma distribuição *a priori* vagamente informativa para a precisão (τ). Para cada um dos 50 modelos ajustados, foram computadas as estimativas médias para cada um dos parâmetros, assim como intervalos de credibilidade de 95%. Com essa medida, calculou-se o que foi denominado como **porcentagem de cobertura** para cada um dos parâmetros, representando a proporção de vezes das 50 em que no intervalo de credibilidade de 95% estava incluso o real valor do parâmetro, definido para a simulação. Além disso, foram calculados os erros médios quadráticos ao comparar as estimativas das médias com os reais valores dos parâmetros. Os resultados para esse primeiro cenário seguem abaixo:

Tabela 3: Estimativas de recuperação do ajuste do modelo para a primeira configuração de distribuições *a priori*.

Estatísticas	β_0	β_1	β_2	τ
Porcentagem de Cobertura I.C. 95%	96%	94%	98%	72%
Erro Médio Quadrático	0,356	0,065	0,142	30,5

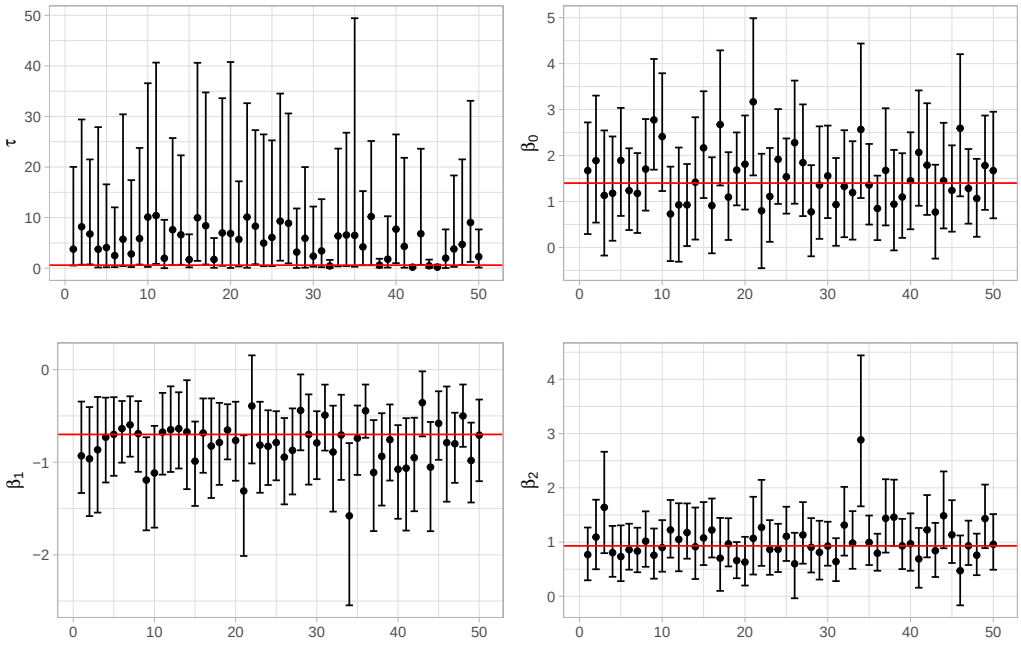

Figura 2: Intervalos de 95% de Credibilidade para cada uma das 50 simulações. Em vermelho tem-se a marcação do real valor do parâmetro e cada ponto representa a média estimada deste parâmetro na simulaçã realizada.

Em geral, percebe-se uma boa capacidade de recuperação dos efeitos das covariáveis, enquanto que a precisão apresenta intervalos muito mais amplos e estimativas pontuais mais distantes do real valor de τ. Também, a porcentagem de cobertura dos intervalos de credibilidade de 95% para τ é consideravelmente baixo e o erro médio quadrático é extremamente divergente em relação aos outros parâmetros.

Visando reduzir a amplitude dos intervalos de credibilidade para a precisão, τ, foram mantidas as distribuições *a priori* utilizadas para $\underset{\sim}{\beta}$ e considerou-se uma distribuição *a priori* Gama

ainda com esperança 1, mas com variância maior. A seguinte configuração foi considerada:

$$\underset{\sim}{\boldsymbol{\beta}} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ 100 I_p)$$

$$\underset{\sim}{\boldsymbol{\alpha}} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ \tau^{-1} I_{N_c})$$

$$\tau \sim Gama(1,\ 1)$$

Equivalentemente, aqui, se tem os seguintes resultados:

Tabela 4: Estatísticas dos ajustes para a segunda configuração considerada na análise de sensibilidade.

Estatísticas	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\tau}$
Porcentagem de Cobertura	96	94	98	100
Erro Médio Quadrático	0,374	0,066	0,139	0,73

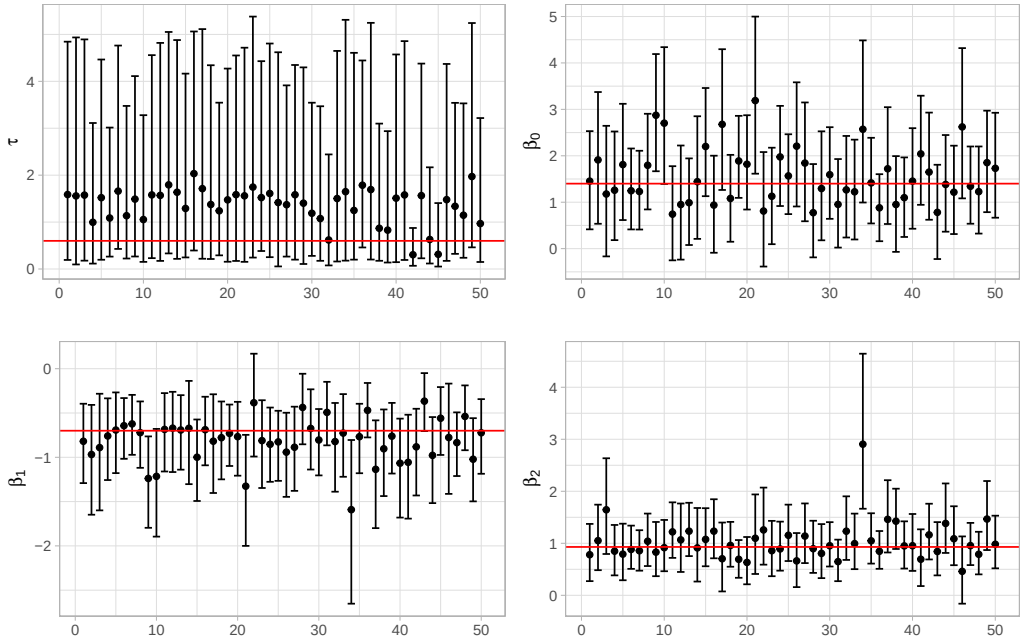

Figura 3: Intervalos de 95% de Credibilidade para cada uma das 50 simulações para a segunda configuração da análise de sensibilidade.

O principal destaque aqui está no fato de que 100% dos intervalos de credibilidade estimados cobriam o real valor do parâmetro τ e de um menor erro médio quadrático para este parâmetro. Para os outros parâmetros, tem-se resultados bastantes similares com a configuração anterior.

A próxima configuração busca investigar os efeitos de considerar uma distribuição *a priori*

mais vagas para os efeitos das covariáveis. A seguinte configuração foi considerada:

$$\underset{\sim}{\boldsymbol{\beta}} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ 100^2 I_p)$$

$$\underset{\sim}{\boldsymbol{\alpha}} \sim \mathcal{N}(\underset{\sim}{\mathbf{0}},\ \tau^{-1} I_{N_c})$$

$$\tau \sim Gama(1,\ 1)$$

Seguem abaixo os resultados considerando as configurações acima definidas:

Tabela 5: Estatísticas de ajustes considerando uma maior variância para a distribuição *a priori* dos efeitos das covariáveis.

Estatísticas	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\tau}$
Porcentagem de Cobertura	96	92	98	100
Erro Médio Quadrático	0,375	0,068	0,142	0,68

Percebe-se que considerar uma distribuição *a priori* menos informativa para os efeitos das covariáveis não causou praticamente nenhum tipo de alteração em relação às estatísticas aqui avaliadas. A proporção de cobertura dos intervalos de credibilidade foram praticamente os mesmos e os erros médios quadráticos não apresentaram diferenças consideráveis.

O estudo de análise de sensibilidade realizado foi crucial para guiar a escolha das distribuições *a priori* a serem utilizadas no modelo descrito na metodologia, o qual é utilizado ao longo do resto do trabalho. Foi apontado que o modelo possui certa sensibilidade à escolha da distribuição *a priori* para o hiperparamêtro da precisão, τ. Os ambientes simulados, que buscam se assemelhar ao conjunto de dados reais e para o qual se aplicará a metodologia proposta, apresentaram melhor cobertura dos parâmetros e menores métricas de erros quando considerada a distribuição *a priori* Gama$(1,\ 1)$, pouco mais informativa que a distribuição Gama$(0{,}1;\ 0{,}1)$.

3.2 Estudos de Simulação

Os estudos de simulação conduzidos têm por principal objetivo verificar a capacidade de recuperação do modelo proposto em ambientes controlados e também de comparar a qualidade de ajuste entre as possíveis abordagens. Busca-se, também, evidenciar o impacto causado devido a diferentes proporções de *singletons* quando utilizado um modelo de efeitos mistos. Todos os dados gerados nos estudos de simulação buscam representar cenários próximos a dados reais em que se desejavam avaliar as metodologias aqui utilizadas.

A primeira parte desta seção busca avaliar os impactos decorridos por conta de grandes

proporções de *singletons* em um estudo envolvendo medidas repetidas irregulares utilizando a abordagem de efeitos mistos, sugerida em Liu e Wu (2003). Em um primeiro momento, a principal hipótese foi a existência de grades vieses na estimação dos efeitos das covariáveis presentes a todos pacientes. Para isso, é feito aqui uma abordagem semelhante ao estudo de Bell, Ferron e Kromrey (2008), em que se avalia o comportamento das estimativas dos efeitos fixos em diferentes cenários simulados. Aproveita-se esta parte para avaliar também como o modelo proposto na recuperação dos reais valores dos parâmetros, avaliando como se dá a presença de viés nesses casos.

Já a segunda parte desta seção avalia, em especial, os impactos de diferentes proporções de *singletons* na análise de curvas ROC em três abordagens diferentes:

1. **Primeira Observação de Cada Paciente:** tornar válida a suposição de independência na análise de curvas ROC usual é manter apenas a primeira observação de cada paciente e se beneficiando de abordagens estatísticas usuais;

2. **GLMM:** utilizar a abordagem de um modelo misto de regressão logística (LIU; WU, 2003) para construir a variável classificadora – probabilidade ajustada – a partir de um conjunto de covariáveis.

3. **Verossimilhança Composta Bayesiana:** utilizar o modelo logístico Bayesiano de verossimilhança composta para, assim como no item anterior, construir a variável classificadora a partir de um conjunto de covariáveis.

Visto que aqui se parte de um conjunto de covariáveis e não apenas de uma variável classificadora, não é possível utilizar a abordagem proposta em Obuchowski (1997). Neste caso, deveria ser utilizada alguma técnica de redução de dimensionalidade para a construção de uma única variável classificadora.

3.2.1 Parte I – Avaliação das Estimativas dos Efeitos Fixos

Foram definidos 12 diferentes cenários para avaliar como as estimativas dos efeitos fixos se dão, em cada uma das abordagens possíveis. Esses 12 cenários surgem da combinação entre as possibilidades:

1. Números de *clusters* variando entre 40 a 80, 180 a 220 e 480 a 520 *clusters*;

2. Proporção de *singletons* variando entre 10%, 30%, 60% e 90%.

Para cada um dos diferentes cenários são gerados 25 bancos de dados, totalizando 300 bancos de dados. Os parâmetros utilizados para a simulação são $\underset{\sim}{\boldsymbol{\beta}} = [-2{,}5,\ 1{,}4]^T$, $\tau = 1$ e os valores da variável explicativa X são gerados de uma distribuição exponencial com esperança $2{,}6$. Em cada simulação, é inicialmente sorteada a quantidade de *clusters* e, depois é definido o número de *singletons*, ambos determinados com base no respectivo cenário. Para cada *cluster* que não é pré-definido como um *singleton*, são geradas as quantidades de observações para este, sendo o mínimo de 2 observações e um máximo de 7, com média de $3{,}78$ observações por *cluster*. O pacote *lme4* (BATES et al., 2015), utilizado no *software R* 4.3.2 (R Core Team, 2023), foi utilizado para realizar os ajustes do modelo de regressão logística misto, enquanto que um código foi desenvolvido no *RStan* (Stan Development Team, 2023) para ajuste do modelo Bayesiano de verossimilhança composta, utilizando como distribuições *a priori* as informadas anteriormente na metodologia. Por efeitos comparativos, foram também exibidos os resultados do modelo que mantém a primeira observação de cada *cluster*.

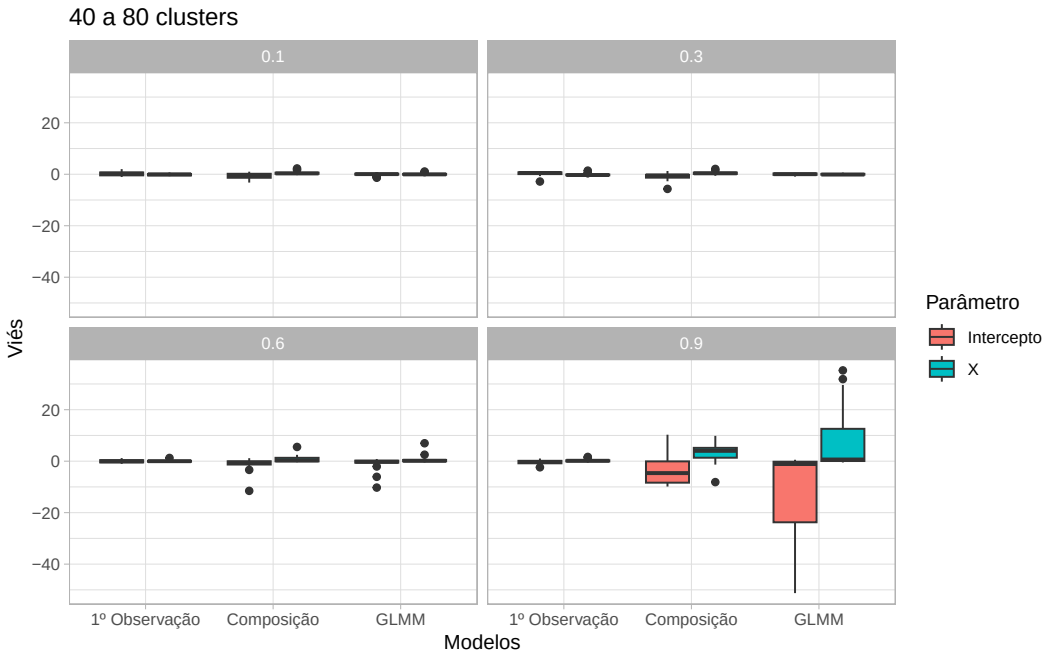

Figura 4: *Box-plots* dos vieses para dados gerados em diferentes proporções de *singletons* para dados com 40 a 80 *clusters*. No título de cada quadro se tem a proporção de *singletons*.

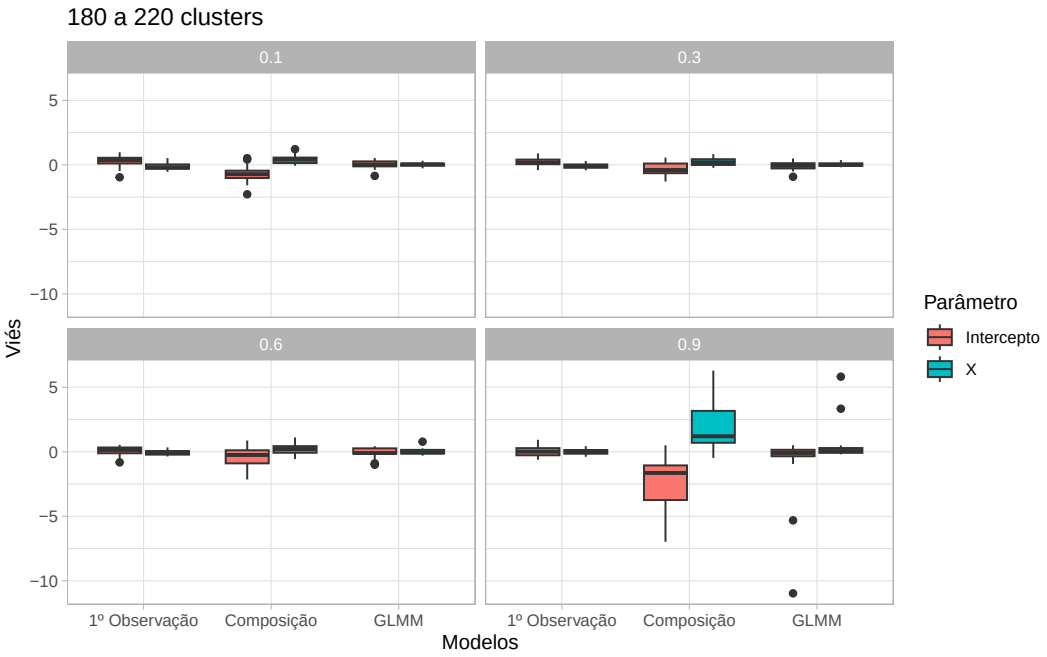

Figura 5: *Box-plots* dos vieses para 180 a 220 *clusters*.

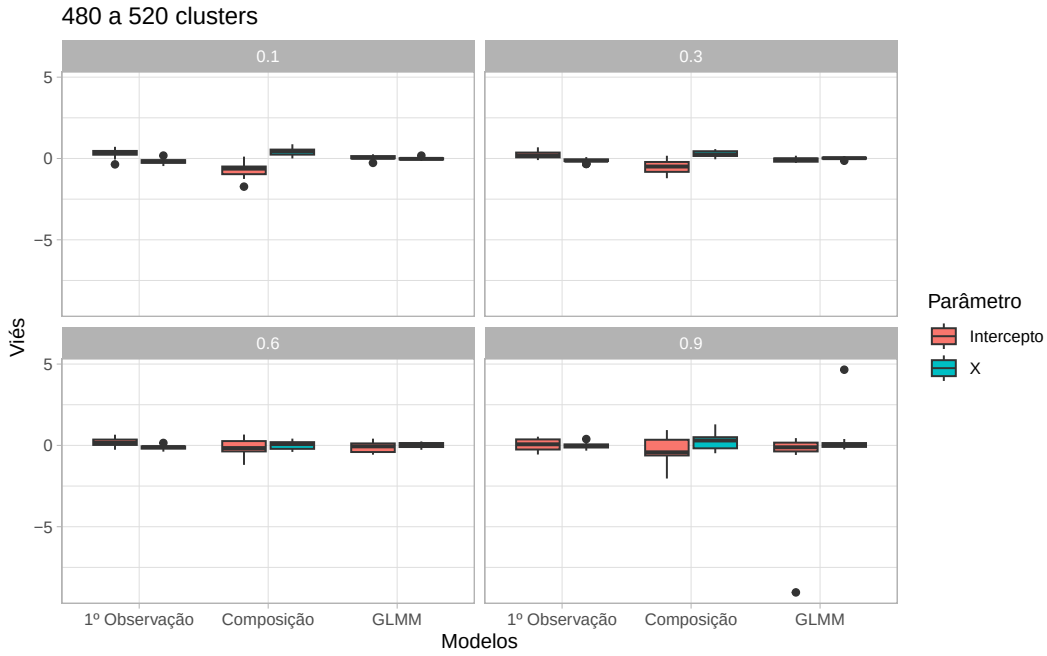

Figura 6: *Box-plots* dos vieses para 480 a 520 *clusters*.

Corroborando com os resultados de Bell, Ferron e Kromrey (2008), percebe-se que o número de *clusters* parece importar muito mais para o viés do que o tamanho dos *clusters* ou a proporção de *singletons*. Em geral, os modelos utilizados apresentam mediana próximas de 0, ou seja, um viés mediano praticamente negligenciável. Em cenários bastante desbalanceados,

com pouco números de observações e grandes proporções de *singletons*, o modelo GLMM – notação utilizada para o modelo logístico de efeitos mistos – apresentou vieses pontuais bastante consideráveis, mas esses tiveram mediana e pelo menos um dos quartis bastante próximos de 0. Principalmente no cenário em que se tem um baixo número de *clusters*, o modelo Bayesiano de verossimilhança composta apresentou viés bem menor que o modelo de efeitos mistos. Apesar disto, este apresentou amplitudes interquartílicas maiores em outros cenários nos quais o modelo logístico performou consistentemente bem.

Desse primeiro estudo de realização, pode-se concluir que o modelo de verossimilhança composta performa tão bem quanto o modelo logístico de efeitos mistos, proposto em Liu e Wu (2003), não apresentando ainda justificativas suficientes para seu uso.

A próxima parte da simulação busca evidenciar como esses modelos, mesmo que sem apresentar consideráveis vieses nas estimativas pontuais dos efeitos das covariáveis, podem afetar a análise de curvas ROC – mais especificamente, como estes se comportam quando se trata da construção da variável classificadora a partir do uso de covariáveis.

3.2.2 Parte II – Análise do Impacto do Modelo na Construção de Curvas ROC

A justificativa para tal estudo se dá pela suspeita de que o modelo logístico de efeitos mistos, quando utilizados dados muito desbalanceados – isto é, com grande presença de *singletons* – pode possuir uma inflação no número de parâmetros desnecessários e que estes, mesmo que não informativos, podem acabar afetando a construção da variável classificadora.

Busca-se simular os dados de forma um pouco diferente aqui, procurando construir cenários simulados mais parecidos ao conjunto de dados reais ao qual se possui o interesse de aplicar a metodologia proposta. Assim, espera-se avaliar o comportamento dos modelos utilizados em cenários controlados nos quais é possível investigar suas propriedades e se estes são adequados ao problema existente.

Para a finalidade descrita acima, é considerado um esquema de simulação de dados um pouco mais elaborado:

- Inicialmente, foram definidos os valores dos parâmetros e aqui 3 variáveis explicativas foram consideradas: $\underset{\sim}{\boldsymbol{\beta}} = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3)^T = (-1{,}3,\ -0.8,\ 1{,}4,\ 0{,}25)^T$. Os valores da variável explicativa X_1 – associada a β_1 – foram gerados de uma distribuição normal com média 0 e variância 1, de X_2 – associada a β_2 – de uma distribuição uniforme contínua com mínimo -2 e máximo $0{,}8$ e de X_3 também geradas de uma uniforme contínua com

mínimo $0{,}5$ e máximo $1{,}4$. Para τ, a precisão dos efeitos de cada *cluster*, considerou-se o valor $0{,}75$. Com os parâmetros definidos e tendo-se os valores das variáveis explicativas, para a i-ésima observação do j-ésimo *cluster* se calcula a probabilidade de sucesso π_{ij} a partir da combinação dos valores das variáveis explicativas e dos efeitos das covariáveis e então é criada a variável resposta $Y_{ij} \sim$ Bernoulli(π_{ij});

- Foram definidos 170 *clusters*, cada um com 10 observações, totalizando uma base de dados com 1700 observações. Esta, é a base de dados completa sobre as quais serão realizadas as simulações;

- Foram consideradas as seguintes proporções de *singletons*: 10%, 20%, 30%, $\ldots$, 90%. Para cada uma das proporções, foram amostradas 100 sub-bases da base de dados completa seguindo o seguinte esquema:

    1. Primeiramente, são sorteados quais *clusters* dos 170 serão definidos como *singletons* – de acordo com a proporção do respectivo cenário – e sorteia-se apenas uma observação das 10 para cada um desses definidos como *singletons*.

    2. Após isso, para cada os *clusters* restantes – que não são definidos como *singletons*, ou seja, possuem mais de uma observação – sorteia-se de uma distribuição Binomial$(10,\ 0{,}25)$ o respectivo número de observações. Caso o valor sorteado seja menor que 2, arredonda-se para 2. É então amostrado entre as 10 observações o número correspondente de amostras sorteado para este *cluster*.

Em resumo para o processo de simulação criado, cria-se uma base de dados balanceada com um número considerável de *clusters* e de observações por *cluster*. Seguindo-se o processo descrito acima, é possível amostrar diferentes bases desbalanceadas partindo desta base maior, balanceada, simulada. Em total, são 9 diferentes cenários de proporções de *singletons* e 100 amostras para cada um dos cenários, totalizando assim 900 bases de dados. Aqui, são ajustados o modelo logístico de efeitos mistos e o modelo Bayesiano de verossimilhança composta e para cada uma das simulações é realizada a análise de curvas ROC a fim de obter a respectiva AUC computada do respectivo cenário.

Tabela 6: Média e desvio-padrão das AUC obtida em cada simulação para o modelo logístico de efeitos Mistos e o modelo Bayesiano de verossimilhança composta.

Proporção de Singletons	**Modelo de Efeitos Mistos**			**Verossimilhança Composta**		
	Média AUC	**Desvio Padrão AUC**	**Proporção AUC = 1**	**Média AUC**	**Desvio Padrão AUC**	**Proporção AUC = 1**
10%	0,932	0,027	0%	0,823	0,018	0%
20%	0,931	0,028	0%	0,811	0,02	0%
30%	0,93	0,028	0%	0,793	0,023	0%
40%	0,929	0,036	0%	0,776	0,026	0%
50%	0,932	0,038	1%	0,756	0,031	0%
60%	0,926	0,049	5%	0,75	0,032	0%
70%	0,928	0,058	10%	0,752	0,04	0%
80%	0,903	0,072	19%	0,733	0,044	0%
90%	0,891	0,079	22%	0,717	0,064	0%

O modelo logístico de efeitos mistos apresentou, em todos os cenários considerados, uma maior consistência entre as AUC estimadas. Destaca-se a proporção de ocorrência $AUC = 1$ pois isso indica que por meio do modelo estatístico, construiu-se uma variável classificadora com capacidade discriminatória perfeita, o que é uma alta evidência de sobreajuste do modelo. Como os pontos de corte avaliados são feitos sobre a probabilidade ajustada fornecida pelo modelo, isso significa que o modelo ajusta probabilidades ou muito próximas de 0 ou muito próximas de 1.Por fins investigativos, foi selecionado um dos casos em que ocorreu $AUC = 1$ no modelo logístico de efeitos mistos e foram avaliadas as estimativas pontuais dos efeitos aleatórios de cada *cluster*, além da probabilidade ajustada.

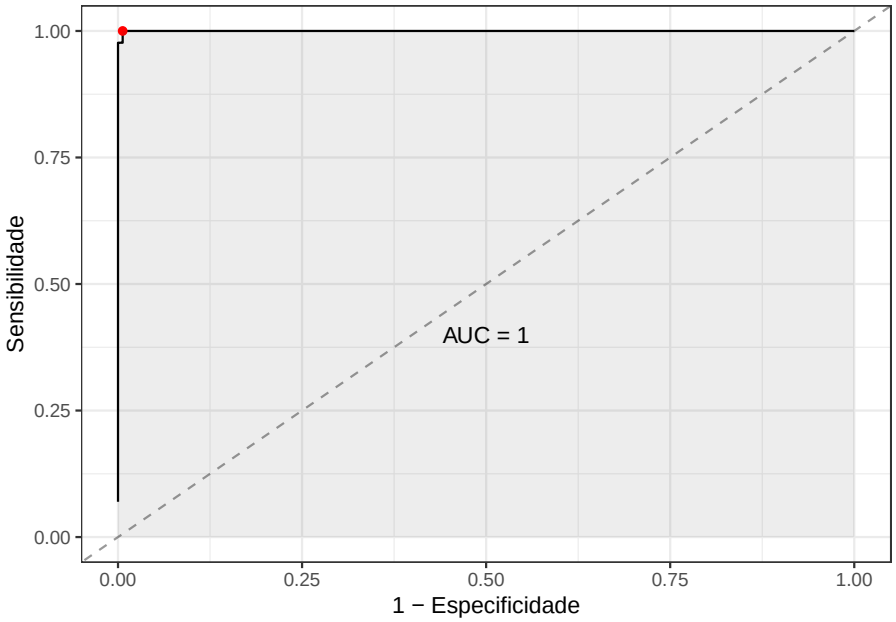

Figura 7: Curva ROC para um dos casos em que $AUC = 1$.

Na Figura 7, percebe-se que apesar de a área abaixo da curva não ser exatamente 1, esta possui sensibilidade e especificidade praticamente 100%, ou seja, uma capacidade perfeita de discriminação entre doentes e não-doentes, por exemplo. Em especial, essa base de dados consiste de 201 observações e investigou-se mais à fundo o motivo deste ajuste perfeito.

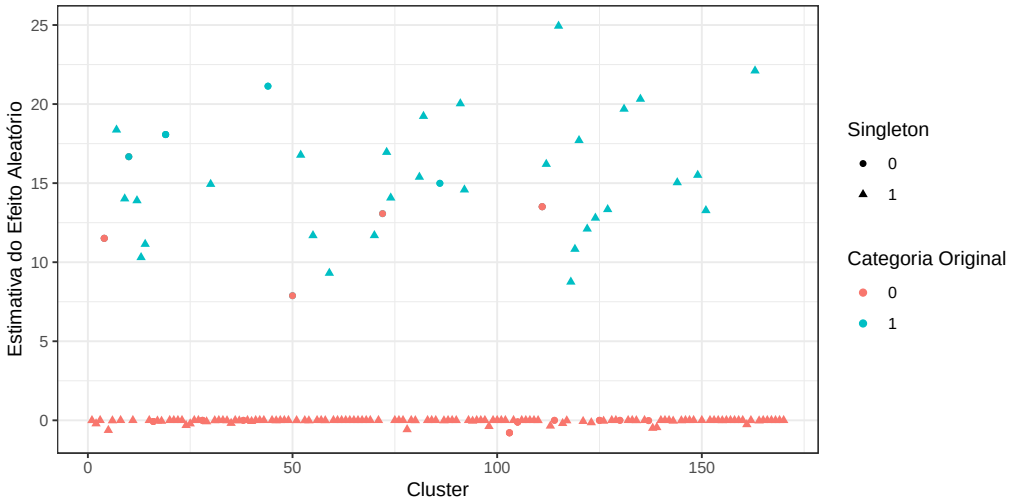

Figura 8: Efeitos estimados para cada *cluster* para um dos casos em que o corre $AUC = 1$.

Em grande maioria, os *singletons* que possuem o desfecho possuem seus efeitos aleatórios estimados em valores extremamente elevados. Aqui, vale reforçar que a variância utilizada para gerar esses efeitos foi de 1,8. Pode-se avaliar diretamente o impacto dessas estimativas na pro-

babilidade ajustada de cada observação:

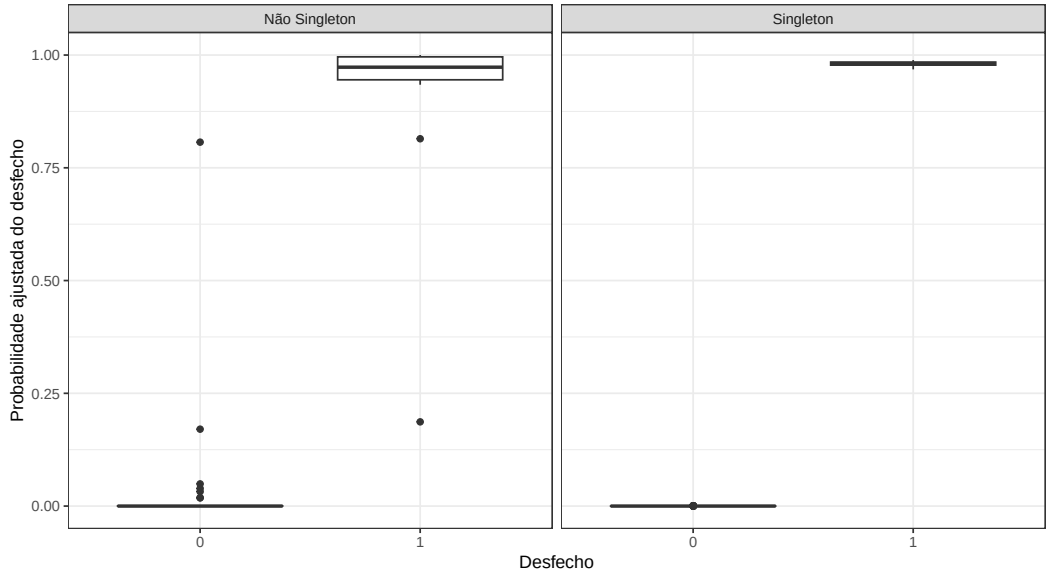

Figura 9: Probabilidades ajustadas das observações para um dos casos em que ocorre $AUC = 1$, separados por se é *singleton* ou não.

Percebe-se na Figura 9 que as probabilidades estimadas quando os *clusters* são *singletons* são basicamente determinísticas, visto que não há desvio ou incerteza atribuída a estas. Essa, por si só, é uma grande evidência de sobreajuste, em que a estimativa atribuída ao efeito aleatório para *singletons* anulam qualquer e todo efeito explicativo das covariáveis devido à sua grandeza. Por efeitos de comparação, foi selecionado um caso em que a AUC estimada é diferente de 1 e que não há indícios de sobreajuste do modelo. Os seguintes resultados são referentes às mesmas características vistas anteriormente, porém em um caso que não há evidências de sobreajuste.

No exemplo visto aqui, a probabilidade ajustada é tão próxima do real valor do desfecho – 0 quando o desfecho é negativo e 1 quando o desfecho é positivo – por conta das estimativas dos efeitos aleatórios, e não por conta dos efeitos das covariáveis, como seria o ideal, Ou seja, o efeito aleatório é incluído no modelo não apenas como uma forma de acomodar a estrutura de dependência mas, por conta do sobreajuste do modelo causado por conta da presença destas, como elementos determinísticos para a classificação do paciente entre com desfecho e sem desfecho.

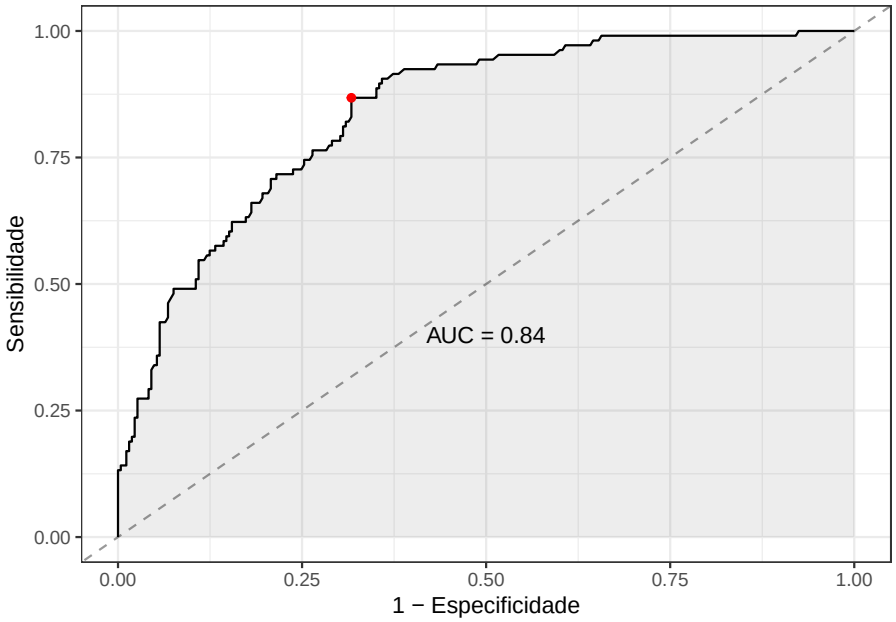

Figura 10: Curva ROC para um caso em que a presença de *singletons* não é prejudicial ao modelo.

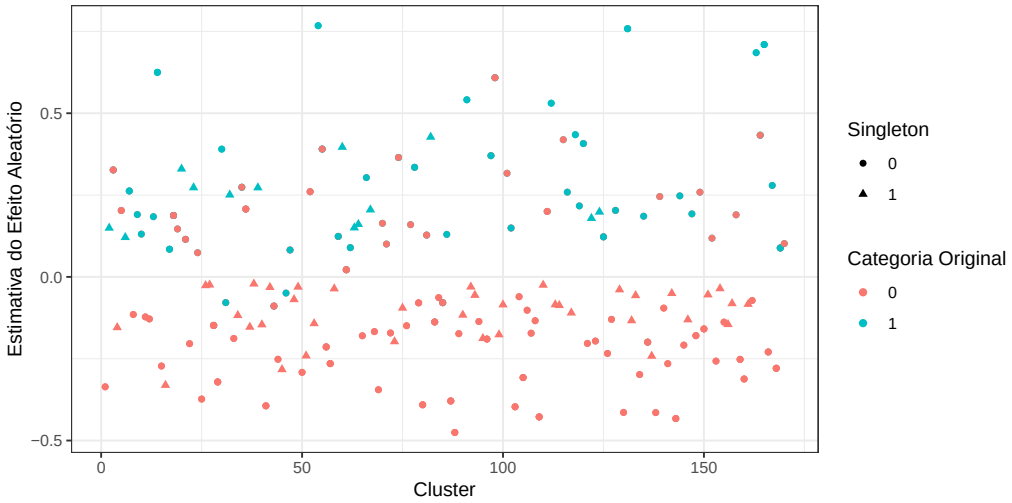

Figura 11: Efeitos estimados para cada *cluster*.

Das figuras anteriores, percebe-se principalmente que a curva ROC construída possui diversos degraus e se assemelha mais com, de fato, uma curva. Além disso, os efeitos estimados se misturam em geral, em que efeitos azuis – referentes aos desfechos positivos – tomam estimativa maior que os efeitos em vermelhos, associados aos desfechos negativos. Além disso, não há mais a separação brusca entre *singletons* e *clusters* com mais de uma observação. Percebe-se, aqui, um melhor comportamento do modelo.

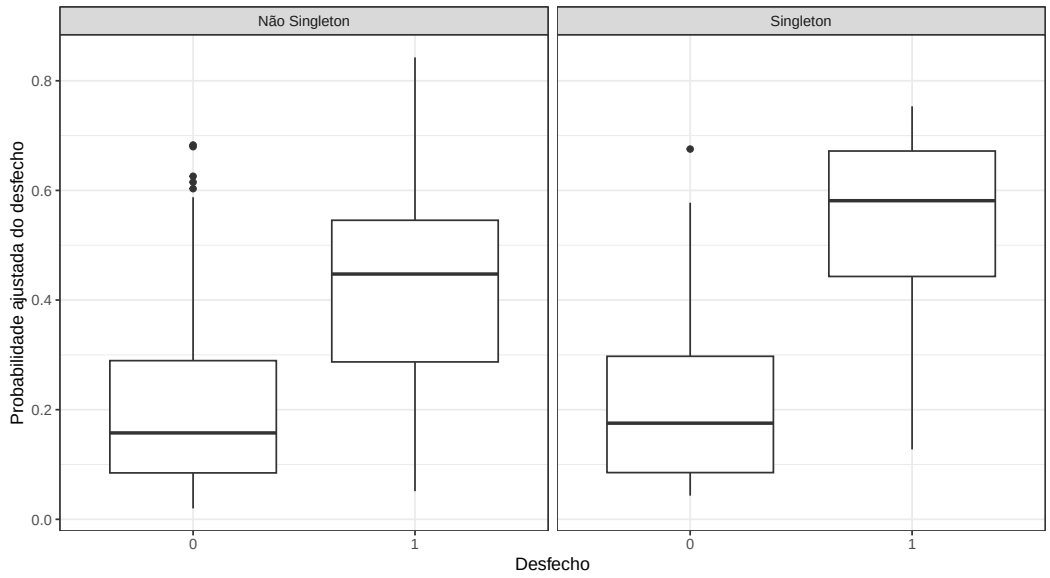

Figura 12: Probabilidades ajustadas das observações separadas por *status* de *singleton*.

Em especial, percebe-se que os *clusters* com mais de uma observação e os *singletons* não apresentam comportamentos distintos. Parece, de fato, que as covariáveis possuem aqui poder preditivo em relação aos efeitos dos *clusters*, e que estes entram na análise de fato para representar a estrutura de dependência entre observações realizadas em uma mesma unidade.

Voltando à Tabela 6, é possível perceber que a AUC média para o modelo de efeitos mistos possui um crescimento do desvio padrão conforme a proporção de *singletons* aumenta. Além disso, espera-se que conjuntos de dados mais desbalanceados, tenha-se uma menor capacidade discriminatória devido à falta de informação. Esse é um comportamento esperado que não é observado por completo no modelo logístico de efeitos mistos, em que a AUC média de cenários simulados com 70% de *singletons* é tão próxima quanto de cenários mais balanceados, como com 10% de *singletons*.

Ao contrário, o modelo Bayesiano de verossimilhança composta não apresenta em nenhuma vez a discriminação perfeita dos dados, sugerindo assim que este não é tão impactado pelo sobreajuste do modelo como o outro modelo. Além disso, o modelo proposto apresenta estimativas da AUC que seguem o comportamento esperado: cenários mais desbalanceados apresentam menor capacidade discriminatória do que àqueles em que os dados são menos desbalanceados. Destaca-se também que o estimador da AUC obtida a partir das estimativas do modelo proposto apresentou um desvio-padrão menor do que com o modelo logístico de efeitos mistos.

Como os cenários simulados buscam se aproximar da aplicação feita em dados reais, é possível perceber a aparição de problemas a partir de uma proporção de 70% de *singletons*, em

que se torna cada vez mais frequente a discriminação perfeita dos dados devido a sobreajuste do modelo. Além disso, do estudo de simulação realizado anteriormente, encontra-se que uma proporção de *singletons* a partir de 60% para dados entre 180 a 220 *clusters* pode implicar em vieses consideráveis nas estimativas dos efeitos fixos.

3.3 Aplicação em Dados Reais

Os dados aqui são oriundos do Projeto PrEP1519, um estudo desenvolvido nas capitais Salvador, Belo Horizonte e São Paulo que tem como público alvo adolescentes de 15 a 19 anos que se identificam como mulheres transexuais ou travestis, homens cis-gêneros gays, bissexuais ou que fazem sexo com outros homens. O objetivo do estudo é avaliar a efetividade a profilaxia pré-exposição – PrEP, uma forma de prevenção ao HIV – ao HIV nesta população. O tratamento consiste na tomada diária de um comprimido de PrEP, em que estes são dispensados em unidades de saúde específicas e a cada ida para consulta, dispensa-se comprimidos suficientes para 1, 2 ou 3 meses, de acordo com a prescrição médica fornecida ao paciente.

A base de dados consta de 302 observações, referentes a 188 *clusters* – pacientes. Do total de 188 pacientes presentes na base de dados, 65 possuem mais de uma observação e 123 possuem apenas uma observação – isto é, são *singletons*. A proporção de *singletons* aqui é de $65{,}43\%$, sendo essa considerada dentro de uma especificação moderada – segundo os cenários simulados. Os 65 *clusters* que não são considerados como *singletons* possuem máximo e média de 7 e $2{,}75$ observações por *cluster*, respectivamente, totalizando 179 observações. Aqui, o número de observações por paciente parece ter sido amplamente impactado pela pandemia, visto que os dados são coletados entre Fevereiro de 2019 e Dezembro de 2020.

O processo de coleta de dados se dá da seguinte forma: os participantes do estudo comparecem a clínicas de saúde para avaliar a aderência ao tratamento e também para a busca de novos comprimidos. Nessa comparecimento, é levado o frasco anterior de comprimidos para contagem de pílulas e é feita a coleta de uma amostra de sangue por meio de DBS – *Dried blood spot*, mancha de sangue seco – na qual é avaliada a concentração dos compostos químicos do medicamento no sangue do paciente. Em especial, verifica-se a concentração do tenofovir difosfato, um dos principais compostos do PrEP e que está diretamente relacionado com a prevenção ao HIV (MOFENSON; BAGGALEY; MAMELETZIS, 2017; CASTILLO-MANCILLA et al., 2019; MATTHEWS et al., 2019). Caso a concentração de tenofovir difosfato esteja acima de 1400 fmol/coleta – aproximadamente 4 comprimidos por semana –, considera-se que o paciente é aderente ao tratamento. Caso contrário, esse paciente é dito não-aderente. Constrói, assim, uma

classificação para cada paciente: 1, caso seja aderente e 0, caso contrário.

Dado que o retorno do paciente à clínica para avaliação médica é altamente impactada pela pandemia, assim como por diversos outros fatores que podem estar associados ao retorno do paciente, esse estudo tem como principal característica a **identificação de fatores alternativos** – que não necessitem do comparecimento do paciente na clínica – que podem ser utilizados para diagnosticar a aderência do paciente ao tratamento com PrEP. Para isso, são utilizadas as seguintes covariáveis:

- **mpr**: uma medida de adesão do paciente que assume valores acima de 0 e é calculada com base no tempo teórico, em dias, que o paciente ficou coberto pela medicação. É calculada com base em quantos medicamentos foram dispensados na última dispensa do paciente e no intervalo de tempo desde a última dispensa. Maiores valores indicam melhores taxas de cobertura e esse valor reduz-se a 0 ao passar do tempo.
- **ad_self**: medida de 0 a 1 que indica o autorrelato do paciente para aderência ao tratamento, fornecido na avaliação clínica. Valores mais próximos de 1 indicam que o paciente se vê como aderente ao tratamento, enquanto valores próximos de 0 indicam que o paciente não se vê como aderente.
- **pillcount**: pontuação acima de 0 referente à contagem de pílulas no momento da coleta de sangue.

Por meio da análise de curvas ROC, busca-se compreender como essas covariáveis podem facilitar a classificação de aderência de um paciente sem necessariamente ter sido feita a coleta de sangue. Para efeitos de comparação, será considerada a abordagem que assume independência, tomando-se somente a primeira observação de cada paciente, e o modelo Bayesiano de verossimilhança composta proposto. Em uma análise prévia realizada com o modelo de efeitos mistos, foram encontradas AUC destoantes – usualmente acima de 0,9 – que sugeriram a ocorrência de sobreajuste do modelo. Considerando que os dados possuem características semelhantes aos cenários em que foram observados maiores problemas em ambientes simulados, desconfia-se de que esse não seja um resultado ilusório que provém por questões de ajuste no modelo.

Em relação às variáveis explicativas, as seguintes combinações foram utilizadas:

- utilizando somente **mpr** de covariável. A base, neste caso, consta com 294 observações válidas.
- utilizando somente **ad_self** de covariável, com 274 observações válidas.

- utilizando **pillcount** de covariável, com 104 observações válidas.
- utilizar **mpr** e **ad_self** de covariáveis – **pillcount** foi eliminada devido ao baixo número de observações válidas –, em que 270 observações são válidas.

Seguem, abaixo, tabelas com algumas estatísticas encontradas nas abordagens realizadas:

Tabela 7: Tabela com estatísticas referentes às análises de curvas ROC realizadas em dados reais, comparando o modelo proposto com a abordagem proposta de *Obuchowski*.

Covariáveis	AUC			Sensibilidade			Especificidade		
	Obuchowski	Composição	Incremento (%)	*Obuchowski*	Composição	Incremento (%)	*Obuchowski*	Composição	Incremento (%)
MPR	0,5913	0,6876	16,29	0,7652	0,8261	7,96	0,4637	0,514	10,11
ad_self	0,7493	0,7804	4,15	0,7818	0,8	2,33	0,6463	0,6707	3,78
pillcount	0,6911	0,7185	3,96	0,6863	0,6863	0	0,6038	0,6226	3,11
MPR + ad_self	-	0,7847	-	-	0,8	-	-	0,675	-

Tabela 8: Tabela com estatísticas referentes às análises de curvas ROC realizadas em dados reais, comparando tomar a primeira observação de cada *cluster* com o modelo proposto.

Covariáveis	AUC			Sensibilidade			Especificidade		
	1º Observação	Composição	Incremento (%)	1º Observação	Composição	Incremento (%)	1º Observação	Composição	Incremento (%)
MPR	0,5499	0,6876	25,04	0,6842	0,8261	20,74	0,4312	0,514	19,20
ad_self	0,7231	0,7804	7,92	0,7671	0,8	4,29	0,5941	0,6707	12,89
pillcount	0,6912	0,7185	3,95	0,5294	0,6863	29,64	0,7941	0,6226	-21,60
MPR + ad_self	0,7422	0,7847	5,73	0,7397	0,8	8,15	0,6465	0,675	4,41

A abordagem de *Obuchowski* não permite calcular a AUC para o cenário de **MPR + ad_self** pois consiste da construção de uma curva ROC a partir de dois diferentes fatores, requerendo a redução de dimensionalidade por meio de um modelo estatístico, por exemplo. Por questão de espaço, o modelo Bayesiano proposto foi referido somente aqui como "Composição". Percebe-se aqui também que a metodologia proposta mantém a estrutura dos dados e acomoda bem a estrutura de dependência, tendo relativo ganho na AUC e nas medidas de performance a utilizar um modelo estatístico. Percebe-se também que considerar a de estrutura de dependência, seja com a abordagem de Obuchowski (1997) ou com o modelo proposto, resulta em maiores AUC estimadas. Devido à baixa quantidade de informações válidas para a variável *pillcount*, não foi avaliada a combinação dessa variável com nenhuma outra.

Nesse caso estudado, a combinação de covariáveis explicativas que melhor discriminou a população entre aderentes e não-aderentes é combinar a medida **MPR** com o autorrelato do

paciente. Tem-se um teste de diagnóstico de aderência com boa à excelente capacidade discriminatória dado seu alto valor da AUC e o ponto de corte ótimo na probabilidade ajustada fornece sensibilidade e especificidade iguais a 80% e $67{,}5\%$, respectivamente. Ou seja, 80% dos pacientes que realmente estão aderentes serão detectados pelo teste como aderentes e $67{,}5\%$ dos pacientes que não estão em adesão correta ao medicamento serão identificados pelo modelo como não-aderentes.

Aqui, a principal motivação é de que a clínica seja capaz de diagnosticar pacientes que não possuem alta adesão ao tratamento, então podem ser feitas modificações na seleção do ponto de corte ótimo tal que a sensibilidade seja reduzida, mas que se obtenha uma maior especificidade. Permite-se assim que os responsáveis pelo estudo sejam capazes de intervirem de forma adequada quando identificado que um paciente não está aderente ao tratamento, com métricas que podem ser obtidas sem que o paciente compareça necessariamente ao tratamento. Por exemplo, Pasipanodya et al. (2018) avalia que o envio de mensagens de texto por celular, como SMS, é uma intervenção de baixo custo e escalável que pode ser tomada e que possui evidências de melhorar a adesão à aderência de PrEP.

Para o modelo logístico Bayesiano de verossimilhança composta, foram consideradas 26000 iterações do Monte-Carlo Hamiltoniano, sendo essas 2000 de aquecimento. Para verificar a convergência do algoritmo, foram amostradas 4 cadeias simultaneamente. A Figura 13 disponibiliza as iterações dos algoritmos para o vetor de coeficientes $\underset{\sim}{\boldsymbol{\beta}} = [\text{intercepto}, \beta_1, \beta_2]^T$ (sendo β_1 um coeficiente referente ao MPR e β_2 referente ao autorrelato de adesão do paciente), a precisão dos efeitos individuais τ e também dois efeitos individuais – do vetor $\underset{\sim}{\boldsymbol{\alpha}}$ – que foram sorteados para visualização. Será disponibilizada uma tabela contendo algumas estatísticas como intervalo de credibilidade de 95%, média, mediana e desvio-padrão.

Tabela 9: Sumário com estatísticas referentes ao ajuste do modelo logístico Bayesiano de verossimilhança composta para o vetor de parâmetros $\underset{\sim}{\boldsymbol{\beta}}$, τ e dois efeitos individuais, α_4 e α_{53}

Variável	**Média**	**Percentil 2,5%**	**Mediana**	**Percentil 97,5%**	**Desvio-Padrão**
Intercepto	-6,86	-12,9	-6,57	-2,42	2,68
β_1	-0,25	-4	-0,3	3,72	1,94
β_2	0,08	0,02	0,07	0,15	0,03
τ	0,98	0,12	0,71	3,33	0,88
α_4	0,02	-3	0,01	2,91	1,42
α_{53}	1,28	-0,69	1,07	4,31	1,28

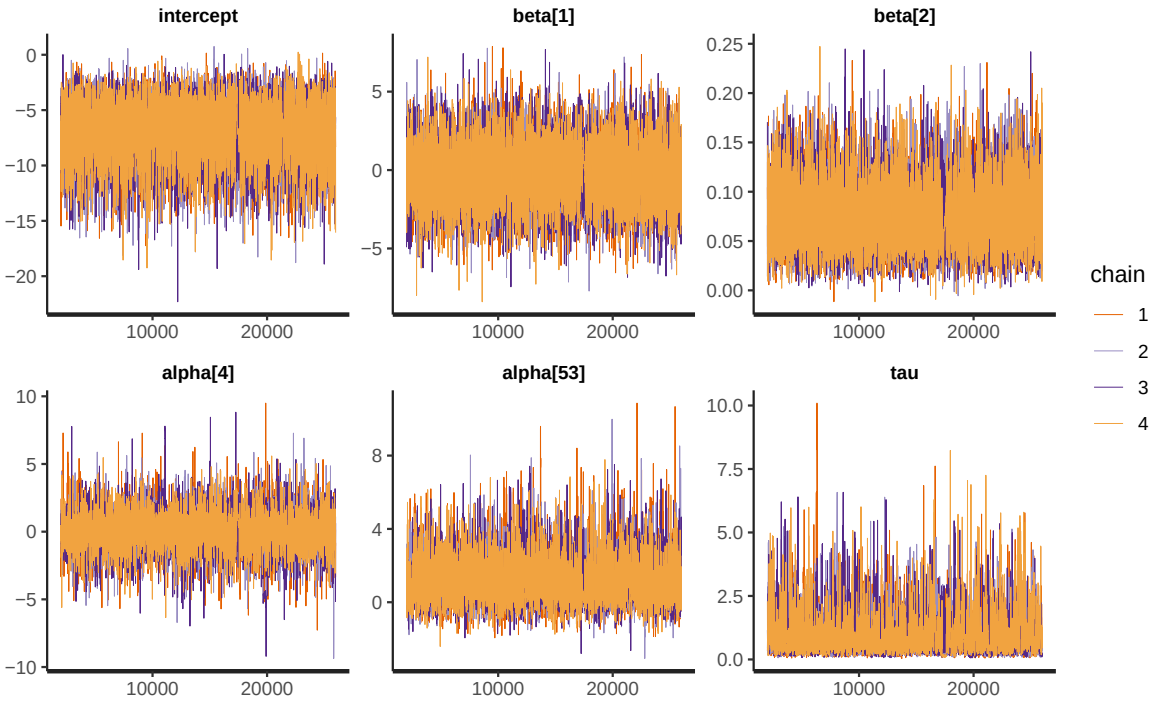

Figura 13: Cadeias do algoritmo Monte-Carlo Hamiltoniano para o ajuste do modelo logístico Bayesiano de verossimilhança composta nos dados do projeto PrEP1519.

Percebe-se que as cadeias claramente convergem para valores de centralidade em cada um dos parâmetros. O parâmetro τ, por ter restrição de valores acima de 0, apresenta um comportamento um pouco menos simétrico que dos demais parâmetros, que possuem suporte em todos reais. O número número de iterações utilizadas no algoritmo foi bem mais alto que o desejado e por conta disso decidiu-se manter um valor a cada 6 iterações, de forma também a eliminar qualquer autocorrelação entre valores das cadeias. Assim, é totalizado 16000 iterações, sendo 4000 para cada uma das 4 cadeias. A Tabela 9 é referente aos valores já espaçados da cadeia.

Se for utilizada como estimativa pontual a média das cadeias de cada um dos parâmetros (e também por ter utilizada a função de ligação logito), a interpretação das estimativas pontuais seria a seguinte:

- **MPR**: cada unidade adicional na medida MPR de um paciente esteve associado, em média, com um decréscimo de 22% na probabilidade de ser aderente ao tratamento;
- **Autorrelato de Adesão**: cada unidade adicional no autorrelato de aderência do paciente esteve associado, em média, com um acréscimo de 8,3% na probabilidade de ser aderente ao tratamento de PrEP.

- O paciente 4 possui um efeito que aumenta em 2% sua probabilidade de ser aderente.
- O paciente 54 possui um efeito individual que aumenta em 260% sua probabilidade de ser aderente ao tratamento, apontando para uma característica de proteção.

O MPR apresentou efeito contrário ao esperado, visto que é uma medida construída a partir da aderência teórica do paciente. Esperava-se, inicialmente, que este estivesse associado com um acréscimo na probabilidade de ser aderente ao tratamento. Isso pode ser efeito de variáveis ausentes na base de dados, atribuindo um excesso de variabilidade à variável MPR e alterando o seu efeito explicativo. Também, é importante notar aqui que a variável MPR tem média e máximo de 0,9 e 2, respectivamente, enquanto que a variável autorrelato de adesão tem média 70 e máximo de 100. Logo, no máximo, um paciente pode ter 44% de decréscimo na probabilidade de ser aderente ao tratamento devido ao MPR, o que é balanceado próximo de 0 com aproximadamente 5 pontos no autorrelato de adesão.

Por fim, o melhor ponto de corte na probabilidade estimada é 0,44, ou seja, classificando todo paciente que possua probabilidade estimada acima de 0,44 como aderente e, caso contrário, não-aderente, garantimos uma sensibilidade de 80% na classificação e uma especificidade de 67,5%.

4 Conclusões

De forma a acomodar a estrutura de dependência entre observações numa base com medidas repetidas irregulares, foram avaliadas diversas metodologias para análise de curvas ROC e proposto um modelo Bayesiano de verossimilhança composta. Tratando-se do modelo proposto, foram realizados estudos de simulação para avaliar a recuperação dos parâmetros deste modelo e também em relação à sensibilidade da especificação da distribuição *a priori*. A princípio, o modelo aparentou ter uma pior recuperação dos parâmetros quando foi considerado uma distribuição *a priori* vaga para o hiperparamêtro da precisão dos efeitos aleatórios, então optou-se por utilizar uma distribuição *a priori* mais informativa para este parâmetro. Em relação à variância dos efeitos fixos das covariáveis, estes não parecem ser sensíveis à escolha da *priori*, apresentando boa capacidade de recuperação para os cenários avaliados.

Em estudos de simulação desenvolvidos também ficou evidenciado o sobreajuste do modelo de efeitos mistos, proposto em Liu e Wu (2003), quando os dados sofrem de grandes proporções de *singletons*, em que os interceptos aleatórios estimados afetam a capacidade preditiva do modelo. Quando ajustado o modelo proposto para os mesmos dados simulados, verificou-se que este não estava sujeito ao sobreajuste. Portanto, percebe-se a importância do desenvolvimento de uma metodologia para análise de curvas ROC em dados de medidas repetidas desbalanceados. Ser capaz de lidar com a estrutura de dependência em cenários como este é importante para se extrair o máximo de informações possíveis em uma análise de curvas ROC.

Notou-se que, em geral, a presença de *singletons* não é algo tão preocupante quanto inicialmente parecia ser: desde que sua prevalência seja moderada, não há viés considerável no ajuste dos modelos aqui vistos. Para cenários moderadamente desbalanceados, o modelo de efeitos mistos apresentou viés praticamente nulo, enquanto que o modelo Bayesiano de verossimilhança composta parece ter melhor performance em casos mais extremos. Não foi identificada dificuldade de o modelo proposto recuperar os efeitos da covariável e nem o intercepto. A principal desvantagem do modelo de efeitos mistos está no fato de que este é sujeito ao sobreajuste, implicando a construção de uma variável classificadora que possui a performance de um teste de diagnóstico padrão-ouro quando essa não é a realidade. Além disso, mantendo uma abordagem

parcimoniosa, não há ganhos de informação ao considerar interceptos aleatórios para *clusters* que possuem apenas uma observação na base de dados, visto que a suposição do modelo é de que os *clusters* são independentes entre si e só existirá uma informação disponível na base de dados para a estimação deste intercepto aleatório.

O desenvolvimento teórico feito aqui é um passo pequeno passo que traz benefícios significativos em algumas situações de análises de curvas ROC, principalmente naquelas em que não se deseja remover dados da análise ou em que o modelo de efeitos mistos sugere um sobreajuste, fornecendo estatísticas ilusórias. Como principais pontos a serem investigados no futuro deste trabalho está sua extensão para metodologias de análises longitudinais, visto que a abordagem de verossimilhança composta garante que a observação anterior sempre existe para observações que possuem estrutura de dependência. Além disso, o grande viés observado do modelo proposto em relação ao modelo de efeitos mistos em cenários com grandes proporções de *singletons* é também um ponto identificado que necessitava de maior investigação e elaboração. Uma outra abordagem possível é utilizar Modelos de Equações de Estimativa Generalizadas – modelos GEE. Como uma principal desvantagem deste modelo, e também justificativa do porquê não ter sido considerado aqui, está na preocupação demonstrada no estudo de Liang, Zeger e Qaqish (1992) quando os *clusters* apresentam diferentes números de observações.

Referências

AUSTIN, P. C.; LECKIE, G. The effect of number of clusters and cluster size on statistical power and type i error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of statistical computation and simulation*, Taylor & Francis, v. 88, n. 16, p. 3151–3163, 2018.

BAMBER, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, Elsevier, v. 12, n. 4, p. 387–415, 1975.

BANOO, S. et al. Evaluation of diagnostic tests for infectious diseases: general principles. *Nature Reviews Microbiology*, Nature Publishing Group UK London, v. 5, n. Suppl 11, p. S21–S31, 2007.

BARNDORFF-NIELSEN, O. *Information and exponential families: in statistical theory*. [S.l.]: John Wiley & Sons, 2014.

BATES, D. et al. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, v. 67, n. 1, p. 1–48, 2015.

BELL, B. A.; FERRON, J. M.; KROMREY, J. D. Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *JSM proceedings, section on survey research methods*, American Statistical Association Alexandria, VA, p. 1122–1129, 2008.

BETANCOURT, M. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

CARDOSO, J. R. et al. What is gold standard and what is ground truth? *Dental press journal of orthodontics*, SciELO Brasil, v. 19, p. 27–30, 2014.

CASTILLO-MANCILLA, J. R. et al. Tenofovir diphosphate in dried blood spots is strongly associated with viral suppression in individuals with human immunodeficiency virus infections. *Clinical Infectious Diseases*, Oxford University Press US, v. 68, n. 8, p. 1335–1342, 2019.

CLARKE, P. When can group level clustering be ignored? multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 62, n. 8, p. 752–758, 2008.

CLARKE, P.; WHEATON, B. Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods. *Sociological methods & research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 35, n. 3, p. 311–351, 2007.

EGAN, J. P. *Signal detection theory and ROC-analysis*. [S.l.]: Academic press, 1975.

EVANS, W. E.; JOHNSON, J. A. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annual review of genomics and human genetics*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 2, n. 1, p. 9–39, 2001.

FOULKES, A. S. et al. Prediction based classification for longitudinal biomarkers. *The annals of applied statistics*, NIH Public Access, v. 4, n. 3, p. 1476, 2010.

GUILLAUME, J. H. et al. Introductory overview of identifiability analysis: A guide to evaluating whether you have the right type of data for your modeling purpose. *Environmental Modelling & Software*, Elsevier, v. 119, p. 418–432, 2019.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398.

KYLE, U. G.; GENTON, L.; PICHARD, C. Low phase angle determined by bioelectrical impedance analysis is associated with malnutrition and nutritional risk at hospital admission. *Clinical nutrition*, Elsevier, v. 32, n. 2, p. 294–299, 2013.

LIANG, K.-Y.; ZEGER, S. L.; QAQISH, B. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 54, n. 1, p. 3–24, 1992.

LIU, D.; ALBERT, P. S. Combination of longitudinal biomarkers in predicting binary events. *Biostatistics*, Oxford University Press, v. 15, n. 4, p. 706–718, 2014.

LIU, H. et al. Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. *Journal of Data Science*, v. 3, n. 3, p. 257–278, 2005.

LIU, H.; WU, T. Estimating the area under a receiver operating characteristic curve for repeated measures design. *Journal of Statistical Software*, v. 8, p. 1–18, 2003.

LUSTED, L. B. Signal detectability and medical decision-making: Signal detectability studies help radiologists evaluate equipment systems and performance of assistants. *Science*, American Association for the Advancement of Science, v. 171, n. 3977, p. 1217–1219, 1971.

MAAS, C. J.; HOX, J. J. Sufficient sample sizes for multilevel modeling. *Methodology*, Hogrefe & Huber Publishers, v. 1, n. 3, p. 86–92, 2005.

MATTHEWS, L. T. et al. Protocol for a longitudinal study to evaluate the use of tenofovir-based prep for safer conception and pregnancy among women in south africa. *BMJ open*, British Medical Journal Publishing Group, v. 9, n. 7, p. e027227, 2019.

MICHAEL, H.; TIAN, L.; GHEBREMICHAEL, M. The roc curve for regularly measured longitudinal biomarkers. *Biostatistics*, Oxford University Press, v. 20, n. 3, p. 433–451, 2019.

MOFENSON, L. M.; BAGGALEY, R. C.; MAMELETZIS, I. Tenofovir disoproxil fumarate safety for women and their infants during pregnancy and breastfeeding. *Aids*, Wolters Kluwer, v. 31, n. 2, p. 213–232, 2017.

MORSE, L. R. et al. Barriers to providing dual energy x-ray absorptiometry services to individuals with spinal cord injury. *American journal of physical medicine & rehabilitation/Association of Academic Physiatrists*, NIH Public Access, v. 88, n. 1, p. 57, 2009.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

OBUCHOWSKI, N. A. Nonparametric analysis of clustered roc curve data. *Biometrics*, JSTOR, p. 567–578, 1997.

PASIPANODYA, E. C. et al. Trajectories and predictors of longitudinal preexposure prophylaxis adherence among men who have sex with men. *The Journal of infectious diseases*, Oxford University Press US, v. 218, n. 10, p. 1551–1559, 2018.

PEPE, M. S. *The statistical evaluation of medical tests for classification and prediction*. [S.l.]: Oxford University Press, USA, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>.

SPACKMAN, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. In: ELSEVIER. *Proceedings of the sixth international workshop on Machine learning*. [S.l.], 1989. p. 160–163.

Stan Development Team. *RStan: the R interface to Stan*. 2023. R package version 2.32.3. Disponível em: <https://mc-stan.org/>.

SWETS, J. A. Measuring the accuracy of diagnostic systems. *Science*, American Association for the Advancement of Science, v. 240, n. 4857, p. 1285–1293, 1988.

VARIN, C.; REID, N.; FIRTH, D. An overview of composite likelihood methods. *Statistica Sinica*, JSTOR, p. 5–42, 2011.

ZHOU, X.-H.; MCCLISH, D. K.; OBUCHOWSKI, N. A. *Statistical methods in diagnostic medicine*. [S.l.]: John Wiley & Sons, 2009.

ZOU, K. H. Receiver operating characteristic (roc) literature research. *On-line bibliography available from:< http://splweb. bwh. harvard. edu*, v. 8000, p. 172, 2002.