Universidade do Minho
Escola de Engenharia

Luís Nuno dos Santos Moncaixa

Regressão logística para dados correlacionados (GEE): Desenvolvimento de aplicações dinâmicas em R

November 2022

Universidade do Minho
Escola de Engenharia

Luís Nuno dos Santos Moncaixa

Regressão logística para dados correlacionados (GEE): Desenvolvimento de aplicações dinâmicas em R

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob a supervisão de:
Professora Doutora Ana Cristina da Silva Braga

November 2022

DIREITOS DE AUTOR E TERMOS DE USO PARA TRABALHO DE TERCEIROS

Esta dissertação descreve um trabalho académico que pode ser utilizado por terceiros desde que sejam respeitadas as normas e boas práticas internacionalmente aceites em matéria de direitos de autor.

Este trabalho pode, posteriormente, ser utilizado nos termos estabelecidos na licença abaixo.

Os leitores que necessitem de autorização não prevista no licenciamento indicado devem contactar o autor através do RepositóriUM da Universidade do Minho.

LICENÇA CONCEDIDA AOS USUÁRIOS DESTA OBRA:

CC BY

https://creativecommons.org/licenses/by/4.0/

AGRADECIMENTOS

Com a conclusão deste trabalho, chega assim, o fim da minha formação académica.
Inicialmente, gostava de agradecer a toda a minha família. Em especial, pai, irmão e avós. Foram incontestávelmente um pilar fundamental durante todo este trajeto. Agradeço pelo carinho, apoio e confiança depositada em mim durante todo este tempo. Sou eternamente grato e vocês sabem o quanto significam para mim.
Agradecer imenso à minha orientadora, Professora Doutora Ana Cristina Braga, por todo o incondicional apoio que me deu durante o desenvolvimento desta dissertação. Toda a paciência, ajuda, disponibilidade e por vezes, repreensões necessárias foram, inquestionávelmente, importantes para todo este trabalho desenvolvido. Desejo-lhe as maiores felicidades e sucessos a todos os níveis.
Aos meus amigos por toda a motivação e entreajuda, que foram, sem dúvida, importantes para cumprir todos os objetivos traçados.
Finalmente, quero dedicar esta dissertação à minha falecida mãe, Cidália Moncaixa, por todo o amor e educação que me deu, que me permitiu caminhar todos os dias na direção correta. Dedico-te a ti minha mãe, és a minha estrela guia e espero que estejas extremamente orgulhosa.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter realizado este trabalho académico com integridade.

Confirmo que não utilizei plágio ou qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Declaro ainda que tomei pleno conhecimento do Código de Conduta Ética da Universidade do Minho.

RESUMO

Os modelos de regressão logística procuram identificar a influência de diferentes variáveis/fatores numa variável resposta de interesse. Estes são normalmente utilizados na àrea da medicina pois permite verificar quais os fatores que influênciam a presença de determinadas patologias.

Os estudos longitudinais, onde se verifica uma repetição de observações ou medições registadas ao longo do tempo, são os mais utilizados para aplicação dos modelos de regressão logística, no entanto grande parte deste modelos não consideram a correlação entre as variáveis em estudo. De modo a ultrapassar este problema foram desenvolvidos os modelos GEE (*Generalized Estimating Equations*) que consideram a correlação existente nos dados, possibilitando assim uma análise mais rigorosa da influência de diferentes factores.

Esta dissertação procura identificar e explorar diferentes *packages* em *R* para aplicação dos modelos GEE através da sua aplicação a um caso de estudo.

Ao longo desta dissertação será desenvolvida uma nova aplicação *web* designada SAGA, desenvolvida utilizando o package *Shiny* na linguagem *R*. Esta aplicação encontra-se disponível no seguinte link: `https://geemodelapp2022.shinyapps.io/Shiny_App/`.

A aplicação SAGA tem como principal objetivo a análise de modelos GEE utilizando um conjunto de dados selecionado pelo utilizador, sendo possível identificar as diferentes variáveis de interesses que serão descritas ao longo da dissertação, bem como validar os modelos desenvolvidos através da validação por curvas *ROC*. Para além dos resultados dos modelos GEE, demonstrados na aplicação, também estão representadas as curvas *ROC* de cada modelo desenvolvido.

PALAVRAS-CHAVE regressão logística, dados correlacionados, *GEE*, *shiny*, *R*, SAGA.

ABSTRACT

Logistic regression models seek to identify the influence of different variables/factors on a response variable of interest. These are normally used in the field of medicine as it allows to verify which factors influence the presence of certain pathologies.

Longitudinal studies, where there is a repetition of observations or measurements recorded over time, are the most used to apply logistic regression models, however most of these models do not consider the correlation between the variables under study. In order to overcome this problem, GEE (Generalized Estimating Equations) models were developed, which consider the existing correlation in the data, resulting in a more rigorous analysis of the influence of different factors.

This dissertation, in a first phase, explores the different existing *R* packages for the application of GEE models, in order to identify differences or similarities between them using a case study for the application of GEE models.

Throughout this dissertation, a new web application called SAGA will be developed, the application will be developed using the *Shiny* package in *R* language. This application is available at the following link: `https://geemodelapp2022.shinyapps.io/Shiny_App/`.

The main purpose of the SAGA application is to develop and analyse GEE models using a dataset selected by the user, where it will be possible to describe all the variables of interest in the development of the model as will be described in the course of this dissertation, as well as to validate the models developed through validation by *ROC* analysis. In addition to the results of the GEE models, shown in the application, the *ROC* curves of each model developed are also represented.

KEYWORDS logistic regression, correlated data, *GEE*, *shiny*, *R*, SAGA.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

1

INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

A utilização de métodos estatísticos e a sua pesquisa associada na investigação humana, especialmente na área da Medicina, tem demonstrado nas últimas décadas ser um elemento fundamental para o desenvolvimento de tratamentos a doenças como para a investigação científica Fitzmaurice et al. (2012).

Os dados longitudinais indicam-nos repetidas observações ao longo do tempo para a mesma amostra/indivíduo, sendo estas observações, à priori, mais idênticas entre um mesmo individuo do que em diferentes indivíduos, indicando assim que estas observações estarão correlacionadas Hedeker and Gibbons (2006).

Este tipo de dados, tendo em conta a correlação existente entre eles, representa uma dificuldade acrescida quando aplicados modelos de regressão para análise, uma vez que uma das principais suposições destes modelos baseia-se na independência das observações, ou seja, a correlação é inexistente. Várias análises biomédicas de dados longitudinais demonstram que a utilização de modelos de regressão desconsiderando a correlação das observações resulta em uma estimativa de erros padrão pouco precisa, influenciando todos os indicadores, tais como intervalos de confiança ou valores p, que permitem efetuar uma boa análise e assim levar a inferências incorretas Kleinbaum and Klein (2002).

Os modelos GEE (*Generalized Estimating Equations*), desenvolvidos por Liang and Zeger (1986), representam uma classe de modelos que são frequentemente utilizados para dados em que as respostas estão correlacionadas. Estes modelos são uma extensão dos GLMs (*Generalized Linear Models*) pois permitem analisar dados longitudinais bem como assumir a correlação existente entre os dados. As GEE poderão ser utilizadas em duas análises distintas, uma análise considerando o individuo ou uma análise considerando uma população. Diferindo dos outros modelos de regressão, os modelos GEE não necessitam de uma distribuição dos dados definida, por sua vez estes necessitam que seja atribuida uma estrutura de correlação dos dados.

Os modelos GEE representam uma ferramenta importante no avanço da investigação biomédica uma vez que permite analisar o impacto de diversas variáveis em individuos singularmente ao longo do tempo, como também em uma população tendo em conta as observações de cada individuo Wang et al. (2022).

Por sua vez, a ferramenta *Shiny* do *R*, permite o desenvolvimento de aplicações dinâmicas que facilita a utilização e análise dos modelos GEE para estudos biomédicos, uma vez que permitie ao utilizador, sem conhecimento prévio de programação ou digitação de código, aplicar estes modelos de modo a obter inferências

estatísticas que lhe permitam realizar uma análise sucinta e válida, quer por via de testes estatisticos quer de representações gráficas.

Face ao exposto levantou-se a seguinte pergunta de investigação:

> Quais as ferramentas disponíveis em R que permitem efetuar a análise utilizando modelos GEE?

1.2 OBJETIVOS

Tendo em conta a questão de investigação enunciada na secção 1.1 delineou-se como principal objetivo desta dissertação efetuar um estudo sobre a metodologia e aplicabilidade dos modelos GEE, explorando os *packages* existentes no *R* de forma a desenvolver uma aplicação *Shiny*. Com base neste objetivo geral foram delineados os seguintes objetivos específicos:

- Estudar e reconhecer os fundamentos importantes, através de literatura relevante, sobre a metodologia dos modelos GEE e a sua aplicação;
- Explorar os diferentes packages existentes em R para aplicação de modelos GEE;
- Desenvolver uma aplicação dinâmica para *web*, utilizando o *Shiny*, que permita:
  - Aplicar os modelos GEE para diferentes tipos de varíaveis;
  - Descrever o modelo utilizado;
  - Identificar a relação dos diferentes parâmetros com a variável resposta;
  - Gerar inferências estatísticas através de intervalos de confiança e testes estastísticos, que podem ser descarregados pelo utilizador;
  - Complementação gráfica de modo a facilitar a análise.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

A estrutura deste documento encontra-se organizada em sete capitulos.

Neste primeiro capítulo é apresentada a parte introdutória ao tema da dissertação apresentando uma breve contextualização e motivação. É também apresentada uma questão de investigação onde elencam os objetivos definidos.

O segundo capítulo corresponde ao estado da arte onde se apresentam os principais conceitos e definições dos modelos GEE, através de revisão bibliográfica. Serão aqui, também abordados alguns algoritmos/*packages* estatisticos de aplicação dos modelos GEE.

No terceiro capítulo está representada a metedologia e abordagens a utilizar no desenvolvimento da aplicação *Shiny* e no desenvolvimento de modelos GEE utilizando *packages* existente no *R*, assim como uma introdução a um caso de estudo que será utilizado para o desenvolvimento dos modelos GEE.

O quarto capítulo descreve todos os processos realizados para desenvolvimento e validação dos modelos GEE, utilizando o caso de estudo apresentado como base de dados.

No quinto capítulo é descrita a aplicação web desenvolvida, com destaque para todas as suas funcionalidades devidamente explicadas, tendo como base os dados provenientes do caso de estudo.

No sexto capítulo estão descritos os resultados obtidos bem como a sua correspondente análise. Numa primeira fase, serão descritos e analisados os resultados dos modelos desenvolvidos bem como a comparação entre os packages utilizados. Os resultados da aplicação web desenvolvida serão comparados com os resultados obtidos pelo desenvolvimento dos packages na fase anterior.

No sétimo capítulo encontram-se representadas as principais conclusões sobre o trabalho desenvolvido bem como uma pequena análise sobre o aproveitamento do mesmo. Neste capítulo encontram-se também sugestões de trabalhos futuros que podem complementar o trabalho desenvolvido.

2

METEDOLOGIA DOS MODELOS GEE

2.1 FUNDAMENTOS SUBJACENTES AOS MODELOS GEE

2.1.1 *GLM (Generalized Linear Models)*

Os modelos lineares generalizados, GLM foram introduzidos por McCullagh and Nelder (1989), sendo esta uma extensão dos modelos clássicos de regressão linear e não linear. Esta classe de modelos lineares permite analisar dados onde as variáveis de resposta assumem uma distribuição normal, como os modelos de regressão linear, mas também dados onde as variáveis de resposta poderão ser discretas ou binárias Myers et al. (2012); Kleinbaum and Klein (2002).

A classe de GLM incluem vários modelos de regressão que apenas diferem na distribuição da variável dependente. Estas distribuições encontram-se agrupadas numa superfamília designada de família exponencial, onde se incluem por exemplo a distribuição binomial, normal, Poisson, exponencial entre outras, ou seja, uma regressão logística normalmente aplica-se quando os dados se encontram numa distribuição binomial, sendo a regressão um dos modelos integrantes dos modelos GLM. Hedeker and Gibbons (2006).

Citado por vários autores (e.g Hardin et al. (2007); Agresti (2015); Dobson and Barnett (2018) as GLMs contêm 3 principais componentes:

- **Componente aleatório** - Consiste na variável resposta (Y), com n observações independentes, que deve de seguir uma distribuição de probabilidade pertencente à família exponencial. Por exemplo se uma variável de resposta seguir uma distribuição de normal, será utilizada uma regressão linear.
- **Componente sistemático/Preditor Linear** - Componente resultante do produto dos conjuntos de parâmetros $\boldsymbol{\beta}$ com a matriz de variáveis independentes $\boldsymbol{X}$, correspondente às observações para cada parâmetro tal como:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} \tag{1}$$

- **Função de Ligação** - Função de ligação que identifica uma função da média, aplicada aos valores ajustados, relacionando-os com o preditor linear. Esta função depende do tipo de variável resposta, por exemplo, se a variável resposta for dicotómica, a função de ligação relacionada será a função logit.

$$g(\mu) = \boldsymbol{X}\boldsymbol{\beta} \tag{2}$$

Para estimar os parâmetros do modelo, os GLM utilizam normalmente o método de máxima verossimilhança ou função de log-verossimilhança, este método procura estimar os valores dos diferentes parâmetros de modo a maximizar a função de verossimilhança. Este método é eficiente quando os dados são independentes, no entanto este necessita da forma de distribuição presente nos dados Kleinbaum and Klein (2002).

A utilização dos GLM seguem o pressuposto de que todas as observações são independentes entre si, pelo que são insuficientes para a análise de dados longitudinais ou dados onde existe correlação entre os dados. Assim, para colmatar esse problema foram desenvolvidos os modelos GEE como uma extensão dos GLM que serão discutidos na secção 2.2.

2.1.2 *Dados longitudinais*

Dados longitudinais são muito utilizados em áreas como a epidemiologia, a pesquisa clínica e a avaliação de terapias ou fármacos, pois procuram identificar a influencia de determinados parâmetros numa resposta obtida por um sujeito, esses parâmetros são regularmente medidos por um período de tempo, sendo alguns destes constantes e outros variáveis ao longo do tempo. A análise de estudos longitudinais pretende, assim, estimar as mudanças que ocorrem em um individuo ao longo do tempo bem como analisar a influência dos fatores dependentes nessa mesma mudança Van Belle et al. (2004); Liang and Zeger (1993); Wang et al. (2022).

A estrutura de dados longitudinais consiste nos Z indivíduos pertencentes ao conjunto de dados, cada individuo terá κ medições ao longo do tempo sendo este número de medições independente de individuo para individuo, a variável tempo também está presente na estrutura uma vez de demonstra que a κ medição do sujeito Z acontece no tempo T com a variável resposta de Y. Uma vez mais a variável tempo é independente de individuo em individuo. Por fim, estão presentes as covariáveis β de interesse no estudo, sendo algumas destas constantes ao longo das medições (ex: sexo) e outras variáveis ao longo do tempo (ex: pressão sanguínea) Kleinbaum and Klein (2002); Hedeker and Gibbons (2006). Na Tabela 1 está representada uma estrutura de dados longitudinais.

A análise de estudos longitudinais apresenta uma grande vantagem que consiste na identificação de padrões de mudança entre indivíduos que posteriormente possibilita retirar conclusões acerca da influência desses mesmos padrões numa população. No entanto, os dados longitudinais apresentam alguns desafios, como por exemplo, a existência de dados omissos, ou seja, medições não realizadas a um ou mais indivíduos durante o estudo, estes dados omissos levam a uma análise mais complexa do estudo de forma a identificar as relações entre as covariáveis e a resposta do indivíduo Hedeker and Gibbons (2006); Dobson and Barnett (2018); Ware (1985).

A correlação das repetidas medições de um sujeito surge como o maior desafio nos estudos longitudinais, uma vez, que os métodos estatísticos para análise destes estudos, normalmente, assumem a independência

Tabela 1: Exemplo de estrutura de dados longitudinais

Indivíduo (Z)	Medição (κ)	Tempo (T)	Resposta (Y)	Covariáveis (β)
1	1	T_{11}	y_{11}	$\beta_{111} \dots \beta_{11x}$
1	2	T_{12}	y_{12}	$\beta_{121} \dots \beta_{12x}$
1	κ	$T_{1\kappa}$	$y_{1\kappa}$	$\beta_{1\kappa1} \dots \beta_{1\kappa x}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Z	κ	$T_{Z\kappa}$	$y_{Z\kappa}$	$\beta_{Z\kappa1} \dots \beta_{Z\kappa x}$

dos dados, ou seja, a correlação é ignorada o que resulta em inferências estatísticas inválidas. Vários estudos comparam a utilização de modelos de regressão quando assumida a correlação dos dados e quando ignorada, revelando diferenças nos resultados obtidos, destacando a importância da correlação na análise de dados longitudinais Glynn and Rosner (2012); Kleinbaum and Klein (2002).

2.2 MODELOS GEE

Os modelos GEE foram desenvolvidos por Liang and Zeger (1986) como uma extensão dos GLM, uma vez que permitem assumir a correlação existente nos dados longitunais em análise Wilson and Lorenz (2015). Ao contrário dos GLM que são baseados na teoria de máxima verosimilhança, a estrutura dos modelos GEE está relacionada com a teoria de *quasi-likelihood* pelo que não necessita de especificar a distribuição do dados, necessitando apenas de especificar a relação entre a variável de resposta e as covariáveis, bem como a relação entre a média e a variância das respostas através de uma estrutura de correlação Christopher (2001); Wang et al. (2022).

2.2.1 *Especificações dos modelos GEE*

Como descrito anteriormente, os modelos GEE são uma extensão das GLM e como tal estes incluem semelhantes especificações, no entanto para os modelos GEE é inserida a estrutura de correlação que descreve a correlação existente nos dados.

Tal como os GLM, nos modelos GEE inicialmente terá que ser definido o preditor linear. Podendo este ser representado por:

$$\boldsymbol{\eta_{it}} = \boldsymbol{X}_{it}\boldsymbol{\beta} \tag{3}$$

onde $\boldsymbol{X_{ij}}$ representa o vetor de covariáveis do individuo i no tempo t.

A função de ligação é determinada consoante o tipo de variável resposta, no caso da variável resposta ser contínua será utilizada a função identidade, quando a variável resposta é binária utiliza-se a função logit,

$$g(\mu_{it}) = \boldsymbol{\eta_{it}} \tag{4}$$

Assim como nos modelos GLM, é definida uma função de variância que relaciona a média e a variância de cada resposta,

$$V(Y_{it}) = \phi * v(\mu_{it}) \tag{5}$$

sendo Y a resposta do individuo i no tempo t, ϕ é um fator que verifica se a variação das respostas é excessiva, se $\phi > 1$, ou baixa, se $\phi < 1$, este fator pode ser definido à priori ou estimado, por fim $v(\mu_{it})$ corresponde a uma função de variância conhecida, dependendo, tal como a função ligação, do tipo de distribuição da variável resposta.

De modo a realizar uma análise através de modelos GEE, para além das especificações demonstradas acima, existe uma última especificação que permite a este tipo de modelos de realizar análises assumindo a correlação dos dados, a estrutura de correlação.

2.2.2 *Estruturas de Correlação*

A partir dos modelos GEE Liang and Zeger (1986) demonstraram que é possível estimar parâmetros consistentes, mesmo quando a estrutura de correlação não é especificada, contudo a eficiência da estimativa será ligeiramente mais baixa. De forma a obter melhores estimativas dos parâmetros, que resultem em inferências estatísticas mais válidas, deve se especificar uma estrutura de correlação para o estudo a realizar. Nesta secção serão descritas algumas das estruturas mais utilizadas para os modelos GEE com base nos autores Ziegler and Vens (2010); Wilson and Lorenz (2015).

Estrutura de correlação independente

Uma estrutura de correlação independente indica que as observações repetidas entre o individuo são independentes, ou seja, não estão correlacionadas. A estrutura de correlação independente será então uma matriz identidade de valor 1 ao longo da diagonal e valor 0 nas restantes posições.

$$Corr(y_{it}, y_{ik}) = \begin{cases} 1, & \text{se } t = k \\ 0, & \text{se } t \neq k \end{cases} \tag{6}$$

Assumindo um conjunto de quatro observações num mesmo paciente, uma estrutura de correlação independente apresenta-se como:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Figura 1: Exemplo de estrutura de correlação independente. Fonte: (Klein, 2002)

Estrutura de correlação intercambiável

A estrutura de correlação intercambiável assume que quaisquer duas ou mais respostas dentro de um mesmo *cluster* (conjuntos de observações que partilhem caracteristicas entre si, como por exemplo, pacientes de uma região em comum, ou observações registada num paciente ao longo do tempo), têm a mesma correlação, sendo a ordem das resposta arbitrária. Quando utilizada esta estrutura na análise por GEE, apenas um parâmetro de correlação é estimado para cada *cluster*.

$$Corr(y_{it}, y_{ik}) = \begin{cases} 1, & \text{se } t = k \\ \rho, & \text{se } t \neq k \end{cases} \tag{7}$$

Assumindo um conjunto de quatro observações α para um mesmo *cluster*, uma estrutura de correlação intercambiável define-se como:

$$\begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

Figura 2: Exemplo de estrutura de correlação intercambiável. Fonte: (Klein, 2002)

Estrutura de correlação autoregressiva

A estrutura de correlação autoregressiva é aplicável a estudos onde a dependência de tempo é assumida, ou seja, a correlação entre as respostas depende do intervalo de tempo entre elas, sendo as respostas com menor intervalo de tempo mais correlacionadas. A estrutura de correlação autoregressiva mais utilizada em análises GEE é designada de AR1. A AR1 é semelhante à estrutura de correlação intercambiável uma vez que apenas possui um parâmetro de correlação assumindo a suposição que quaisquer duas respostas do mesmo sujeito é igual à correlação elevada a uma potência igual à diferença entre intervalos de tempo.

$$Corr(y_{it}, y_{ik}) = \begin{cases} 1, & \text{se } t = k \\ \rho^{|t-k|}, & \text{se } t \neq k \end{cases} \tag{8}$$

Assumindo um *cluster* com quatro observações em diferentes intervalos de tempo , uma estrutura de correlação autoregressiva define-se como:

$$\begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{bmatrix}$$

Figura 3: Exemplo de estrutura de correlação autoregressiva. Fonte: (Klein, 2002)

Estrutura de correlação não estruturada

Uma estrutura de correlação não estruturada difere de todas as outras descritas anteriormente uma vez que cada parâmetro de correlação é diferente em relação a outros, dentro do mesmo *cluster*, por isso, contrariamente às outras estruturas a sua ordem não é arbitrária. Esta estrutura segue uma regra geral, ou seja, para um conjunto de n observações, o número de parâmetros estimados devem de ser iguais a n(n-1)/2, no entanto poderá gerar problema se o número de parâmetros estimados for elevado, resultando num modelo instável e com resultados inválidos. Contudo quando analisados poucos parâmetros a utilização desta estrutura poderá enriquecer a análise uma vez que cada parâmetro permite retirar diferentes conclusões.

Assumindo um conjunto de quatro observações α e seguindo a regra geral de parâmetros, uma estrutura de correlação não estruturada com seis parâmetros estimados define-se como:

$$\begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{bmatrix}$$

Figura 4: Exemplo de estrutura de correlação autoregressiva. Fonte: (Klein, 2002)

Embora descritas apenas estas estruturas existem outras também importantes para a análise GEE, a escolha da estrutura de correlação fica ao critério do investigador, de forma a obter uma estrutura que melhor descreva a correlação existente no estudo.

2.2.3 *Seleção da estrutura de correlação*

Apesar de Liang and Zeger (1986) demonstrarem que os modelos GEE apresentam estimativas de parâmetros robustas, mesmo quando a estrutura de correlação não é especificada, quando introduzida corretamente pelo investigador esta permite uma maior eficiência na estimação de parâmetros.

Conforme sugerido por Horton and Lipsitz (1999), dependendo dos dados que serão utilizados na análise podem ser selecionadas diferentes estruturas de correlação. Este autor recomenda a utilização de uma matriz de correlação não estruturada caso o número de observações por *cluster* seja pequeno e que os dados estejam balanceados, ou seja, todos os *clusters* contém o mesmo número de observações. A utilização de uma estrutura de correlação autoregressiva é sugerida caso existam observações incorretas ou em falta e uma estrutura de correlação intercambiável caso a ordem das observações não seja relevante.

Por vezes os dados apresentam várias caracteristicas anteriormente apresentadas por Horton and Lipsitz (1999) e dificultam a escolha da estrutura de correlação que melhor se identifica com a natureza do dados, de forma a auxiliar na tomada de decisão da estrutura de correlação Pan (2001) desenvolveu um método designado *quasi-verossimilhança sob o critério do modelo de independência*, QIC.

O método de QIC desenvolvido por Pan (2001), é uma adaptação do método de *Akaike information criterion*, método utilizado para estimar a qualidade de diferentes modelos permitindo a seleção do melhor modelo, utilizado nos modelos GLM, uma vez que o método AIC é baseado nas propriedades da máxima verossimilhança e estas não se adequam aos modelos GEE, o método de QIC foi desenvolvido com base no método de quasi-verossimilhança desenvolvido por Wedderburn (1974) permitindo assim a sua aplicabilidade nos modelos GEE. O método de QIC define-se por,

$$QIC(R) = -2Q(\widehat{\beta}(R); I, D) + 2\mathrm{Tr}(\widehat{\Omega}_I \widehat{V}_r) \tag{9}$$

sendo $\boldsymbol{Q}$ a quasi-verossimilhança, $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{R}})$ é o vetor dos parâmetros estimados utilizando a estrutura de correlação $\boldsymbol{R}$, $\boldsymbol{I}$ representa a matriz identidade, $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{I}}$ é obtido por

$$\widehat{\boldsymbol{\Omega}_{\boldsymbol{I}}} = \frac{-\delta^2 \boldsymbol{Q}(\boldsymbol{\beta}; \boldsymbol{I}, \boldsymbol{D})}{\delta \boldsymbol{\beta} \delta \boldsymbol{\beta}'} \mid_{\beta=\widehat{\beta}} \tag{10}$$

e $\widehat{\boldsymbol{V}}_{\boldsymbol{r}}$ representa o estimator de covariâncias gerado pelo modelo com a estrutura de correlação $\boldsymbol{R}$.
Para diferentes modelos GEE com diferentes estruturas de correlação, são gerados vários valores de QIC, pelo que o melhor modelo e, por sua vez, a melhor estrutura de correlação será o modelo com o valor de QIC menor Pan (2001).

2.2.4 *Quasi-verossimilhança*

Como descrito na secção 2.1.1 o método de máxima verossimilhança é utilizado para estimar os parâmetros nos GLM, contudo este necessita que lhe seja atribuida uma distribuição dos dados, e nos modelos GEE essa distribuição dos dados não é conhecida Kleinbaum and Klein (2002).
Através da proposta de função de quasi-verossimilhança desenvolvida por Wedderburn (1974), McCullagh and Nelder (1989) demonstra que é possível estimar parâmetros de forma semelhante à função de máxima

verossimilhança, assumindo a independência entre as variáveis, especificando apenas a média e a variância de cada observação. A função de quasi-verossimilhança para cada observação é definida por

$$Q(\mu; y) = \int_y^{\mu} \frac{y - t}{\sigma^2 V(t)} dt \tag{11}$$

sendo μ as médias das observações e V as variâncias das mesmas.
A função de quasi-verossimilhança pode ser ajustada, dependendo do tipo de função da variância, McCullagh and Nelder (1989) descrevem alguns exemplos de funções de quasi-verossimilhança para diferentes funções de variância. Na Tabela 2 estão descritas algumas dessas funções.

Tabela 2: Funções de Quasi-verossimilhança descritas por (McCullagh,Nelder 1989)

Distribuição	Função de variância $V(\mu)$	Quasi-verossimilhança $Q(\mu; y)$	Restrições
Normal	1	$-(y-\mu)^2/2$	_
Binomial	$\mu(1-\mu)$	$y\log(\frac{\mu}{1-\mu}) + \log(1-\mu)$	$0 < \mu < 1; 0 \leq y \leq 1$
Poisson	μ	$y \log \mu - \mu$	$\mu > 0; y \leq 0$

Uma vez que a função de quasi-verossimilhança assume a independência das observações, esta não é adequada para análise assumindo a correlação entre observações, contudo Liang and Zeger (1986) adaptaram a função de quasi-verossimilhança, introduzindo-lhe a matriz de correlação especifica de modo a ser possivel a sua utilização para estimar parâmetros.

2.2.5 *Estimativa dos parâmetros*

Liang and Zeger (1986) demonstram, através de teoremas, que as estimativas geradas pelos modelos GEE são consistentes, ou seja, à medida que o número de *clusters* aumenta, as estimativas dos parâmetros aproximam-se dos seus valores reais e que as estimativas assumem uma distribuição assintoticamente normal, significando que N *clusters* se aproximam de infinito. O facto das estimativas seguirem uma distribuição assintoticamente normal permite a realização de inferências estatisticas, como por exemplo, intervalos de confiança e testes de hipóteses. As equações de estimativa de parâmetros nos modelos GEE destinam-se quando existem *clusters* com correlação existente entre eles. As equações para além dos parâmetros de regressão, também existente nos GLMs contêm

também parâmetros de correlação que poderão ser estimados, tal como os parâmetros de regressão.
De modo a criar a equação de estimativa de parâmetros para modelos GEE, inicialmente é determinada a matriz de covariância, como descrita por Liang and Zeger (1986),

$$V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}} / \phi \tag{12}$$

sendo A uma matriz diagonal, onde a úlitma diagonal representada pela ultima observação do ultimo individuo é a função da variância, $R(\alpha)$ representa a estrutura de correlação escolhida onde α representa os parâmetros de correlação.
Definida a matriz de covariância a equação de estimativa de parâmetros é representada por,

$$\sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0 \tag{13}$$

D_i representa a derivada parcial, tal que $D_i = \frac{\delta \mu}{\delta \beta}$, V_i representa a matrix de covariância definida anteriormente e S_i representa os resíduos sendo $S_i = (y_i - \mu_i)$.
Analisando a equação de estimativa de parâmetros dos modelos GEE, vários autores Fitzmaurice et al. (2012); Hedeker and Gibbons (2006); Liang and Zeger (1986); Wang et al. (2022), identificam que a mesma permite não só estimar os parâmetros de regressão (definidos como β) como também os parâmetros de correlação (definidos como α), uma vez que a matriz de covariância relaciona ambos os parâmetros, ou seja, as GEE permitem estimar ambos os parâmetros. Como ambos os parâmetros estão relacionados entre si, existe a necessidade de utilizar um algoritmo que permita iterar a equação que estima os parâmetros β e um método robusto para estimar α em função de β. O procedimento do algoritmo divide-se em duas etapas repetidas até ocorrer convergência. Na primeira etapa são estimados os parâmetros de covariância β através da equação de estimativa de parâmetros descrita anteriormente, onde os valores de α e ϕ são pré-definidos na primeira iteração. Na segunda etapa são estimados os parâmetros de correlação α e o fator ϕ, através do cálculo dos residuos de Pearson.

$$r_{it} = (y_{it} - \widehat{\mu}_{it}) / \sqrt{v(\widehat{\mu}_{it})} \tag{14}$$

Existem diferentes formas de estimar α dependendo do tipo de estrutura de correlação, conforme descrito por Liang and Zeger (1986). Na Tabela 3 estão representadas algumas das estimativas descritas relacionadas à estrutura de correlação, bem como a estimativa de ϕ.

2.2.6 *Inferências e testes estatísticos*

De modo a obter intervalos de confiança válidos e efetuar testes estatísticos, os parâmetros estimados por via da equação de estimativa de parâmetros e a sua distribuição normal assintótica, não são suficientes para gerar

Tabela 3: Estimativas de α e ϕ segundo Liang and Zeger (1986)

Parâmetro	Estrutura de correlação	Estimativa
ϕ	...	$\phi = \sum_{i=1}^{K} \sum_{t=1}^{K_i} r_{it}^2/(N-p)$
α	Independente	...
	Intercambiável	$\alpha = \phi \sum_{i=1}^{K} \sum_{t \neq j} r_{it} r_{ij} / \sum_{i=1}^{K} \frac{1}{2} n_i (n_i - 1) - p$
	AR1	$\alpha = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i - 1} \sum_{t \leq n_{i-1}} r_{it} r_{it+1}$
	Não Estruturada	$\alpha_{tj} = \frac{1}{K} \sum_{i=1}^{K} r_{it} r_{ij}$

inferências estatísticas de forma a retirar conclusões sobre o efeitos desses parâmetros na variável resposta Fitzmaurice et al. (2012); Kleinbaum and Klein (2002); Hedeker and Gibbons (2006). A estimativa da variância dos parâmetros estimados permite obter os erros-padrão necessários para a construção de intervalos de confiança.

Métodos de estimação da variância

A estimativa da variância dos parâmetros estimados a partir dos modelos GEE pode ser realizada através de dois métodos, o **estimador baseado no modelo (*Naive Estimator*)** e o **modelo empírico (robusto)**.

O estimador baseado no modelo é bastante utilizado em modelos GLM, uma vez que é baseado no método de máxima verossimilhança, no entanto, apesar de o método de máxima verossimilhança não se aplicar aos modelos GEE, se o modelo para a média das observações bem como a estrutura de correlação estiverem corretamente especificados, este método garante uma estimativa consistente Fitzmaurice et al. (2012); Kleinbaum and Klein (2002). O estimador baseado no modelo é definido por:

$$\mathbf{V}(\widehat{\beta}) = \left[\sum_{i}^{N} \mathbf{D_i'} \mathbf{V_i^{-1}} \mathbf{D_i} \right]^{-1} \tag{15}$$

Na generalidade das análises estatísticas por modelos GEE, a estrutura de correlação muitas das vezes é incorreta ou inexistente, pelo que Liang and Zeger (1986) desenvolveram um método robusto de estimativa da variância de modo a garantir uma estimativa consistente mesmo quando a estrutura de correlação é incorreta Kleinbaum and Klein (2002). O modelo robusto é definido por:

$$\mathbf{V}(\widehat{\beta}) = \mathbf{M_0^{-1}M_1M_0^{-1}} \tag{16}$$

onde,

$$\mathbf{M_0} = \sum_{i}^{N} \mathbf{D_i'V_i^{-1}D_i} \tag{17}$$

e,

$$\mathbf{M_1} = \sum_{i}^{N} \mathbf{D_i'V_i^{-1}}(y_i - \widehat{\mu}_i)(y_i - \widehat{\mu}_i)'\mathbf{V_i^{-1}D_i} \tag{18}$$

O método robusto de estimativa da variância é por norma o mais utilizado em análises estatisticas utilizando os modelos GEE uma vez que possibilita estimativas robustas da variância mesmo com a estrutura de correlação não especificada, no entanto demonstra alguns problemas de eficiência quando utilizados para dados não balanceados, ou com poucas observações entre indivíduos, nestes casos a estimativa baseado no modelo permite obter estimativas mais consistentes Fitzmaurice et al. (2012).

Testes de Hipóteses

O testes de hipóteses aplicados aos modelos GEE procuram testar a hipótese nula de que os parâmetros são iguais a zero. De modo a aceitar ou rejeitar a hipótese nula, é utilizada a estatística de Wald uma vez que os parâmetros estimados por os modelos GEE são assintoticamente normais. A estatística de Wald foi desenvolvido por Rotnitzky and Jewell (1990) e define-se por:

$$T_w = K(\widehat{\gamma}_\mathbf{G} - \gamma_\mathbf{0})^T\widehat{V}_\gamma^{-1}(\widehat{\gamma}_\mathbf{G} - \gamma_\mathbf{0}) \tag{19}$$

sendo $\widehat{\gamma}_G$ um vetor com os primeiros parâmetros estimados e V_γ a principal submatriz da variância estimada.

A estatística de Wald segue uma distribuiçao qui-quadrado com graus de liberdade iguais ao número de parâmetros testados. Esta estatística poderá ser utilizada para testar um ou vários parâmetros estimados em simultâneo Kleinbaum and Klein (2002).

No entanto, a estatística de Wald apresenta algumas lacunas como descrito porRotnitzky and Jewell (1990), uma das lacunas é a sua dependência da escala de medição dos parâmetros de regressão, como alternativa, o autor descreve um novo teste estatístico, sendo este o teste estatístico de *score*, definido por:

$$T_s = K^{-1}\mathbf{U_{fl}}\left[\gamma_0, \widetilde{\delta}(\gamma_0)\right]^T\mathbf{\Sigma_{fl}^{-1}U_{fl}}\left[\gamma_0, \widetilde{\delta}(\gamma_0)\right] \tag{20}$$

onde, $\boldsymbol{U}_\gamma$ representa a equação de estimação de parâmetros dos modelos GEE e $\boldsymbol{\Sigma}_\gamma^{-1}$ representa a estimativa da variância.

Ambos os testes estatísticos descritos são válidos para os modelos GEE, no entanto quando os dados não se encontram balanceados pode gerar problemas nos resultados gerados. Rotnitzky and Jewell (1990) apresentam adaptações destes dois testes estatísticos que permitem ultrapassar esse problema.

2.3 PROGRAMAS/ALGORÍTMOS ESTATÍSTICOS PARA MODELOS GEE

Programas estatísticos para modelos GEE estão presentes nas diferentes linguagens de análise estatística como *Stata*, *R*, *SPSS* entre outros.

Vários autores descrevem a utilização e comparam os resultados obtidos nas diferentes linguagens, Horton and Lipsitz (1999) comparou diferentes *packages*, em *S-plus, Stata, SAS e SUDAAN* para modelos GEE. O autor verificou que o *package* proveniente do SUDAAN apesar de apresentar resultados semelhantes a todos os outros *packages*, este apresentava limitações nas estruturas de correlação e necessitava que todas as variáveis fossem números (limitação encontrada também no *package* proveniente do *Stata*). Os outros *packages* apresentavam uma codificação mais geral para as matrizes de trabalho o que permitia efetuar análises mais complexas.

De modo a verificar limitações entre diferentes *packages*, Nooraee et al. (2014) utilizou um conjunto de *packages* disponíveis me diferentes linguagens, os *packages* utilizados foram *GENMOD, SAS*, *GENLIN, SPSS*, *repolr, multgee* e *geepack, R*. Este autor verificou que os *packages*, *multgee* e *repolr* apresentam mais funcionalidades que os restantes, no entanto todos eles demonstraram limitações na estimação dos parâmetros quando o número de individuos é muito baixo.

O *R* apresenta no seu repositório vários *packages* referentes aos modelos GEE, como por exemplo:

- *gee* realiza todo o processo de estimação de parâmetros dos modelos GEE, permite definir várias estruturas de correlação e ultrapassa a limitação dos dados omissos;
- *geepack* permite estimar os parâmetros utilizando a metedologia dos modelos GEE, gerando tambem inferências estatísticas;
- *multgee* utiliza duas funções dependendo do tipo de varíavel resposta, seja ela nominal ou ordinal, no entanto as estruturas de correlação são mais limitadas quando comparado com outros *packages*.

3

PLANEAMENTO E METODOLOGIA

Como descrito nas secções anteriores, os modelos GEE são utilizados para identificar a influência/relação de vários parâmetros com a variável resposta. A aplicação proposta para esta dissertação, será desenvolvida recorrendo à técnica de programação por objetos, utilizando os *packages* existentes em *R*, pelo que neste capítulo irão ser abordados alguns conceitos associados a este tipo de linguagem de programação.

3.1 PROGRAMAÇÃO EM R

A linguagem *R* foi desenvolvida por Ihaka and Gentleman (1996), sendo esta baseada na linguagem *S*. O *R* propociona vários recursos computacionais para análise estatística, manipulação de dados orientados a objetos, bem como representações gráficas detalhadas, para além de fornecer vários recursos para análise estatística, a linguagem *R* permite criar interfaces com outras linguagens de modo a facilitar a sua utilização Morandat et al. (2012); Team (2000).

Contrariamente a outras linguagens de programação, que descrevem os vários *outputs* gerados ao longo de uma análise, o *R* armazena os resultados obtidos em vários objetos que podem ser utilizados em outras funções do *R*, minimizando assim a extensão de *outputs* obtidos Venables et al. (2009).

A utilização da linguagem *R* para o utilizador é interativa, uma vez que lhe permite analisar os seus dados de uma forma rápida e intuitiva através de expressões simples para análises e representações gráficas. Esta caracteristica da linguagem *R* facilita a sua utilização, não necessitando de formação extensiva por parte do utilizador, uma vez que os seus conceitos são de aprendizagem simples Morandat et al. (2012).

O *R* fornece também uma plataforma para o desenvolvimento de funções ou algoritmos, tal como outras linguagens de programação, facilitando a modelação estatistica bem como as representações gráficas, evitando assim a repetição de várias expressões simples tornando o processo mais eficiente. Team (2000); Wilson and Lorenz (2015)

O desenvolvimento de novas funções e algoritmos em *R*, permitiu a criação de mais de 4000 *packages* disponíveis em vários repositórios como o CRAN, garantindo aos vários utilizadores uma vasta opção de métodos/algoritmos de análise estatística. Para além dos diversos *packages* existentes, o *R*, por via do *Shiny*, possibilita também a criação de aplicações *web* , utilizando várias funções de análise estatistica do *R* Morandat et al. (2012).

A versatilidade do *R* permite ao utilizador o uso de funções contendo algoritmos presentes no *R*, assim como o uso de funções desenvolvidas em outras linguagens, como por exemplo *C*, e o desenvolvimento de novas funções utilizando os diversos algoritmos existentes no *R* Wilson and Lorenz (2015).

3.2 SHINY

O desenvolvimento de aplicações web para análises estatísticas, apresentava ser um desafio, uma vez que necessitava de algum conhecimento profundo sobre outras linguagens, tais como JavaScript, HTML ou CSS para desenvolver a interface da aplicação e estabelecer uma ligação com a linguagem R para processamento e análise estatística Wickham (2021).

O *Shiny* é um *package* do *R*, que possibilita o desenvolvimento de aplicações *web*, sem necessitar de conhecimento prévio de outras linguagens de programação, utilizando apenas a linguagem *R*. A aplicação desenvolvida em *Shiny* é composta por dois componentes:

- **Interface:** Componente que corresponde à estrutura da aplicação e será o componente observável por o utilizador, também é responsável por receber todas as instruções fornecidas por o utilizador que serão utilizadas para realizar as tarefas desejadas. A interface permite ao utilizador visualizar, após realizadas as análises, os resultados obtidos;
- **Servidor:** Componente responsável por todo o processamento das instruções fornecidas pelo utilizador de modo a realizar a análise estatística desejada, através de funções/algoritmos desenvolvidos em *R*.

O *Shiny* utiliza programação reativa, que permite atualizar em tempo real todos os resultados e processos efetuados, quando ocorre uma alteração nas intruções fornecidas pelo utilizador. Após desenvolvidos os dois componentes essencias a aplicação encontra-se disponível na web para qualquer utilizador de forma gratuita.

Na Fig. 5 encontra-se ilustrada uma estrutura que pretende explicar o funcionamento de uma aplicação *Shiny*.

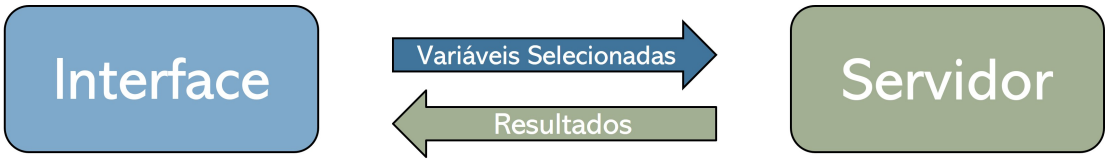


Figura 5: Exemplo ilustrativo da aplicação *Shiny*

3.3 CASO DE ESTUDO

De modo a verificar quais as ferramentas existentes em *R* para a análise de modelos GEE, bem como numa fase posterior exemplificar a utilização da aplicação *Shiny* desenvolvida, foi utilizado um conjunto referente à identificação da presença de pseudoexfoliação ocular em vários pacientes sujeitos a cirurgia de remoção de

cataratas, no Departamento de Oftalmologia do Centro Hospitalar de Vila Nova de Gaia/Espinho, no período de 1 de Junho a 31 de Dezembro de 2016.

3.3.1 *Pseudoexfoliação Ocular*

A pseudoexfoliação ocular (PEX) é uma patologia relacionada com a idade e caracteriza-se por depósitos de material fibrilar, normalmente identificados por uma cor branca, presentes na estruturas oculares do segmento anterior banhadas pelo humor aquoso Ritch and Schlötzer-Schrehardt (2001); Schlötzer-Schrehardt and Naumann (2006). Esta doença está associada a alterações e comportamentos anormais do gene LOXL1, responsável por sintetizar o material fibrilar presente na pseudoexfoliação ocular Sangal and Chen (2014); Govetto et al. (2015); Jammal et al. (2021), apesar desta associação, existem também fatores externos associados à pseudoexfoliação ocular, como a idade, o sexo e a área geográfica onde reside Sangal and Chen (2014); Shazly et al. (2011).
A presença de pseudoexfoliação ocular pode se apresentar unilateralmente ou bilateralmente e identifica-se como um fator de risco para o desenvolvimento de glaucoma e cataratas densas, pelo que são estas as implicações clínicas mais importantes desta patologia. A presença de pseudoexfoliação ocular contribuí para um maior risco de complicações cirurgicas na remoção de cataratas uma vez que são apresentadas alterações, nomeadamente na dilatação pupilar, fraquezas zonulares que podem contribuir para luxações na lente intraocular, entre outros Govetto et al. (2015); Schlötzer-Schrehardt and Naumann (2006), Sangal and Chen (2014) apresentam um conjunto de riscos associados nas cirurgias de remoção de cataratas associados à presença de pseudoexfoliação ocular em pacientes.

3.3.2 *Modelos de Regressão associados à PEX*

Vários estudos da pseudoexfoliação (PEX) ocular procuram identifcar associações com o desenvolvimento e presença de glaucoma, cataratas e outras patologias, bem como verificar a influência de outros fatores na presença da PEX. De modo a identificar essas relações são muitas vezes utilizados modelos de regressão que permitem analisar a influência de vários fatores para a presença da pseudoexfoliação, como para analisar a influencia da pseudoexfoliação no desenvolvimento de outras patologias.
Vários autores utilizaram modelos de regressão logística de modo a identifcar a associação entre a pseudoexfoliação e o glaucoma, Mitchell et al. (1999) procurou verificar se a pseudoexfoliação teria influência em fatores de risco para o desenvolvimento do glaucoma. Através dos modelos de regressão logística verificou que a presença de pseudoexfoliação ocular encontra-se relacionada com o glaucoma em indivíduos mais velhos. No entanto através de modelos GEE verificou que os fatores de risco associados à glaucoma são independentes da presença da pseudoexfoliação, Jammal et al. (2021) verificaram que o desenvolvimento, tanto de glaucoma como cataratas é mais evidente em pacientes com bilateralidade de pseudoexfoliação ocular. Identificou também que a frequência de complicações nas cirurgias de remoção de cataratas é relativamente pouco significativa entre pacientes com e sem a pseudoexfoliação. Shazly et al. (2011) verificaram a associação entre a PEX e o desenvolvimento ou presença de glaucoma e cataratas de um ponto de vista geográfico. Verificou-se à

semelhança de outros estudos uma associação relevante entre a PEX e a presença de glaucoma e cataratas, no entanto a associação é mais elevada do que em outros estudos considerando outras zonas geográficas, no entanto relativamente às complicações nas cirurgias de remoção de cataratas não foram identificadaslimitações que permitam identificar qualquer relação.
Govetto et al. (2015) centraram-se em pacientes agendados para cirurgias de remoção de cataratas, de modo a verificar a influência da presença de pseudoexfoliação considerando diferentes fatores. Através de regressões logísticas, identificou-se uma relação entre a presença de pseudoexfoliação e diferentes fatores, como a idade, a redução pupilar, presença de glaucoma e a pressão intraocular. Este estudo permitiu reforçar que pacientes com pseudoexfoliação ocular são mais propensos a complicações nas cirurgias uma vez que a pseudoexfoliação influência fatores chave como a redução pupilar ou pressão intraocular.

A amostra utilizada neste estudo é proveniente do Hospital de Gaia referente a pacientes sujeitos a cirurgia de remoção de cataratas, onde foram retirados ao longo do tempo os dados clinicos, bem como a idade e sexo, do paciente, sendo o conjunto de dados dividido nas diferentes colunas:

- **Processo**: Número de identificação do paciente.
- **Data**: Data da consulta associada ao paciente.
- **Idade e Sexo**: Idade e Sexo referentes ao paciente.
- **Lat**: Olho analisado na respetiva consulta, representado numa escala binomial sendo o valor 1 correspondente ao olho direito e o valor 2 correspondente ao olho esquerdo.
- **PEX**: Presença de pseudoexfoliação no respetivo olho analisado, representado numa escala binomial sendo o valor 1 a presença da PEX e o valor 0 a ausência da mesma.
- **Dil pp (mm)**: Medição da dilatação pupilar dividida em 5 diferentes graus, grau 1 correspondente a uma dilatação superior ou igual a 8 mm, grau 2 correspondente a uma dilatação entre 7 e 8 mm, grau 3 correspondente a uma dilatação entre 6 e 7 mm, grau 4 correspondente a uma dilatação entre 5 e 6 mm e grau 5 correspondente a uma dilatação menor que 5 mm.
- **LIO (Dp)**: Dioptrias da lente intraocular do paciente.
- **Complicações**: Complicações existentes durante a cirurgia sendo o local onde esta ocorreu identificado com um valor, sendo o valor 1 correspondente à ausência de complicações durante a cirurgia, 2 corresponde a uma rutura da cápsula posterior, 3 corresponde a problemas na deiscência zonular e 4 indica que um problema ocorreu na zona da vitrectomia anterior.
- **Local da LIO**: Local onde inserida a lente intraocular, sendo atribuido o valor 1 a lente inserida no saco, o valor 2 a lente inserida no sulco, o valor 3 na câmara anterior da irís, o valor 4 indica a ausência da lente intraocular.

3.4 ANÁLISE EXPLORATÓRIA DO CONJUNTO DE DADOS PARA CASO DE ESTUDO

Previamente ao desenvolvimento dos modelos GEE aplicados ao caso de estudo descrito anteriormente e de modo a descrever o conjunto de dados foi realizada uma análise exploratória dos mesmos. Para a realização da análise exploratória o conjunto de dados foi alterado de modo a identificar as colunas que apresentam variáveis categóricas e as que apresentam variáveis contínuas. O conjunto de dados utilizado para a análise exploratória encontra-se no seguinte link de *GitHub*: https://github.com/LuisMoncaixa1996/Thesis/blob/main/GEE%20Models/data_cat.xlsx
Inicialmente foi analisado o conjunto de dados, de modo identificar valores omissos presentes. Através da análise foi possível verificar que o conjunto de dados apresenta apenas 2% de valores omissos, sendo a coluna "LIO(DP)"a que contém um maior número de valores omissos conforme representado na Fig. 6.

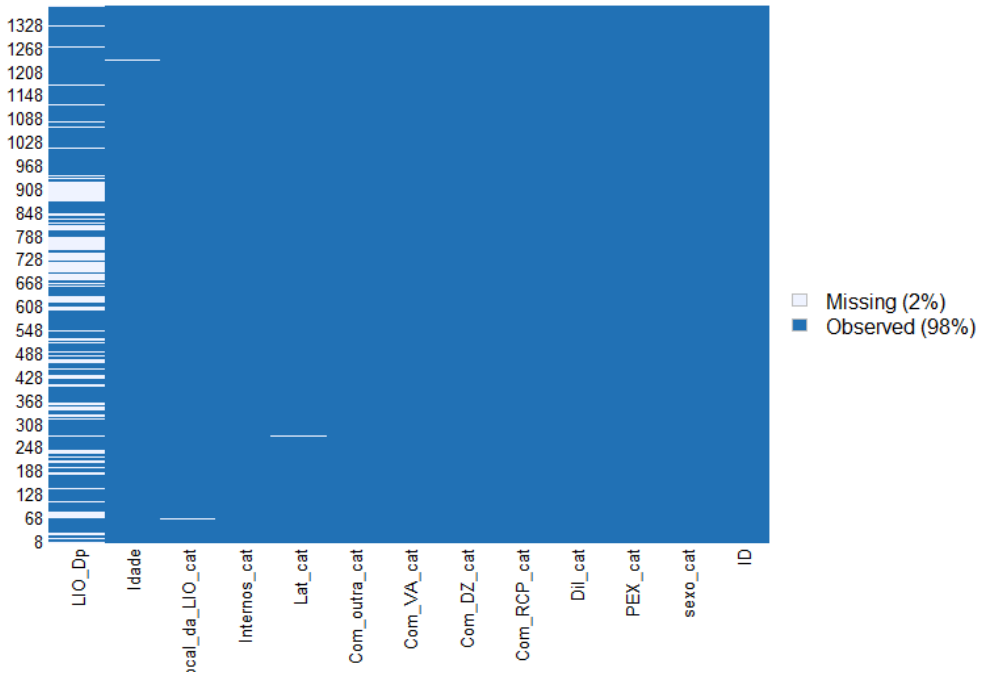


Figura 6: Quantificação de valores omissos presentes no conjunto de dados.

O conjunto de dados utilizado para o caso de estudo apresenta apenas duas variáveis contínuas, sendo estas a Idade e a LIO (DP). De modo a descrever ambas as variáveis, inicialmente foram obtidas as estatísticas descritivas e posteriormente representadas através de um histogramas para cada uma das variáveis. Relativamente às idades dos pacientes, verifica-se que a idade média dos pacientes apresenta-se nos 72 anos, sendo a idade do paciente mais velho de 103 anos e do mais novo 31 anos, quanto às dioptrias presentes nas lentes intraoculares dos pacientes, verifica-se uma média de 21,1 dioptrias, sendo a lente com 36 dioptrias a lente com o valor máximo e a lente com -3 dioptrias com o valor mínimo, estes valores encontram-se descritos na tabela 4.

Tabela 4: Estatísticas descritivas das variáveis contínuas

Variáveis Contínuas	Máximo	Mediana	Média	Minímo
Idade	103	73	71,9	39
LIO DP	36,0	21,5	21,1	-3,0

Através dos histogramas representados na Fig.s 7 e 8 para as duas variáveis contínuas é possível verificar que a existe um maior número de pacientes com idades entre os 70 e 80 anos. Quanto às dioptrias utilizadas nas lentes intraoculares existe um elevado número de pacientes que possuem lentes com dioptrias entre os 20 e 25.

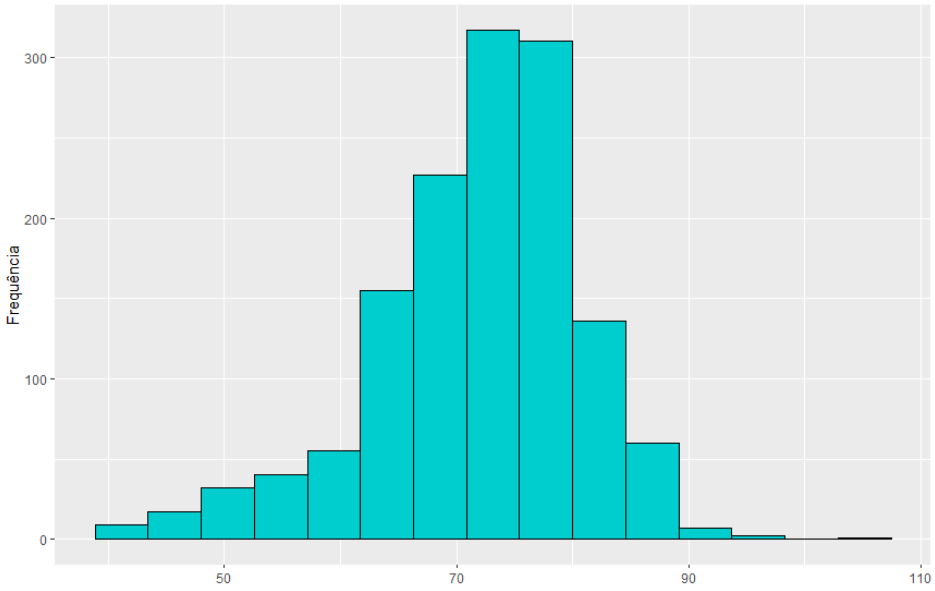


Figura 7: Histograma relativo às idades dos pacientes.

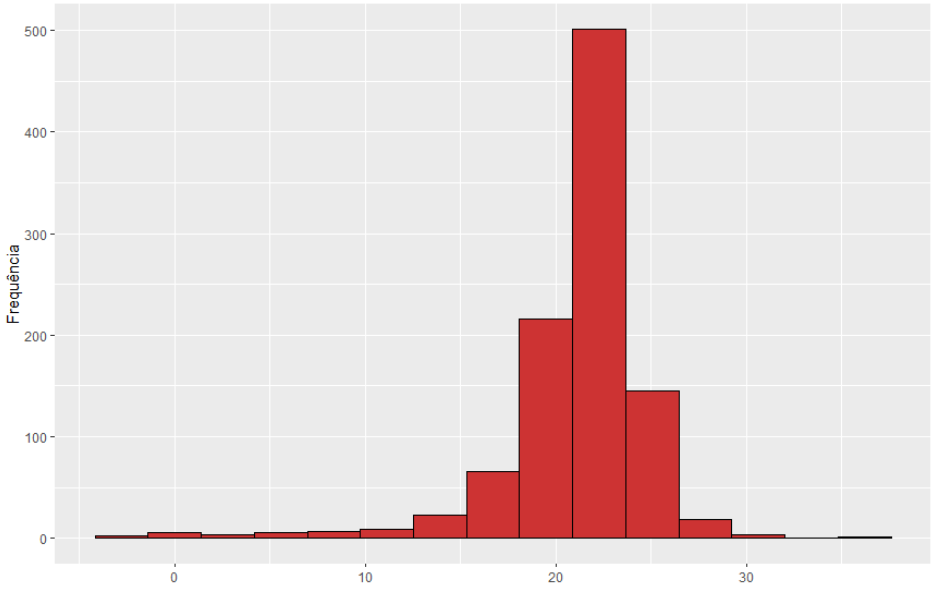


Figura 8: Distribuição segundo o valor das dioptrias introduzidas nas lentes intraoculares.

Na amostra em estudo é possível verificar, quanto ao género, que exite uma maior predominâcia de pacientes do sexo feminino, cerca de 59% (800) dos pacientes, em relação com o sexo masculino, cerca de 41% (600) dos pacientes, conforme representado na Fig. 9.

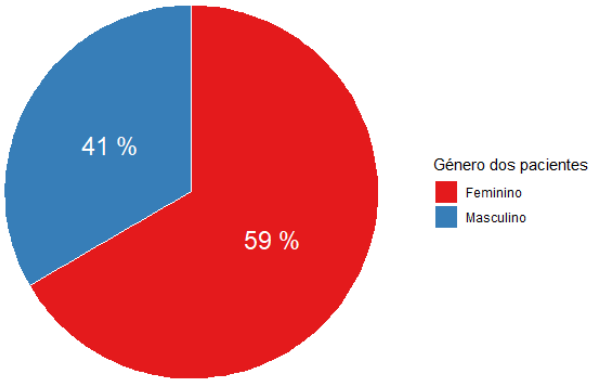


Figura 9: Distribuição dos pacientes segundo o género.

A presença da pseudoexfoliação pode ocorrer em um olho singular ou nos dois do mesmo paciente, ou em diferentes pacientes. Na Fig. 10 verifica-se que a pseudoexfoliação manifesta-se mais no olho direito dos diferentes pacientes em relação ao olho esquerdo.

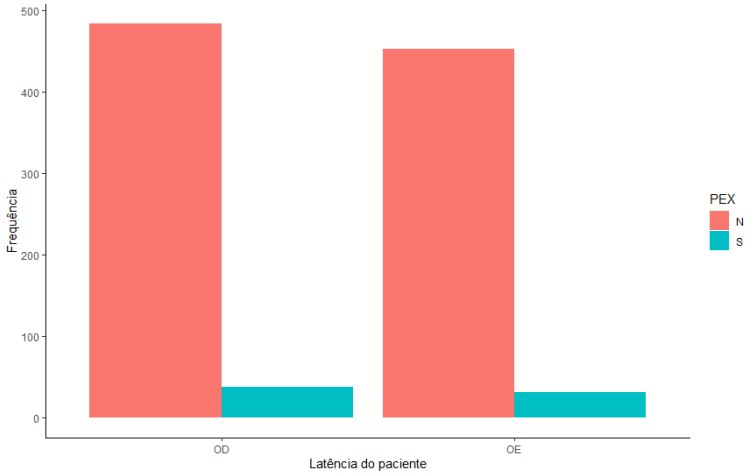


Figura 10: Presença da pseudoexfoliação segundo os diferentes olhos do paciente.

Como descrito por Govetto et al. (2015), a dilatação pupilar apresenta-se como um factor importante que pode influenciar a presença de pseudoexfoliação em pacientes, com base nesta permissa analisou-se a dilatação pupilar registada nos pacientes de modo a identificar qual a dilatação mais observada e posteriormente analisou-se qual a classe de dilatação onde está mais representada a presença da PEX. Na Fig. 11 verifica-se que a classe de dilatação superior a 8mm encontra-se muito mais representada que as restantes, considerando a presença de pseudoexfoliação ou ausência da mesma, a Fig. 12 desreve que a maior percentagem de presença de pseudoexfoliação ocular ocorre quando a dilatação pupilar encontra-se entre os 5-6 mm, considerando a classe mais representada, segundo a Fig. anterior, verifica-se que a proporção de presença da PEX é muito reduzida.

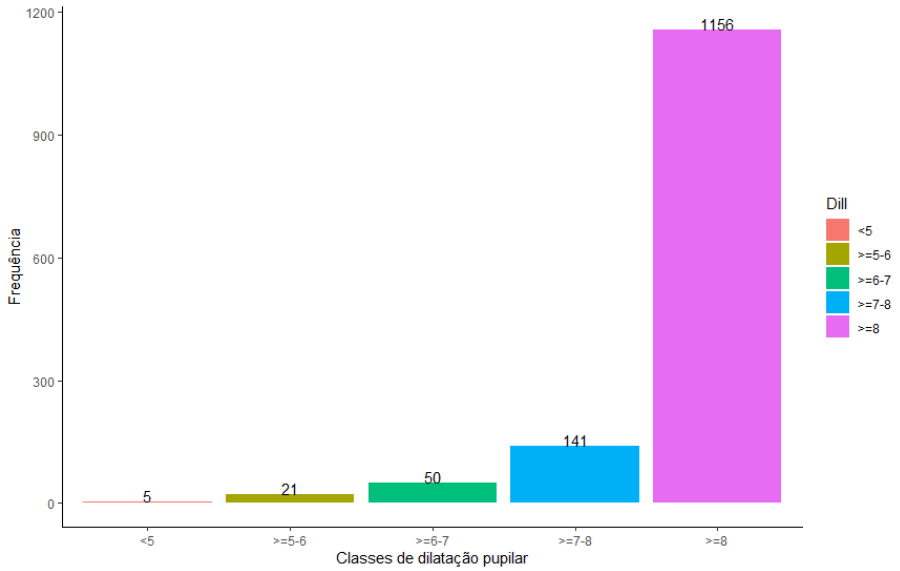


Figura 11: Distribuição segundo as diferentes classes de dilatação pupilar.

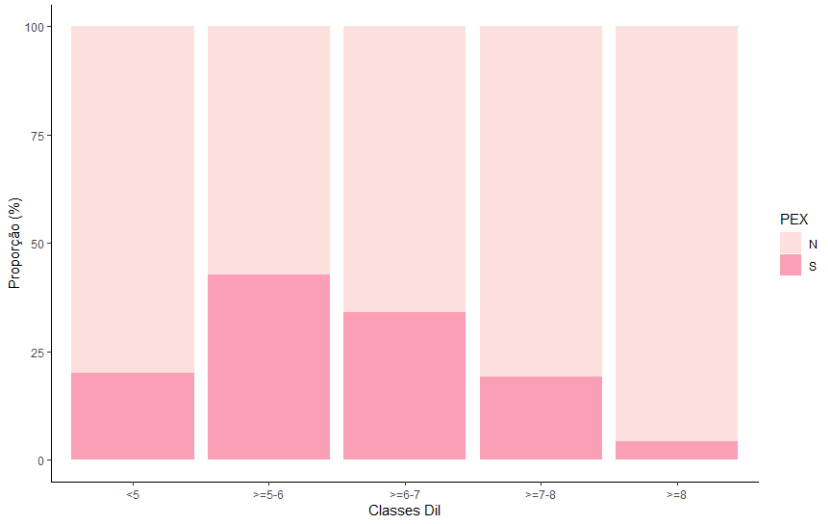


Figura 12: Distribuição da dilatação pupilar segundo a presença da PEX.

Uma vez que os pacientes observados no caso de estudo foram sujeitos a cirurgía de remoção de cataratas, estes podem possuir uma lente intraocular em diferentes zonas do olho. Posto isto foi analisado para cada olho, esquerdo ou direito, aquele que onde foram implementadas lentes intraoculares e qual o local de implementação mais utilizado. Conforme representado na Fig. 13, o número de lentes intraoculares implementadas no olho direito é ligeiramente superior quando comparadas com o olho esquerdo. Verifica-se também que grande parte das lentes encontram-se implementadas no saco.

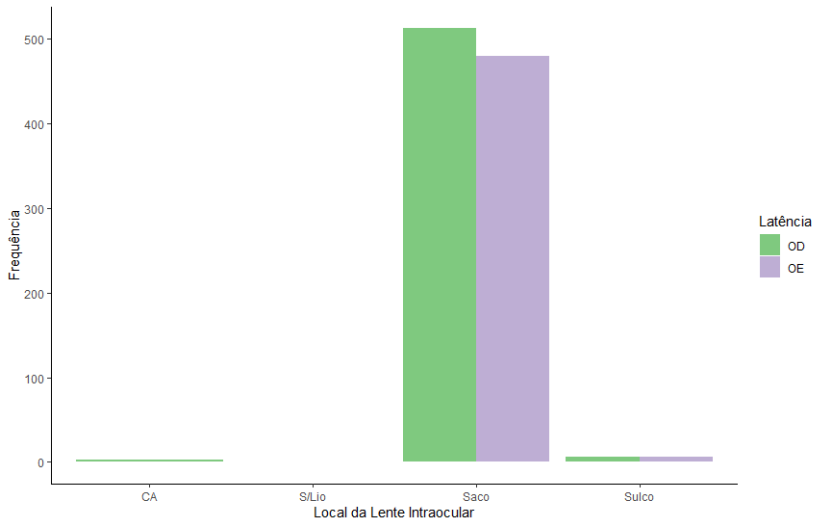


Figura 13: Distribuição segundo as diferentes zonas de implantação da lente por olho.

3.5 ABORDAGEM

Nos estudos apresentados verificou-se uma tendência na utilização de modelos de regressão logística para identificar a influência da pseudoexfoliação em vários fatores. Como tal numa primeira abordagem e de forma a identificar as ferramentas existentes em *R*, serão aplicados modelos GEE no conjunto de dados descrito na secção anterior, de forma a verificar relações e associações entre a pseudoexfoliação e os diferentes fatores associados a pacientes sujeitos a cirurgia de remoção de cataratas. Durante o desenvolvimento dos modelos serão utilizados diferentes *packages* de modo a verificar as diferenças entre eles e as suas funcionalidades.

Apesar de existirem diversos *packages* em diferentes linguagens de programação para aplicação da análise de modelos GEE, não existem efetivamente aplicações que permitam a utilização destes modelos de forma intuitiva e que permitam uma análise mais eficiente por parte do utilizador.

De modo a facilitar a utilização dos modelos GEE e melhorar a análise dos resultados gerados, será desenvolvida uma aplicação, através do *Shiny*, que permita ao utilizador:

- Carregar o conjunto de dados relevante para a análise;
- Selecionar o tipo de variáveis, a estrutura de correlação, bem como editar os dados;
- Importar/Exportar modelos;
- Analisar os resultados obtidos de forma intuítiva;
- Exportar as representações gráficas geradas;
- Obter ajuda no funcionamento da aplicação.

Na Fig. 14 encontra-se representado um esquema simplificado da aplicação *Shiny* a desenvolver. Inicialmente o utilizador terá que carregar a sua estrutura de dados, posteriormente irá selecionar todas as variáveis e covariáveis de interesse para o modelo, bem como a estrutura de correlação associada, de seguida será possivel visualizar o modelo gerado e editar caso ocorram erros na seleção das variáveis, após selecionadas as variáveis todas as informações serão enviadas para o servidor que irá gerar o modelo e validar através de curvas ROC. Por fim serão apresentados todos os resultados obtidos do modelo gerado, sendo estes estatísticos e gráficos.

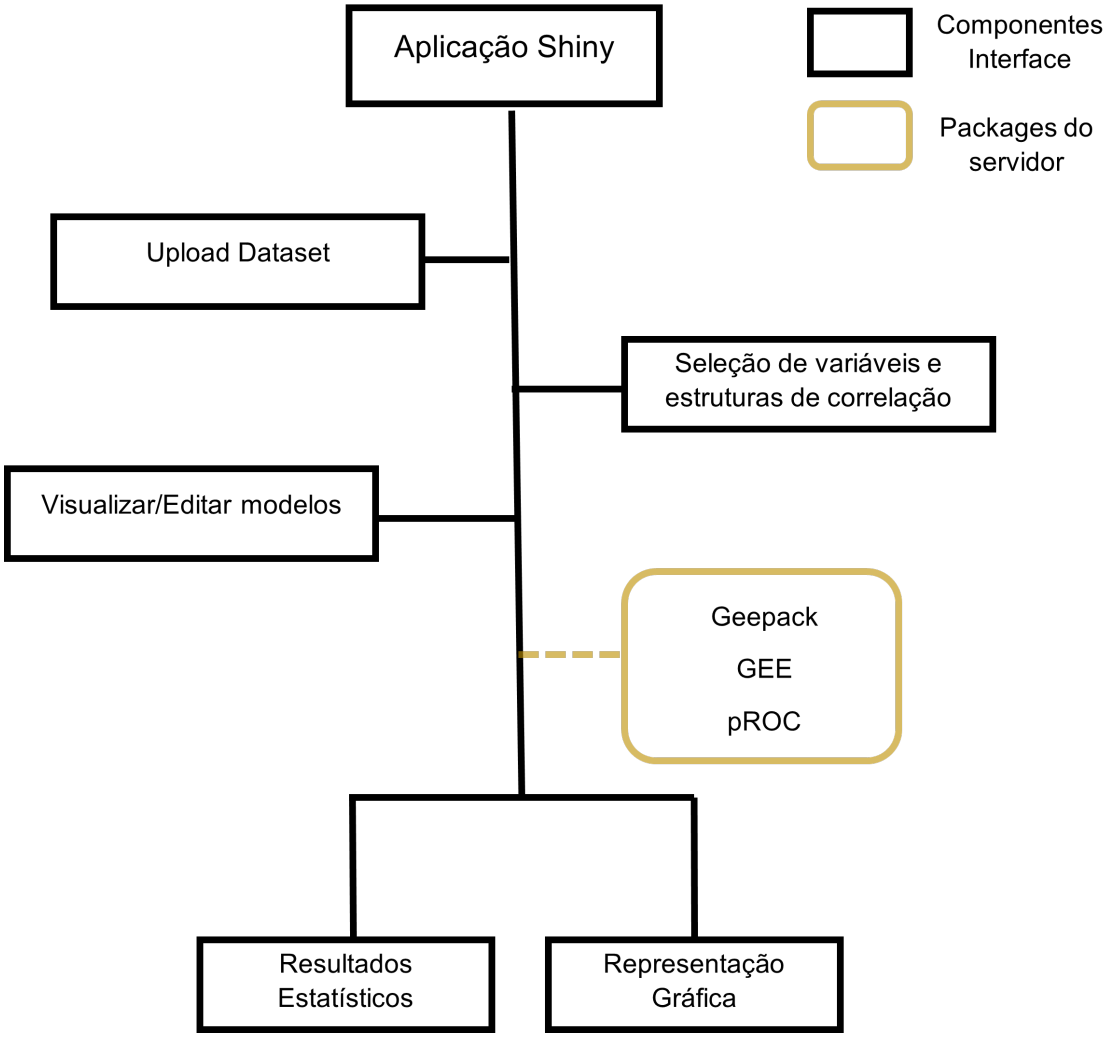


Figura 14: Estrutura da aplicação *Shiny*

4

MODELOS GEE - PSEUDOEXFOLIAÇÃO OCULAR

Tendo como base a dependência existente entre medições recolhidas ao longo do tempo a pacientes, sujeitos a cirurgia de remoção de cataratas, e de modo a analisar a influência de vários fatores na presença de pseudoexfoliação ocular, são desenvolvidos vários modelos GEE que procuram analisar essas associações. Paralelamente ao desenvolvimento dos modelos GEE será comparada a *performance* de dois *packages*, na linguagem de programação *R*, responsáveis pelo desenvolvimento dos modelos.

A Fig. 15 representa o processo de desenvolvimento e validação dos modelos GEE. A primeira fase corresponde inicialmente a um processo de divisão do conjunto de dados em dados de treino e dados de teste, posteriormente, utilizando os dados de treino serão desenvolvidos os modelos GEE, por fim os modelos serão validados através de métricas resultantes da análise ROC. Na segunda fase serão desenvolvidos diferentes modelos GEE utilizando várias combinações de covariáveis para a váriavel resposta (presença de pseudoexfoliação ocular), os modelos desenvolvidos serão validados, à semelhança da primeira fase, através da análise ROC. O código desenvolvido para criação e validação dos modelos podem ser visualizados a partir da seguinte página do GitHub: https://github.com/LuisMoncaixa1996/Thesis/tree/main/GEE%20Models.

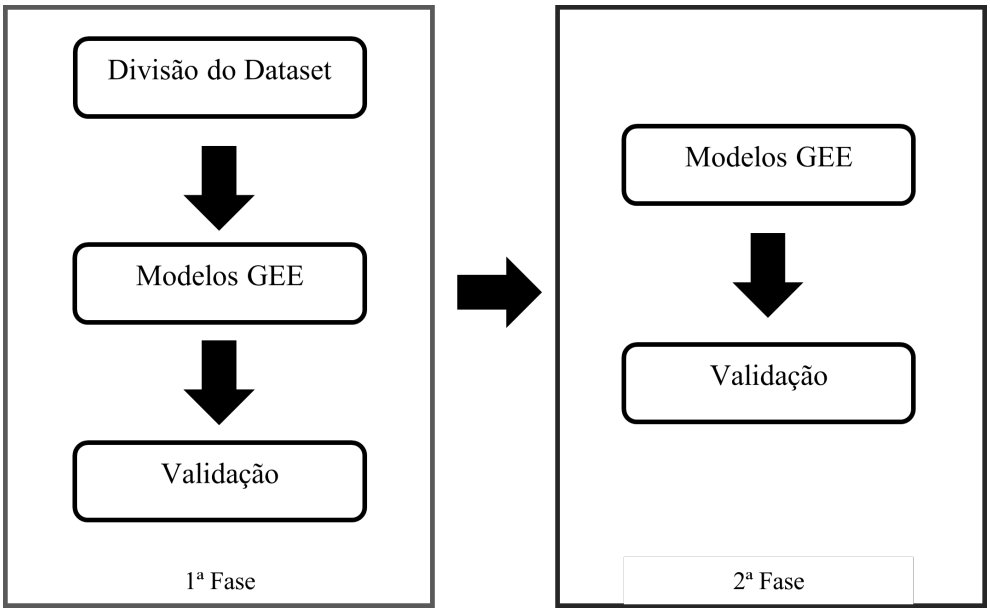


Figura 15: Processo de desenvolvimento dos modelos GEE

4.1 DESENVOLVIMENTO E VALIDAÇÃO DOS MODELOS GEE: 1ª FASE

Para realizar uma análise mais rigorosa sobre os modelos GEE desenvolvidos e os seus respetivos resultados, inicialmente a base de dados foi dividida em dados de treino e dados de teste, foram realizadas 3 diferentes divisões dos dados:

- 70% treino e 30% teste;
- 75% treino e 25% teste;
- 80% treino e 20% teste;

Após a divisão dos dados, para cada uma das divisões, foram desenvolvidos dois modelos GEE, cada modelo foi desenvolvido utilizando dois *packages* distintos, *geepack* e GEE, uma vez que neste caso de estudo especifico existe apenas uma variável resposta o *package multgee* não foi utilizado. Em ambos os modelos desenvolvidos, foi definida como variável resposta a presença de pseudoexfoliação ocular (PEX) e como covariaveis as colunas Idade, Sexo, Lat, Dilpp, Internos, LIO_DP, Local da LIO e as Complicações. Cada covariável foi categorizada de modo a validar a influência de cada categoria na variável resposta.

No desenvolvimento dos modelos GEE, de forma a verificar qual a melhor estrutura de correlação a utilizar, foi utilizado o método de QIC para as seguintes estruturas de correlação:

- Estrutura de correlação intercambiável;
- Estrutura de correlação autoregressiva;
- Estrutura de correlação independente;

De modo a selecionar a divisão dos dados que permite desenvolver um modelo GEE mais preciso para uma posterior análise foi considerada a métrica correspondente à *accuracy* do modelo, bem como realizada uma análise ROC para cada um dos modelos desenvolvidos, considerando as divisões treino-teste diferentes. A escolha da análise ROC remete-se para o facto da variável resposta em estudo ser do tipo binária, permitindo assim selecionar o modelo pelo qual é maximizado o número de classificações corretas.

4.1.1 Accuracy

Uma vez que variável resposta (presença de pseudoexfoliação ocular) do caso de estudo é binária, a *accuracy* irá verificar a proporção de classificações corretas por parte do modelo. No entanto esta métrica apresenta algumas limitações quando utilizada isoladamente para validação dos modelos. Metz (1978) demonstra que a *accuracy* num contexto de diagnóstico de doença, a prevalência da patologia pode influenciar o aumento do valor da mesma. Devido às limitações que o autor demonstra, a interpretação do valor da *accuracy* deve ser realizada com alguma cautela, pois dependerá do objetivo principal do estudo e da prevalência apresentada pela patologia no contexto diagnóstico. De modo a complementar o processo de validação dos modelos, para além da métrica de *accuracy* será também utilizada a área abaixo da curva resultante da análise *ROC*.

4.1.2 *Análise* ROC

A análise através de curvas *ROC* permite ao investigador, classificar uma determinada amostra mediante um pressuposto de interesse, bem como comparar diferentes classificações consoante a margem pretendida, de forma a maximizar o número de classificações corretas Fawcett (2006).

A curva *ROC* empírica é uma representação gráfica bidimensional da FVP (Sensibilidade), que representa a probabilidade de um determinado pressuposto ser corretamente classificado, em função da FFP (1 – Especificidade), que representa a probabilidade de um pressuposto contrário ao representado pela "Sensibilidade"ser incorretamente classificado, num plano unitário de eixos coordenados. A curva *ROC* surge assim, pela união de diferentes pares (FFP, FVP) para valores de corte diferentes ao longo do eixo de decisão Braga (2000).

Área abaixo da curva (AUC)

De modo a avaliar o desempenho do teste diagnóstico/classificador, utilizando a curva ROC empírica, existe um valor determinante para esta avaliação, a área abaixo da curva (AUC). Estando esta área incluída na representação gráfica da curva *ROC*, esta terá um valor entre 0 e 1, sendo que 0,5 corresponde à área do triângulo definido pela diagonal (igualdade entre a FVP e a FFP) no espaço, o que demonstra uma não existência de capacidade discriminante, ou seja, um modelo cujo valor de AUC seja inferior a 0,5 terá um fraco desempenho Bradley (1997); Hanley and McNeil (1982). Este indicador é determinante para a comparação de duas curvas ROC empíricas, sendo que o modelo, cuja curva com AUC maior é aquele que apresenta melhor desempenho.

O uso da AUC simplifica as análises estatisticas em dados onde existem amostras repetidas ao longo do tempo, sem necessitar de ignorar algumas dessas amostras, uma vez que com a AUC apenas são realizadas as comparações estatisticas entre os clusters, esta também resulta em uma diminuição do número de comparações estatísticas realizadas entre grupos. A AUC apresenta-se também uma métrica alternativa, quando o intervalo de tempo entre as medições é diferente entre grupos Fekedulegn et al. (2007). Sendo o caso de estudo do tipo longitudinal, uma vez que apresenta medições ao longo do tempo, a AUC apresenta-se como uma métrica de interesse para a validação dos modelos.

4.2 DESENVOLVIMENTO E VALIDAÇÃO DOS MODELOS GEE: 2ª FASE

Após definida a divisão do conjunto de dados a partir das métricas de validação definidas anteriormente, são desenvolvidos três modelos GEE, utilizando a divisão para treino e teste selecionada na fase anterior, considerando diferentes covariáveis. O diferente conjunto de covariavéis selecionadas em cada um dos modelos tem como objetivo diferenciar a análise e a interpretação da inflência das mesmas na variável resposta. Os modelos desenvolvidos apresentam a seguinte estrutura:

- Modelo Total
  - **Variável Resposta:** PEX;
  - **Covariáveis:** Idade, Sexo, Lat, Dilpp, Internos, LIO_Dp, Local da LIO, Complicações;

- Modelo de Caracteristicas Pessoais
  - **Variável Resposta:** PEX;
  - **Covariáveis:** Idade, Sexo, Lat;
- Modelo de Fatores Oculares
  - **Variável Resposta:** PEX;
  - **Covariáveis:** Local da LIO, Dilpp, LIO_Dp, Internos, Complicações;

Os modelos desenvolvidos, à semelhança da fase anterior descrita, serão validados de modo a selecionar o melhor modelo para análise GEE e interpretação dos seus resultados. As métricas para validação são as utilizadas na fase anterior, ou seja, *accuracy* e *AUC*.

Para complementar a análise e interpretação do modelo GEE, selecionado e validado a partir das duas fases anteriores, serão gerados intervalos de confiança das estimativas geradas para cada uma das covariáveis de interesse, assim como será adicionado o valor correspondente ao teste de Wald.

5

APLICAÇÃO SHINY PARA ANÁLISE E VALIDAÇÃO DE MODELOS GEE

5.1 SAGA: *shiny application for gee analysis*

De modo a automatizar e simplificar a utilização de modelos GEE, bem como, realizar a sua validação e análise dos seus resultados foi desenvolvida uma aplicação utilizando o *package Shiny* do *R*. A aplicação é denominada de *SAGA* que corresponde à soma das letras iniciais de *Shiny Application for GEE Analysis*.
Esta aplicação tem como principal objetivo a formação de modelos GEE a partir de um conjunto de dados fornecido pelo utilizador permitindo visualizar o seus resultados e validar os mesmos considerando a análise *ROC*. A linguagem utilizada será o inglês de modo a facilitar a sua utilização entre vários utilizadores. O código desenvolvido para a aplicação encontra-se no seguinte link do *GitHub:* `https://github.com/LuisMoncaixa1996/Thesis/tree/main/Shiny_App`.
A aplicação divide-se em 3 fases conforme descrito na Fig.16 sendo essas fases:

A. Introdução e edição do conjunto de dados;

B. Seleção das variáveis para o modelo GEE;

C. Validação e resultados dos modelos gerados;

No decorrer deste capítulo serão descritos todos os menus e secções da aplicação SAGA, de modo a exemplificar cada uma das fases foi utilizado um conjunto de dados disponível online no seguinte link `http://static.lib.virginia.edu/statlab/materials/data/depression.csv`. Este conjunto de dados demonstra o efeito de dois diferentes medicamentos ao longo do tempo para uma determinada patologia.

5.2 SAGA: INTRODUÇÃO DO CONJUNTO DE DADOS

Previamente à construção dos modelos GEE, será necessário por parte do utilizador descarregar o conjunto de dados e, se necessário, efetuar algumas alterações no mesmo para posterior desenvolvimento dos modelos GEE. Esta primeira fase da aplicação será executada no menu *"Import Data"*, este menu divide-se em duas etapas distintas conforme representado na Fig. 17:

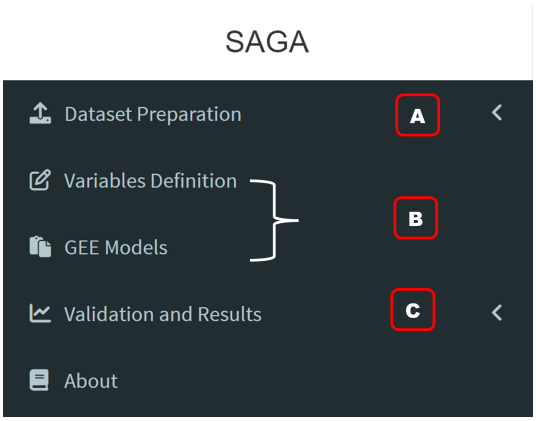


Figura 16: Menu da aplicação SAGA.

- ***"Upload Dataset"***, menu responsável por descarregar o dataset e preparar a sua estrutura;
- ***"Data Changes"***, menu utilizado para efetuar algumas alterações no *dataset* descarregado anteriormente (opcional);

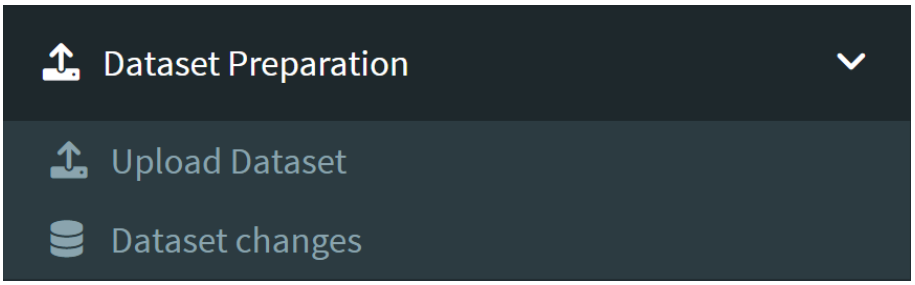


Figura 17: Representação dos menus de introdução e edição do conjunto de dados.

5.2.1 *SAGA: Upload Dataset*

Neste menu o utilizador irá introduzir o conjuto de dados executando o botão "*Browse...*", uma nova janela irá surgir de modo a ser possível selecionar o conjunto de dados encontrando-se este limitado apenas a ficheiros EXCEL, com a extensão *.xlsx*, e CSV , *.csv*, qualquer outro tipo de estrutura de dados irá gerar um erro quando importado. O conjunto de dados irá surgir na janela de fundo. Para além da descarregar o conjunto de dados também existem algumas funcionalidades que permitem visualizar melhor o *dataset* conforme representados na Fig.18.

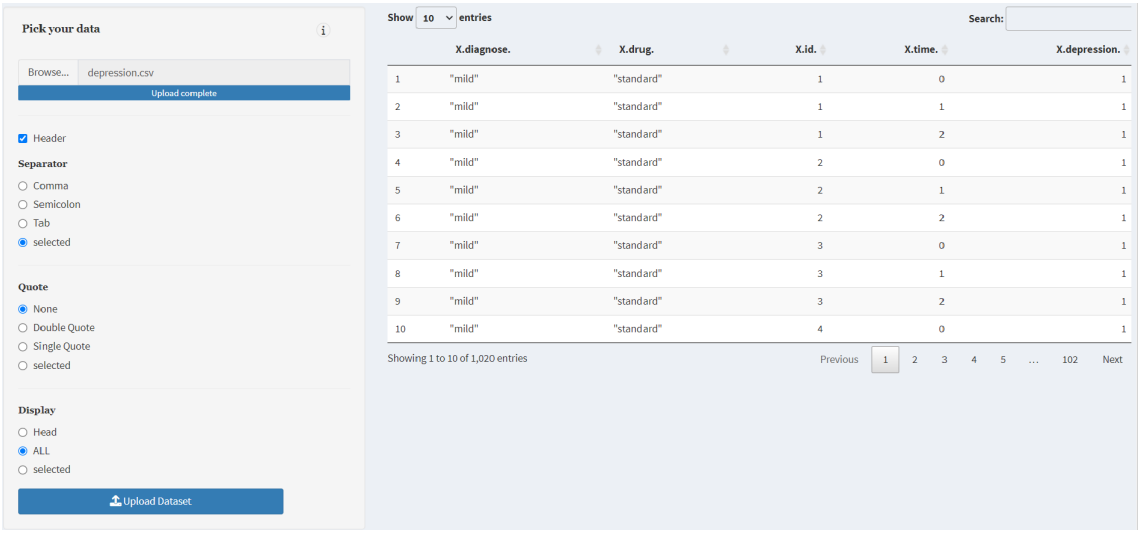


Figura 18: Processo de importação de um ficheiro *.csv*.

5.2.2 *SAGA: Data Changes*

Após a importação do conjunto de dados existem, por vezes, algumas alterações que podem ser feitas no conjunto de dados, nomeadamente o nome das respetivas colunas existentes. No caso dos modelos GEE, deve-se evitar nomes de columas que contenham espaços entre os nomes, muitas das vezes esses nomes geram erros quando a análise do modelo é realizada. A alteração dos nomes das columas poderá ser realizada na secção "*Columns Names*", nesta secção estão representados todos os nomes das columas existentes no conjunto de dados descarregado anteriormente e uma caixa de texto onde será introduzido o novo nome da coluna. De modo a efetuar a alteração será selecionada a columa desejada e colocar o seu novo nome na caixa de texto, a alteração será efetuada automaticamente. Na Fig.19 está representado um exemplo de utilização desta secção.

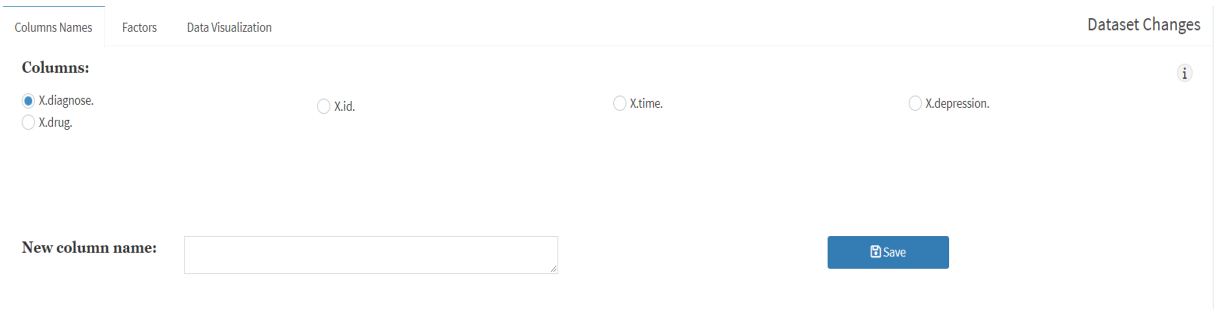


Figura 19: Exemplo de alteração do nome de uma coluna na secção "Columns Names"

Para além de alterar os nomes das colunas do conjunto de dados o utilizador poderá fatorizar as colunas cujo os seus valores sejam variáveis categóricas. Com a fatorização estes valores serão divididos por níveis, um nível para cada variável correspondente na coluna selecionada, facilitando assim a sua utilização no desenvolvimento

dos modelos GEE. Este processo pode ser realizado na secção denominada de "*Factors*", de modo similiar à secção anterior, estão apresentadas todas as columas do conjunto de dados descarregado e o utilizador terá que selecionar qual a coluna que deseja fatorizar. Este processo será automáticamente executado assim que o botão seja pressionado.

Por fim o conjunto de dados com as alterações já efetuadas poderá ser visualizado na secção "*Data Visualization*"concluindo assim a primeira fase da aplicação. Na Fig.20 está representado o conjunto de dados apresentado na secção 5.2.1 com algumas alterações nos nomes das colunas.

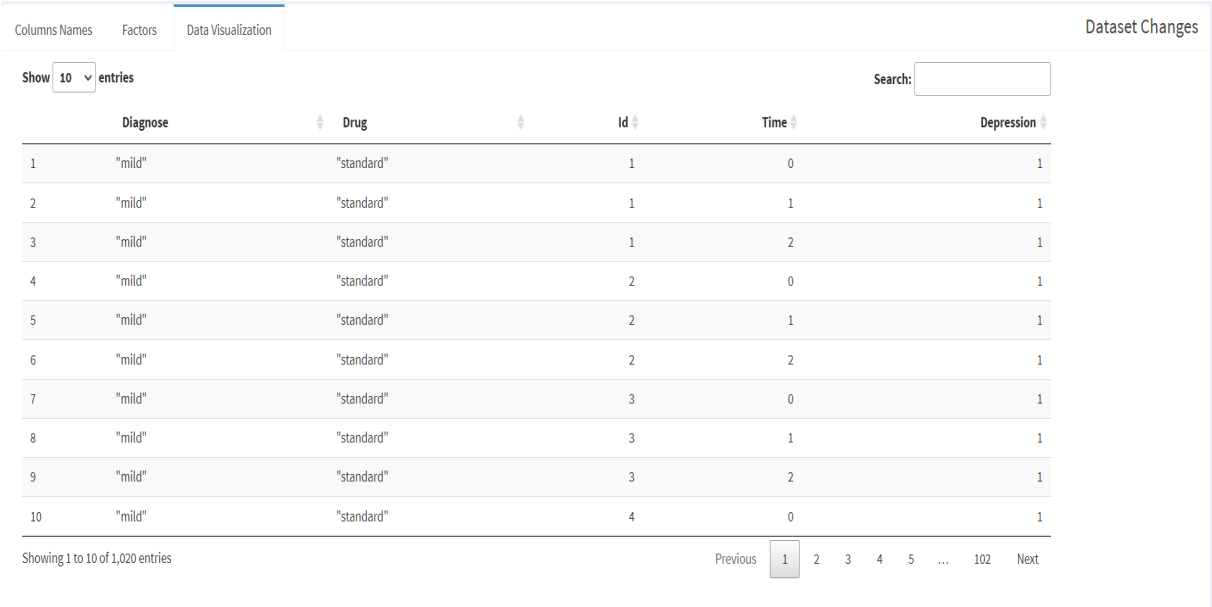


Figura 20: *Dataset* com alterações no nome das colunas.

5.3 SAGA: SELECÇÃO DE VARIÁVEIS PARA OS MODELOS GEE

Como descrito anteriormente, o menu de selecção de váriaveis corresponde à fase 2 da aplicação, sendo este um dos mais fundamentais, pois serão aqui definidas todas as variáveis necessárias para o desenvolvimento dos modelos GEE. Conforme descrito durante a secção 2.2 serão definidas nesta fase:

- Variável Resposta;
- Covariáveis;
- Coluna de Identificação;
- Família de Distribuição;
- Estrutura de Correlação;

5.3.1 *SAGA: Variables*

A selecção das variáveis inicia-se com a seleção da variável resposta do modelo GEE e as suas respetivas covariáveis, na seleção da variável resposta estão presente todas as colunas correspondentes ao *dataset*,descarregado anteriormente, e será selecionada a coluna que corresponde à variável resposta de interesse. A selecção das covariáveis e a relação entre elas poderá ser definida de duas formas:

- **Seleção das covariáveis:** Representado como **A** na Fig.21. Serão apresentadas, à semelhança da variável resposta, todas as colunas do *dataset* sendo que o utilizador poderá selecionar as colunas que desejar, através deste processo de seleção a relação das cováriaveis entre si será de incrementação. (Exemplo: Idade + Sexo)
- **Introdução da relação das covariáveis:** Representado como **B** na Fig.21. Outro método de seleção das covariáveis será por introdução das mesmas e da sua relação entre si, o utilizador deve introduzir o nome das colunas de interesse e especificar qual a relação entre as covariáveis selecionadas.

A secção de seleção da variável resposta e covariáveis encontra-se representada na Fig.21.

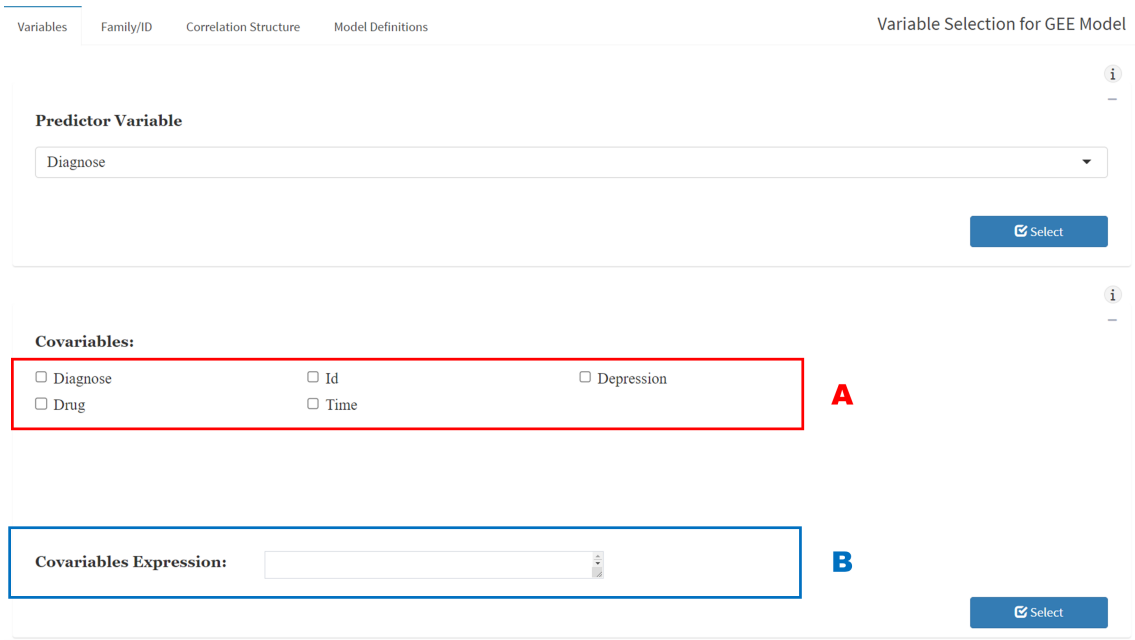


Figura 21: Seleção da variável resposta e processos de selecção das covariáveis.

O utilizador apenas poderá escolher apenas um dos processos de seleção das covariáveis, caso utilize ambos os processos será assumido o processo de Introdução. Após o clique no botão "Select"na variável resposta e covariáveis estas serão automaticamente adicionadas ao modelo sendo possível seguir para a próxima secção.

5.3.2 *SAGA: Family / ID*

Nesta secção são definidas a família de distribuição da variável resposta e a coluna correspondente à identificação da amostra/observação. O processo de seleção da coluna de identificação é idêntico ao descrito para a variável resposta, ou seja, serão apresentadas todas as colunas do *dataset* descarregado, onde o utilizador deverá selecionar a coluna de interesse.
A selecção da família de distribuição procede-se de forma semelhante à selecção da coluna de identificação, sendo possível selecionar a familia de distribuição que mais se identifica com a variável resposta de entre oito possíveis famílias de distribuição disponíveis. De modo semelhante à secção anterior, após o clique no botão na seleção da família de distribuição e da coluna de identificação estas informações serão adicionadas automaticamente ao modelo GEE.

5.3.3 *SAGA: Correlation Structure*

A selecção da estrutura de correlação será a última etapa de desenvolvimento do modelo GEE, o utilizador poderá selecionar uma das três estruturas de correlação disponíveis sendo elas:

- Estrutura de correlação Independente;
- Estrutura de correlação Intercambiável;
- Estrutura de correlação Autoregressiva;

Estas estruturas encontram-se descritas na secção 2.2.2. Na situação de o utilizador desconhecer qual a estrutura de correlação que mais se identifica com o modelo, existe a opção de visualizar os valores de QIC, selecionando a opção "*Display QIC values*". Através desta opção será apresentada uma tabela com os respetivos valores de QIC e QICu correspondentes a cada estrutura de correlação, considerando as variáveis anteriormente selecionadas, conforme representado na Fig.22.

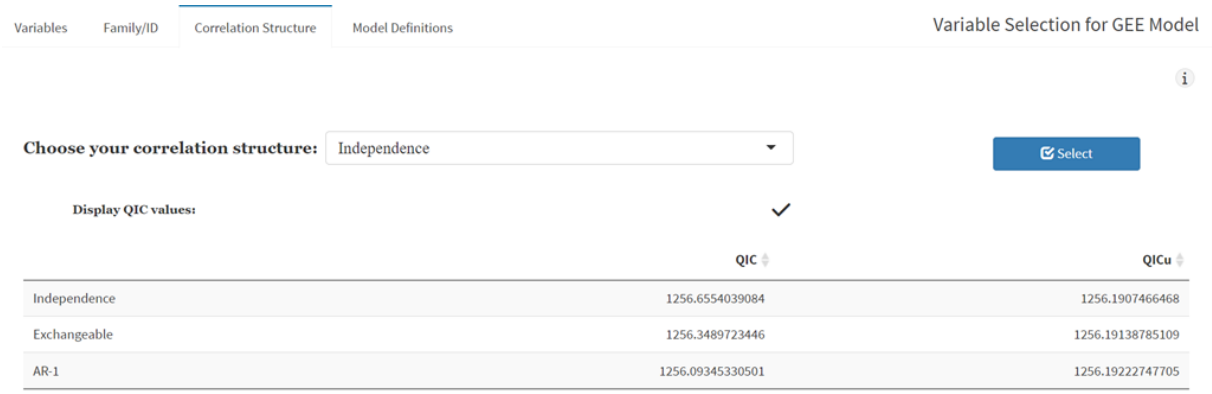


Figura 22: Seleção da estrutura de correlação utilizando os valores de QIC.

O critério de seleção da estrutura de correlação utilizando os valores de QIC encontra-se descrito na secção 2.2.3. Esta opção só pode ser utilizavél se todas as variáveis anteriormente descritas estiverem selecionadas. A estrutura de correlação selecionada será automáticamente adicionada ao modelo.

5.3.4 *SAGA: Model definitions*

Nesta última secção será atribuido um nome ao novo modelo GEE desenvolvido e uma breve descrição do modelo. Esta secção não irá influênciar o modelo desenvolvido, apenas será necessário para rotular os resultados gerados. Após terminados todos os processos anteriores, o modelo gerado poderá ser observado no menu *GEE Models*.

5.4 SAGA: GEE MODELS

Neste menu apresentam-se os modelos GEE criados descrevendo as variáveis, descritas anteriormente, selecionadas. Apresentam-se 3 funcionalidades neste menu:

- **Add:** Esta funcionalidade permite construir um novo modelo GEE adicional ao primeiro modelo criado. Quando executado o botão, o utilizador irá voltar ao processo de selecção de variáveis, na secção *Model Definitions*, pelo que deverá selecionar como secção inicial a "*Variables*"e continuar o procedimento descrito anteriormente, o novo modelo será adicionado ao menu quando executado o botão "*Select*"na secção de "*Model definitions*. Existe um limite de 3 modelos criados, sendo que o botão ficará inativo após a criação desse número de modelos.

- **Edit:** De modo a executar este botão, um dos modelos criados deve estar selecionado, o botão estará inativo quando existirem dois ou mais modelos selecionados. Ao executar este botão, o utilizador poderá alterar as variáveis que selecionou no procedimento anterior. A alteração do modelo efetua-se quando o botão "*Save*"na secção *Model definitions* for executado pelo que sugere-se que se verifique todas as secções correspondentes às variáveis anteriormente descritas.

- **Run:** A execução deste botão ocorre quando estão selecionados um ou mais modelos. Através deste botão serão consideradas as variáveis selecionadas pelo utilizador, não existindo possibilidade de alteração das mesmas. As variáveis consideradas permitem a realização da análise GEE, bem como a sua utilização na fase de validação ROC, que pode ser visualizada numa secção posterior.

Na Fig.23 encontra-se um exemplo da visualização de três modelos criados bem como os três botões com as funcionalidades descritas anteriormente, no menu *GEE Models*.

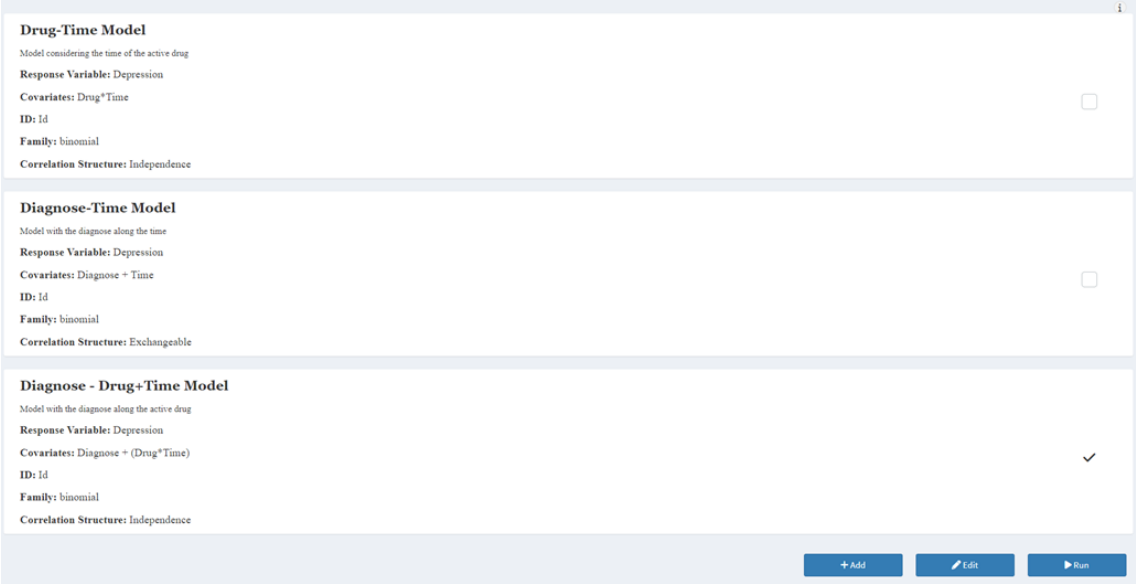


Figura 23: Representação de três modelos com as suas variáveis como exemplo.

5.5 SAGA: VALIDAÇÃO E RESULTADOS DOS MODELOS GEE DESENVOLVIDOS

Após o desenvolvimento dos modelos GEE por parte do utilizador e posterior execução do botão "*Run*"descrito anteriormente, atinge-se a fase final da aplicação SAGA.
A fase final da aplicação consiste na validação dos modelos desenvolvidos através da análise *ROC* e análise aos resultados obtidos para cada modelo GEE.

5.5.1 *SAGA: ROC Analysis*

A validação através da análise *ROC* dos modelos desenvolvidos, é uma funcionalidade adicional e opcional, ou seja, o utilizador não necessita obrigatóriamente de executar a validação para aceder aos resultados da análise GEE. No entanto através desta validação o utilizador pode obter informação adicional sobre o modelo, ou comparar os mesmos através das métricas obtidas durante a análise. Por via da validação *ROC* serão obtidos resultados gráficos, sendo estes as curvas *ROC* geradas e o valor de duas métricas, *accuracy* e àrea abaixo da curva (descritas na secções 4.1.1 e 4.1.2 respetivamente). Esta fase de validação encontra-se no menu "*Validation and Results*".

SAGA: Data Split

Previamente à análise *ROC*, o utilizador deverá selecionar uma divisão possível para o seu conjunto de dados, de modo a possibilitar a validação *ROC*. O utilizador pode optar por dividir o conjunto de dados em treino e teste considerando três possibilidades, 70% treino - 30% teste, 75% treino - 25% teste, 80% treino e 20% teste. Existe também a divisão pré-definida para o caso de o utilizador não definir uma divisão específica. De modo a

apresentar uma validação mais precisa e rigorosa, a divisão do conjunto de dados definida pelo utilizador será utilizada para todos os modelos selecionados na secção anterior, permitindo assim ao utilizador verificar através dos resultados obtidos qual o/os modelos mais significativos e representativos para o caso de estudo. Após definida a divisão do conjunto de dados o utilizador deve selecionar o botão "*Run ROC*".

SAGA: ROC Curves

Inicializada a análise *ROC* na secção anterior, nesta fase são apresentadas as representações gráficas das diferentes curvas *ROC* geradas, para cada um dos modelos selecionados nas secções anteriores. Abaixo de cada uma das curvas geradas encontra-se um botão de "*download*"onde o utilizador poderá descarregar para o seu computador local as curvas obtidas, o ficheiro encontra-se no formato "*.png*". Na Fig.24 estão representadas as respetivas curvas *ROC* considerando três modelos diferentes e os respetivos botões de *download*.

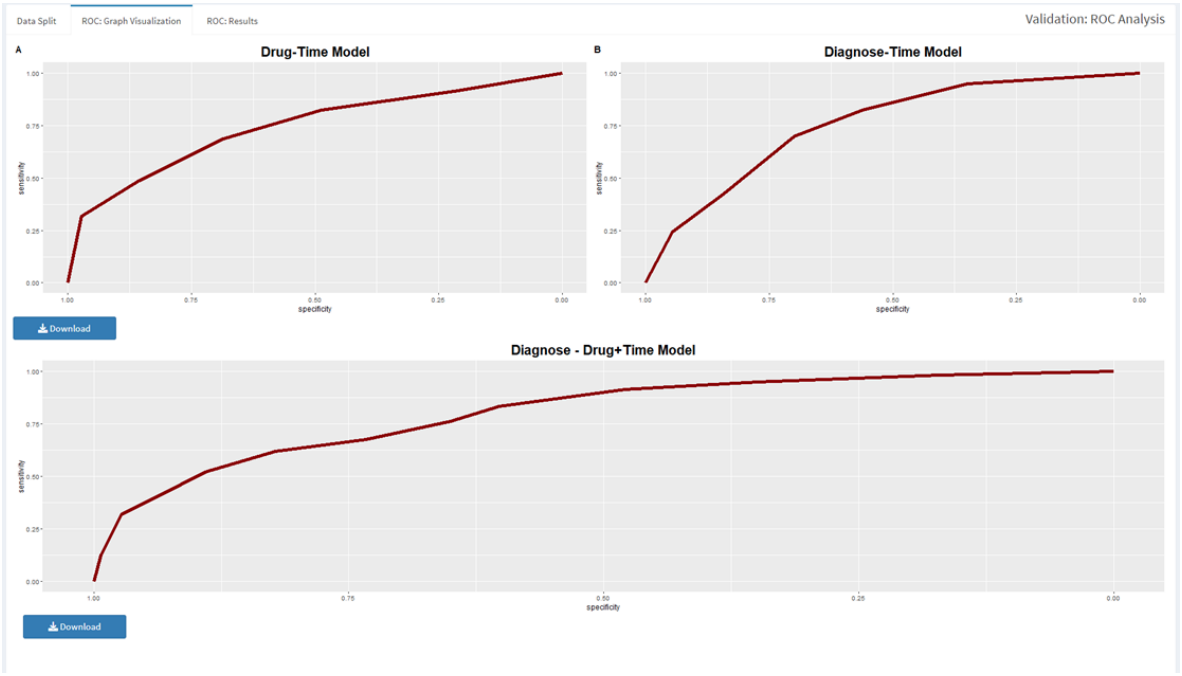


Figura 24: Representação gráfica das diferentes curvas ROC de três modelos diferentes.

SAGA: ROC Results

De modo complementar à representação gráfica das curvas ROC geradas para os diferentes modelos, nesta secção são apresentados os resultados obtidos na análise *ROC* considerando as duas métricas referenciadas nas secções anteriores, *Accuracy* e *AUC*. Para cada modelo selecionado anteriormente serão apresentados os valores das métricas descritas, permitindo ao utilizador verificar, através dos resultados obtidos, o modelo que apresenta melhores resultados de validação.

5.5.2 *SAGA: GEE Results*

Os resultados obtidos nos modelos criados nas secções anteriores, são apresentados nesta secção. Para cada modelo desenvolvido será apresentada uma tabela que descreve os resultados obtidos da análise GEE, na tabela obtida poderá ser analisada a influência das covaráveis para com a variável resposta, sendo acompanhada por valores de intervalos de confiança e o teste de Wald. Estes resultados permitem ao utilizador realizar uma análise rigorosa sobre o modelo desenvolvido ao longo de todas as etapas. A tabela de resultados poderá ser descarregada para o computador pessoal no formato *.csv*, permindo o botão de *download* para visualização/análise posterior por parte do utilizador. Na Fig. 25 encontra-se representada a tabela com os resultados obtidos de um dos modelos criados ao longo da descrição de todas as secções anteriores.

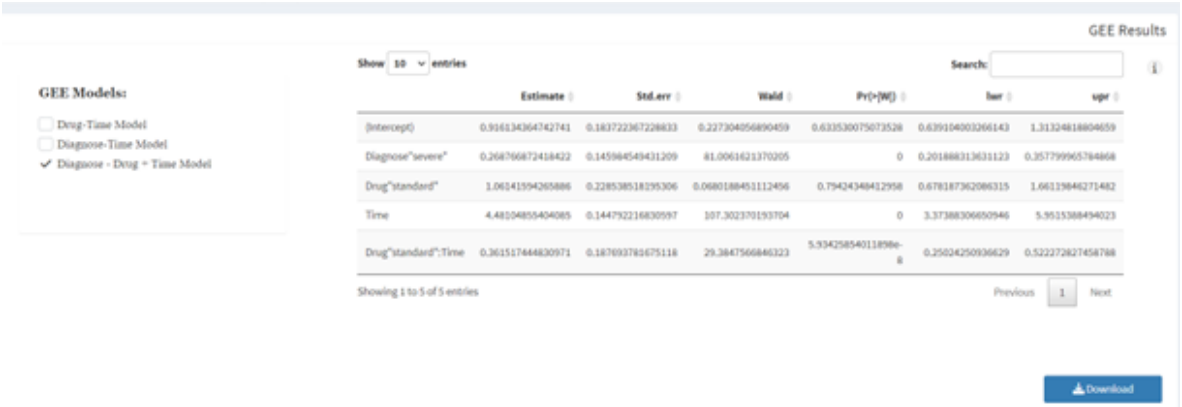


Figura 25: Resultados obtidos para um dos modelos criados nas secções anteriores.

6

ANÁLISE DOS RESULTADOS

O caso de estudo, descrito no Capítulo 3, encontra-se no formato .xls e será utilizado para desenvolver os modelos GEE, segundo o processo referenciado no Capitulo 4, bem como validar esses mesmos resultados utilizando a aplicação *Shiny* desenvolvida. Durante o desenvolvimento dos modelos GEE, foi desenvolvido um script em *R* utilizando dois *packages* de interesse, GEE e *geepack*, de forma a conseguir identificar as suas diferenças e obter as métricas da análise GEE para interpretação dos resultados. A validação dos modelos por a análise *ROC* será efetuada por o *package ROCR* e apenas realizada para os modelos GEE gerados com o *package geepack*, assim como também será utilizado o método de QIC para este tipo de modelos de modo a verificar a melhor estrutura de correlação. Posteriormente, e de modo a identificar a eficiência da aplicação Shiny desenvolvida, o caso de estudo foi aplicado através da aplicação SAGA sendo comparados os resultados obtidos na aplicação e no *script* desenvolvido.

6.1 DIVISÃO TREINO-TESTE DO DATASET E VALIDAÇÃO

Conforme descrito no capítulo 4.1, foram definidas três possibilidades de divisão do conjunto de dados para treino e teste. Em cada uma das possibilidades foi gerado, para o conjunto de treino, um modelo GEE considerando a variável resposta como sendo a PEX, cujo o valor 0 representa a ausência da patologia e o valor 1 a presença da mesma, e como covariáveis a Idade, Sexo, Lat, Dilpp, Local da LIO , Complicações e Internos.

Para auxiliar a seleção da estrutura de correlação que melhor se ajusta ao modelo GEE, foi obtido o valor de *QIC* para cada uma das estruturas avaliadas, sendo a estrutura que apresentar um valor de *QIC* mais baixo a melhor estrutura. Para além do valor de *QIC*, também será apresentado o valor de *QICu* que pode diferenciar as diferentes estruturas caso os valores de *QIC* sejam semelhantes. Este processo de seleção foi aplicado em um modelo GEE considerando todas as covariáveis descritas anteriormente. Os resultados obtidos encontram-se descritos na Tabela 5.

A partir dos resultados obtidos quanto à estrutura de correlação, verifica-se que os valores de *QIC* são muito semelhantes em todas as estruturas, no entanto, a estrutura de correlação independente apresenta um valor ligeiramente inferior. Posto isto a estrutura selecionada para o desenvolvimento dos modelos GEE será a estrutura de correlação independente.

Tabela 5: Valores de *QIC* e *QICu* aplicados a cada estrutura de correlação.

	Independente	Intercambiável	AR-1
QIC	405,037	405,045	405,045
QICu	433,758	433,758	433,758

Posteriormente á selecção da estrutura de correlação, foi avaliada através da análise ROC qual a divisão do conjunto de dados que permite obter melhores resultados nos modelos GEE por desenvolver. As divisões do conjunto de dados avaliadas são as descritas na secção anterior. A Tabela 6 apresenta os resultados das métricas de validação dos modelos, *accuracy* e *AUC* obtidas na análise ROC aplicada ao modelo GEE considerando todas as covariáveis descritas.

Tabela 6: Métricas de validação dos modelos considerando diferentes divisões do dataset.

Modelos	ACC	AUC
Modelo 70/30	0,934	0,746
Modelo 75/25	0,940	0,727
Modelo 80/20	0,930	0,767

Analisando os resultados das métricas apresentadas, foi definido como melhor modelo aquele cujo a divisão é de 80% para treino e 20% para teste, uma vez que, apesar de se verificar que a métrica de *ACC* é superior no modelo 70% treino e 30% teste, a diferença registada no valor da *AUC* é maior do que no modelo com divisão de 80% treino e 20% teste.

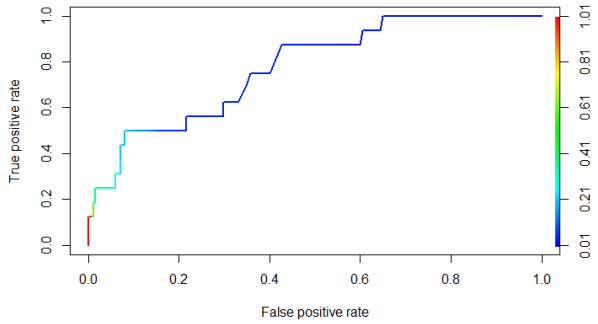
6.2 MODELOS GEE CONSIDERANDO DIFERENTES COVARIÁVEIS

Os modelos GEE descritos na secção 4.2 foram desenvolvidos considerando a divisão do dataset como 80% treino e 20% teste e a sua estrutura de correlação definida como independente, tal como foi explicado durante a análise na secção 6.1. Estes modelos foram validados considerando as mesmas métricas aplicadas na secção 6.1. Os resultados obtidos da validação nos diferentes modelos encontram-se descritos na Tabela 7.

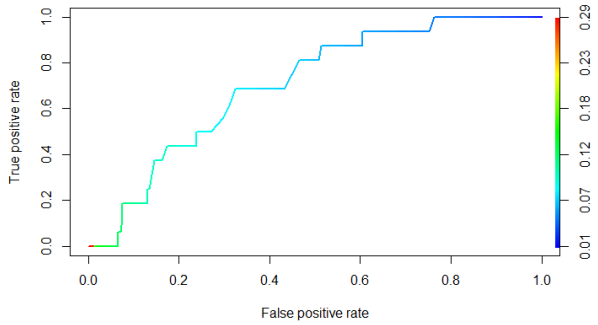
Tabela 7: Métricas de validação para os modelos GEE desenvolvidos.

Modelos	ACC	AUC
Modelo Total	0,937	0,776
Modelo CP	0,920	0,706
Modelo FO	0,930	0,660

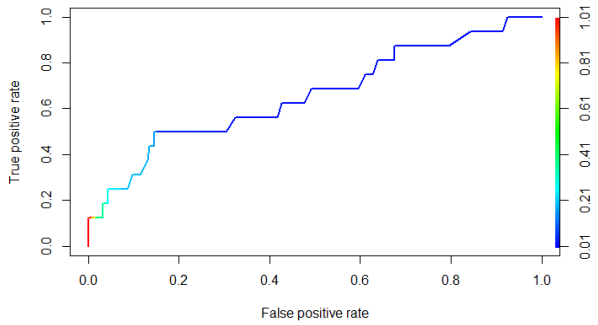
Adicionalmente ás metricas de validação, e de modo a complementar a interpretação dos resultados e seleção dos modelos, foram geradas as curvas ROC correspondentes a cada modelo. Na Fig.26 estão representadas cada uma dessas curvas para cada modelo GEE desenvolvido.



(a) Modelo Total



(b) Modelo Características Pessoais



(c) Modelo Fatores Oculares

Figura 26: Representação gráfica das curvas *ROC* geradas.

A análise das métricas de validação permitem concluir que o melhor modelo gerado baseia-se no modelo que considera ambas as váriaveis correspondentes às características pessoais e fatores oculares como covariáveis, indicando que estas covariáveis têm mais influência sobre a variável resposta.

6.3 ANÁLISE GEE DO MODELO CONSIDERANDO O MODELO TOTAL

As primeiras duas fases permitiram selecionar o melhor modelo GEE, sendo o modelo que considera todas as covariáveis. De modo a identificar diferenças nos resultados obtidos na análise GEE do modelo, este foi analisado utilizando os dois *packages* descritos anteriormente. Uma vez que existem fatores nas covariáveis, "Local da LIO", "Internos" e "Complicações" que apresentam pouca variabilidade nas suas observações, não serão consideradas para análise do modelo utilizando o *package* gee, uma vez que resulta em diversos erros de aplicação. No contexto do caso de estudo serão analisados os resultados do modelo GEE de modo a identificar a influência que as covariáveis têm sobre a variável resposta. Todas as estimativas e intervalos de confiança são exponenciados de forma a obter os *odd ratios*. Nas tabelas 8 e 9 estão representados os resultados obtidos do modelo segundo os diferentes *packages*.

Tabela 8: Modelo GEE utilizando o *package* gee

	Estimativas	Naive S.E	Naive Z	Robust S.E	Robust Z	IC - 95%
Intercept	0.0004	1,4380	-5,427	1,2103	-6,448	[3,81e-05 - 0,0044]
Idade	1,06	0,0169	3,183	0,0159	3,380	[1,02 - 1,0886]
Sexo	1,94	0,2900	2,287	0,2913	2,277	[1,10 - 3,440]
Dilpp (>=7-8)	7,08	0,3216	6,086	0,3207	6,101	[3,77 - 13,2691]
Dilpp ($\geq$6-7)	12,2	0,4213	5,945	0,4200	5,962	[5,37 - 27,8771]
Dilpp ($\geq$5-6)	17,5	0,5839	4,900	0,5727	4,997	[5,69 - 53,7206]
Lat (OE)	0,835	0,2686	-0,673	0,2756	-0,656	[0,486 - 1.4324]
LIO Dp	0,982	0,0356	-0,522	0,0319	-0,584	[0,922 - 1.0449]

Através dos resultados obtidos do modelo GEE, utilizando o *package* gee, verifica-se que a semelhança de valores entre os erros-padrão robusto e naive demonstra que a estrutura de correlação selecionada é adequada para o modelo GEE. A partir dos intervalos de confiança é possível analisar que a dilatação pupilar apresenta-se como uma covariável influente para a variável resposta, sendo esta influência positiva, conforme verificado através dos valores de estimativa positivos, ou seja, à medida que a gravidade na classe de dilatação aumenta, maior é a probabilidade de desenvolvimento de PEX. Note-se a classe de referência é Dil $\geq$ 8. Assim, verifica-se que a dimensão da dilatação observada influência na presença da pseudoexfoliação.

Tabela 9: Modelo GEE utilizando o *package* geepack

	Estimativas	Erro-Padrão	Wald-Test	Pr(>|W|)	IC - 95%
Intercept	0,000399	1,2242	41,0000	1,49e-10	[3,56e-05 - 4,32e-03]
Idade	1,05	0,0152	10,100	$\mathbf{0,001}$*	[1,02 - 1,08]
Sexo	1,77	0,3022	3,5000	0,059	[0,978 - 3,20]
Lat (OE)	0,941	0,2832	0,0461	0,830	[0,540 - 1,640]
LIO DP	1,000	0,0277	1,09e-03	0,974	[0,948 - 1,06]
Local da LIO (Sulco)	2,15	0,7856	0,9520	0,329	[0,462 - 10,000]
Local da LIO (CA)	**9,65e+17**	1,3457	946,8802	0,0	[6,90e+16 - 1,35e+19]
Local da LIO (S/LIO)	**4,60e-17**	1,9434	374,7035	0,0	[1,02e-18 - 2,07e-15]
Dilpp ($\geq$7-8)	7,08	0,3301	35,1415	$\mathbf{3,07e-09}$*	[3,71 - 13,5]
Dilpp ($\geq$6-7)	12,4	0,4438	32,2105	$\mathbf{1,38e-08}$*	[5,20 - 29,6]
Dilpp ($\geq$5-6)	17,0	0,6142	21,2715	$\mathbf{3,99e-06}$*	[5,10 - 56,6]
Internos (S)	0,877	0,4464	0,0861	0,769	[0,366 - 2,10]
Comp DZ (S)	10,8	0,8027	8,7671	$\mathbf{3,07e-03}$*	[2,23 - 51,90]
Comp Outra (S)	14,7	0,7482	12,9163	$\mathbf{3,26e-04}$*	[3,40 - 63,8]
Comp VA (S)	0,365	1,3729	0,5400	0,462	[2.47e-02 - 5,38]
Comp RCP (S)	4,19	0,8333	2,9544	0,086	[0,818 - 21,4]

* Valor estatisticamente significativo para $\boldsymbol{\alpha = 0,05}$

Neste modelo, utilizando o *package geepack*, não são apresentados os erros-padrão Naive e Robusto sendo apenas demonstrado o erro-padrão *standard*, no entanto é apresentado o valor do teste de Wald de forma a complementar a análise. Comparativamente ao modelo anterior é mais uma vez identificada a covariável correspondente à dilatação pupilar como sendo uma covariável influente na presença da pseudoexfoliação, no entanto, adicionalmente à dilatação pupilar, as covariáveis correspondentes à idade e complicações DZ (deiscência zonular) e outro tipo de complicação existentes durante a cirurgia de remoção de cataratas, também apresentam influência positiva na presença da pseudoexfoliação.

Verifica-se ainda que as estimativas apresentadas para as diferentes classe do local da lente intraocular, levam a crer que este parâmetro se encontra sobrestimado, não devendo ser tido em consideração por falta de informação.

6.4 COMPARAÇÃO ENTRE GEE *package* E GEEPACK *package*

Os dois *packages* utilizados para análise de modelos GEE conseguiram identificar a mesma covariável que inflência positivamente a presença de pseudoexfoliação, no entanto, o verificou-se que o package *geepack* identificou outras covariáveis que podem influenciar a presença da patologia. Os resultados obtidos para as covariáveis comuns nos dois *packages* apresentaram resultados semelhantes. Conforme descrito no desenvolvimento e validação dos modelos gerados, existiram algumas semelhanças entre os dois *packages* utilizados para desenvolvimento dos modelos GEE, no entanto estes também apresentam algumas diferenças. Para o caso

de estudo utilizado os dois *packages* corresponderam com a análise GEE realizada demonstrando o mesmo resultado em ambos os modelos.

Tabela 10: Comparação entre os *packages* gee e geepack.

	gee	geepack
Estruturas de Correlação	7	5
Método QIC	✗	✓
Dados Omissos	✓	✗
Familias	8	8
Objeto para validação (ex.Análise ROC)	✗	✓
Teste de Wald	✗	✓
Erros-Padrão Robusto e Naive	✓	✗

6.5 SAGA: APLICAÇÃO AO ESTUDO DA PSEUDOEXFOLIAÇÃO OCULAR

De modo a validar os resultados obtidos e a eficiência da aplicação SAGA desenvolvida, foi aplicado o caso de estudo, utilizado para comparar os diferentes *packages* de desenvolvimento e análise dos modelos GEE, através da aplicação. Ao longo do processo de construção dos modelos GEE foram utilizadas as mesmas covariáveis, estruturas de correlação e outros fatores que as definidas no *script* de comparação e estudo dos *packages*. Inicialmente na construção dos modelos foram analisados os valores obtidos de *QIC* e *QICu* para a selecção das estruturas de correlação em cada um dos modelos desenvolvidos. O quadro seguinte compara os valores obtidos na aplicação SAGA e no script anteriormente descrito.

Tabela 11: Comparação de métricas de selecção da estrutura de correlação.

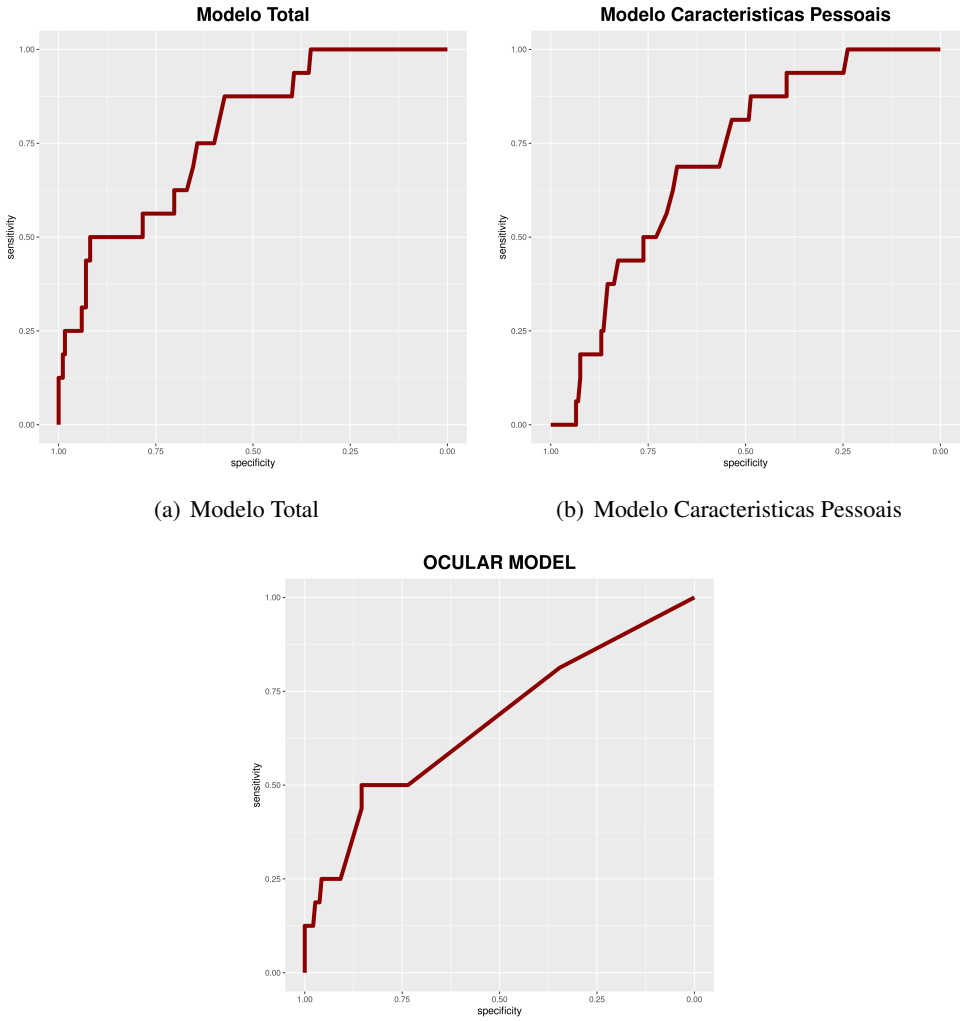
Métricas de seleção da EC	SAGA			Script		
	Ind	Exch	AR-1	Ind	Exch	AR-1
QIC	405,042	405,046	405,046	405,037	405,045	405,045
QICu	433,222	433,222	433,222	433,758	433,758	433,758

A validação dos modelos desenvolvidos por análise *ROC* ocorreu utilizando a divisão 80% treino - 20% teste e foram obtidos os seguintes resultados, conforme descrito na tabela 12:

Tabela 12: Métricas de validação dos modelos GEE na aplicação e *script* desenvolvido.

Métricas *ROC*	Modelo Total		Modelo CP		Modelo FO	
	SAGA	Script	SAGA	Script	SAGA	Script
ACC	0,930	0,937	0,920	0,920	0,930	0,930
AUC	0,785	0,776	0,706	0,706	0,666	0,666

As curvas *ROC* geradas para cada um dos modelos encontram-se descritas na Fig. 27.



(a) Modelo Total

(b) Modelo Caracteristicas Pessoais

(c) Modelo Fatores Oculares

Figura 27: Representação gráfica das curvas *ROC* geradas para cada um dos modelos GEE.

Na fase de validação através da análise *ROC* podemos verificar uma ligeira diferença nos valores das métricas de validação em um dos modelos desenvolvidos, no entanto, em ambos os casos o modelo que considera todas

as covariáveis, sejam pessoais ou fatores oculares, continua a ser o modelo que apresenta melhores resultados de validação, continuando a ser este o modelo a analisar na análise GEE. Relativamente às curvas *ROC* geradas verifica-se que nas diferentes abordagem não existiu qualquer modificação na estrutura da curva, demonstrando assim uma ambiguidade em ambas as abordagens para a validação *ROC*.

Finalmente na análise GEE para o modelo com os melhores resultados na validação *ROC*, modelo considerando as covariáveis oculares e pessoais, verificou-se uma ligeira subida na ordem decimal dos valores obtidos utilizando a aplicação SAGA. Os resultados obitdos encontram-se descritos na tabela 13.

Tabela 13: SAGA: Resultados da análise GEE

	Estimativas	Erro-Padrão	Wald-Test	Pr(>|W|)	IC - 95%
Intercept	0,000399	1,253	39,005	4,23e-10	[3,42e-05 - 0,005]
Idade	1,049	0,0147	10,700	0,0010	[1,020 - 1,080]
Sexo	1,769	0,3013	3,582	0,0584	[0,980 - 3,192]
Lat (OE)	0,941	0,2829	0,0464	0,8295	[0,540 - 1,638]
LIO DP	1,000	0,0277	1,09e-03	0,974	[0,948 - 1,06]
Local da LIO (Sulco)	2,152	0,7856	0,9520	0,3292	[0,46 - 10,04]
Local da LIO (CA)	9,65e+17	1,3457	946,8749	0,0	[6,90e+16 - 1.35e19]
Local da LIO (S/LIO)	4,60e-17	1,9434	374,7035	0,0	[1,012e-18 - 2,07e-15]
Dilpp ($\geq$7-8)	7,078	0,3301	35,1415	3,07e-09	[3,706 - 13,519]
Dilpp ($\geq$6-7)	12,414	0,4438	32,2105	1,38e-08	[5,20 - 29,63]
Dilpp ($\geq$5-6)	16,999	0,6142	21,2715	3,99e-06	[5,10 - 56,62]
Internos (S)	0,877	0,4464	0,0861	0,7692	[0,37 - 2,10]
Comp RCP (S)	4,188	0,8333	2,9544	0,0856	[0,82 - 21,45]
Comp DZ (S)	10,769	0,8027	8,7671	0,0031	[2,23 - 51,93]
Comp VA (S)	0,365	1,3729	0,5400	0,4625	[0,02 - 5,38]
Comp outra (S)	14,718	0,7482	12,9163	0,0003	[3,40 - 63,79]

De modo geral, verifica-se através da análise GEE, utilizando a aplicação SAGA, que à semelhança das analises de comparação entre *packages*, as covariáveis correspondentes à idade, dilatação pupilar e duas complicações que surgiram durante a cirurgia de remoção das cataratas, apresentam influência positiva para a presença de pseudoexfoliação ocular. A validação desta influência é realizada com base nos valores obtidos de *p-value* menor que 0,05, o valor das estimativas, para as covariáveis consideradas influentes, apresentam valores positivos e os intervalos de confiança, que uma vez que apresentam valores superior a 1 demonstram uma significância nas covariáveis. Conforme já abordado por outros autores, e demonstrado nos resultados obtidos, a idade e a dilatação pupilar continuam a apresentar influência para a presença ou desenvolvimento da pseudoexfoliação ocular. Também, através dos resultados, verifica-se que complicações durante a cirurgia de remoção de cataratas, observadas quando ocorre uma deiscência zonular, pode indicar a presença da pseudoexfoliação ocular, uma vez que pode surgir o aparecimento de infecções resultantes da deiscência registada, sendo uma dessas infecções a própria pseudoexfoliação. A covariável que distingue os diferentes

locais de implementação da lente intraocular, apesar de apresentar valores p inferiores a $\mathbf{0,05}$ para duas diferentes localizações da lente, o valor da estimativa pontual e dos intervalos de confiança, indicam problemas de sobrestimação, isto pode ser justificado por uma reduzia variabilidade de pacientes, onde se observe a presença da lente intraocular nas especificas localizações. Através dos resultados obtidos pela utilização da aplicação desenvolvida, verifica-se uma semelhança de resultados, em todas as etapas percorridas, com aqueles obtidos através do script para estudo e aplicação de diferentes *packages*.

7

CONCLUSÕES E TRABALHOS FUTUROS

7.1 VISÃO GLOBAL DO TRABALHO DESENVOLVIDO

O objetivo principal deste trabalho consistiu na identificação e desenvolvimento de ferramentas para a construção e análise de modelos GEE, possibilitando, assim, a análise de estruturas de dados longitudinais onde exista uma correlação entre os dados. Foi demonstrado durante fase de identificação de ferramentas/*packages* existentes para a análise GEE que existem alguns *packages* que possibilitam a construção de modelos GEE bem como garantem uma análise rigorosa, através dos resultados obtidos e através da validação *ROC*. O caso de estudo sobre a pseudoexfoliação ocular permitiu expor as diferentes etapas na construção de modelos GEE assim como identificar as diferentes dificuldades que surgem durante o processo, apesar de contribuir para demonstrar a utilidade dos modelos GEE para este tipo de casos específicos, permitiu também identificar algumas diferenças que existem nos diferentes *packages*.

Posteriormente, e de modo a colmatar a inexistência de uma aplicação destinada exclusivamente a este tipo de análise para dados longitudinais, foi desenvolvida e concluída a aplicação *Shiny* SAGA. A aplicação permite o desenvolvimento de diferentes modelos GEE considerando diferentes cenários, realiza uma validação, através de curvas *ROC* para cada um dos modelos gerando as curvas obtidas e as respetivas métricas possibilitando uma análise mais rigorosa sobre os modelos desenvolvidos, por fim descreve os resultados da análise GEE para cada um dos modelos permitindo ao longo dos diversos processos guardar os resultados obtidos no computador pessoal.

A aplicação SAGA foi desenvolvida na sua totalidade em *R* com uso do *package Shiny* e tendo como base os *packages* de construção e análise GEE, bem como o *package ROCR* para a validação por análise *ROC*. Apesar de toda a estruturação da aplicação ser desenvolvida em *R* a interface da mesma implicou adquirir algum conhecimento de diferentes linguagens de programação como *HTML* e *CSS*.

De um modo geral é possível verificar que no desenvolvimento do trabalho foi possível demonstrar, através de um caso de estudo, a importância da utilização dos modelos GEE, desenvolvidos por Liang and Zeger (1993), assim como as ferramentas existentes para realizar uma análise GEE eficiente e rigorosa. A aplicação SAGA permite realizar uma análise GEE de modo mais interativo e rápido, necessitando apenas de descarregar o conjunto de dados a utilizar.

7.2 PERSPETIVAS FUTURAS

As linhas de orientação para próximos trabalhos futuros, recaem sobre a aplicação SAGA que pode sofrer algumas melhorias como a adição de outras ferramentas de validação dos modelos, de modo a existir variabilidade na validação dos modelos desenvolvidos. Poderá também existir uma integração de uma componente gráfica para representação dos resultados GEE para conjunto de dados onde exista uma causa-efeito.
A aplicação exige que o conjunto de dados introduzido apresente-se pré-processado, ou seja, não existem muitas ferramentas na aplicação que possibilitem o tratamento do conjunto de dados antes da construção do modelos GEE, como por exemplo, a remoção de dados omissos.
É possível acessar todos os menus existentes na aplicação, mesmo que não estejam concluídas algumas etapas ou não esteja introduzido o conjunto de dados, a introdução de uma limitação ao acesso destes mesmos menus possibilita uma melhor análise e reduz a probabilidade de erro ou de parâmetros não definidos.
Por fim a aplicação SAGA está sempre sujeita a *updates* que ocorram nos *packages* utilizados bem como alterações na interface de forma a melhorar estéticamente e permitir aumentar a interatividade da mesma.

REFERÊNCIAS BIBLIOGRÁFICAS

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Braga, A. C. (2000). *Curvas ROC: Aspectos funcionais e aplicações [Universidade do Minho]*. PhD thesis, Tese de Doutorado. Braga.

Christopher, Z. (2001). Generalized estimating equation models for correlated data: a review with applications. *American Journal of Political Science*, 45(2):470–490.

Dobson, A. J. and Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC press.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fekedulegn, D. B., Andrew, M. E., Burchfiel, C. M., Violanti, J. M., Hartley, T. A., Charles, L. E., and Miller, D. B. (2007). Area under the curve and other summary indicators of repeated waking cortisol measurements. *Psychosomatic medicine*, 69(7):651–659.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Glynn, R. J. and Rosner, B. (2012). Regression methods when the eye is the unit of analysis. *Ophthalmic epidemiology*, 19(3):159–165.

Govetto, A., Lorente, R., de Parga, P. V., Rojas, L., Moreno, C., Lagoa, F., and Lorente, B. (2015). Frequency of pseudoexfoliation among patients scheduled for cataract surgery. *Journal of Cataract & Refractive Surgery*, 41(6):1224–1231.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hardin, J. W., Hardin, J. W., Hilbe, J. M., and Hilbe, J. (2007). *Generalized linear models and extensions*. Stata press.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis.* Wiley-Interscience.

Horton, N. J. and Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician*, 53(2):160–169.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

Jammal, H., Abu Ameera, M., Al Qudah, N., Aldalaykeh, M., Abukahel, A., Al Amer, A., and Al Bdour, M. (2021). Characteristics of patients with pseudoexfoliation syndrome at a tertiary eye care center in jordan: a retrospective chart review. *Ophthalmology and Therapy*, 10(1):51–61.

Kleinbaum, D. G. and Klein, M. (2002). Logistic regresion for correlated data: Gee. *Logistic regression: A self-learning text*, pages 327–375.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liang, K.-Y. and Zeger, S. L. (1993). Regression analysis for correlated data. *Annual review of public health*, 14(1):43–68.

McCullagh, P. and Nelder, J. (1989). Generalized linear models ii.

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

Mitchell, P., Wang, J. J., and Hourihan, F. (1999). The relationship between glaucoma and pseudoexfoliation: the blue mountains eye study. *Archives of Ophthalmology*, 117(10):1319–1324.

Morandat, F., Hill, B., Osvald, L., and Vitek, J. (2012). Evaluating the design of the r language. In *European Conference on Object-Oriented Programming*, pages 104–131. Springer.

Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*, volume 791. John Wiley & Sons.

Nooraee, N., Molenberghs, G., and van den Heuvel, E. R. (2014). Gee for longitudinal ordinal data: comparing r-geepack, r-multgee, r-repolr, sas-genmod, spss-genlin. *Computational Statistics & Data Analysis*, 77:70–83.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Ritch, R. and Schlötzer-Schrehardt, U. (2001). Exfoliation syndrome. *Survey of ophthalmology*, 45(4):265–315.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3):485–497.

Sangal, N. and Chen, T. C. (2014). Cataract surgery in pseudoexfoliation syndrome. In *Seminars in ophthalmology*, volume 29, pages 403–408. Taylor & Francis.

Schlötzer-Schrehardt, U. and Naumann, G. O. (2006). Ocular and systemic pseudoexfoliation syndrome. *American journal of ophthalmology*, 141(5):921–937.

Shazly, T. A., Farrag, A. N., Kamel, A., and Al-Hussaini, A. K. (2011). Prevalence of pseudoexfoliation syndrome and pseudoexfoliation glaucoma in upper egypt. *BMC ophthalmology*, 11(1):1–6.

Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*.

Van Belle, G., Fisher, L. D., Heagerty, P. J., and Lumley, T. (2004). *Biostatistics: a methodology for the health sciences*, volume 519. John Wiley & Sons.

Venables, W. N., Smith, D. M., Team, R. D. C., et al. (2009). An introduction to r.

Wang, Y.-G., Fu, L., and Paul, S. (2022). Analysis of longitudinal data with examples.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, 39(2):95–101.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

Wickham, H. (2021). *Mastering shiny*. "O'Reilly Media, Inc.".

Wilson, J. R. and Lorenz, K. A. (2015). *Modeling binary correlated responses using SAS, SPSS and R*, volume 9. Springer.

Ziegler, A. and Vens, M. (2010). Generalized estimating equations. *Methods of information in medicine*, 49(05):421–425.