UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

ANTÔNIO ARTHUR SILVA DE LIMA
ROMULO BARROS DE FREITAS

MODELAGEM MARGINAL PARA RESPOSTAS AGRUPADAS CORRELACIONADAS

FORTALEZA
2024

ANTÔNIO ARTHUR SILVA DE LIMA
ROMULO BARROS DE FREITAS

MODELAGEM MARGINAL PARA RESPOSTAS AGRUPADAS CORRELACIONADAS

Trabalho apresentado ao curso de Bacharelado em Estatística do Centro de Ciências da Universidade Federal do Ceará, como parte dos requisitos para a aprovação na disciplina de Análise de Dados Categorizados no semestre de 2024.1.
Prof. Dr.: Gualberto Segundo Agamez Montalvo.

FORTALEZA
2024

Sumário

1 Introdução **5**

2 Modelos Marginais para Repostas Binárias Clusterizadas **6**

3 Generalized Estimating Equations (GEE) **6**

4 Modelos Marginais para Repostas Multinomiais Clusterizadas **8**

5 Modelagem Transicional **9**

6 Dados Faltantes **9**

7 Exemplo - Confiança nas Instituições dos EUA **10**

7.1 Ajustando o Modelo GEE 11

7.2 Adicionando Preditores ao Modelo 14

Lista de Tabelas

1 Confiança nas Instituições dos EUA 10

1 Introdução

Diversos estudos monitoram a variável resposta para cada indivíduo de forma recorrente, em diferentes instantes (como em pesquisas longitudinais) ou sob distintas circunstâncias. Esse fato é comum em pesquisas no campo da saúde, como quando um médico avalia pacientes em intervalos regulares de tempo quanto ao sucesso de um tratamento medicamentoso. Um grande contingente das pesquisas sobre modelagem de dados agrupados se concentrou em uma única variável resposta. Entre os autores que estudaram modelagem conjunta de respostas discretas e contínuas estão Catalano e Ryan (1992), Fitzmaurice e Laird (1995), Regan e Catalano (1999) e Rochon (1996).

Observações repetidas em um indivíduo geralmente estão positivamente correlacionadas [1]. Ademais, correlações positivas também costumam ocorrer quando a variável de resposta é observada para conjuntos de sujeitos emparelhados. Um conjunto de observações pareadas é chamado de *cluster*, e análises estatísticas nesses conjuntos devem considerar a correlação positiva intrínseca [4].

Modelos para dados multivariados agrupados são complexos, porque devem considerar dois tipos de correlações: entre medições em diferentes variáveis para cada grupo e entre medições em diferentes sujeitos dentro de um grupo [4]. Uma dificuldade na construção de modelos paramétricos para a modelagem conjunta de respostas contínuas e discretas é a falta de uma distribuição multivariada (distribuição conjunta) natural. Este relatório se concentra em modelos marginais, ajustados pela resolução de equações de estimação generalizada (GEE). Este é um método multivariado que, para dados discretos, é computacionalmente muito mais simples do que a máxima verossimilhança (MV) e mais prontamente disponível em softwares.

Como citado anteriormente, dados agrupados são comuns em pesquisas de saúde pública. Para ilustrar um problema em que os dados agrupados surgem, a seguir é apresentado brevemente um problema de pesquisa com respostas dependentes.

- Estudo de pesquisa ocular de Baltimore, onde mais de 5000 pessoas com 40 anos ou mais receberam um exame visual como parte de um estudo de prevalência baseado na população de distúrbios oculares [6]. O objetivo é identificar variáveis demográficas, como idade, raça, nível de educação e acesso aos cuidados médicos, que estão associadas à perda de visão. Os dados estão disponíveis para ambos os olhos de todos os sujeitos. Um único modelo de regressão expressando a deficiência visual em termos de variáveis demográficas aborda os objetivos científicos. No entanto, os dois olhos da mesma pessoa provavelmente não são independentes, porque muitas causas de deficiência são binoculares. Essa associação deve ser considerada.

Nessa perspectiva, este relatório visa apresentar de forma clara e sucinta algumas das principais técnicas estatísticas aplicadas em problemas de pesquisa que envolvem modelos marginais para respostas agrupadas e correlacionadas. Por fim, será apresentado um exemplo de aplicação que envolve uma pesquisa realizada sobre o nível de confiança nas instituições dos EUA.

2 Modelos Marginais para Repostas Binárias Clusterizadas

Para pares emparelhados em uma resposta binária, as respostas (y_1, y_2) têm probabilidades marginais para a tabela de contingência que faz a interseção entre y_1 e y_2. Modelos marginais descrevem como os logitos das probabilidades marginais dependem de variáveis explicativas.

Abordagens ingênuas para a análise de dados binários agrupados ignoram a correlação entre subunidades. Em um modelo marginal, a regressão de Y em x e a dependência dentro do cluster são modeladas separadamente. Para o primeiro, modelamos a esperança marginal $E(Y_{ij})$ como uma função das variáveis explicativas. A esperança marginal é a resposta média sobre a população de indivíduos com um valor comum de x, assim como no caso univariado quando $n_i = 1$ para cada cluster.

O modelo marginal é estimado usando um modelo de equação de estimativa generalizada (GEE). Se a intervenção for binária, o efeito da intervenção (razão de chances logarítmica) é interpretado como o efeito médio em todos os indivíduos, independentemente do grupo ou cluster ao qual possam pertencer. (Esta estimativa é sensível aos tamanhos relativos dos clusters.)

À medida que a variação entre os clusters aumenta, também aumenta a discrepância entre os modelos condicional e marginal. Usar um modelo linear generalizado que ignora completamente o agrupamento fornecerá a estimativa pontual correta (marginal), mas subestimará a variância subjacente (e os erros padrão), desde que haja variação entre os clusters. Se não houver variação entre os clusters, o modelo GLM deve ser adequado.

3 Generalized Estimating Equations (GEE)

O método GEE surgiu como uma alternativa à estimação por máxima verossimilhança, que pode ser muito mais complexa e intensa computacionalmente quando utilizada para repostas categóricas multivariadas agrupadas. De maneira geral, ao invés de assumir uma distribuição multivariada para o cluster (distribuição conjunta), este método conecta cada média marginal da variável resposta a um preditor linear, gerando estimativas para a matriz de variância-covariância da variável resposta.

Para isso, as estimativas da variância, erro padrão e covariância são obtidas empiricamente, isto é, por meio da amostra obtida, e o procedimento geral tem a seguinte ordem:

1. Definir um modelo marginal para a média da variável resposta $E(Y)$ escolhendo uma função de ligação e formando um preditor linear;

2. Fazer suposição sobre como a variância da variável reposta $V(Y)$ depende da sua média $E(Y)$, baseada na distribuição que modela a variável;

3. Prover um chute inicial sobre a correlação dentro do cluster (*working correlation matrix*), isto é, sobre $\{Y_T\}$, que envolve as múltiplas respostas para um mesmo indivíduo;

4. Obter as estimativas do método por meio do modelo marginal e da estrutura suposta para a correlação e variância;

5. Utilizar as estimativas para ajustar os erros padrão e obter valores mais robustos, mesmo com uma suposição errônea para a matriz de variância-covariância.

No passo 3 do método, há quatro possíveis formas que podem ser utilizadas para supor o comportamento da matriz de correlação entre as múltiplas respostas, que são:

- Forma **Independente**: assume que a correlação para cada par combinado das respostas, seja nula, resultando em estimativas idênticas ao método da máxima verossimilhança;
- Forma **Permutável**: assume que a correlação para cada par combinado das respostas, seja igual e desconhecida em todos os pares;
- Forma **Autoregressiva**: assume que a correlação para cada par combinado das respostas, seja dada por ρ^{t-s}, onde t e s são as respostas para a variável. Esta abordagem é muito utilizada em séries temporais;
- Forma **Desestruturada**: assume que a correlação para cada par combinado das respostas, seja diferente e desconhecida em todos os pares.

Tomando como exemplo um cluster qualquer de tamanho 3, teríamos a seguinte estrutura em cada forma considerada para a correlação entre as respostas:

- Independente: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- Permutável: $\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$
- Autoregressiva: $\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$
- Desestruturada: $\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$

Como o método não faz nenhuma suposição sobre a distribuição conjunta dos clusters, não há aqui a função de verossimilhança, fazendo dele um método de *quase-verossimilhança*. As principais vantagens residem na simplicidade de construção e disponibilidade em pacotes computacionais para dados categóricos, além de prover uma estimação consistente, mesmo com a matriz de correlação escolhida não sendo a ideal. As desvantagens do método são consequência da falta da função de verossimilhança: não há como medir a qualidade do ajuste, nem realizar inferências dos parâmetros estimados ou comparar modelos. A inferência limita-se, nesse caso, ao uso de estatísticas como a estatística de Wald. Ademais, exceto sejam utilizadas grandes amostras, os erros padrão irão subestimar os verdadeiros valores.

4 Modelos Marginais para Repostas Multinomiais Clusterizadas

O método GEE é estendido para variáveis respostas clusterizadas que podem assumir mais de uma categoria. Desta forma, para modelar respostas categóricas nominais, o modelo logit é utilizado, descrevendo a razão de odds da categoria em relação a outra fixa. Já para variáveis categóricas ordinais, modelos logit acumulados entram em cena, descrevendo as probabilidades acumuladas daquela categoria ocorrer na resposta.

De maneira geral, para modelar a associação de variáveis categóricas, faz mais sentido usar razão de odds em detrimento das correlações entre tais variáveis (working correlation), principalmente no caso em que tais variáveis são multinomiais. Com isso, uma versão alternativa do método GEE usa a razão de odds para mensurar as associações, ao invés da correlação, porém, ainda utilizando as estruturas matriciais mostradas anteriormente.

5 Modelagem Transicional

Quando há observações que ocorrem em um intervalo de tempo, pode-se modelar o comportamento das variáveis respostas em detrimento dos seus valores passados observados continuamente no tempo, junto às suas respectivas variáveis explicativas. Tal abordagem é mais comum para os casos nos quais se quer fazer previsões da variável resposta com base em valores passados da mesma.

Um modelo de Markov de primeira ordem é um modelo transicional que considera que para todo tempo t, a distribuição condicional da resposta no tempo considerado, dado y_{t-1}, seja independente de outros tempos.

Considerando como exemplo uma variável resposta binária y, e p variáveis explicativas, um modelo de regressão logística de Markov tem a forma

$$\text{logit}\left[P(Y_t = 1)\right] = \alpha + \beta_0 y_{t-1} + \beta_1 x_{1t} + \cdots + \beta_p x_{pt},$$

onde x_{it} é a variável explicativa no tempo t. Este tipo de modelo trata repetições de um sujeito de maneira independente. Portanto, algoritmos comuns para modelos lineares generalizados podem ser usados para ajustar esses modelos, tratando cada observação separadamente.

Também é possível, em alguns casos, tratar a primeira variável resposta como uma covariável, fixando a segunda variável resposta como sendo univariada e ajustando um modelo para y_1.

6 Dados Faltantes

Em diversas ocasiões, é muito comum ao realizar estudos com dados pareados, que ao menos uma observação esteja indisponível no cluster, seja por conta da natureza da pesquisa/estudo, ou por fatores externos incontroláveis.

Análises estatísticas que consigam tratar de forma adequada esses dados faltantes, em geral, sobressaem-se em relação àquelas que não consideram tal situação, ou que fazem um tratamento inadequado. Algumas abordagens frequentemente utilizadas são: deletar todos os registros que possuem ao menos uma variável sem dado; e realizar imputação — com base em algum algoritmo — para preencher as lacunas vazias.

O primeiro procedimento, chamado de "análise de caso completo", possui muitas desvantagens, pois faz com que se perda muita informação, principalmente se a quantidade de valores faltantes for grande, além de gerar maiores erros padrão. Além disso, é possível também que seja introduzido nos dados e nos parâmetros estimados, um forte viés.

Já no segundo procedimento, é realizada uma predição de qual seria o valor faltante, de acordo com alguma abordagem definida, e então, o novo conjunto é utilizado para realizar a modelagem e análises futuras. Um método capaz de realizar predições nesse contexto é o de Monte Carlo, que substitui as lacunas com base em simulações das suas distribuições condicionais, dadas as observações existentes. Por exemplo, para realizar imputações de uma

variável explicativa x_1, pode-se criar um modelo de regressão para fazer a predição da variável, com base na variável resposta e as outras variáveis explicativas, e para todo dado faltante em x_1, gera-se um valor aleatório de uma distribuição normal com média dada pela predição do modelo ajustado, e desvio padrão igual ao desvio estimado com os resíduos do ajuste. Se a variável x_1 for binária, considera-se então uma regressão logística para o ajuste, e números aleatórios gerados por uma distribuição binomial ao invés da normal.

Com essa abordagem, é possível quantificar a variação das estimativas dos parâmetros do modelo em cada simulação, em relação ao modelo que seria criado sem nenhum tipo de tratamento nos dados faltantes.

7 Exemplo - Confiança nas Instituições dos EUA

Desde 1972, a Pesquisa Social Geral, "General Social Survey" (GSS) em inglês [2] tem monitorado as mudanças sociais e estudado a crescente complexidade da sociedade americana. O GSS Data Explorer, do NORC na Universidade de Chicago, torna mais fácil do que nunca utilizar os dados coletados pela GSS. A tabela 1 contém informações sobre uma pesquisa realizada sobre o nível de confiança nas instituições americanas. Para as três perguntas do GSS sobre confiança na educação, medicina e na comunidade científica, as respostas originais foram divididas em três categorias: "Muito", "Quase nenhuma" e "Somente algumas".

Tabela 1: Confiança nas Instituições dos EUA

	Confiança		
Instituição	Muito	Quase nenhuma	Somente algumas
Educação	370	276	821
Medicina	525	194	748
Comunidade Científica	665	95	707

Para esta análise, o foco estará nas proporções de "Muito" e as outras duas categorias serão combinadas. Assim, temos respostas binárias com medidas repetidas em cada um dos 1467 sujeitos ou grupos, onde Y_{ij} representa a resposta binomial para o i-ésimo sujeito e a j-ésima pergunta.

Para referência, se ignorarmos temporariamente a correlação entre as respostas dentro dos sujeitos e, em vez disso, tratarmos todas Y_{ij} como independentes, podemos calcular estimativas das chances de "Muito" de confiança na educação, medicina e comunidade científica, respectivamente, como

$$\pi_{edu} = \frac{370}{276 + 821} = 0,337; \quad \pi_{med} = \frac{525}{194 + 748} = 0,557; \quad \pi_{cie} = \frac{665}{95 + 707}.$$

Podemos estimar o erro padrão para as diferenças nessas probabilidades (logarítmicas). Por exemplo, o erro padrão estimado para comparar as probabilidades logarítmicas de "Muito" para educação *versus* medicina seria:

$$SE = \sqrt{\frac{1}{370} + \frac{1}{(276+821)} + \frac{1}{525} + \frac{1}{(194+748)}} = 0,0811$$

Mas esperaríamos que isso fosse menor se a correlação entre essas respostas fosse positiva e levada em consideração. Isso é o que será explorado a seguir com uma análise no *software R* [3].

7.1 Ajustando o Modelo GEE

O modelo para as respostas agrupadas como uma função apenas do tipo de pergunta seria

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + Med_{ij}\beta_1 + Sci_{ij}\beta_2$$

onde π_{ij} é a probabilidade de que o *i-ésimo* sujeito responda "Muito" para a *j-ésima* pergunta. O coeficiente β_1 é interpretado como o logaritmo da razão de chances para "Muito" de confiança na medicina, em relação à educação.

A abordagem ingênua assumiria que todas as respostas de todos os sujeitos são independentes e estimaria os parâmetros e erros padrão com o modelo logístico usual, o que pode ser feito com a função `glm` do R.

```
# pacotes
if(!require(tidyverse)){install.packages("tidyverse");library(tidyverse)}
if(!require(gee)){install.packages("gee");library(gee)}
if(!require(geepack)){install.packages("geepack");library(geepack)}

# base de dados
gss = read.csv("gss.confidence.csv")
gss$c.age = gss$age - mean(gss$age)
head(gss)
table(gss$question, gss$conf)

# considerando as observacoes independentes
fit.glm = glm(greatly~question, data=gss, family=binomial(link='logit'))
summary(fit.glm)
```

Listing 1: Modelo logístico usual

Por padrão, o R tratará a pergunta sobre educação (a primeira em ordem alfabética entre os tipos de pergunta) como a referência.

```
> summary(fit.glm)

Call:
glm(formula = greatly ~ question, family = binomial(link = "logit"),
    data = gss)
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.08683    0.06012 -18.078  < 2e-16 ***
questionconmedic   0.50222    0.08112   6.191 5.98e-10 ***
questionconsci     0.89951    0.07978  11.275  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5722.8  on 4400  degrees of freedom
Residual deviance: 5591.4  on 4398  degrees of freedom
AIC: 5597.4

Number of Fisher Scoring iterations: 4
```

Listing 2: Summary do modelo logístico usual

Se, em vez disso, estimarmos os parâmetros e os erros padrão de acordo com a abordagem GEE, usamos a função `geeglm` do pacote geepack [5] e temos várias opções para a estrutura da matriz de correlação intragrupo (*working correlation matrix*). Mesmo com a estrutura da matriz de correlação especificada como independente, como está abaixo com a opção `corstr`, as correlações amostrais entre as três respostas dentro dos sujeitos continuam sendo incorporadas às estimativas dos erros padrão.

```
# gee com estrutura de correlacao independente
fit.ind = geeglm(greatly~question, data=gss, id=id, family=binomial,
                 corstr="independence", scale.fix=T)
summary(fit.ind)
anova(fit.ind)
```

Listing 3: Modelo logístico com estrutura de correlação independente

A opção `id = id` é onde especificamos os clusters dentro dos quais temos observações repetidas que podem estar correlacionadas, e a opção `scale.fix = T` é para evitar a abordagem padrão do R de introduzir um parâmetro de escala de sobredispersão. Em comparação com a saída da função `glm` acima, as estimativas são idênticas, mas os erros padrão são ligeiramente menores. A função `ANOVA` também pode ser usada para gerar uma tabela ANOVA para um efeito geral devido ao tipo de pergunta. Em comparação com a distribuição qui-quadrado de referência com dois graus de liberdade, o valor de teste de 149 é uma forte evidência de que a confiança dos sujeitos difere entre as três instituições, com a maior proporção de "Muito" de confiança aplicando-se à comunidade científica.

```
> summary(fit.ind)

Call:
geeglm(formula = greatly ~ question, family = binomial, data = gss,
```

```
    id = id, corstr = "independence", scale.fix = T)

 Coefficients:
                 Estimate  Std.err   Wald Pr(>|W|)
(Intercept)      -1.08683  0.06012 326.81  < 2e-16 ***
questionconmedic  0.50222  0.06966  51.98 5.61e-13 ***
questionconsci    0.89951  0.07365 149.15  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1

Correlation structure = independence
Scale is fixed.

Number of clusters:   1467  Maximum cluster size: 3
```

Listing 4: Summary do modelo logístico com estrutura de correlação independente

```
> anova(fit.ind)
Analysis of 'Wald statistic' Table
Model: binomial, link: logit
Response: greatly
Terms added sequentially (first to last)

         Df  X2 P(>|Chi|)
question  2 149    <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1
```

Listing 5: ANOVA do modelo logístico com estrutura de correlação independente

Outras opções para a estrutura de correlação de trabalho podem produzir resultados ligeiramente diferentes, mas, em geral, os resultados não são muito sensíveis à estrutura de correlação de trabalho especificada porque as correlações empíricas entre as respostas dos dados são dominantes nos cálculos do GEE. Com isso em mente, um bom compromisso entre ajuste do modelo e parcimônia de parâmetros é a estrutura permutável, que permite um único parâmetro de correlação para todas as respostas em pares em um sujeito. Os resultados abaixo são para a estrutura permutável; estes são particularmente semelhantes aos da estrutura de correlação independente neste exemplo porque a correlação estimada nos dados é pequena.

```
# gee com estrutura de correlacao exchangeable
fit.exch = geeglm(greatly~question, data=gss, id=id, family=binomial,
                  corstr="exchangeable", scale.fix=T)
summary(fit.exch)
```

Listing 6: Modelo logístico com estrutura de correlação exchangeable

```
> summary(fit.exch)

```

```
Call:
geeglm(formula = greatly ~ question, family = binomial, data = gss,
    id = id, corstr = "exchangeable", scale.fix = T)

 Coefficients:
                 Estimate Std.err Wald Pr(>|W|)
(Intercept)       -1.0868  0.0601  327  < 2e-16 ***
questionconmedic   0.5022  0.0697   52  5.6e-13 ***
questionconsci     0.8995  0.0737  149  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1

Correlation structure = exchangeable
Scale is fixed.

  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.235  0.0173
Number of clusters:   1467  Maximum cluster size: 3
```

Listing 7: Summary do modelo logístico com estrutura de correlação exchangeable

7.2 Adicionando Preditores ao Modelo

Com a idade (centralizada em torno de sua média), o modelo pode ser estendido para

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + Med_{ij}\beta_1 + Sci_{ij}\beta_2 + C.age_i\beta_3$$

Observação: No caso da idade, ou de qualquer preditor que seja medido no nível do sujeito, um índice para a pergunta (tipo de instituição) não é necessário.

Adicionar preditores adicionais, incluindo interações, não altera a abordagem GEE para estimativa, mas as interpretações dos parâmetros são ajustadas para outros efeitos. Por exemplo, β_0 agora representa as log-odds de ter "Muito" de confiança no sistema educacional (referência) para um sujeito com a idade média destes dados de pesquisa, que acontece ser 48,1 anos. O coeficiente β_1 é a mudança nas log-odds de ter "Muito" de confiança no sistema educacional para cada aumento de 1 ano na idade do sujeito. Além disso, esse efeito é comum para as outras instituições também, a menos que a interação seja levada em consideração.

```
# gee com a covariavel idade
fit.age = geeglm(greatly~question+age, data=gss, id=id, family=binomial,
                 corstr="exchangeable", scale.fix=T)
```

Listing 8: Modelo logístico com estrutura de correlação exchangeable com a covariável idade

Observamos que, além da instituição em questão, a confiança de uma pessoa também está relacionada com a idade. Para cada ano adicional de idade, as probabilidades estimadas de ter "Muito" de confiança no sistema educacional (ou em qualquer outro) são multiplicadas por

$$e^{-0.00504} = 0,995.$$

Portanto, as proporções de pessoas que têm "Muito" de confiança nessas instituições tendem a ser menores para idades mais avançadas, mas esse efeito é muito pequeno em termos práticos. A significância estatística provavelmente se deve ao grande tamanho da amostra.

```
summary(fit.age)
> summary(fit.age)

Call:
geeglm(formula = greatly ~ question + age, family = binomial,
    data = gss, id = id, corstr = "exchangeable", scale.fix = T)

 Coefficients:
                 Estimate  Std.err   Wald Pr(>|W|)
(Intercept)       -0.84631  0.12304  47.31  6.1e-12 ***
questionconmedic   0.50306  0.06979  51.95  5.7e-13 ***
questionconsci     0.90109  0.07381 149.06  < 2e-16 ***
age               -0.00504  0.00228   4.91    0.027 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1

Correlation structure = exchangeable
Scale is fixed.

  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.233  0.0174
Number of clusters:   1467  Maximum cluster size: 3
```

Listing 9: Summary do modelo logístico com estrutura de correlação exchangeable com a covariável idade

Referências

[1] Alan Agresti. *An introduction to categorical data analysis*. Third edition. Wiley series in probability and statistics. Hoboken, NJ: John Wiley & Sons, 2019. ISBN: 9781119405276 9781119405283.

[2] Michael Davern et al. *General Social Survey 1972-2024. [Machine-readable data file]*. Ed. por Michael Davern. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. Sponsored by National Science Foundation. NORC ed. Chicago: NORC, 2024: NORC at the University of Chicago [producer and distributor]. Data accessed from the GSS Data Explorer website at gssdataexplorer.norc.org. Chicago: National Opinion Research Center (NORC) at the University of Chicago, 2024. URL: `https://gssdataexplorer.norc.org` (acesso em 25/05/2024).

[3] *Download R-4.3.1 for Windows. The R-project for statistical computing.* URL: `https://cran.r-project.org/bin/windows/base/` (acesso em 25/10/2023).

[4] Ralitza V Gueorguieva e Alan Agresti. "A correlated probit model for joint modeling of clustered binary and continuous responses". Em: *Journal of the American Statistical Association* 96.455 (2001), pp. 1102–1112.

[5] Ulrich Halekoh, Søren Højsgaard e Jun Yan. "The *R* Package **geepack** for Generalized Estimating Equations". en. Em: *Journal of Statistical Software* 15.2 (2006). ISSN: 1548-7660. DOI: `10.18637/jss.v015.i02`. URL: `http://www.jstatsoft.org/v15/i02/` (acesso em 25/05/2024).

[6] James M. Tielsch. "Blindness and Visual Impairment in an American Urban Population: The Baltimore Eye Survey". en. Em: *Archives of Ophthalmology* 108.2 (fev. de 1990), p. 286. ISSN: 0003-9950. DOI: `10.1001/archopht.1990.01070040138048`. URL: `http://archopht.jamanetwork.com/article.aspx?doi=10.1001/archopht.1990.01070040138048` (acesso em 25/05/2024).