Universidade de Brasília
Departamento de Estatística

Aplicação do Modelo Logístico Multinomial para Identificação de Fatores Associados à Qualidade do Sono dos Residentes de Ouro-Preto/MG Durante a Pandemia da COVID-19.

Eduardo Souza Moita da Silva

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Eduardo Souza Moita da Silva

Aplicação do Modelo Logístico Multinomial para Identificação de Fatores Associados à Qualidade do Sono dos Residentes de Ouro-Preto/MG Durante a Pandemia da COVID-19.

Orientador:

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Sumário

Resumo

A pandemia de COVID-19 teve um impacto significativo na saúde mental e física das pessoas em todo o mundo. Aspectos como, as medidas de isolamento social, preocupações com a saúde, mudanças na rotina diária e incertezas econômicas contribuíram para um aumento nos níveis de estresse e ansiedade. Esses fatores, por sua vez, desempenharam um papel crucial na qualidade do sono das pessoas. O presente estudo tem como objetivo analisar os fatores de risco para a qualidade do sono dos habitantes de Ouro Preto/MG durante a pandemia de COVID-19, por meio de modelos de regressão logística para respostas politômicas. Foram comparados modelos cumulativos, nos quais a ordem da variável resposta é levada em consideração e o modelo multinomial, no qual a variável resposta é considerada nominal. Os resultados mostraram que a depressão, a baixa escolaridade e transtorno de ansiedade são fatores fortemente associados à baixa qualidade do sono. O modelo multinomial mostrou melhor ajuste em relação aos modelos cumulativos e, portanto, a ordem da variável resposta foi considerada somente para interpretação dos resultados.

Palavras-chaves: COVID-19, Qualidade do sono, Depressão, Modelo Multinomial, Modelo cumulativo.

Abstract

The COVID-19 pandemic has had a significant impact on the mental and physical health of people around the world. Factors like, social isolation measures, health concerns, changes in daily routine, and economic uncertainty have developed into an increase in stress and anxiety levels. These factors, in turn, played a crucial role in the quality of people's sleep. The present study aims to analyze the risk factors for the quality of sleep of the inhabitants of Ouro Preto/MG during the COVID-19 pandemic, using logistic regression models for polytomous responses. Cumulative models were compared, in which the order of the outcome variable is taken into account, and the multinomial model, in which the response variable is considered nominal. The results demonstrated that depression is a factor strongly associated with poor sleep quality, as well as low education and anxiety disorders. The multinomial model showed the best adjustment about the cumulative models and, therefore, the order of the response variable was considered only for the interpretation of the results.

Keywords: COVID-19, Sleep quality, Depression, Multinomial model, Cumulative model.

1 Introdução

O sono representa um importante período do cotidiano de qualquer ser humano. É neste período em que o corpo restaura e reabastece as proteínas por meio de processos bioquímicos e metabólicos, incluindo seus benefícios do efeito restaurador em nossa capacidade de nos sentirmos bem durante o dia. As necessidades individuais do sono variam de sujeito para sujeito. Geralmente recomenda-se que tenhamos entre 6 e 10 horas por dia de sono. De uma forma geral, o número ideal de horas é aquele que faz com que a pessoa sinta que dormiu o suficiente e acorde com uma sensação de bem-estar e de bom funcionamento do corpo. Apesar da maioria das pessoas dormirem à noite, muitas precisam dormir durante o dia para conciliarem o descanso com os seus horários de trabalho, uma situação que pode provocar transtornos do sono-vigília. Fatores hormonais, sociais e ambientais podem influenciar na qualidade do sono e desencadear distúrbios do sono afetando negativamente as funções física, emocional, cognitiva e social, como citado por Shukla und Basheer (2016).

Uma má qualidade do sono pode implicar em diversos problemas na vida do indivíduo, como a diminuição do rendimento nas atividades cotidianas e o aumento das chances de desenvolver alguns problemas de saúde. Segundo Kim u. a. (2015) cita-se como exemplos a ansiedade, obesidade, diabetes e resistência à insulina, além da desregulação de vários hormônios, o que impacta negativamente na saúde . Simões u. a. (2019); Drager u. a. (2022) Apontam outros fatores que podem estar associados à baixa qualidade do sono, como o uso de equipamentos eletrônicos, o consumo excessivo de bebidas alcoólicas, o uso de substâncias ilegais, a obesidade e a depressão, entre outros .

Com a descoberta do primeiro caso do novo coronavírus, denominado de COVID-19 causado pelo vírus da síndrome respiratória aguda grave 2 (SARS-CoV-2), a rotina das pessoas foi muito alterada. Por exemplo, como forma de desacelerar a propagação do vírus e reduzir a taxa de mortalidade, dirigentes de todo o mundo adotaram uma série de medidas para controlar esse problema de saúde pública, a saber, o distanciamento social, o fechamento temporário de estabelecimentos comercias, locais de entretenimento e sociabilidade,cinemas, restaurantes e igrejas, ocasionando em perda da fonte de renda para muitas famílias como citado por Haleem u. a. (2020). Estes e outros impactos da pandemia na rotina de vida da população são de amplo espectro e podem ter consequências prolongadas nos campos da saúde, da economia e social.

Um estudo realizado na China por Lin u. a. (2021), em uma grande amostra de adultos, mostrou o impacto agudo da pandemia da COVID-19 no sono e nos sintomas psicológicos. Os principais achados do estudo, como esperado, revelaram taxas muito altas de insônia clinicamente significativa, estresse agudo, ansiedade e depressão. Os

entrevistados foram classificados em quatro grupos de acordo com seu nível de exposição e ameaça à infecção pela COVID-19 .

Estudos similares conduzidos no Brasil, como os de Richter u. a. (2021); Souza u. a. (2021) apontaram que as mudanças de rotinas repentinas e transferências de aulas (ou atividades) para modelos virtuais proporcionaram uma flexibilização na rotina da população, impactando assim no ritmo circadiano das pessoas e, consequentemente, na qualidade do sono . Entretanto, em ambos os estudos, a análise se restringiu a uma simples descrição do banco de dados.

Neste contexto, faz-se necessário investigar se a pandemia pode ter influenciado na qualidade do sono dos residentes da região de Ouro Preto/Minas Gerais (MG), durante o período do pico da pandemia da COVID-19, isto é, entre Outubro a Dezembro de 2020, por meio de uma modelagem estatística considerada robusta em comparação às técnicas usadas nas pesquisas citadas anteriormente.

2 Objetivos

2.1 Objetivo Geral

Investigar o impacto de um conjunto de fatores de risco na qualidade do sono dos residentes de Ouro-Preto/MG durante a pandemia de COVID-19.

2.2 Objetivos Específicos

- Descrever os principais fatores sociodemográficos, comportamentais e clínicos.
- Aplicar o modelo logístico multinomial em um banco de dados reais.
- Comparar o desempenho dos modelos cumulativos com o modelo multinomial usual.
- Obter estimativas para a proporção de indivíduos que reportaram ter uma boa qualidade de sono em função das variáveis sexo e faixa etária.
- Avaliar a qualidade do ajuste dos modelos obtidos.

3 Revisão da Literatura

3.1 Regressão Logística

O modelo de regressão logística pertence à classe dos modelos lineares generalizados e é utilizado quando a variável resposta pode assumir apenas dois valores possíveis. Neste caso, a regressão linear usual não pode ser utilizada porque os resíduos não têm distribuição normal e não são homoscedásticos, já que a variância da distribuição binomial depende do parâmetro da mesma. Apesar do modelo de regressão logística não ser linear, é possível transformá-lo em linear aplicando a função logito na probabilidade de sucesso da variável resposta, de modo que

$$\text{logito}(p) = \frac{p}{1-p}. \tag{3.1.1}$$

Assim, pode se estabelecer uma relação linear entre uma variável resposta dicotômica Y e uma variável explicativa X, como expresso a seguir:

$$\text{logito}\left(\mathbb{P}\left(Y=1|x\right)\right) = \log\left(\frac{\mathbb{P}(Y=1|x)}{\mathbb{P}(Y=0|x)}\right) = \beta_0 + \beta x. \tag{3.1.2}$$

Uma vantagem de utilizar a função logito é a facilidade de interpretação dos coeficientes da regressão β. Tomando a exponencial dos dois lados da Equação 3.1.2, temos

$$\frac{\mathbb{P}(Y=1|x)}{\mathbb{P}(Y=0|x)} = \exp\{\beta_0 + \beta x\} = e^{\beta_0}\left(e^{\beta}\right)^x. \tag{3.1.3}$$

O termo à esquerda da equação 3.1.3 é chamado de chance relativa (ou simplesmente `chance`), e representa o quão provável é a ocorrência de um evento em relação ao seu complementar. Desta forma, a relação exponencial aplicada no coeficiente β, isto é, e^{β} fornece uma interpretação para o coeficiente de regressão, indicando as chances entre $Y=1$ e $Y=0$, para cada variação unitária de x (quando x for quantitativa), ou a relação de chances para os níveis $x=1$ e $x=0$, quando o fator for binário.

3.1.1 Regressão Logística Multinomial

A regressão logística multinomial pode ser vista como uma extensão da regressão logística binária. Este modelo é adequado em situações em que a variável reposta é qualitativa ou quantitativa agrupada e possui mais de duas categorias. Ou seja, é conve-

niente olhar para o modelo multinomial como uma coleção de diversos modelos logísticos primários em um só, de acordo com as várias categorias assumidas pela variável dependente, como destaca Agresti (2018). Desta forma, uma variável resposta com m categorias implica em $m-1$ modelos binários, em que cada modelo compara uma categoria da variável resposta com uma categoria de referência. A Equação 3.1.4 mostra a forma do j-ésimo modelo.

$$\ln\left(\frac{\mathbb{P}(Y=j|x)}{\mathbb{P}(Y=m|x)}\right)=\beta_{j0}+\beta_j x, \quad j=1,...,m-1. \tag{3.1.4}$$

Neste caso, $Y=m$ denota a categoria de referência. Assim, para cada $j\in\{1,2,\cdots,m\}$, seja $p_j=\mathbb{P}(Y_i=j|x)$ a probabilidade da variável resposta Y_i assumir a j-ésima categoria. Então, a coleção de probabilidades $\{p_1,p_2,\cdots,p_m\}$, é tal que, $\sum_{j=1}^{m} p_j=\sum_{j=1}^{m}\mathbb{P}(Y_i=j|x)=1$. Sob suposição de independência das observações, a distribuição de probabilidade para o número de categorias da variável resposta segue o modelo multinomial.

Denota-se por $\mathbf{X}$ uma matriz experimental de dimensão $n\times p$, tal que, a i-ésima linha é dada por $\mathbf{x}_i=(x_0,x_1,\cdots,x_n)^\top$, com p representando o número de colunas em $\mathbf{X}$. Observe que, no contexto envolvendo o modelo de regressão logística binária, temos que:

$$\text{logito}\left(p_j\right)=\log\left(\frac{\mathbb{P}\left(Y_i=j|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i=m|\mathbf{x}_i\right)}\right)=\mathbf{x_i}^\top\boldsymbol{\beta}, \tag{3.1.5}$$

sendo $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)^\top$, o vetor de parâmetros desconhecidos, comumente designados por coeficientes de regressão. A quantidade $x_{0i}=1$ para todo $i\in\{1,2,\cdots,n\}$ é incluído em $\mathbf{x_i}^\top$ para acomodar o intercepto. Seguindo a mesma analogia do modelo apresentado na Equação (3.1.5), para o contexto do modelo multinomial, a chance da categoria j em relação à categoria de referência é dada por:

$$\text{OR}_j=\frac{\mathbb{P}\left(Y_i=j|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i=m|\mathbf{x}_i\right)}, \tag{3.1.6}$$

com $j\in\{1,2,\cdots,m-1\}$. Seguindo a mesma notação apresentada em Hosmer Jr u. a. (2013), é possível obter diferentes modelos para as categorias $1,2,\cdots,m-1$ usando a m-ésima categoria como a de referência. Desta forma, podemos obter as seguintes funções:

$$
\begin{aligned}
g_1(\mathbf{x}_i) &= \log\left[\frac{\mathbb{P}\left(Y_i=1|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i=m|\mathbf{x}_i\right)}\right] = \beta_{10}+\beta_{11}x_{1i}+\cdots+\beta_{1p}x_{pi} = \mathbf{x_i}^\top\boldsymbol{\beta}_1 \\
g_2(\mathbf{x}_i) &= \log\left[\frac{\mathbb{P}\left(Y_i=2|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i=m|\mathbf{x}_i\right)}\right] = \beta_{20}+\beta_{21}x_{2i}+\cdots+\beta_{2p}x_{pi} = \mathbf{x_i}^\top\boldsymbol{\beta}_2 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
g_{m-1}(\mathbf{x}_i) &= \log\left[\frac{\mathbb{P}\left(Y_i=m-1|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i=m|\mathbf{x}_i\right)}\right] = \beta_{(m-1)0}+\beta_{(m-1)1}x_{2i}+\cdots+\beta_{(m-1)p}x_{pi} \\
&= \mathbf{x_i}^\top\boldsymbol{\beta}_{m-1}. \qquad (3.1.7)
\end{aligned}
$$

Ou seja, a Expressão (3.1.5) agora é apresentada na forma de $m-1$ novas expressões que calculam as chances de ocorrência de determinada categoria de Y_i em relação à categoria de referência. De igual forma, com base nas expressões apresentadas em (3.1.7), segue que as probabilidades condicionais para cada categoria da variável resposta, condicionais ao vetor de covariáveis são dadas por:

$$
\begin{aligned}
\mathbb{P}\left(Y_i=j|\mathbf{x}_i\right) &= \exp\{g_j(\mathbf{x}_i)\}\left(1+\sum_{k=1}^{m-1}\exp\{g_k(\mathbf{x}_i)\}\right)^{-1}, \text{ com } j=1,\cdots,m-1 \text{ e} \\
\mathbb{P}\left(Y_i=m|\mathbf{x}_i\right) &= \left(1+\sum_{k=1}^{m-1}\exp\{g_k(\mathbf{x}_i)\}\right)^{-1}, \quad \text{para a categória de referência.}
\end{aligned}
$$

Cada expressão de probabilidade é uma função do vetor dos $2(p+1)$ coeficientes de regressão. Observe que, as probabilidades apresentadas na expressão anterior podem ser resumidas na seguinte expressão:

$$
\mathbb{P}\left(Y_i=j|\mathbf{x}_i\right) = \exp\{g_j(\mathbf{x}_i)\}\left(\sum_{k=1}^{m-1}\exp\{g_k(\mathbf{x}_i)\}\right)^{-1} \quad \forall j\in\{1,\cdots,m\}, \qquad (3.1.8)
$$

para $g_m(\mathbf{x}_i)=0$.

3.1.2 Modelos Cumulativos

Quando as categorias da variável resposta possuem uma ordem estabelecida, pode-se utilizar modelos que consideram esta ordem na análise de regressão. Neste cenário, os modelos mais comuns são os chamados modelos cumulativos, que consideram os logitos das probabilidades acumuladas ao invés das probabilidades propriamente ditas Agresti (2018). O modelo mais utilizado entre os modelos cumulativos é chamado de modelo de

chances proporcionais, que tem a seguinte forma:

$$\log\left(\frac{\mathbb{P}(Y \leq j|\mathbf{x})}{\mathbb{P}(Y > j|\mathbf{x})}\right) = \beta_{0j} + \sum_{k=1}^{p} \beta_k x_{kj},\ j = 1, ..., m-1. \tag{3.1.9}$$

É importante observar que os coeficientes angulares da regressão, β, não possuem um subíndice j. Isto significa que este modelo assume que o efeito das variáveis independentes sobre os logitos acumulados são iguais em todas as categorias da variável resposta. Portanto, ao utilizar esta formulação, esta suposição deve ser verificada e, se não for atendida, pode ser necessário utilizar modelos com coeficientes distintos para cada categoria, como o modelo de chances não-proporcionais ou o modelo de chances proporcionais parciais.

Para efeitos de interpretação dos coeficientes de regressão, a notação da razão de chances (notação `OR` do inglês) apresentada na seção 3.1, pode ser utilizada para o propósito no contexto do modelo de regressão multinomial ou logóstica binária ou em modelos cumulativos. Para tal, usando a expressão 3.1.9, denote por $x_{kj}^{'} = x_{kj} + 1$ um incremento unitário da variável x_{kj}, mantendo-se constante os demais fatores. Assim, pode-se obter as seguintes chances:

$$\begin{aligned}
\text{chance}_{x_{kj}^{'}} &= \frac{P\left(Y \leq j|x_{kj}^{'}\right)}{P\left(Y > j|x_{kj}^{'}\right)} = \exp\left(\boldsymbol{\beta}_k + \boldsymbol{\beta}_k x_{kj}\right),\ \text{e} \\
\text{chance}_{x_{kj}} &= \frac{P\left(Y \leq j|x_{kj}\right)}{P\left(Y > j|x_{kj}\right)} = \exp\left(\beta_k x_{kj}\right).
\end{aligned}$$

Nota-se que o logaritmo das chances apresentadas anteriormente vão fornecer o logito cumulativos. Portanto, a razão de chances será dada por: $\frac{\text{chance}_{x_{kj}^{'}}}{\text{chance}_{x_{kj}}} = e^{\beta_k}$, o que significa dizer que, para cada variação unitária da variável x_{kj} a chance associada a j-ésima categoria de Y_i, com relação à categoria de referência $Y_i = m$, tem o seguinte comportamento: (i) aumenta em e^{β_k}, se $\beta_k > 0$; (ii) mantem-se constante se $\beta_k = 0$ e (iii) diminui em e^{β_k} unidades se $\beta_k < 0$. Interpretações análogas podem ser feitas nos casos em que a variável x_{kj} for categórica.

Apesar da distribuição multinomial fazer parte da família exponencial, a natureza multidimensional da variável resposta não permite a utilização dos mesmos algoritmos de estimação utilizados nos modelos lineares generalizados nos softwares estatísticos. Na tentativa de abranger tais modelos, Yee (2015) definiu uma classe mais geral de modelos, chamada de Modelos Lineares Generalizados Vetoriais (MLGV). Estes modelos são uma generalização dos Modelos Lineares Generalizados (MLG), que permite substituir o vetor de coeficientes com elementos β_j por uma matriz de coeficientes com elementos β_{kj}.

Esta notação é útil para modelos com respostas politômicas porque permite computar coeficientes distintos para cada um dos modelos apresentados em 3.1.7, além de incluir também os modelos cumulativos. A Equação 3.1.10, apresentada a seguir, mostra a parte sistemática deste modelo.

$$\eta_j(\mathbf{x}) = \beta_{j0} + \beta_{j1}x_{1i} + ... + \beta_{(m-1)p}x_{pi},\ j = 1, ..., m-1 \tag{3.1.10}$$

Observa-se que se $m = 1$, então há apenas um modelo, e a relação 3.1.10 retorna aos GLMs..

Estes modelos estão implementados no software RStudio a partir da versão 4.0 no pacote VGAM através da função vglm(). Em situações envolvendo amostragem complexa, Lumley (2023) propôs o pacote svyVGAM para incluir pesos amostrais nestes modelos com a função svy_vglm(). A ideia de Lumley foi incorporar os pesos amostrais como se fossem os pesos de uma tabela de frequências. Desta forma, os parâmetros do modelo são estimados com base nos dados censitários. A justificativa para esta metodologia é que o estimador da média é o mesmo quando usa-se pesos amostrais ou pesos de frequência. No entanto, a variância nesses dois casos é calculada de formas diferentes, sendo necessário utilizar técnicas de reamostragem ou linearização de Taylor para calcular a variância correta dos estimadores.

3.2 Função de verossimilhança

A função de verossimilhança pode ser construída codificando a variável resposta em 0's e 1's para indicar o grupo em que uma dada observação pertence, conforme indicado por Hosmer Jr u. a. (2013). Desta forma, para cada observação, utiliza-se uma variável indicadora que determina em qual categoria da variável resposta aquela observação pertence.

Usando a notação apresentada anteriormente, a função de verossimilhança condicional para uma amostra de n observações independentes é dada por:

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{x}) &= \prod_{i=1}^{n} p_1(\mathbf{x}_i)^{y_{1i}} p_2(\mathbf{x}_i)^{y_{1i}} \times \cdots \times p_m(\mathbf{x}_i)^{y_{mi}} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} p_j(\mathbf{x}_i)^{y_{ji}} \right\}, \end{aligned}$$

sendo $p_j(\mathbf{x}_i)$ a probabilidade da i-ésima observação assumir a j-ésima categoria da variável resposta, dado o vetor de covariáveis $\mathbf{x}_i$. Esta formulação pode ser usada tanto para o modelo multinomial quanto para os modelos cumulativos, sendo que no modelo multino-

mial as probabilidades $p_j(\mathbf{x}_i)$ tem a forma da Equação (3.1.8) e no modelo cumulativo são utilizadas as probabilidades acumuladas. Usando o fato de que $\sum_{j=1}^{m} y_{ji} = 1$ para cada i, o logaritmo da função de verosimilhança é dado por:

$$\ell(\boldsymbol{\beta}|\mathbf{x}) = \sum_{i=1}^{n} \left\{ \sum_{j=0}^{m-1} y_{ji} g_j(\mathbf{x}_i) - \log \left(\sum_{k=0}^{m-1} \exp\{g_k(\mathbf{x}_i)\} \right) \right\}. \tag{3.2.1}$$

Observa-se que $g_j(\mathbf{x}_i)$ para $j \in \{1, 2, \cdots, m\}$ é uma função que é igual ao preditor linear $\mathbf{x_i}^\top \boldsymbol{\beta}_j$. Então, as equações escore são obtidas derivando-se $\ell(\boldsymbol{\beta}|\mathbf{x})$ em função de β_{jk} e igualando-as a zero, com $j \in \{0, 1, \cdots, m-1\}$ e $k \in \{0, 1, \cdots, p\}$.

É importante observar que esta definição usual da função de verossimilhança não considera os pesos decorrentes do delineamento de amostragem. Uma forma de incorporar estes pesos é estender o banco de dados utilizando os pesos como se fossem frequências. Por exemplo, um indivíduo que possui peso $w_i = 2$ pode ser considerado como representante de dois indivíduos na população e será duplicado e outro indivíduo que possui peso $w_i = 3$ será triplicado, e assim por diante. Dessa forma, pode-se calcular a função de verossimilhança utilizando este banco de dados estendido. Quando os pesos são calibrados de forma que a soma equivale ao tamanho da população, este banco de dados é chamado na literatura de dados censitários e a função de verossimilhança calculada sobre estes dados é chamada de pseudo-verossimilhança, já que é uma aproximação da verossimilhança verdadeira.

3.2.1 Estimação dos Parâmetros

Os coeficientes de regressão $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ são comumente estimados através da maximização do logaritmo da função pseudo-verossimilhança na expressão (3.2.1). Para isto, tal função é derivada com relação ao vetor $\boldsymbol{\beta}$, e igualada à zero. Isto é,

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta}|\mathbf{x})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{3.2.2}$$

Os valores de $\boldsymbol{\beta}$ que satisfazem a relação na Equação (3.2.2) são chamados de estimadores de máxima verossimilhança, ou seja, são os valores que maximizam a probabilidade de se observar os dados obtidos pela amostra.

Segundo Steven G. Heeringa (2010), o cálculo da matriz de variância-covariância dos estimadores $\hat{\boldsymbol{\beta}}$ em estudos envolvendo amostragem complexa pode ser feito de duas formas: Por reamostragem ou Linearização. A reamostragem consiste em replicar os pesos amostrais e usá-los para produzir amostras aleatórias da amostra original. A linearização consiste em aproximar a variância pela Série de Taylor de primeira ordem utilizando as

funções de influência dadas por

$$h_i(\beta) = -\hat{I}_{w_i}^{-1}\ell_i(\boldsymbol{\beta}), \tag{3.2.3}$$

em que $\hat{I}$ é a estimativa ponderada da matriz de informação de Fisher populacional, $\ell_i(\boldsymbol{\beta})$ é a contribuição da i-ésima observação para a pseudo-verossimilhança e w_i é o peso amostral. As funções de influência possuem a seguinte propriedade:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 = \sum_i w_i h_i(\boldsymbol{\beta}^0) + R(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0|). \tag{3.2.4}$$

sendo $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{\beta}^0$ as estimativas dos parâmetros incluindo ou não a i-ésima observação, respectivamente e $R(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0|)$ indica os termos de ordem superiores para a aproximação de Taylor. Desta forma, a variância do estimador $\hat{\boldsymbol{\beta}}$ é assintoticamente igual à variância do total populacional das funções de influência.

3.3 Inferência

3.3.1 Testes de Hipóteses

Estimados os parâmetros da regressão, o próximo passo envolve testar cada parâmetro individualmente. Se a estimativa do parâmetro é muito próxima de zero, opta-se por retirar a variável correspondente do modelo, desde que esta retirada faça sentido no contexto do problema. Hosmer und Lemeshow (2000) recomendam evitar testes baseados na função de verossimilhança em contextos de amostragem complexa, já que esta função é aproximada e pode resultar em erros de decisão. Uma alternativa é utilizar o teste de Wald, que consiste em particionar o vetor de coeficientes do modelo β em $(\boldsymbol{\beta^{(0)}}, \boldsymbol{\beta^{(1)}})^{\boldsymbol{T}}$, com $\boldsymbol{\beta^{(0)}}$ sendo o conjunto de coeficientes que se deseja testar se são iguais a um valor fixo c. O teste envolve as seguintes hipóteses:

H_0: $\boldsymbol{\beta} = \boldsymbol{\beta^{(0)}}$.

H_1: $\boldsymbol{\beta} \neq \boldsymbol{\beta^{(0)}}$.

A estatística do teste é dada por

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^{(0)}})^{\boldsymbol{T}} \hat{Var}[\hat{\boldsymbol{\beta}}]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^{(0)}}),$$

que tem distribuição aproximada χ^2 com q graus de liberdade em que q é a dimensão do vetor de parâmetros a serem testados. Alternativamente, $\frac{W}{q}$ tem distribuição F com q graus de liberdade no numerador e m no denominador, em que m é a diferença entre o

número de unidades primárias da amostragem e o número de estratos.

3.3.2 Intervalo de Confiança

Uma outra forma de testar a significância dos coeficientes da regressão é através do intervalo de confiança. Como estes coeficientes possuem distribuição assintoticamente normal, o intervalo de confiança de Wald para um determinado coeficiente de regressão ($\hat{\beta}_s$) é dado por

$$IC(\beta_s; (1-\alpha)\%) = (\hat{\beta}_s - Z_{1-\frac{\alpha}{2}}\text{ep}(\hat{\beta}_s); \hat{\beta}_s + Z_{1-\frac{\alpha}{2}}\text{ep}(\hat{\beta}_s)),$$

em que $Z_{1-\frac{\alpha}{2}}$ é o quantil da distribuição normal padrão relativo ao nível de significância desejado e $\text{ep}(\hat{\beta}_s)$ é a raiz quadrada da variância obtida na Equação (3.2.4).

O intervalo de confiança para a razão de chances pode ser obtido aplicando-se a exponencial nos limites do intervalo de confiança de β de modo que:

$$IC(OR; (1-\alpha)\%) = (\exp\{\hat{\beta}_s - Z_{1-\frac{\alpha}{2}}\text{ep}(\hat{\beta}_s)\}; \exp\{\hat{\beta}_s + Z_{1-\frac{\alpha}{2}}\text{ep}(\hat{\beta}_s))\}.$$

4 Metodologia

4.1 Conjunto de dados

O banco de dados que será utilizado nesta pesquisa está vinculado a um estudo sorológico de natureza transversal, conduzido pelo Grupo de Pesquisa e Ensino em Nutrição e Saúde Coletiva (GPENSC/UFOP), entre Outubro e Dezembro de 2020 em duas regiões de médio porte da região centro-sul de Minas Gerais. A pesquisa foi realizada em três momentos com intervalos de 21 dias, considerando o período de incubação do vírus SARS-CoV-2.

O processo de coleta dos dados cumpriu com todos os princípios e normas éticas que regem a condução de pesquisas envolvendo seres humanos. Os dados foram obtidos por meio de entrevistas presenciais de um questionário, visando obter dos participantes (maiores de idade), informações de caráter sociodemográfico e clínico. Como os dados foram coletados durante o período da pandemia da COVID-19, o monitoramento da saúde dos entrevistadores foi realizado por meio de avaliação periódica, antes do início de cada etapa da pesquisa. Para o processo de amostragem, a cidade foi dividida em três estratos de acordo com a renda média familiar, considerando os dados disponíveis no censo demográfico de 2010. No total, cerca de 1529 indivíduos responderam a uma série de perguntas que faziam parte do questionários de Menezes-Júnior u. a. (2022).

4.2 Amostragem

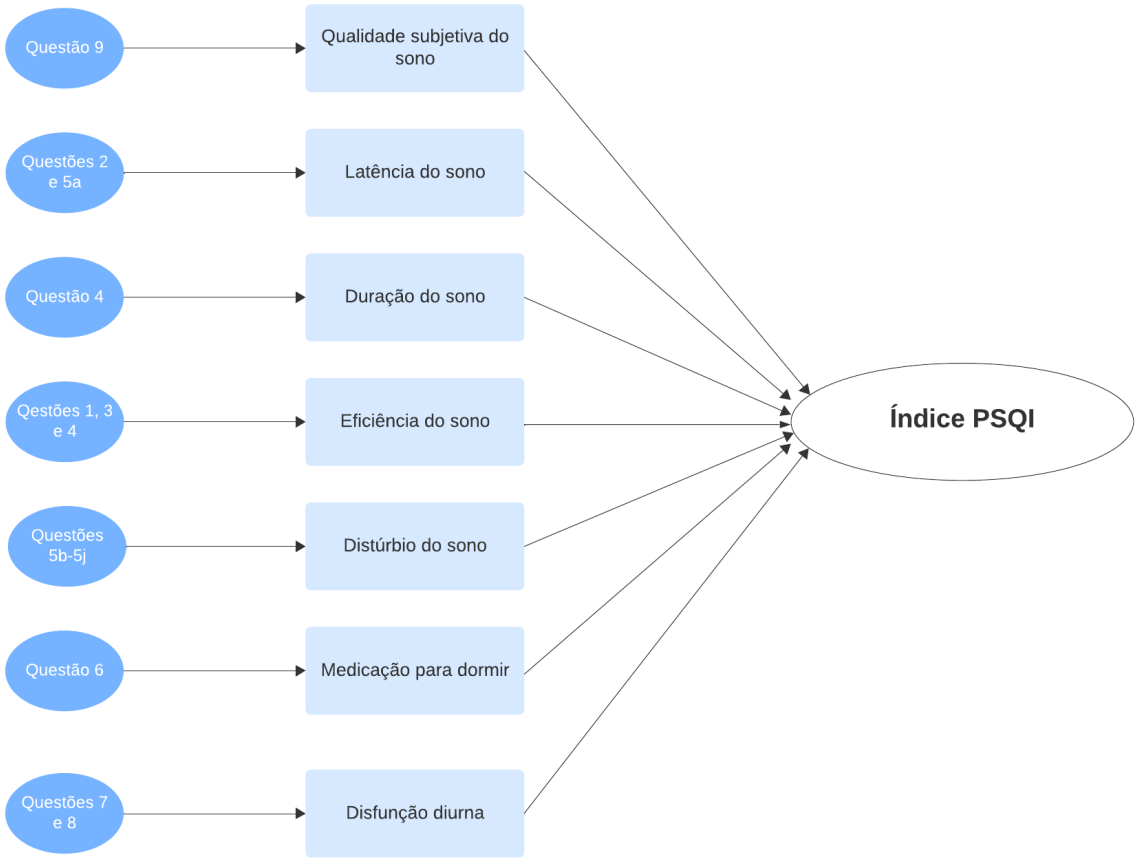
Para a obtenção dos dados, foi utilizado um processo de amostragem complexa em três estágios. Inicialmente, o território de interesse foi dividido em três estratos de acordo com a renda média familiar obtida pelo Censo Demográfico de 2010. Em sequência, cada estrato foi subdividido em conglomerados considerando os setores censitários estabelecidos pelo IBGE. Em cada conglomerado, foi utilizada amostragem sistemática para selecionar os domicílios. Por fim, no terceiro estágio, um indivíduo adulto dentro de cada domicílio foi sorteado por amostra aleatória simples para responder à pesquisa.

4.2.1 Descrição das variáveis

Como ficou claro dos parágrafos anteriores, nosso desfecho de interesse no presente estudo é a qualidade de sono. Utilizou-se a versão brasileira do Pittsburgh Sleep Quality Index (PSQI) traduzida pelos autores em Bertolazi u. a. (2011) para a construção da variável resposta, apresentado no final do texto. Em suma, uma seção do questionário

supracitado é dedicada exclusivamente para se obter informações que sejam úteis para a construção do índice PSQI. De modo geral, o instrumento é composto por 19 perguntas categorizadas em sete componentes com pontuações que variam de 0 a 3, como mostra a Figura 1. O desfecho de interesse (Y) será definido conforme as recomendações apresentadas em Wang u. a. (2020): $Y = 1$ se PSQI ≤ 3 (boa qualidade do sono), $Y = 2$ se o PSQI estiver entre $4 - 7$ (definindo o nível moderado da qualidade do sono) e $Y = 3$ se PSQI > 7 (para o nível baixo de qualidade do sono). Assim, o índice representa a soma dos escores obtidos para cada respondente e a qualidade de sono é considerada melhor quanto menor o valor do índice obtido, com a melhor classificação sendo de índices menores ou iguais a 3.

Figura 1 – Componentes do Índice PSQI



As principais variáveis avaliadas foram: a presença de doenças crônicas não-transmissíveis (DCNT), que é obtida com base no auto-relato do entrevistado, conforme a sua situação clínica no período de avaliação. As doenças crônicas consideradas foram: as diabetes, asma, hipertensão arterial, doenças pulmonares, transtornos de ansiedade, doenças cardíacas ou da tireoide. Demais fatores de risco incluem: ausência ou presença

de depressão, e seus níveis foram obtidos usando como base um índices de classificação frequentemente usado pelos pesquisadores da área. Gênero (1: se masculino, 0: feminino); faixa etária (1: para 18–34 anos, 2: para 35–59 anos e 3: para $>$ 60 anos); estado civil (1: se casado, 0: caso contrário); educação (1: até o ensino básico completo, 2: até o ensino fundamental completo e 3: ensino médio ou superior); renda familiar (1: se $<$ 2, 2: 2–4 e $>$ 4 salários mínimos); ocupação (1: se trabalhador e 0: desempregado); cor da pele (1: se branco, 0: caso contrário).

4.3 Seleção de Variáveis

A seleção de variáveis envolve escolher um conjunto de variáveis que consiga explicar relativamente bem a variável resposta usando o menor número possível de variáveis explicativas, já que quanto mais variáveis presentes no modelo, maiores os erros padrões estimados e maior a dependência do modelo em relação aos dados observados. Hosmer e Lemeshow recomendam iniciar o processo de seleção avaliando a significância das variáveis explicativas separadamente através de modelos univariados. Bendel und Afifi (1977) mostraram que o nível de significância tradicionalmente utilizado de $\alpha = 5\%$ para selecionar as variáveis significantes pode não incluir variáveis importantes para o modelo e indicam a utilização de um nível maior, de $\alpha = 25\%$ ou $\alpha = 35\%$. Após a retirada das variáveis não significativas do modelo, deve-se utilizar um teste apropriado para modelos hierárquicos, como o teste de Wald, para avaliar se alguma das variáveis remanescentes pode ser retirada sem prejudicar o poder descritivo do modelo.

Para comparar modelos não aninhados, foi utilizada uma versão do Critério de Informação de Akaike (AIC) para amostras complexas proposto por Lumley und Scott (2015), dado por

$$dAIC = 2\Lambda - 2p\hat{\bar{\delta}},$$

em que Λ é a diferença entre as pseudo-verossimilhanças dos dois modelos e $\hat{\bar{\delta}} = \frac{tr(\mathcal{I}V)}{p}$ é o efeito médio do delineamento da amostragem. Esta medida, assim como o AIC, indica que o melhor modelo apresenta o menor dAIC.

Por fim, foram testadas as interações entre as variáveis explicativas. Todos esses passos devem levar em consideração o efeito de confundimento, já que a relação entre duas variáveis pode ser alterada pela inclusão de uma terceira variável no modelo. Isso implica que uma variável significativa no modelo univariado pode deixar de ser significatvia no modelo completo e vice-versa.

4.4 Diagnóstico do Modelo

Com o melhor conjunto possível de variáveis selecionado, o próximo passo é avaliar a qualidade do ajuste, por meio de medidas de ajustamento global, análise de resíduos e detecção de observações influentes. No entanto, muitas das ferramentas usualmente utilizadas no diagnóstico de modelos lineares generalizados ainda não possuem formas confiáveis de acomodar pesos amostrais e delineamento de amostragem Steven G. Heeringa (2010).

Uma forma de se avaliar a qualidade do ajuste desses modelos é por métodos de validação cruzada. No entanto, esta técnica permite avaliar apenas a capacidade preditiva do modelo, e um modelo com boas propriedades de predição não necessariamente implica em um bom modelo para explicar o fenômeno em estudo. Apesar disso, de uma forma geral, um modelo com alta acurácia pode indicar bom ajustamento e a sensibilidade e a especificidade podem ajudar a detectar algum nível de assimetria nos resíduos.

4.5 Validação do Modelo

A técnica de validação é utilizada para avaliar a capacidade preditiva de um modelo. A ideia principal é dividir o banco de dados em duas partes para que a avaliação do modelo ocorra de forma separada dos dados que foram utilizados para construir o modelo. Esta abordagem evita um problema conhecido como sobreajuste, que ocorre quando o modelo se ajusta de maneira artificial aos dados e superestima o poder preditivo do modelo, que não consegue se adaptar bem a novos dados. O processo de validação se inicia selecionando-se aleatoriamente um subconjunto dos dados, chamado dados de treinamento, para construir o modelo. Após o modelo ser ajustado, os valores preditos pelo modelo são comparados com os valores dos dados remanescentes, chamados de dados de teste, e os resultados são colocados em uma matriz quadrática de dimensão m chamada de matriz de confusão, em que m é a quantidade de categorias da variável resposta do modelo. Em geral, quando $m > 2$, é útil separar a matriz mxm em m-1 matrizes dicotomizadas $2x2$ para facilitar a interpretação, seguindo a mesma metodologia proposta em Hosmer und Lemeshow (2000).

Quadro 1 – Matriz de Confusão

Atual	Predito	
	Positivo	Negativo
Positivo	Verdadeiro Positivo	Falso Negativo
Negativo	Falso Positivo	Verdadeiro Negativo

A partir desta matriz, podem ser calculados diversos índices relacionados à qualidade preditiva do modelo.

- Acurácia: Indica a proporção de acertos na previsão do modelo;

$$A_c = \frac{VP + VN}{TOTAL}$$

- Sensibilidade: Indica, entre os valores positivos, a proporção de acertos na previsão;

$$S_t = \frac{VP}{VP + FN}$$

- Especificidade: Indica, entre os valores negativos, a proporção de acertos na previsão;

$$S_p = \frac{VN}{VN + FP}$$

- Precisão: Indica a proporção de acertos entre os valores preditos como positivos;

$$P = \frac{VP}{VP + FP}$$

- Valor de predição negativa: Indica a proporção de acertos entre os valores preditos como negativos;

$$NPV = \frac{VN}{VN + FN}$$

- Curva ROC: Método gráfico para avaliar a sensitividade e especificidade do modelo. O eixo Y representa a Sensitividade e o eixo X representa o complementar da Taxa de Especificidade. Quanto maior a área sob a curva, melhor o modelo consegue classificar.

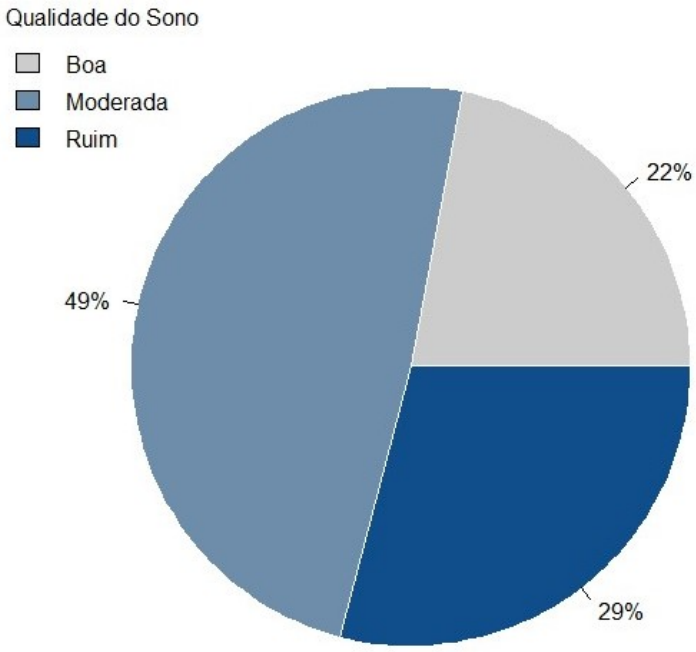
em que VP, VN, FP e FN representam, respectivamente, os valores verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos apresentados no Quadro 1.

5 Resultados

5.1 Análise Exploratória

Inicialmente, verifica-se a distribuição dos indivíduos em cada categoria da variável qualidade do sono.

Figura 2 – Distribuição da qualidade do sono



Com base no gráfico de setores apresentado na Figura 2, é possível observar que, quase metade dos indivíduos apresentaram qualidade do sono regular, 29% apresentaram baixa qualidade e 22% boa qualidade do sono. A Figura 3 mostra a distribuição da qualidade do sono por gênero. A proporção de mulheres com qualidade do sono ruim é significativamente maior que a proporção com sono de boa qualidade, relação que não é observada entre os homens.

Outro fator que parece impactar a qualidade do sono, é a quantidade de DCNT. O gráfico da Figura 4 mostra um aumento significativo da proporção de indivíduos com qualidade de sono ruim conforme aumentam o número de doenças crônicas não-transmissíveis, de forma que os indivíduos que não apresentaram nenhuma DCNT possuem praticamente a mesma proporção de pessoas com qualidade do sono boa e ruim. Em contrapartida, entre os indivíduos com 5 ou mais doenças crônicas, aproximadamente 95% deles apresentaram uma qualidade de sono ruim. Além disso, é notável que a partir de 2 doenças crônicas, a proporção de indivíduos com baixa qualidade de sono é significativamente maior que a

proporção de indivíduos com boa qualidade do sono.

Figura 3 – Distribuição da Qualidade do Sono por Gênero com Intervalos de Confiança de 95%

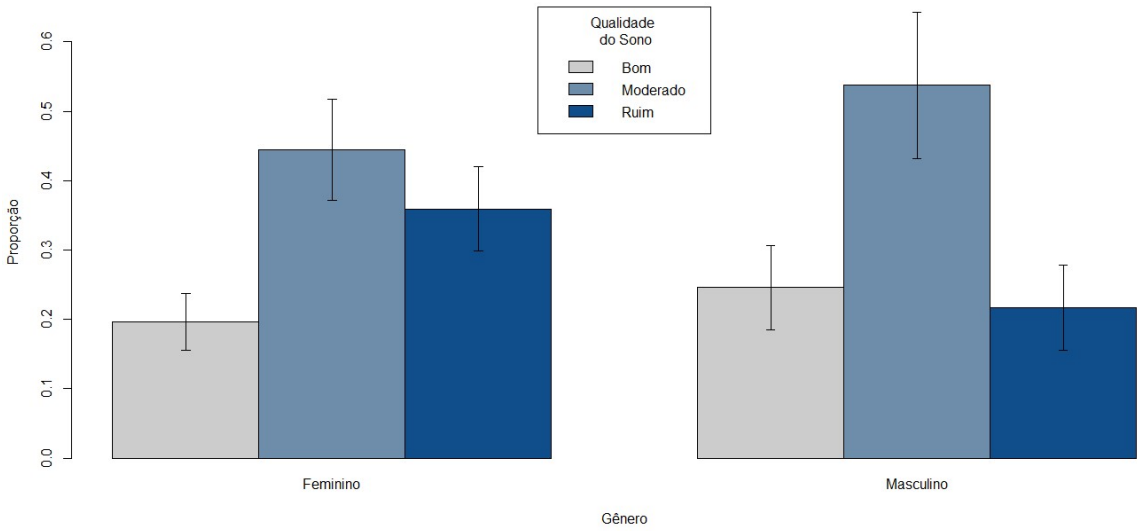
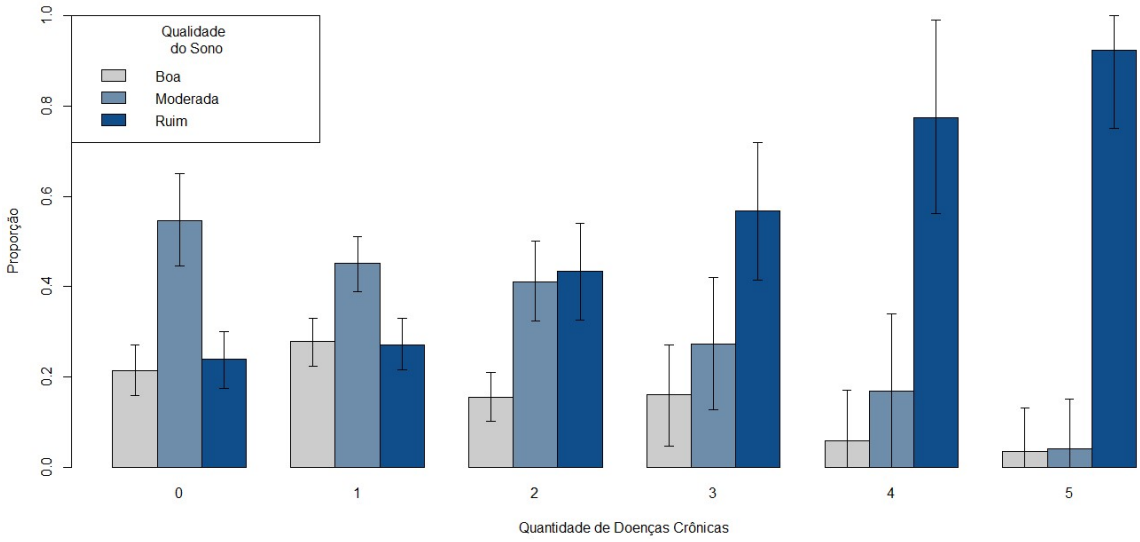
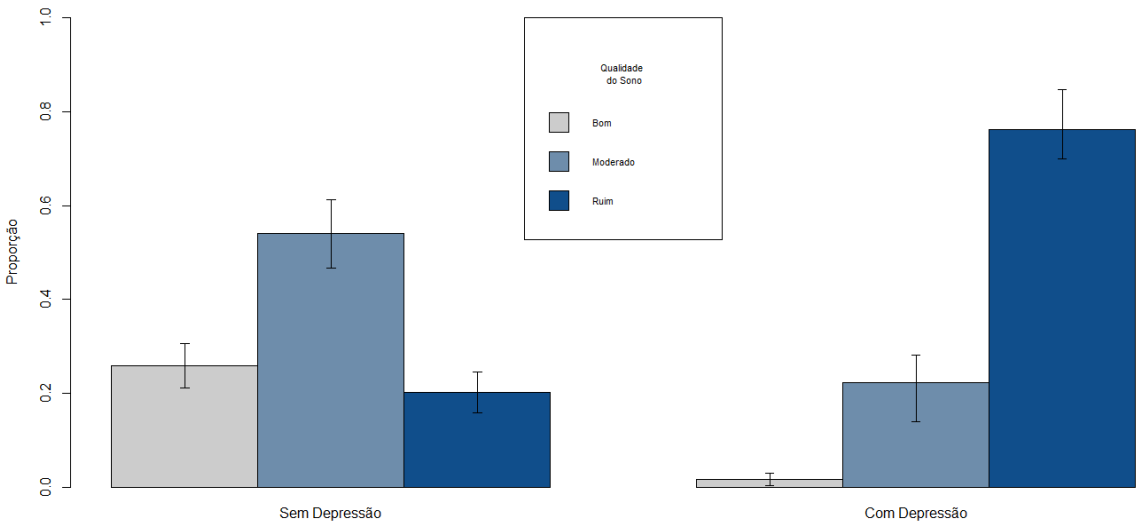


Figura 4 – Distribuição da Qualidade do Sono por Quantidade de Doenças Crônicas não-transmissíveis com Intervalos de Confiança de 95%



A Figura 5 mostra que, entre as pessoas sem depressão, não há diferença significativa entre a proporção de indivíduos com qualidade do sono boa e ruim. Já, entre as pessoas com depressão, a maior parte apresentou baixa qualidade de sono, enquanto que, uma fração muito pequena dos indivíduos apresentou boa qualidade do sono, e a diferença entre essas proporções foi estatisticamente significativa.

Figura 5 – Qualidade do sono por presença ou ausência de depressão com Intervalos de Confiança de 95%



A relação entre o grau de escolaridade e qualidade do sono mostrou-se significativa apenas em um dos níveis da escolaridade, relativo ao ensino básico, como mostra a Figura 6. Os níveis mais altos de educação não parecem ter relação com a qualidade do sono, já que a distribuição é semelhante entre os dois níveis mais altos. Por outro lado, no primeiro nível de escolaridade, quase todos os indivíduos reportaram sono com qualidade ruim.

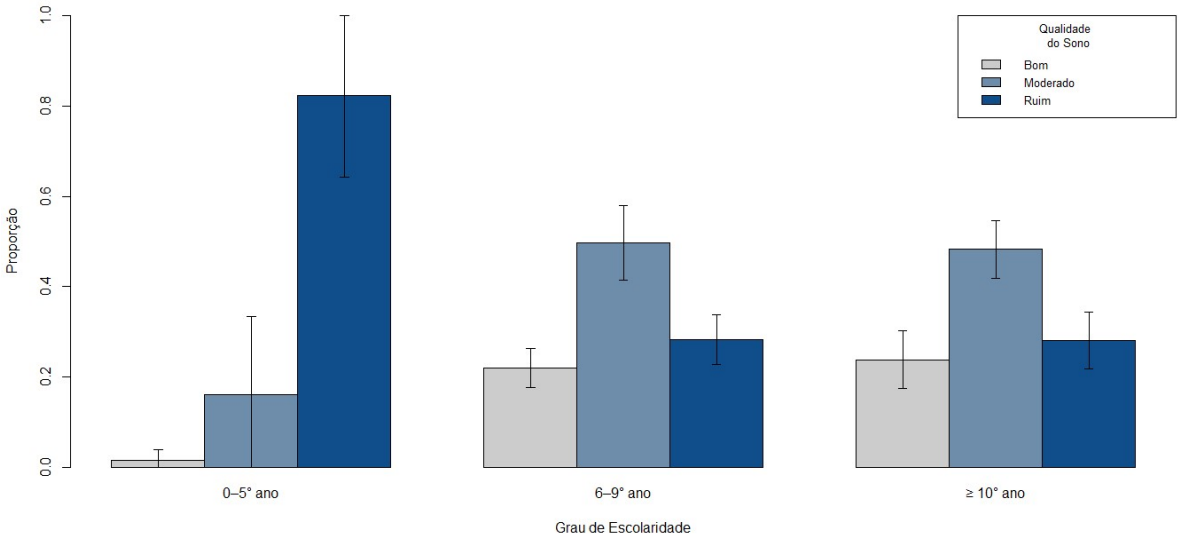


Figura 6 – Distribuição da Qualidade do Sono por Grau de Escolaridade com Intervalos de Confiança de 95%

As variáveis gênero e depressão apresentaram forte associação, com a maior parte

dos indivíduos com depressão sendo mulheres. Como a depressão aparenta ser, pelo menos a princípio, um dos fatores que mais influenciam na qualidade do sono, essa relação com o gênero pode provocar uma associação espúria entre gênero e qualidade do sono. Essa é uma das razões para se utilizar modelos de regressão que consideram o efeito de todas as variáveis conjuntamente.

5.2 Ajuste do Modelo Multinomial

5.2.1 Seleção de Variáveis

Como destacado na metodologia, o primeiro passo na seleção das variáveis é avaliar a significância das variáveis isoladas nos modelos univariados. As Tabelas 1 e 2 mostram os coeficientes de regressão do modelo ajustado e os respectivos desvios padrão e p-valores. Os termos entre parênteses indicam a categoria que está sendo comparada com a categoria de referência.

Tabela 1 – Modelos univariados para o nível moderado da qualidade do sono ($Y_i = 2$)

Fatores de risco	Estimativas	Erro-padrão	Z	valor-p
Gênero (feminino)	-0.80	0.17	-4.66	<0.01
Estado Civil (casado)	0.49	0.25	1.93	0.05
Cor da Pele (branca)	-0.35	0.27	-1.27	0.20
Número de Doenças Crônicas	-0.39	0.08	-4.92	<0.01
Consumo de Álcool (sim)	0.63	0.25	2.50	0.01
Ocupação (trabalhador)	0.47	0.20	2.31	0.02
Consumo de Cigarros (sim)	-0.32	0.34	-0.95	0.34
Ansiedade (sim)	-2.32	0.33	-6.85	<0.01
Depressão (sim)	-4.13	0.46	-8.92	<0.01
Faixa Etária [35–59 anos]	0.29	0.27	1.07	0.28
Faixa Etária [>60 anos]	-0.14	0.31	-0.47	0.63
Grau de Escolaridade (0–5° ano)	-3.68	0.88	-4.15	<0.01
Grau de Escolaridade (6–9° ano)	0.05	0.22	0.23	0.81
Renda [2-4] (Sim)	0.19	0.25	0.75	0.44
Renda [>4] (Sim)	<0.01	0.31	0.02	0.98

Tabela 2 – Modelos univariados para a qualidade do sono ruim ($Y_i = 3$)

Fatores de risco	Estimativas	Erro-padrão	Z	valor-p
Gênero (feminino)	-0.69	0.28	-2.47	0.01
Estado Civil (casado)	0.15	0.28	0.52	0.59
Cor da Pele (branca)	-0.18	0.32	-0.55	0.57
Número de Doenças Crônicas	-0.52	0.11	-4.70	<0.01
Consumo de Álcool (sim)	0.89	0.29	3.00	<0.01
Ocupação (trabalhador)	0.36	0.21	1.73	0.08
Consumo de Cigarros (sim)	-0.07	0.37	-0.20	0.83
Ansiedade (sim)	-0.86	0.27	-3.13	<0.01
Depressão (sim)	-2.28	0.27	-8.46	<0.01
Faixa Etária [35–59 anos]	0.02	0.25	0.08	0.93
Faixa Etária [>60 anos]	-0.51	0.22	-2.31	0.02
Grau de Escolaridade (0–5° ano)	-2.09	0.67	-3.12	<0.01
Grau de Escolaridade (6–9° ano)	0.18	0.22	0.83	0.40
Renda [2-4] (Sim)	0.01	0.34	0.04	0.96
Renda [>4] (Sim)	0.08	0.24	0.34	0.73

As variáveis renda e consumo de cigarros se mostraram pouco significativas e foram retiradas do modelo. A variável educação apresentou significância apenas no nível mais baixo, justificando a retirada dos níveis altos do modelo. Em seguida, foram retiradas as variáveis por ordem crescente de significância, uma de cada vez, resultando em um modelo apenas com variáveis significativas, utilizando nível de significância de $\alpha = 5\%$. As estimativas dos coeficientes destes modelos são apresentadas nas Tabelas 3 e 4.

Tabela 3 – Modelo Multinomial - coeficientes estimados para o nível moderado da qualidade do sono

Fatores de risco	Estimativa	Erro-padrão	Z	valor-p
Intercepto	0.54	0.16	3.18	<0.01
Grau de Escolaridade (0–5° ano)	1.92	0.84	2.26	0.02
Doenças Crônicas	-0.13	0.10	-1.25	0.20
Consumo de Álcool (sim)	0.19	0.23	0.86	0.38
Ansiedade (sim)	1.28	0.34	3.71	<0.01
Depressão (sim)	1.23	0.47	2.61	<0.01

Tabela 4 – Modelo Multinomial - coeficientes estimados para o nível ruim da qualidade do sono ($Y_i = 3$)

Fatores de risco	Estimativa	Erro-padrão	Z	valor-p
Intercepto	-0.35	0.19	-1.82	0.06
Grau de Escolaridade (0–5° ano)	2.78	0.84	3.29	<0.01
Doenças Crônicas	0.28	0.11	2.54	0.01
Consumo de Álcool (sim)	-0.48	0.23	-2.05	0.03
Ansiedade (sim)	1.24	0.36	3.41	<0.01
Depressão (sim)	3.48	0.48	7.17	<0.01

Em seguida, foram testadas as possíveis interações entre as variáveis do modelo final. Os resultados das Tabelas 5 e 6 mostram que as interações significativas foram entre faixa etária e doenças crônicas e entre ansiedade e tabagismo. A significância dessas interações indica que o efeito das variáveis dependem da ocorrência ou não de uma outra variável.

Tabela 5 – Modelo com Interações - nível moderado da qualidade do sono ($Y_i = 2$)

Fatores de risco	Estimativa	Erro-padrão	Z	valor-p
Intercepto	0.60	0.20	2.93	<0.01
Grau de Escolaridade (0–5° ano)	1.58	0.87	1.81	0.06
Faixa Etária [>60]	-0.79	0.49	-1.61	0.10
Doenças Crônicas	-0.34	0.15	-2.22	0.02
Consumo de Álcool (sim)	0.18	0.23	0.78	0.43
Ansiedade (sim)	1.52	0.40	3.72	<0.01
Depressão (sim)	1.26	0.46	2.70	<0.01
Consumo de Cigarros (sim)	0.36	0.34	1.07	0.28
Faixa Etária:Doenças crônicas	0.72	0.26	2.73	<0.01
Ansiedade:Consumo de Cigarros	-1.46	0.67	-2.16	0.03

Tabela 6 – Modelo com Interações - nível ruim da qualidade do sono ($Y_i = 3$)

Fatores de risco	Estimativa	Erro-padrão	Z	valor-p
Intercepto	-0.34	0.21	-1.59	0.11
Grau de Escolaridade (0–5° ano)	2.49	0.80	3.09	<0.01
Faixa Etária [>60]	-0.59	0.47	-1.26	0.20
Doenças Crônicas	0.05	0.14	0.39	0.69
Consumo de Àlcool (sim)	-0.51	0.23	-2.18	0.02
Ansiedade (sim)	1.51	0.41	3.64	<0.01
Depressão (sim)	3.49	0.48	7.15	<0.01
Consumo de Cigarros (sim)	0.60	0.47	1.26	0.20
Faixa Etária:Doenças Crônicas	0.66	0.29	2.26	0.02
Ansiedade:Consumo de Cigarros	-1.38	0.69	-1.99	0.04

O modelo final então é dado pelas seguintes equações:

$$\begin{aligned} Modelo1 = \log\left[\frac{\mathbb{P}\left(Y_i = 2|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i = 1|\mathbf{x}_i\right)}\right] &= 0.605 + 1.586x_1 - 0.798x_2 - 0.344x_3 \\ &+0.183x_4 + 1.524x_5 + 1.265x_6 \\ &+0.369x_7 + 0.725x_2x_3 - 1.464x_4x_7 \end{aligned}$$

$$\begin{aligned} Modelo2 = \log\left[\frac{\mathbb{P}\left(Y_i = 3|\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i = 1|\mathbf{x}_i\right)}\right] &= -0.343 + 2.497x_1 - 0.596x_2 - 0.059x_3 \\ &-0.513x_4 + 1.512x_5 + 3.494x_6 \\ &+0.603x_7 + 0.664x_2x_3 - 1.385x_4x_7, \end{aligned}$$

onde

- x_1: Baixa Escolaridade
- x_2: Faixa Etária Maior que 60 Anos
- x_3: Doenças Crônicas
- x_4: Consumo de Álcool
- x_5: Ansiedade
- x_6: Depressão

- x_7: Consumo de Cigarros

As Tabelas 7 e 8 mostram as estimativas das razões de chances e os respectivos limites inferior (LI) e superior (LS) dos intervalos de 95% de confiança. A chance de uma pessoa que tem depressão ter um sono de baixa qualidade é quase 33 vezes maior em relação a uma pessoa que não tem depressão, ajustando-se pelas demais variáveis. Além disso, a chance de uma pessoa que não completou o ensino básico ter um sono ruim é 12 vezes maior do que de uma pessoa com maior escolaridade. O consumo de álcool se mostrou ser um fator de proteção. Ou seja, diminui a chance de ter qualidade do sono ruim, sendo quase 2 vezes maior a chance de ter uma boa qualidade do sono pra quem consome álcool em relação a quem não consome, apesar do limite superior do intervalos e confiança estar muito próximo de 1. O fator ansiedade, quando o indivíduo é fumante, tem uma razão de chances estimada próxima de 1, indicando que a ansiedade, neste caso, não é um fator de risco para a qualidade do sono. Mas, entre os indivíduos que não consomem cigarros a razão de chances estimada é próxima de 4,5, ou seja, a chance de uma pessoa que tem ansiedade e não fuma ter um sono de baixa qualidade é 4,54 vezes maior do que a de uma pessoa que fuma, e esta chance pode ser até 10 vezes maior, considerando um intervalo de confiança de 95%.

Tabela 7 – Razão de chances para o nível moderado da qualidade do sono ($Y_i = 2$)- modelo multinomial

Fatores de Risco	Interações	OR	LI	LS
Grau de Escolaridade (0-5 ano)	-	4.88	0.87	27.15
Consumo de Álcool (Sim)	-	1.20	0.76	1.89
Depressão (Sim)	-	3.54	1.41	8.86
Doenças Crônicas	Faixa Etária >60 anos	1.46	0.52	4.05
	Faixa Etária <60 anos	0.70	0.52	0.95
Faixa Etária >60 anos	Doenças Crônicas = 0	0.45	0.17	1.18
	Doenças Crônicas = 1	0.93	0.33	2.57
Ansiedade (Sim)	Consumo de Cigarros = 0	4.59	2.06	10.23
	Consumo de Cigarros = 1	1.06	0.38	2.93
Consumo de Cigarros (Sim)	Ansiedade = 0	1.44	0.73	2.83
	Ansiedade = 1	0.33	0.12	0.92

Tabela 8 – Razão de chances para o nível ruim da qualidade do sono ($Y_i = 3$) - modelo multinomial

Fatores de Risco	Interações	OR	LI	LS
Grau de Escolaridade (0-5 ano)	-	12.15	2.49	59.18
Consumo de Álcool (Sim)	-	0.59	0.37	0.94
Depressão (Sim)	-	32.93	12.64	85.80
Doenças Crônicas	Faixa Etária >60 anos	2.06	0.70	6.00
	Faixa Etária <60 anos	1.06	0.79	1.42
Faixa Etária >60 anos	Doenças Crônicas = 0	0.55	0.21	1.39
	Doenças Crônicas = 1	1.07	0.36	3.11
Ansiedade (Sim)	Consumo de Cigarros = 0	4.54	2.01	10.23
	Consumo de Cigarros = 1	1.13	0.33	3.88
Consumo de Cigarros (Sim)	Ansiedade = 0	1.82	0.71	4.64
	Ansiedade = 1	0.45	0.13	1.56

5.3 Diagnóstico do modelo

O processo que consiste em avaliar desempenho de modelos envolvendo amostras complexas pode ser uma tarefa desafiadora. Hosmer und Lemeshow (2000) recomendam fazer o diagnóstico do modelo utilizando as versões binárias separadamente, já que elas produzem resultados semelhantes ao modelo completo e possuem mais ferramentas disponíveis na literatura por fazerem parte dos Modelos Lineares Generalizados. Outra recomendação é preferir análises gráficas a testes de hipótese já que a inclusão dos pesos amostrais pode tornar as distribuições instáveis e dificultar testes de hipóteses.

5.3.1 Análise de Resíduos

Uma forma de incorporar os pesos amostrais nos resíduos de Pearson é multiplicar pela raiz do pesos

$$r_j = \sqrt{w_j}\frac{y_j - \hat{p}_j}{\sqrt{\hat{p}_j\left(1-\hat{p}_j\right)}} \tag{5.3.1}$$

em que w_j é o peso amostral da j-ésima observação, y_j é o valor da variável resposta binária e $\hat{p}_j = \text{logito}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$ é a probabilidade estimada pelo modelo. As Figuras 7 e 10 mostram esses resíduos para os níveis moderado e ruim da qualidade do sono em relação ao nível bom.

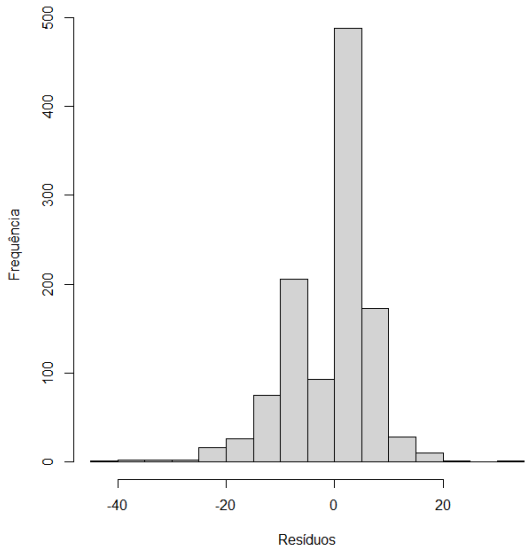


Figura 7 – Resíduos de Pearson para qualidade do sono moderada em relação à boa

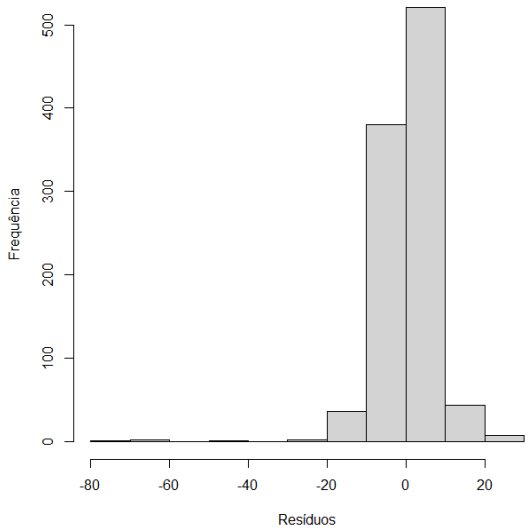


Figura 8 – Resíduos de Pearson para qualidade do sono ruim em relação à boa

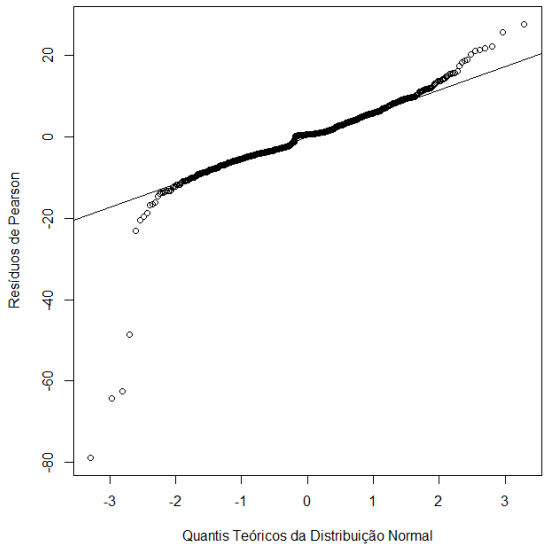


Figura 9 – Gráfico Q-Q de normalidade dos resíduos para $P(Y_i = 2)$

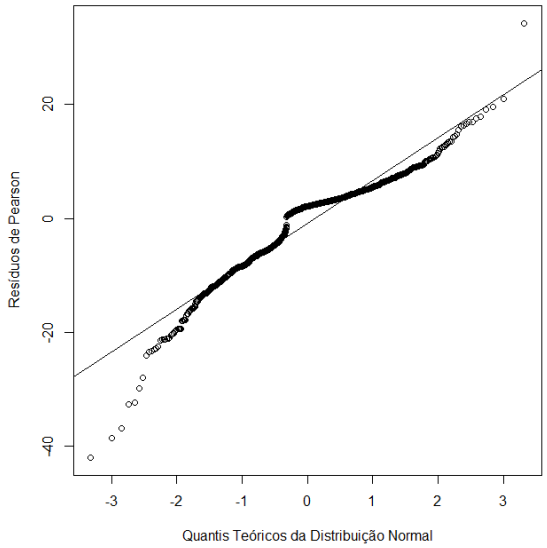


Figura 10 – Gráfico Q-Q de normalidade dos resíduos para $P(Y_i = 3)$

É perceptível no gráfico de normalidade dos dois níveis um ajuste ruim nos primeiros quantis, indicando que a distribuição possui caudas pesadas, e essa falta de ajustamento é ainda mais perceptível no nível de qualidade do sono ruim. Este comportamento indica alto distanciamento entre alguns valores preditos pelo modelo e os valores reais da qualidade do sono. Ao identificar os indivíduos correspondentes aos maiores desvios dos resíduos de Pearson, foi constatado que todos apresentavam boa qualidade do sono, o que indica deficiência no modelo para prever esse nível da variável reposta.

Outra ferramenta útil para o diagnóstico do modelo é o teste de Hosmer-Lemeshow, que consiste em separar os dados em grupos e comparar, em cada grupo, a proporção esperada do evento em questão com a proporção realmente observada naquele grupo. Apesar de não haver generalização da estatística do teste para dados complexos, a análise gráfica deste teste pode auxiliar na detecção de ajustamento ruim em algum nível. Os gráficos a seguir mostram a posição de cada grupo em relação à proporção prevista pelo modelo e a observada. Pontos próximos à linha indicam bom ajustamento do modelo.

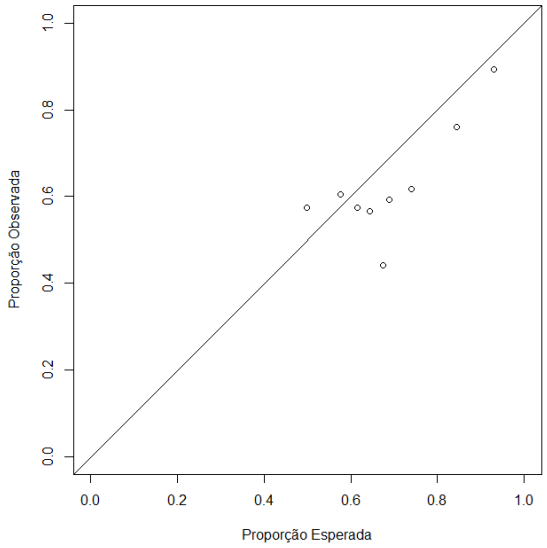


Figura 11 – Gráfico do teste de Hosmer-Lemeshow para a qualidade moderada do sono

A Figura 11 aponta alguns pontos um pouco afastados da linha, próximos à proporção esperada de 70%.

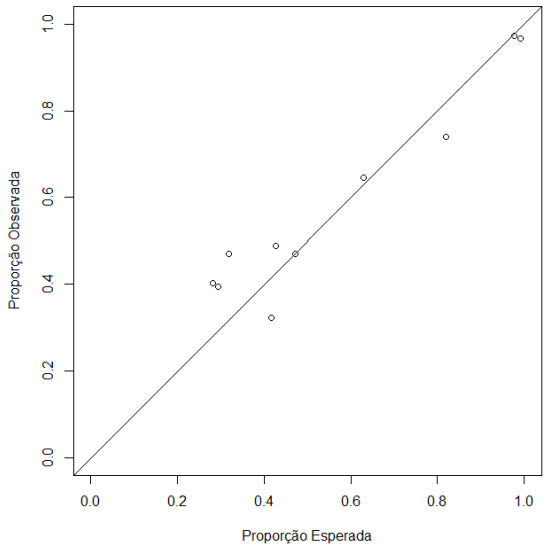


Figura 12 – Gráfico do teste de Hosmer-Lemeshow para a qualidade ruim do sono

A Figura 12 mostra um ajustamento um pouco melhor com alguns pontos sobre a linha de 45 graus, e um certo grau de afastamento nos níveis mais baixos.

5.3.2 Validação

Os modelos apresentados anteriormente foram construídos a partir de uma amostra aleatória simples de 80% dos dados originais. Para considerar os pesos amostrais e evitar que observações muito representativas tenham o mesmo peso na reamostragem que observações pouco representativas, a amostra foi selecionada dos dados censitários. Em seguida, os 20% restantes dos dados foram utilizados para calcular as probabilidades de cada observação pertencer a cada nível da variável resposta. A partir das probabilidades estimadas pelos modelos, os dados foram classificados utilizando as proporções da qualidade do sono (apresentados na Figura 2) como pontos de corte, como a seguir:

$$Y_i = \begin{cases} 1, & \text{se } P(Y_i = 1) \geq 0.22 \\ 2, & \text{se } P(Y_i = 2) \geq 0.49 \; e \; P(Y_i = 1) < 0.22 \\ 3, & \text{caso contrário} \end{cases}$$

Com os dados da amostra de teste devidamente classificados, é possível construir a matriz de classificação que compara os dados classificados com os dados reais.

Quadro 2 – Quadro dos acertos de previsão da qualidade do sono - modelo multinomial

Atual	Predito		
	1	2	3
1	138	185	109
2	229	348	184
3	36	127	242

É notável o nível crescente no acerto de previsões, com o nível 3 acertando a maior parte das previsões. O nível 2, apesar de não acertar a maior parte, ainda previu corretamente boa parte das observações, enquanto o nível 1 acertou apenas uma pequena parcela das previsões. Os Quadros 3 e 4 mostram as respectivas matrizes de confusão dos modelos binários separados. Essa separação permite o cálculo de algumas medidas de qualidade preditiva do modelo, apresentadas na Tabela 9.

Quadro 3 – Quadro dos acertos de previsão da qualidade do sono - modelo binário para o nível regular em relação ao nível bom

Atual	Predito	
	Positivo	Negativo
Positivo	139	186
Negativo	264	474

Quadro 4 – Quadro dos acertos de previsão da qualidade do sono - modelo binário para o nível ruim em relação ao nível bom

Atual	Predito	
	Positivo	Negativo
Positivo	354	246
Negativo	49	289

Tabela 9 – Medidas da Qualidade Preditiva dos Modelos Binários

Medida	$Y_i = 2$ (%)	$Y_i = 3$ (%)
Acurácia	57,67	68,55
Sensibilidade	42,77	59,00
Especificidade	64,23	85,50
Precisão	34,49	87,85
Valores Preditos Negativos	71,93	54,01
Área sob a curva roc	53,50	72,25

As medidas referentes ao nível ruim da qualidade do sono são melhores em quase

todos os indicadores, com exceção dos valores preditos negativos. Este resultado indica que uma parte considerável dos indivíduos com boa qualidade do sono não foram previstos corretamente, o que reforça o resultado obtido pela análise dos resíduos.

5.4 Modelo cumulativo

Existem muitas formas de checar a validade da suposição de proporcionalidade das razões de chances. Harrell (2015) recomendam utilizar abordagens gráficas para detectar diferenças significativas entre os efeitos individuais de cada variável nas probabilidades cumulativas da variável resposta. Agresti (2018) recomenda não utilizar este modelo quando há violação desta suposição, já que a utilização de coeficientes distintos nos logitos acumulados pode, em alguns casos, levar a estimativas viesadas.

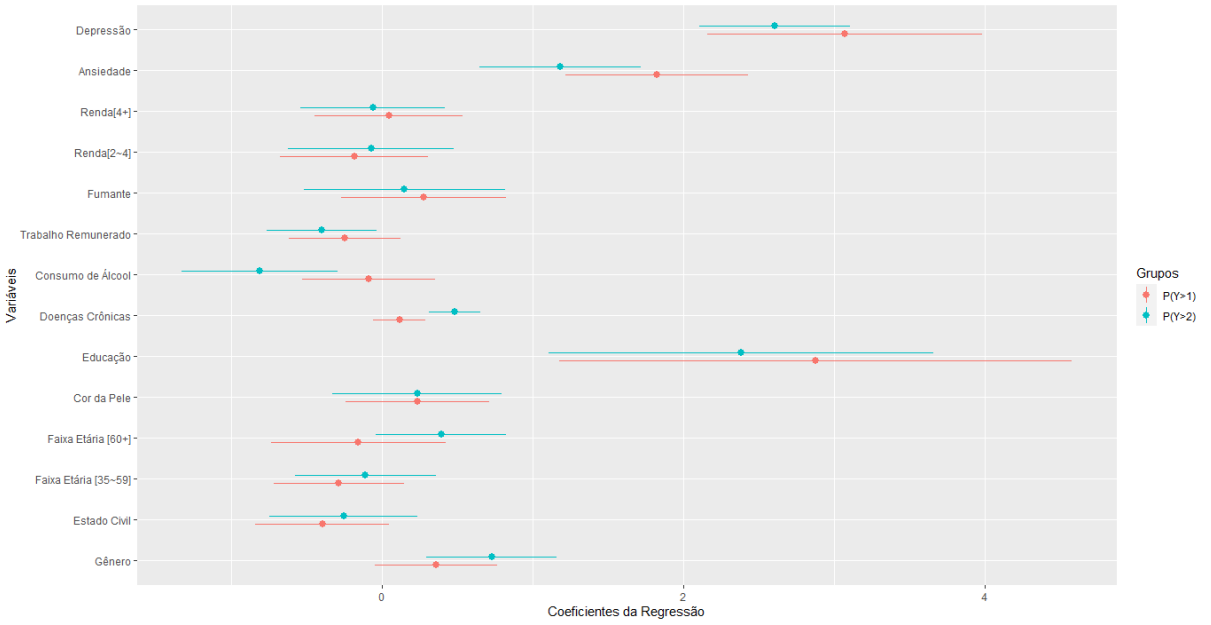


Figura 13 – Comparação dos coeficientes da regressão nos dois níveis da qualidade do sono

O gráfico da Figura 13 mostra que as variáveis consumo de álcool, DCNT e ansiedade apresentam algum grau de afastamento, indicando necessidade de utilizar coeficientes distintos para essas variáveis no modelo cumulativo. As tabelas 10 e 11 mostram os coeficientes estimados deste modelo para os níveis $Y_i \geq 2$ e $Y_i \geq 3$ e os respectivos erros-padrão e p-valores.

Tabela 10 – Modelo cumulativo final - nível moderado ou ruim ($Y_i \geq 2$)

Fatores de risco	Estimativa	Erro-padrão	valor-p
Intercepto	0.90	0.17	<0.001
Grau de Escolaridade (0–5° ano)	1.19	0.43	<0.01
Faixa Etária (>60)	-0.44	0.37	0.23
Doenças Crônicas	-0.34	0.15	0.02
Consumo de Álcool (sim)	-0.01	0.20	0.95
Ansiedade (sim)	1.31	0.30	<0.01
Depressão (sim)	2.42	0.28	<0.01
Consumo de Cigarros (sim)	0.40	0.29	0.17
Faixa Etária:Doenças crônicas	0.39	0.20	0.05
Ansiedade:Consumo de Cigarros	-0.48	0.49	0.32

Tabela 11 – Modelo cumulativo final - nível moderado ou ruim ($Y_i \geq 3$)

Fatores de risco	Estimativa	Erro-padrão	valor-p
Intercepto	0.90	0.17	<0.01
Grau de Escolaridade (0–5° ano)	1.19	0.43	<0.01
Faixa Etária (>60)	-0.44	0.37	0.23
Doenças Crônicas	0.21	0.10	0.04
Consumo de Álcool (sim)	-0.61	0.22	<0.01
Ansiedade (sim)	0.33	0.28	0.24
Depressão (sim)	2.42	0.28	<0.01
Consumo de Cigarros (sim)	0.40	0.29	0.17
Faixa Etária:Doenças crônicas	0.39	0.20	0.05
Ansiedade:Consumo de Cigarros	-0.48	0.49	0.32

As tabelas 12 e 13 mostram as razões de chances estimadas para este modelo e os respectivos intervalos de 95% de confiança.

Tabela 12 – Razões de chances do modelo cumulativo $P(Y_i \geq 2)$

Fatores de Risco	Interações	Razão de Chances	LI	LS
Grau de Escolaridade (0-5 ano)	-	3.30	1.40	7.77
Consumo de Álcool (Sim)	-	0.98	0.65	1.48
Depressão (Sim)	-	11.25	6.38	19.84
Doenças Crônicas	Faixa Etária >60 anos	1.35	0.59	3.07
	Faixa Etária <60 anos	0.90	0.72	1.14
Faixa Etária >60 anos	Doenças Crônicas = 0	0.64	0.30	1.33
	Doenças Crônicas = 1	0.95	0.42	2.16
Ansiedade (Sim)	Consumo de Cigarros = 0	3.71	2.03	6.78
	Consumo de Cigarros = 1	2.28	0.96	5.43
Consumo de Cigarros (Sim)	Ansiedade = 0	1.49	0.83	2.67
	Ansiedade = 1	0.92	0.38	2.19

Tabela 13 – Razões de chances do modelo cumulativo $(Y_i \geq 3)$

Fatores de Risco	Interações	Razão de Chances	LI	LS
Grau de Escolaridade (0-5 ano)	-	3.30	1.40	7.77
Consumo de Álcool (Sim)	-	0.54	0.35	0.83
Depressão (Sim)	-	11.25	6.38	19.84
Doenças Crônicas	Faixa Etária >60 anos	1.84	0.81	4.14
	Faixa Etária <60 anos	1.23	1.00	1.53
Faixa Etária >60 anos	Doenças Crônicas = 0	0.64	0.30	1.33
	Doenças Crônicas = 1	0.95	0.42	2.16
Ansiedade (Sim)	Consumo de Cigarros = 0	1.39	0.79	2.46
	Consumo de Cigarros = 1	0.86	0.35	2.08
Consumo de Cigarros (Sim)	Ansiedade = 0	1.49	0.83	2.67
	Ansiedade = 1	0.92	0.38	2.19

Assim como no modelo multinomial, os fatores Depressão e grau de escolaridade foram significativos para a qualidade do sono. A chance de uma pessoa que tem depressão ter um sono de baixa qualidade é 11 vezes maior em relação às que não tem depressão, ajustando-se pelas demais variáveis. Da mesma forma, a chance de uma pessoa com no máximo o ensino básico completo ter uma baixa qualidade do sono é 3 vezes maior em relação às que possuem grau de escolaridade maior.

5.4.1 Análise de Resíduos

As Figuras 14 a 17 mostram os gráficos dos resíduos para estes modelos, de forma análoga aos procedimento dos modelos nominais.

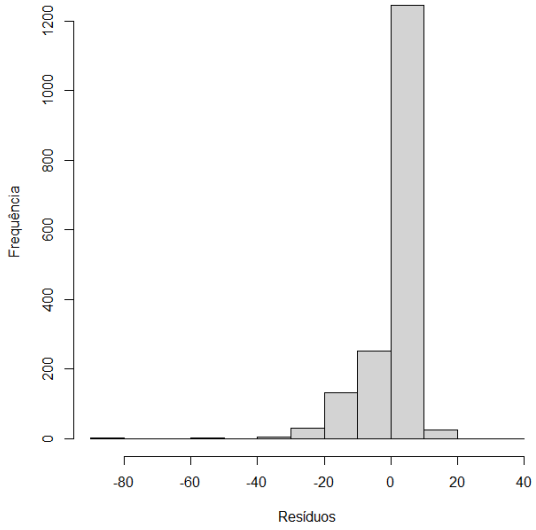


Figura 14 – Histograma do resíduo de Pearson para $P(Y_i \geq 2)$

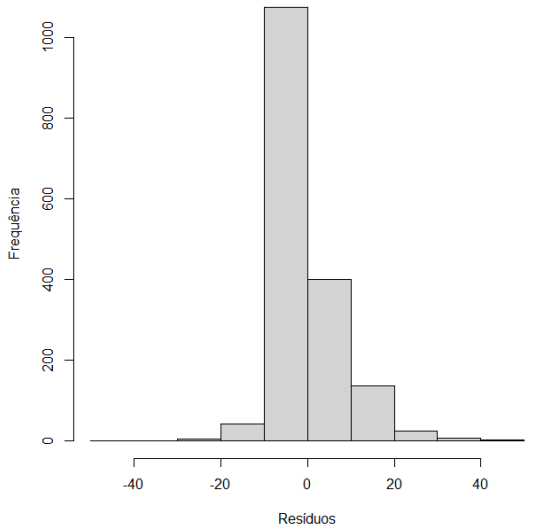


Figura 15 – Histograma do resíduo de Pearson para $P(Y_i \geq 3)$

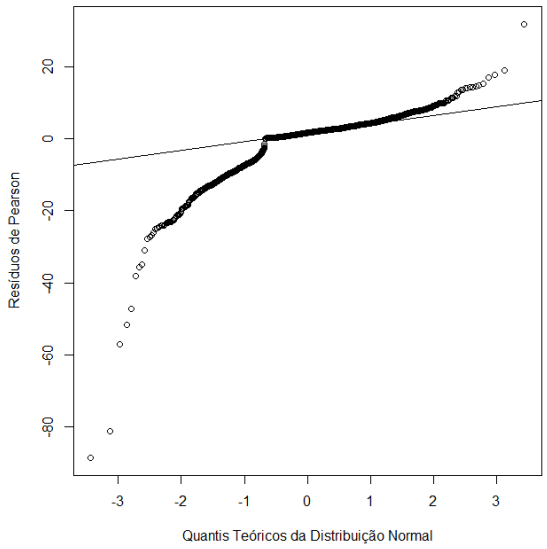


Figura 16 – Gráfico Q-Q de normalidade dos resíduos de Pearson para $P(Y_i \geq 2)$

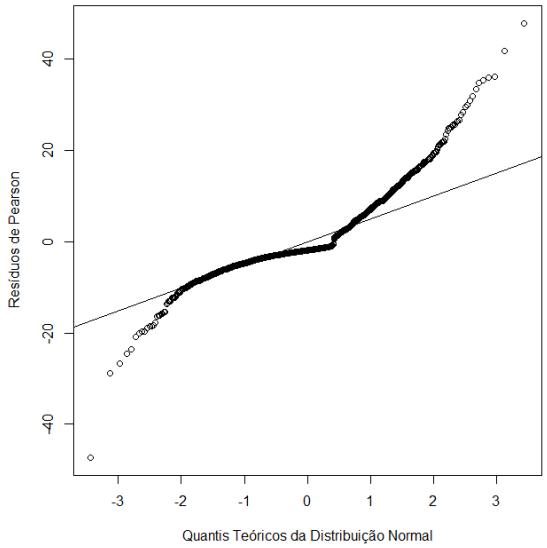


Figura 17 – Gráfico Q-Q de normalidade dos resíduos de Pearson para $P(Y_i \geq 3)$

É perceptível que a qualidade do ajuste em ambos os níveis aparenta ser pior do que nos gráficos correspondentes do modelo multinomial.

5.4.2 Validação do Modelo

Os Quadros 5 e 6 mostram as matrizes de confusão para $Y_i \geq 2$ e $Y_i \geq 3$, respectivamente

Quadro 5 – Quadro dos acertos de previsão da qualidade do sono - modelo binário para o nível regular ou superior

Atual	Predito	
	Positivo	Negativo
Positivo	138	294
Negativo	265	901

Quadro 6 – Quadro dos acertos de previsão da qualidade do sono - modelo binário para o nível ruim ou superior

Atual	Predito	
	Positivo	Negativo
Positivo	900	287
Negativo	163	248

A Tabela 14 resume algumas medidas das matrizes de confusão para os dois níveis do modelo cumulativo.

Tabela 14 – Medidas de Qualidade Preditiva dos Modelos Cumulativos Separados

Medida	$Y_i \geq 2$ (%)	$Y_i \geq 3$ (%)
Acurácia	65,02	71,84
Sensibilidade	31,94	75,82
Especificidade	77,27	60,34
Precisão	34,24	84,67
Valores Preditos Negativos	75,39	48,36
Área sob a curva roc	54,61	68,08

Ao contrário do modelo nominal, o modelo cumulativo não apresentou índices melhores no nível mais alto. Apesar deste modelo indicar uma especificidade muito maior no terceiro nível do que o modelo nominal, a sensitividade se mostrou muito mais baixa em relação ao modelo nominal. Esta diferença entre os modelos pode indicar que o modelo nominal consegue prever melhor o nível de qualidade do sono ruim, enquanto o modelo cumulativo consegue prever os níveis mais baixos com mais precisão. Como o objetivo do estudo é identificar fatores de risco para a qualidade do sono ruim, esta comparação entre os modelos justifica a escolha do modelo nominal.

6 Conclusão

No final, foi constatado que fatores como depressão, baixo nível de instrução e ansiedade em não-fumantes são fatores de risco para a baixa qualidade do sono. A chance de uma pessoa com depressão ter um sono de baixa qualidade é cerca de 30 vezes maior do que entre pessoas sem depressão. Da mesma forma, a chance de uma pessoa que não concluiu o Ensino Médio ter um sono de baixa qualidade é cerca de 12 vezes maior do que de uma pessoa de nível de instrução mais alto. O consumo de álcool se mostrou ser um fator benéfico para a qualidade do sono, com as pessoas que consomem álcool apresentando chances 3 vezes maior de terem uma boa qualidade do sono em relação às que não consomem.

Alguns fatores de risco que foram consideradas em outros estudos como fatores de risco para a qualidade do sono não foram significativas para o modelo, como gênero, renda, trabalho remunerado e estado civil. Estas diferenças podem estar associadas à pandemia de COVID-19, que alterou as relações sociais e as condições sociais das pessoas, e também ao contexto geográfico, já que o estudo está limitado à região de Ouro Preto.

Apesar do grande tamanho de amostra utilizado no estudo, algumas variáveis consideradas importantes tiveram poucas observações, como a variável Ensino Fundamental incompleto, implicando em intervalos de confiança muito amplos e impossibilitando incluir interações envolvendo esta variável no modelo. O estudo também mostrou que uma melhor explicação da qualidade do sono pode ser feita incluindo variáveis que ajudam a prever os níveis mais altos da qualidade do sono, como a prática de exercícios físicos e a padronização dos horários do sono.

Índice da qualidade do sono de Pittsburgh

As seguintes perguntas são relativas aos seus hábitos de sono **durante o ultimo mês somente**. Suas respostas devem indicar a lembrança mais exata da maioria dos dias e noites do ultimo mês. Por favor, responda a todas as perguntas.

Nome:

Idade:

Data:

1. Durante o último mês, quando você geralmente foi para a cama a noite?

   hora usual de deitar:

2. Durante o último mês, quanto tempo (em minutos) você geralmente levou para dormir a noite?

   número de minutos:

3. Durante o último mês, quando você geralmente levantou de manhã?

   hora usual de levantar?

4. Durante o último mês, quantas horas de sono você teve por noite? (Esta pode ser diferente do número de horas que você ficou na cama)

   Horas de sono por noite:

5. Durante o último mês, com que frequência você teve dificuldade para dormir porque você:

   A) não conseguiu adormecer em até 30 minutos
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

   B) acordou no meio da noite ou de manhã cedo
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

   C) precisou levantar para ir ao banheiro
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

   D) não conseguiu respirar confortavelmente
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

   E) tossiu ou roncou forte
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

F) Sentiu muito frio
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

G) sentiu muito calor
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

H) teve sonhos ruins
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

I) teve dor
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

J) outras razões, por favor descreva: _____
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

6. Durante o último mês como você classificaria a qualidade do seu sono de uma maneira geral:
Muito boa Boa Ruim Muito ruim

7. Durante o último mês, com que frequencia você tomou medicamento (prescrito ou por conta própria) para lhe ajudar
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

8. No último mês, que frequencia você teve dificuldade para ficar acordado enquanto dirigia, comia ou participava de uma atividade social (festa, reunião de amigos)
1 = nenhuma no último mês 2 = menos de uma vez por semana
3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

9. Durante o último mês, quão problemático foi pra você manter o entusiasmo (ânimo) para fazer as coisas (suas atividades habituais)?
Nenhuma dificuldade Um problema leve
Um problema razoável Um grande problema

10. Você tem um parceiro (a), esposo (a) ou colega de quarto?

A) Não
B) Parceiro ou colega, mas em outro quarto

C) Parceiro no mesmo quarto, mas em outra cama
D) Parceiro na mesma cama

Se você tem um parceiro ou colega de quarto pergunte a ele com que freqüência, no ultimo mês você apresentou:

E) Ronco forte
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

F) Longas paradas de respiração enquanto dormia
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

G) contrações ou puxões de pernas enquanto dormia
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

D) episodios de desorientação ou confusão durante o sono
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

E) Outras alterações (inquietações) enquanto você dorme, por favor descreva:_____
   1 = nenhuma no último mês 2 = menos de uma vez por semana
   3 = uma ou duas vezes por semana 4 = três ou mais vezes na semana

Código em R

```
library(survey)
library(caret)
library(stats)
library(MASS)
library(VGAM)
library(svyVGAM)
extdata<-expandRows(data, 'pesofinalcal')
newdatatest<-extdata
trainIndex <- createDataPartition(extdata$psqi_index, p = .8,
                                  list = FALSE,
                                  times = 1)

Train <- newdatatest[ trainIndex,]
Test  <- newdatatest[-trainIndex,]
delintrain<-svydesign(id=~setor_censitario,strata=~painel,
weights=~pesofinalcal,data=df_merge,nest=TRUE)

#modelo multinomial#
svy_vglm(factor(psqi_index, ordered = FALSE) ~
    education0 + age_cat3 * cncds + alcohol
    + depression + anxiety * smoke,
    family = multinomial(refLevel = 1),
    design = delintrain)

#modelo cumulativo#
svy_vglm(psqi_index ~ education0 + cncds * age_cat3 +
    alcohol + anxiety * smoke + depression,
    family = cumulative(parallel = F ~
    anxiety + cncds + alcohol, reverse = T), delintrain)

#modelos dicotomicos#
svyglm(formula = I(psqi_index == 2) ~ education0 + age_cat3 *
    cncds + alcohol + anxiety + depression + alcohol + anxiety *
    smoke, design = subset(delintrain, psqi_index %in% c(1, 2)),
    family = "quasibinomial")
svyglm(formula = I(psqi_index == 3) ~ education0 + age_cat3 *
```

```
    cncds + alcohol + anxiety + depression + alcohol + anxiety *
    smoke, design = subset(delintrain, psqi_index %in% c(1, 3)),
    family = "quasibinomial")
svyglm(formula = I(psqi_index >= 2) ~ education0 + age_cat3 *
    cncds + alcohol + anxiety + depression + alcohol + anxiety *
    smoke, design = delintrain,
    family = "quasibinomial")
svyglm(formula = I(psqi_index >= 3) ~ education0 + age_cat3 *
    cncds + alcohol + anxiety + depression + alcohol + anxiety *
    smoke, design = delintrain,
    family = "quasibinomial")

#Estimativas das medias e proporcoes#
svyby(~psqi_index,~cncds,delintotal,svymean)
svyby(~psqi_index,~depression,delintotal,svymean)
svyby(~psqi_index,~education,delintotal,svymean)
svyby(~gender,~depression,delintotal,svymean)

#graficos dos residuos#
ri2<-sqrt(df_merge[df_merge$psqi_index!=3,]$pesofinalcal)*
((df_merge[df_merge$psqi_index!=3,]$psqibase2-fitted(modbasef2))
/sqrt(fitted(modbasef2)*(1-fitted(modbasef2))))
qqnorm(ri2, ylab="Residuos de Pearson",
       xlab="Quantis Teoricos da Distribuicao Normal",main=NULL)
qqline(ri2)
```

Referências

Agresti, Alan: *An introduction to categorical data analysis*. John Wiley & Sons, 2018

Bendel, Robert B. ; Afifi, A. A.: Comparison of Stopping Rules in Forward "Stepwise"Regression. In: *Journal of the American Statistical Association* 72 (1977), Nr. 357, S. 46–53

Bertolazi, Alessandra N. ; Fagondes, Simone C. ; Hoff, Leonardo S. ; Dartora, Eduardo G. ; Silva Miozzo, Ilsis C. da ; Barba, Maria Emília F. de ; Barreto, Sérgio Saldanha M.: Validation of the Brazilian Portuguese version of the Pittsburgh sleep quality index. In: *Sleep medicine* 12 (2011), Nr. 1, S. 70–75

Drager, Luciano F. ; Pachito, Daniela V. ; Morihisa, Rogerio ; Carvalho, Pedro ; Lobao, Abner ; Poyares, Dalva: Sleep quality in the Brazilian general population: A cross-sectional study. In: *Sleep Epidemiology* 2 (2022), S. 100020

Haleem, Abid ; Javaid, Mohd ; Vaishya, Raju: Effects of COVID-19 pandemic in daily life. In: *Current medicine research and practice* 10 (2020), Nr. 2, S. 78

Harrell, SFrank E.: *Regression Modeling Strategies*. Springer New York, NY, 2015

Hosmer, David W. ; Lemeshow, Stanley: *Applied logistic regression*. Wiley New York, 2000

Hosmer Jr, David W. ; Lemeshow, Stanley ; Sturdivant, Rodney X.: *Applied logistic regression*. Bd. 398. John Wiley & Sons, 2013

Kim, Tae W. ; Jeong, Jong-Hyun ; Hong, Seung-Chul: The impact of sleep and circadian disturbance on hormones and metabolism. In: *International journal of endocrinology* 2015 (2015)

Lin, Li-yu ; Wang, Jie ; Ou-Yang, Xiao-Yong ; Miao, Qing ; Chen, Rui ; Liang, Feng-xia ; Zhang, Yang-pu ; Tang, Qing ; Wang, Ting: The immediate impact of the 2019 novel coronavirus (COVID-19) outbreak on subjective sleep status. In: *Sleep medicine* 77 (2021), S. 348–354

Lumley, Thomas: *svyVGAM: Design-Based Inference in Vector Generalised Linear Models*, 2023. – URL ⟨https://CRAN.R-project.org/package=svyVGAM⟩. – R package version 1.2

Lumley, Thomas ; Scott, Alastair: AIC and BIC for modeling with complex survey data. In: *Journal of Survey Statistics and Methodology* 3 (2015), Nr. 1, S. 1–18

Menezes-Júnior, Luiz Antônio A. de ; Souza Andrade, Amanda C. de ; Coletro, Hillary N. ; Deus Mendonça, Raquel de ; Menezes, Mariana C. de ; Machado-Coelho, George Luiz L. ; Meireles, Adriana L.: Food consumption according to the level of processing and sleep quality during the COVID-19 pandemic. In: *Clinical Nutrition ESPEN* 49 (2022), S. 348–356

Richter, S ; Schilling, L ; Camargo, N ; Taurisano, M ; Fernandes, N ; Silva, LE u. a.: How COVID-19 quarantine might affect the sleep of children and adolescents. In: *Residência Pediátrica* 11 (2021), S. 1–5

Shukla, Charu ; Basheer, Radhika: Metabolic signals in sleep regulation: recent insights. In: *Nature and science of sleep* 8 (2016), S. 9

Simões, Naiane D. ; Monteiro, Luiz Henrique B. ; Lucchese, Roselma ; Amorim, Thiago Aquino d. ; Denardi, Tainara C. ; Vera, Ivânia ; Silva, Graciele C. ; Sverzut, Carolina: Quality and sleep duration among public health network users. In: *Acta Paulista de Enfermagem* 32 (2019), S. 530–537

Souza, Luiz Felipe Ferreira d. ; Paineiras-Domingos, Laisa L. ; Melo-Oliveira, Maria Eduarda de S. ; Pessanha-Freitas, Juliana ; Moreira-Marconi, Eloá ; Lacerda, Ana Cristina R. ; Mendonça, Vanessa A. ; SÁ-Caputo, Danubia da C. ; Bernardo-Filho, Mario: The impact of COVID-19 pandemic in the quality of sleep by Pittsburgh Sleep Quality Index: A systematic review. In: *Ciência & Saúde Coletiva* 26 (2021), S. 1457–1466

Steven G. Heeringa, Patricia A. B.: *Applied survey data analysis*. Taylor and Francis Group, 2010

Wang, Peng ; Song, Lin ; Wang, Kaili ; Han, Xiaolei ; Cong, Lin ; Wang, Yongxiang ; Zhang, Lei ; Yan, Zhongrui ; Tang, Shi ; Du, Yifeng: Prevalence and associated factors of poor sleep quality among Chinese older adults living in a rural area: a population-based study. In: *Aging clinical and experimental research* 32 (2020), Nr. 1, S. 125–131

Yee, T. W.: *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA : Springer, 2015