UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA
ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

ANÁLISE DE DADOS DISCRETOS

Predição de níveis de obesidade baseados em hábitos alimentares e atividades físicas

Vitória Linda da Silva Oliveira - RA212826

Brenda Luiza Correa - RA216037

Paula Liserre Calabrez - RA242782

CAMPINAS

2024

1. Introdução

A obesidade é um dos principais desafios de saúde pública enfrentados globalmente, com implicações significativas para a saúde individual e os sistemas de saúde em todo o mundo. Países como México, Peru e Colômbia não estão imunes a esse problema crescente, que é influenciado por uma série de fatores, incluindo hábitos alimentares, atividade física e estilo de vida.

Assim, para analisar os níveis de obesidade em indivíduos desses países, explorando um conjunto de dados abrangente que reúne informações sobre hábitos alimentares, condição física e outras características, foi utilizada a regressão logística.

A regressão logística é uma extensão da regressão linear que é especialmente adequada para lidar com variáveis dependentes binárias ou categóricas ordinais. Ao contrário da regressão linear, que produz valores contínuos como saída, a regressão logística estima a probabilidade de um evento com base nas variáveis independentes. Essa probabilidade é restrita a valores entre 0 e 1, tornando-a ideal para problemas de classificação, onde há interesse em prever a probabilidade de uma observação pertencer a uma determinada classe, como por exemplo, estudado neste trabalho, a probabilidade de uma observação estar em algum nível de obesidade.

2. Objetivos

Tendo em vista que a obesidade, responsável por problemas físicos e mentais, representa um desafio de saúde global com consequências sérias. Este trabalho tem como objetivo utilizar a regressão logística como uma ferramenta analítica para investigar e prever os níveis de obesidade com base no banco de dados estudado. A análise se torna necessária pois, compreender os fatores associados à obesidade e prever os níveis de obesidade em uma população específica são passos essenciais para o desenvolvimento de estratégias eficazes de prevenção e intervenção.

3. Metodologia

3.1 Regressão Logística

A Regressão Logística Ordinal modela a relação entre variáveis independentes e uma variável resposta ordinal, estimando as probabilidades cumulativas de cada categoria. Assim, para analisar esses dados utilizaremos Regressão Logística Politômica Ordinal, modelo com logito acumulado para respostas ordinais com mais de dois níveis, uma vez que nossa variável resposta, o Nível de Obesidade, possui quatro categorias com ordem estabelecida.

O modelo de odds proporcional para predizer $P(Y \leq j)$ é:

$$logito[P(Y \leq j)] = \alpha_j + \beta x, \; j = 1, \ldots, J-1.$$

Então, a função log odds é dada por:

$$P(Y \leq j) = \pi_1 + \pi_2 + \ldots + \pi_j = \frac{exp(\alpha_j + x\beta)}{1 + exp(\alpha_j + x\beta)}.$$

No qual, β descreve o efeito de x (covariável) na log odds da resposta na categoria j ou abaixo.

Este modelo assume que o efeito de x é o mesmo para todos os $J-1$ logitos acumulados, assim, quando bem ajustado, precisamos de um único parâmetro para descrever o efeito de x.

Na interpretação do modelo utilizaremos razões de chances (odds) para examinar a relação entre as variáveis independentes e as probabilidades acumuladas (ou complementares).

3.2 Conjunto de Dados

O conjunto de dados inclui dados do México, Peru e Colômbia, com base nos seus hábitos alimentares e atividades físicas de diversos indivíduos. O conjunto de dados contém 17 variáveis e 498 observações. A variável NObeyesdad (Nível de Obesidade) permite a classificação dos indivíduos, de acordo com seus IMCs, em Peso Insuficiente, Peso Normal, Sobrepeso Nível I, Sobrepeso Nível II, Obesidade Tipo I, Obesidade Tipo II e Obesidade Tipo III.

Para facilitar a análise descritiva e a modelagem a variável resposta foi agrupada da seguinte maneira: Peso Insuficiente, Peso Normal, Sobrepeso, contendo indivíduos com Sobrepeso Nível I e Sobrepeso Nível II, e Obesidade, contendo indivíduos com Obesidade Tipo I, Obesidade Tipo II e Obesidade Tipo III. Os valores do IMC com suas respectivas classificações estão presentes na Figura I.

Os dados foram coletados a partir de uma base de dados pública que agrupava respostas de uma enquete focada em hábitos alimentares e outras características.

- **Gender** (gênero): variável categórica (Feminino, Masculino);
- **Age** (idade): variável contínua (em anos);
- **Height** (altura): variável contínua (em metros);
- **Weight** (peso): variável contínua (em kg);
- **family_history_with_overweight** (se algum membro da família sofreu ou sofre de excesso de peso): variável binária (Sim, Não);
- **FAVC** (se a pessoa come alimentos com alto teor calórico com frequência): variável binária (Sim, Não);
- **FCVC** (frequência do consumo de vegetais): variável ordinal (1 = nunca, 2 = ocasionalmente, 3 = sempre);
- **NCP** (quantas refeições faz diariamente): variável ordinal (1 = entre uma e duas, 2 = três, 3 = mais de três, 4 = sem resposta);
- **CAEC** (frequência que se come algum alimento entre as refeições): variável ordinal (Não, Ocasionalmente, Frequentemente, Sempre);
- **SMOKE** (O indivíduo fuma?): variável binária (Sim, Não);

- **CH2O** (consumo diário de água): variável ordinal (1 = menos de um litro, 2 = entre um e dois litros, 3 = mais de três litros);
- **SCC** (se monitora as calorias ingeridas diariamente): variável binária (Sim, Não);
- **FAF** (frequência em que pratica atividades físicas): variável ordinal (1 = nunca, 2 = uma ou duas vezes por semana, 3 = três vezes por semana, 4 = quatro ou mais vezes por semana);
- **TUE** (tempo de uso diário de dispositivos tecnológicos): variável ordinal (0 = nenhum, 1 = menos de uma hora, 2 = entre uma e três horas, 3 = mais de três horas);
- **CALC** (frequência em que consome bebidas alcoólicas): variável ordinal (Não, Ocasionalmente, Frequentemente, Sempre);
- **MTRANS** (tipo de transporte utilizado): variável categórica (1 = caminhada e bicicleta, 2 = transporte público, 3 = automóvel e moto);
- **NObeyesdad** (nível de obesidade de acordo com o IMC): variável ordinal (Peso.Insuficiente, Peso.Normal, Sobrepeso, Obesidade).

3.3 Análise Descritiva

Para a análise descritiva utilizamos o IMC dos indivíduos para criar gráficos de boxplot para analisar visualmente o efeito das variáveis independentes com a variável resposta. O IMC é a sigla para Índice de Massa Corpórea, parâmetro adotado pela Organização Mundial de Saúde para calcular o peso ideal de cada pessoa e classificá-la de acordo com o valor obtido, o índice é calculado dividindo o peso do paciente pela sua altura elevada ao quadrado.

Tabela 1: Valores do IMC: pessoas de 20 a 60 anos

Valor do IMC	Classificação
Menor que 18,5	Peso insuficiente
De 18,5 a 24,99	Normal
De 25 a 29,99	Sobrepeso
Maior que 30	Obesidade

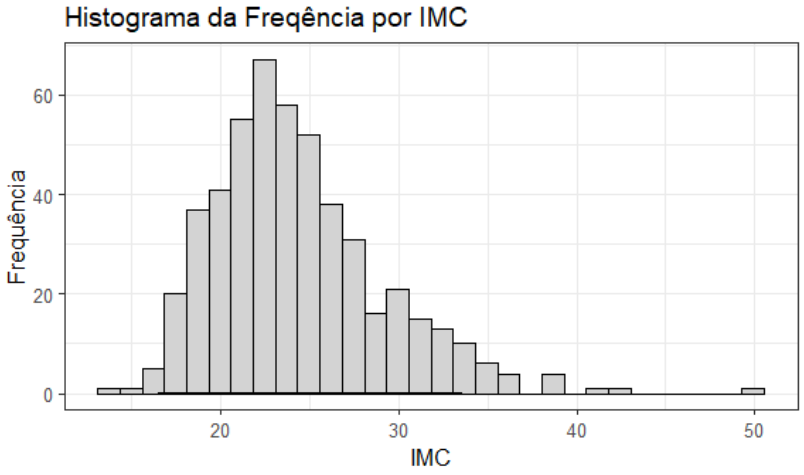

Figura 1: Histograma da frequência por IMC

Na Figura 1, nota-se que uma assimetria à direita, com uma concentração em valores entre 20 e 30, indicando uma presença maior de indivíduos com peso normal e sobrepeso, condizente com a Tabela 2:

Tabela 2: Contagem de indivíduos por categoria

Peso Insuficiente	Peso Normal	Sobrepeso	Obesidade
34	287	116	61

Tabela 3: Estatísticas sumárias do IMC

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Desvio Padrão
13,29	21,00	23,69	24,31	26,67	49,47	4,77

A Tabela 3 apresenta as estatísticas sumárias do IMC.Observa-se que o IMC varia significativamente na amostra, com um mínimo de 13,29 e um máximo de 49,47. Isso indica uma ampla gama de valores de IMC entre os indivíduos estudados, refletindo diferentes composições corporais. A mediana sugere que metade da amostra possui um IMC abaixo de 23,69, enquanto a média é ligeiramente mais alta, indicando uma possível influência de valores mais altos.

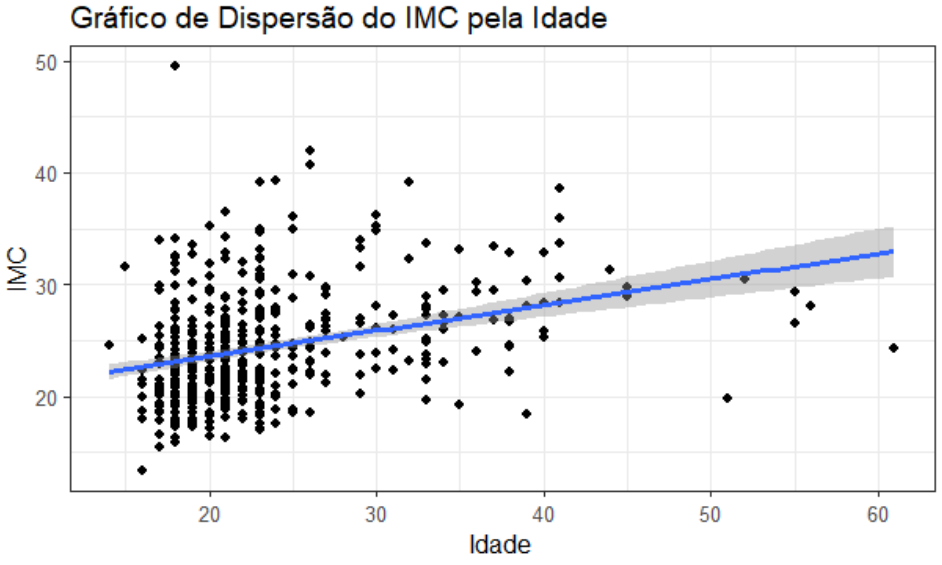

Figura 2: Gráfico de dispersão Idade x IMC

Na Figura 2, para a ver o efeito da idade no IMC foi feito um gráfico de dispersão das duas variáveis, nota-se pelo gráfico uma pequena relação positiva, ou seja, conforme a idade aumenta o IMC também aumenta, porém há uma grande concentração de pontos nas idades próximas de 20 anos. Além do gráfico, foi calculada também a correlação entre as variáveis e obteve-se um valor de 0.32, ou seja, ele possui uma correlação positiva como indicado na análise visual, mas a correlação é baixa em termos de magnitude, ou seja, apesar de ter efeito no IMC, ele é muito pequeno. Portanto, a idade não possui um efeito forte no valor do IMC.

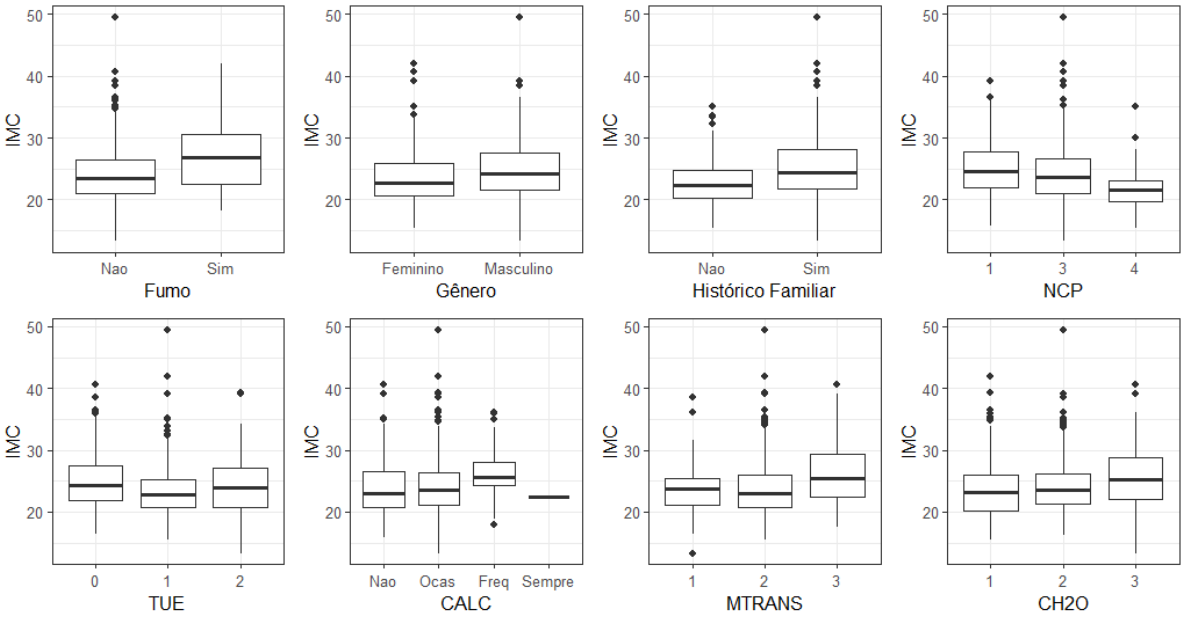

Figura 3: Boxplots das variáveis que aparentam ter efeito na variável resposta

Na Figura 3 é apresentado os gráficos com as variáveis que aparentam ter efeito na variável resposta. Nas variáveis binárias, os fumantes, o gênero masculino e indivíduos com histórico familiar de sobrepeso apresentaram valores de IMC maiores. Nas outras variáveis categóricas, pelo menos uma categoria se apresentou maior ou menor que as outras. Assim, as variáveis que visualmente aparentam ter efeito foram aquelas que notou-se certa diferença entre os boxplots observados.

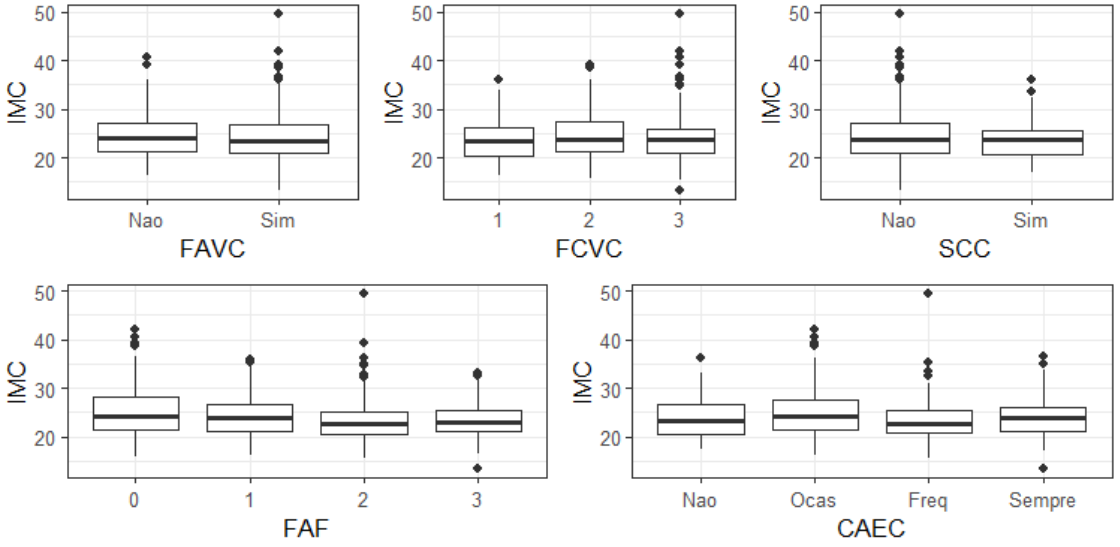

Figura 4: Boxplots das variáveis que não aparentam ter efeito na variável resposta

Na figura 4 é apresentado os gráficos com as variáveis que aparentam não ter efeito na variável resposta, ou seja, não foi possível notar diferença entre os boxplots observados, com os valores do IMC não se alterando de acordo com as categorias.

4. Aplicação

Para a modelagem utilizou-se Regressão Logística Ordinal Politômica devido a natureza da variável resposta.

Suposições do modelo:

1. Variável dependente ordinal;
2. Independência das observações (sem medidas repetidas);
3. Ausência de multicolinearidade, ou seja, as variáveis independentes não possuem uma alta correlação entre elas;
4. Chances proporcionais.

Com o objetivo de diminuir o número de variáveis preditoras, uma vez que o banco de dados possui 17 variáveis, utilizou-se o método de seleção backward com base nos valores do critério de informação de Akaike (AIC) e as seguintes variáveis foram selecionadas: Gênero, idade, família com histórico de sobrepeso, NCP, CAEC, SMOKE, CH2O e FAF, o modelo com todas as variáveis foi testado, mas não se mostrou satisfatório no teste de chances proporcionais, com p-valor do modelo geral de 0.03, rejeitando a hipótese de chances proporcionais. Assim, outros modelos foram testados até chegar no melhor ajustado, com as variáveis gênero, família com histórico de sobrepeso, NCP, CAEC, SMOKE, CH2O , FAF e MTRANS, sendo incluída apenas uma variável que considerou-se não ter efeito no nível de obesidade, a FAF.

Na checagem de pressupostos, o primeiro e o segundo são teóricos e válidos para o banco de dados utilizado. Já para verificar a multicolinearidade foi usado o valor de inflação VIF, caso este fosse maior do que 10, existiria multicolinearidade entre as variáveis, mas todos os valores apresentados foram próximos de um. Para testar as chances proporcionais, utilizou-se um teste com distribuição Qui Quadrado que apresentou um valor-p de 0.09, assim, com um nível de significância de 5%, não rejeitou a hipótese nula de chances proporcionais. Desse modo, todos os pressupostos são válidos nesse modelo.

Os testes de significância dos efeitos globais do modelo e os valores com os coeficientes estão apresentados na Tabela 4 e Tabela 5, respectivamente.

Tabela 4: Testes de significância dos efeitos globais

	Qui Quadrado	Graus de Liberdade	Pr (> Qui Quadrado)
Gênero	9,5100	1	0,0020436
Histórico familiar de sobrepeso	15,8800	1	0,0000000
Número de refeições diárias	8,5231	2	0,0141007
Frequência que come algum alimento entre as refeições	11,3623	3	0,0099200
Se fuma	5,4264	1	0,0198343
Consumo diário de água	8,0358	2	0,0179903
Frequência em que pratica atividades físicas	11,3098	3	0,0101634
Tipo de transporte utilizado	14,6515	2	0,0006584

Pela Tabela 4, observa-se que todos os efeitos globais são significativos para o modelo, sob um nível de significância de 5%.

Tabela 5: Coeficientes dos efeitos específicos

	Valor	Desvio Padrão	t-valor	p-valor
Peso Insuficiente | Peso Normal	-1,5556	0,3617	-4,3006	0,0000
Peso Normal | Sobrepeso	2,0834	0,3597	5,7914	0,0000
Sobrepeso | Obesidade	3,6619	0,3866	9,4715	0,0000
Gênero Masculino	0,5793	0,1891	3,0632	0,0022
Histórico familiar com sobrepeso	0,7713	0,1961	3,9326	0,0001
Mais de 3 refeições diárias	-0,7533	0,2815	-2,6757	0,0075
Sem resposta sobre refeições diárias	-0,0305	0,1782	-0,1711	0,8642
Às vezes come algum alimento entre as refeições	0,0350	0,3744	0,0934	0,9256
Frequentemente come algum alimento entre as refeições	-0,0912	0,2941	-0,3102	0,7564
Sempre come algum alimento entre as refeições	0,5521	0,1910	2,8906	0,0038
Fuma	0,8510	0,3627	2,3464	0,0190
Consumo diário de água entre 1L e 2L	0,4985	0,2007	2,4836	0,0130
Consumo diário de água maior que 3L	0,2398	0,1499	1,5994	0,1097
Pratica atividade física 1 ou 2 vezes por semana	-0,5741	0,2181	-2,6321	0,0085
Pratica atividade física 3 vezes por semana	0,2380	0,1980	1,2024	0,2292
Pratica atividade física 4 ou mais vezes por semana	0,1366	0,1836	0,7439	0,4570
Utiliza transporte público	0,1944	0,2898	0,6709	0,5023
Utiliza automóvel ou moto	0,9743	0,3250	2,9980	0,0027

Na Tabela 5, é possível observar os valores dos interceptos dos três modelos utilizados e os valores dos coeficientes dos efeitos específicos, que são fixos para os modelos, ou seja, apenas os interceptos que se alteram. Também está presente os valores-p dos testes de significância de cada efeito.

Assim, observando-se os resultados dos testes tem-se que os efeitos específicos: Gênero Masculino, Histórico familiar com sobrepeso, Mais de 3 refeições diárias, Sempre come algum alimento entre as refeições, Fumante, Consumo diário de água entre 1L e 2L, Pratica atividade física 1 ou 2 vezes por semana e Utiliza automóvel ou moto são significativos e portanto podem ser interpretados utilizando a razão de chances com relação à categoria de referência e seus respectivos intervalos de confiança.

A partir da Tabela 5, analisou-se os efeitos específicos de cada variável independente utilizando a razão de chance com IC(Intervalo de confiança) 95% (usando a exponencial da log-verossimilhança). Os efeitos não significativos não foram interpretados, uma vez que, por não terem, estatisticamente analisando, influência sobre o nível de obesidade, interpretá-los não faria sentido para a análise. As informações sobre a razão de chances e intervalos de confiança são apresentados na Tabela 6.

Tabela 6: Razão de chances e IC 95%

Característica	Razão de Chances	IC 95%	p-valor
Gênero			0,0020
● Masculino	1,785	1,234; 2,592	
Histórico Familiar com sobrepeso			0,0001
● Sim	2,163	1,477; 3,188	
Número de refeições diárias			0,0141
● Mais de 3	0,471	0,270; 0,814	
Frequência que come algum alimento entre as refeições			0,0099
● Sempre	1,737	1,198; 2,535	
Se fuma			0,0198
● Sim	2,342	1,146; 4,776	
Consumo diário de água			0,0180
● Entre 1L e 2L	1,646	1,111; 2,443	
Frequência em que pratica atividades físicas			0,0102
● 1 ou 2 vezes por semana	0,563	0,366; 0,861	
Tipo de transporte utilizado			0,0007
● Automóvel ou moto	2,649	1,408; 5,040	

Antes de analisar cada categoria individualmente, vale ressaltar que:

- RC < 1 indica que a categoria de referência tem menores chances de estar em uma categoria superior comparada à categoria específica.
- RC > 1 indica que a categoria específica tem maiores chances de estar em uma categoria superior comparada à categoria de referência.

Analisando os intervalos de confiança de 95%, um intervalo mais estreito geralmente indica uma estimativa mais precisa da razão de chances e além disso, nota-se que nenhum intervalo de confiança contém o valor 1, o que condiz com o teste de significância apresentado na Tabela 5, uma vez que, o intervalo incluir o valor 1 sugere que não há uma associação estatisticamente significativa.

Para os indivíduos do sexo masculino em comparação com os do sexo feminino, razão de chances de 1.785 sugere que a probabilidade de pertencer a uma categoria de peso superior é 1.785 vezes maior para homens do que para mulheres, mantendo todas as outras variáveis constantes.

Indivíduos com histórico familiar de sobrepeso têm razão de chances de 2.163 em comparação com aqueles sem histórico familiar. Isso significa que a probabilidade de estar em uma categoria de peso superior é 2.163 vezes maior para indivíduos com histórico familiar de sobrepeso, ajustado para outras variáveis no modelo.

Quando fazem mais de três refeições diárias a razão de chances é de 0,471, em comparação com aqueles que fazem menos do que três refeições por dia, ou seja a probabilidade de estar com o peso elevado reduz em 0,529 vezes. Enquanto comer sempre

entre as refeições, comparado com comer às vezes ou frequentemente, aumenta a probabilidade de pertencer a uma categoria de peso superior em 1,737 vezes.

Para os indivíduos fumantes em comparação com não fumantes, a razão de chances de 2.342 sugere que a probabilidade de pertencer a uma categoria de peso superior é 2.342 vezes maior para fumantes do que para não fumantes, mantendo todas as outras variáveis constantes.

Para consumo diário de água, a RC de 1.646 indica que, comparado à categoria de referência menos de um litro por dia , as chances de estar em uma categoria superior na variável dependente são multiplicadas por 1.646 quando o preditor é de um a dois litros por dia, mantendo outros fatores constantes. Isso sugere um aumento nas chances de pertencer a uma categoria superior na variável dependente quando a categoria específica é comparada com a de referência.

As pessoas que praticam atividades físicas de uma a duas vezes por semana tem a razão de chances de 0,563, ou seja a probabilidade de um alto peso reduz em 0,437 vezes comparada a nunca fazer atividades físicas, quando mantido os outros fatores constantes.

Por fim, utilizar automóvel ou moto como transporte aumenta a probabilidade de estar em uma categoria superior de peso em 2,649 vezes, comparada a pessoas que utilizam bicicleta ou fazem caminhada.

5. Conclusão

Portanto, através do processo backward e de testes para checar as suposições, o modelo utilizado foi selecionado e suas variáveis testadas. Nos testes globais todas variáveis foram significativas para um modelo, adotando um nível de significância de 5%, mas nos testes para as categorias específicas apenas algumas foram significativas e interpretadas, umas vez que, se o efeito é não significativo ele não influencia no nível de obesidade, não sendo necessário interpretá-lo.

Para a interpretação das razões de chances e intervalos de confiança das categorias foi utilizado que se a razão de chances é menor do que 1, isso indica que a categoria de referência tem menores chances de estar em uma categoria superior comparada à categoria específica, se for maior do que uma categoria específica tem maiores chances de estar em uma categoria superior comparada à categoria de referência. No intervalo de confiança foi observado seu tamanho e se possuía o valor 1, que indica que não há uma associação estatisticamente significativa entre as categorias.

6. Bibliografia

Obesity Levels: <https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels/data>. Acesso em 29 de abril de 2024.

Ordinal Logistic Regression | R data Analysis Examples. UCLA: Statistical Consulting Group. <https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/>. Acesso em 29 de jun. de 2024.

Parry, Stephen. Ordinal Logistic Regression models and Statistical Software: What You Need to Know. Cornell Statistical Consulting Unit, 2020. <https://cscu.cornell.edu/wp-content/uploads/91_ordlogistic.pdf>. Acesso em 29 de jun. de 2024.