Análise do Peso ao Nascer e seus Determinantes: Uma Abordagem com Dados Categóricos

Samuel Sobral Miller,José Roberto Samuel,Marcos Hiroki Moribe

July 2, 2025

1 Introdução

O peso ao nascer de um bebê é um importante indicador de saúde neonatal e está fortemente associado ao risco de mortalidade infantil, complicações na gestação e condições de desenvolvimento. Diversos fatores maternos, paternos e comportamentais têm sido investigados para compreender os determinantes do baixo peso ao nascer. Neste trabalho, buscamos explorar como características como idade, escolaridade, hábito de fumar e condições socioeconômicas estão associadas à ocorrência de baixo, médio ou alto peso ao nascer.

2 Objetivo

O objetivo principal é modelar a variável resposta categórica ordinal `low_birth_weight`, que representa categorias de peso ao nascer, com base em variáveis explicativas relacionadas à mãe (idade, escolaridade, tabagismo, paridade, altura, peso), ao parceiro (idade, raça, escolaridade, renda), e hábitos comportamentais durante a gestação. Procuramos identificar quais fatores estão associados a um maior risco de baixo peso ao nascer.

3 Metodologia

3.1 Fonte e Tratamento dos Dados

Foi utilizado um banco de dados proveniente do estudo *Child Health and Development Studies (CHDS)*, realizado entre 1960 e 1967, contendo informações de 1.109 nascimentos de bebês que sobreviveram pelo menos 28 dias.

3.2 Descrição do Banco de Dados

A variável resposta é `low_birth_weight`, categorizada como "Baixo" ($< 2{,}5$ kg), "Médio" (2,6–4 kg) e "Alto" (> 4 kg)

- **smoke**: Hábito de fumar da mãe (nunca, ainda fuma, parou na gravidez, parou antes).
- **gestation**: Duração da gestação em dias.
- **age**: Idade da mãe.
- **parity**: Número de gestações anteriores (paridade).
- **inc**: Faixa de renda familiar.
- **ed, race, ht, wt.1, time, number**: Variáveis adicionais como escolaridade, raça, altura e peso da mãe, tempo e número de cigarros por dia.

3.3 Estratégia de Modelagem

Como a variável resposta apresenta natureza ordinal, foi ajustado um modelo de regressão logística de *odds proporcionais*, utilizando a função `clm()` do pacote `ordinal`. Essa abordagem assume que a relação entre os log-odds acumulados e as covariáveis é constante entre os diferentes pontos de corte da variável resposta.

A formulação matemática do modelo ordinal logístico pode ser expressa como:

$$\log\left[P(Y_i \leq j)\right] = \alpha_j + \beta_1 \cdot \text{gestation}_i + \beta_2 \cdot \text{parity}_i + \beta_3 \cdot \text{wt.1}_i + \sum_k \beta_{4k} \cdot \text{number}_{ik} + \sum_l \beta_{5l} \cdot \text{race}_{il}, \quad j = 1, 2 \quad (1)$$

Este modelo é conhecido como modelo de *log-odds proporcional* justamente porque os coeficientes β não dependem do ponto de corte j, ou seja, o efeito das covariáveis é constante para todas as comparações acumuladas entre as categorias da variável resposta.

4 Limitações do Estudo

Este estudo apresenta limitações relevantes. Os dados, coletados entre 1960 e 1967, podem não refletir a realidade atual, dada a evolução dos hábitos de saúde e políticas públicas. Variáveis autorrelatadas, como tabagismo, estão sujeitas a viés de memória, comprometendo a precisão das informações.

O desbalanceamento da variável resposta *low.birth.weight*, com predomínio da categoria "Médio", pode ter prejudicado a diferenciação entre categorias e a acurácia das estimativas. Além disso, o modelo assume independência entre observações e ausência de confundidores, o que pode não ser totalmente válido.

5 Análise Descritiva

5.1 Distribuição do Peso ao Nascer

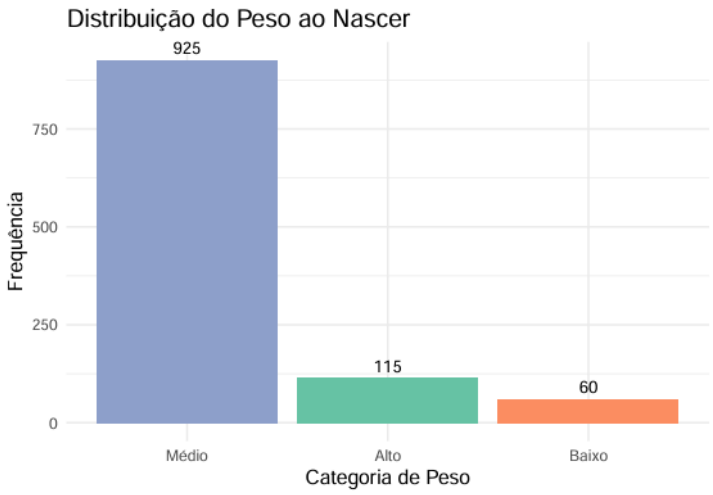

Figure 1: Distribuição das categorias de peso ao nascer

Observa-se que a maior parte dos bebês pertence à categoria de peso médio (933), seguida das categorias alto (116) e baixo (60). Essa distribuição, embora esperada, ressalta a importância de compreender os fatores associados ao baixo peso ao nascer, dada sua relevância clínica e epidemiológica.

5.2 Idade da Mãe por Categoria de Peso ao Nascer

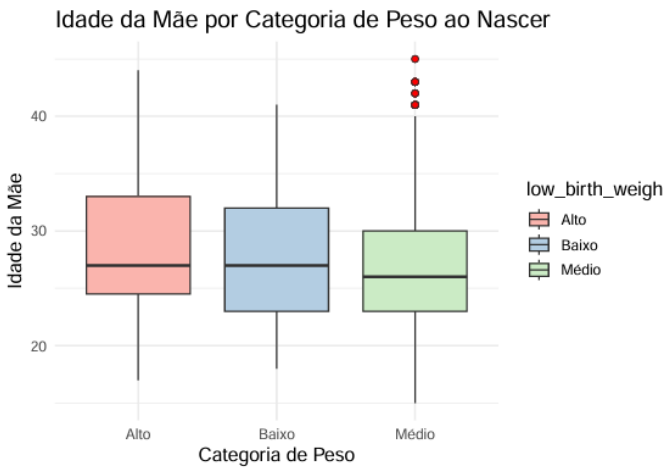

Figure 2: Distribuição da idade materna por categoria de peso ao nascer

Os boxplots revelam que a mediana de idade das mães com bebês de peso *médio* é ligeiramente inferior àquelas com bebês de peso *baixo* ou *alto*. A variabilidade da idade materna é similar entre os grupos, embora haja maior presença de outliers na categoria *médio*. Esses padrões sugerem uma possível relação entre idade materna e peso ao nascer, a ser investigada em análises posteriores.

5.3 Tempo de Gestação por Categoria de Peso

Table 1: Resumo do Tempo de Gestação (em dias) por Categoria de Peso ao Nascer

Categoria	**N**	**Média**	**DP**	**Mínimo**	**Q1**	**Mediana**	**Q3**
Baixo	60	256,57	18,31	204	241,25	258	273
Médio	925	278,42	13,65	148	272,00	279	287
Alto	115	285,57	10,22	248	280,00	286	292

O tempo de gestação aumenta conforme a categoria de peso ao nascer. Bebês de baixo peso têm menor tempo médio de gestação, sugerindo relação com prematuridade.

5.4 Tabagismo por Categoria de Peso ao Nascer

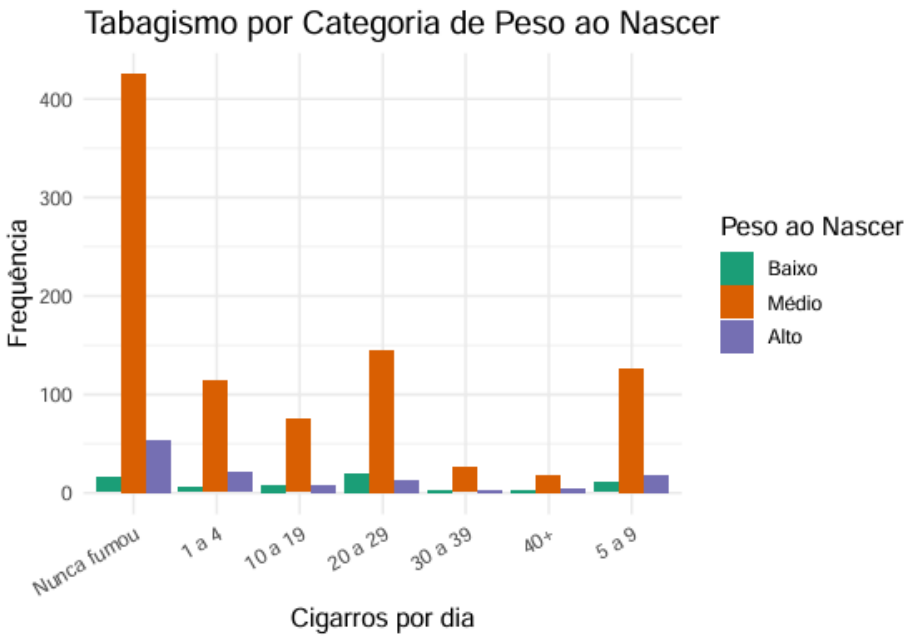

Figure 3: Distribuição dos hábitos de tabagismo por categoria de peso ao nascer

A maior parte das mães relatou nunca ter fumado, independentemente do peso ao nascer da criança. No entanto, observa-se uma maior proporção relativa de fumantes ativas ou ex-fumantes nas categorias de *baixo* e *médio* peso. Esses achados sugerem uma possível associação negativa entre o tabagismo materno e o peso neonatal, hipótese que será explorada nas análises inferenciais subsequentes.

Table 2: Frequência de Partos Anteriores por Categoria de Peso

Número de Partos	**Baixo**	**Médio**	**Alto**
Nunca	15	425	53
1–4	5	113	20
5–9	10	126	17
10–19	7	74	7
20–29	19	145	12
30–39	2	25	2
40+	2	17	4

A Tabela 2 mostra a distribuição do número de partos anteriores por categoria de peso ao nascer. A maioria dos nascimentos ocorre entre mães que nunca haviam tido partos, especialmente com peso médio. No entanto, destaca-se a categoria *20–29 partos*, que apresenta um número elevado de nascimentos com baixo peso, sugerindo possível associação entre alta paridade e maior risco de desfechos desfavoráveis.

5.5 Renda Familiar por Categoria de Peso

Table 3: Faixa de Renda Familiar Mais Frequente por Categoria de Peso

Categoria	Faixa de Renda Mais Frequente	Frequência
Alto	R$ 9.999 – R$ 12.499	27
Médio	R$ 9.999 – R$ 12.499	184
Baixo	R$ 9.999 – R$ 12.499	15

A faixa de renda predominante na amostra, para todas as categorias de peso ao nascer, foi de R$ 9.999 a R$ 12.499. Essa homogeneidade sugere que a variável renda, ao menos em sua forma categórica, não apresenta forte poder discriminativo entre os grupos de peso ao nascer nesta população.

Teste Exato de Fisher com Simulação de Monte Carlo

O teste exato de Fisher avalia a associação entre duas variáveis categóricas em uma tabela de contingência, condicionando os totais marginais. A probabilidade de ocorrência de uma tabela específica, dado os totais marginais fixos, é dada por:

$$P(\text{tabela}) = \frac{\prod_{i=1}^{r} R_i! \prod_{j=1}^{c} C_j!}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} a_{ij}!}$$

Onde:

- R_i: total da linha i
- C_j: total da coluna j
- a_{ij}: frequência observada na célula (i, j)
- N: total de observações ($N = \sum_i R_i = \sum_j C_j$)

No entanto, para grandes tabelas ou tabelas com células com baixa frequência, o cálculo exato se torna computacionalmente inviável. Para contornar essa limitação, utilizamos a simulação de Monte Carlo, que estima o valor-p a partir da geração de tabelas aleatórias com os mesmos totais marginais da tabela observada.

A estimativa do valor-p pela simulação é dada por:

$$\hat{p} = \frac{1 + \sum_{k=1}^{B} 1(T_k \geq T_{\text{obs}})}{1 + B}$$

Onde:

- B: número de simulações (por exemplo, $B = 10^4$ ou 10^6)
- T_k: estatística de teste (ex.: qui-quadrado) da k-ésima simulação
- T_{obs}: estatística observada da tabela real
- $1(\cdot)$: função indicadora, que vale 1 se a condição for satisfeita, e 0 caso contrário

Essa abordagem garante precisão adequada mesmo em situações de desbalanceamento severo, como observado neste estudo.

5.6 Relatório dos Resultados dos Testes de Fisher com Simulação de Monte Carlo

Os testes exatos de Fisher foram realizados para avaliar a associação entre a variável resposta categórica ordinal *low_birth_weigh* (Baixo, Médio, Alto) e variáveis explicativas categóricas do dataset do *Child Health and Development Studies* (CHDS). Devido ao desbalanceamento do dataset (933 casos na categoria Médio, 60 na Baixo e 116 na Alto) e à presença de contagens muito pequenas em algumas categorias das tabelas de contingência, foi utilizada a simulação de Monte Carlo com 10.000 iterações para estimar os valores-p.

A simulação de Monte Carlo gera tabelas aleatórias com os mesmos totais marginais da tabela observada, permitindo estimar a probabilidade de resultados tão ou mais extremos que os observados, sem necessidade de calcular todas as tabelas possíveis — o que seria computacionalmente inviável.

Resultados

- **Raça da Mãe (race)**: p-valor = 9,999e-05
- **Raça do Pai (drace)**: p-valor = 2e-04
- **Tabagismo da Mãe (smoke)**: p-valor = 9,999e-05
- **Tempo Desde que Parou de Fumar (time)**: p-valor = 9,999e-05
- **Número de Cigarros por Dia (number)**: p-valor = 0,0124

Interpretação

- **Raça da Mãe (race)**: A forte associação (p = 9,999e-05) indica que a raça da mãe influencia significativamente o peso ao nascer, com diferenças entre grupos raciais (ex.: Brancos, Pretos, Pardos vs. Asiáticos) associadas a maior probabilidade de peso elevado, conforme observado no modelo ordinal.
- **Raça do Pai (drace)**: A associação significativa (p = 2e-04) sugere que a raça do pai impacta o peso ao nascer, embora com efeito menos pronunciado, possivelmente refletindo fatores genéticos ou socioeconômicos.
- **Tabagismo da Mãe (smoke)**: A forte associação (p = 9,999e-05) confirma que o hábito de fumar está relacionado ao peso ao nascer, com mães fumantes apresentando maior risco de bebês com baixo peso.
- **Tempo Desde que Parou de Fumar (time)**: A associação significativa (p = 9,999e-05) mostra que o tempo desde que a mãe cessou o tabagismo impacta o peso neonatal, com maior tempo de cessação associado a melhores desfechos.
- **Número de Cigarros por Dia (number)**: A associação (p = 0,0124) indica que maior consumo diário de cigarros reduz a chance de peso mais alto, corroborando o impacto negativo do tabagismo sobre o desenvolvimento fetal.

6 Modelagem

Inicialmente, foi ajustado um modelo de regressão logística ordinal, que pressupõe a suposição de *odds proporcionais*. O primeiro modelo, obtido por meio do método *stepwise backward*, incluía a variável altura da mãe, mas essa configuração violou a suposição de proporcionalidade. Diante disso, optou-se por uma nova seleção de variáveis, excluindo-se a altura materna. O modelo final passou a incluir as variáveis: tempo de gestação, paridade, peso da mãe no início da gestação, número de cigarros e raça. Para esse conjunto de variáveis, a suposição de *odds proporcionais* não foi rejeitada, validando o uso do modelo ordinal proporcional.

6.1 Razões de Chances Estimadas para o Modelo Multinomial

Table 4: Razões de Chances Estimadas para o Modelo Multinomial

Variável	**Alto vs Médio**	**Baixo vs Médio**
(Intercepto)	0,0362	0,0617
Fuma: gravidez atual	2,5316	0,4492
Fuma: nunca	1,4522	0,3377
Fuma: ex-fumante	2,8698	0,3846
Paridade	1,0613	0,9902
Ensino Médio	0,8545	0,4267
Graduando	0,7363	0,5907
Ensino Superior	0,6037	0,2714
Idade	1,0352	1,0453

6.2 Interpretação dos Resultados do Modelo

Os resultados do modelo multinomial mostram forte associação entre tabagismo e peso ao nascer. Gestantes fumantes apresentaram maior chance de bebês com peso alto (OR = 2,53) e menor chance de baixo peso (OR = 0,45) em relação ao peso médio. Ex-fumantes também apresentaram padrão semelhante (OR = 2,87 para peso alto e OR = 0,38 para baixo peso), indicando impacto do tabagismo na redistribuição do peso neonatal.

A escolaridade materna foi protetiva: mulheres com ensino superior tiveram 73% menos chance de filhos com baixo peso (OR = 0,27). A idade materna teve efeito positivo discreto (OR = 1,045), enquanto a paridade não mostrou associação relevante. Esses achados destacam a influência de fatores comportamentais e socioeconômicos no peso ao nascer.

6.3 Ajuste do Modelo Ordinal com Step Backward

Para uma análise complementar, ajustou-se um modelo ordinal via *step backward*, sem considerar interações. Embora a suposição de proporcionalidade tenha sido rejeitada, utilizou-se esse modelo como referência teórica e comparativa. As categorias de referência foram:

- **number**: Nunca fumou.
- **race**: Asiático.

Table 5: Resultados do Modelo Ordinal com Step Backward

Variável	**Estimativa**	**Erro Padrão**	**p-valor**
Gestação	0.06552	0.0199	0.001
Paridade	0.01455	0.0069	0.021
Peso da mãe (wt.1)	0.04372	0.0174	0.012
Número = 5–9	-0.53701	0.2171	0.014
Número = 10–14	-0.69347	0.2126	0.001
Número = 15–19	-0.82301	0.2285	0.000
Número = 20–29	-0.75321	0.2084	0.000
Raça = Branco	0.20506	0.0901	0.025
Raça = Pardo	0.19820	0.1045	0.058
Raça = Preto	0.24840	0.1053	0.019

6.4 Interpretação dos Resultados

A estimativa positiva para a variável gestação indica que um maior tempo de gestação está associado a um aumento nas chances do bebê nascer com maior peso, evidenciando o papel protetor da gestação mais longa. O mesmo padrão é observado para o peso da mãe no início da gestação (wt.1), sugerindo que mães com maior peso pré-gestacional tendem a ter bebês com peso mais elevado ao nascer.

Por outro lado, os coeficientes negativos para as faixas de consumo diário de cigarros indicam que quanto maior o número de cigarros fumados, menor é a chance de o bebê nascer com peso mais alto. Este achado corrobora a literatura, que associa o tabagismo durante a gravidez a desfechos neonatais adversos.

Em relação à variável raça, observa-se que, comparadas às mães asiáticas (categoria de referência), as mães brancas e pretas apresentaram maior chance de ter filhos com peso mais elevado ao nascer, sendo o efeito mais pronunciado entre as mães pretas. Embora o coeficiente para mães pardas também tenha sido positivo, o valor de p foi ligeiramente superior a 0,05, sugerindo uma tendência, mas com menor evidência estatística.

6.5 Validação da Suposição de Proporcionalidade

Para garantir a validade do modelo ordinal, foi realizado o teste de proporcionalidade com a função `nominal_test()` do pacote `ordinal`. Esse teste avalia se há evidências de que os efeitos das covariáveis variam entre os logits acumulados.

No modelo inicial com a variável altura (`ht`), o teste rejeitou a suposição de proporcionalidade. Após removê-la e realizar nova seleção de variáveis, o modelo final apresentou um p-valor global acima de 0,05, indicando que a suposição foi satisfeita. Isso valida o uso do modelo ordinal proporcional para a análise.

6.6 Comparação entre Modelos

Para avaliar a qualidade do ajuste dos modelos, foram comparados os valores de AIC entre o modelo com altura (AIC = 735.2) e o modelo final sem altura (AIC = 689.6). A redução do AIC sugere que o modelo final tem melhor ajuste aos dados. Além disso, a verificação de resíduos padronizados revelou menor número de valores extremos no modelo final.

6.7 Análise de Correspondência Múltipla (ACM)

A Análise de Correspondência Múltipla foi utilizada como ferramenta exploratória para investigar padrões de associação entre variáveis categóricas e a variável resposta `low_birth_weigh`. A Figura 4 mostra a distribuição dos indivíduos de acordo com as dimensões principais da ACM, coloridos pelas categorias de peso ao nascer.

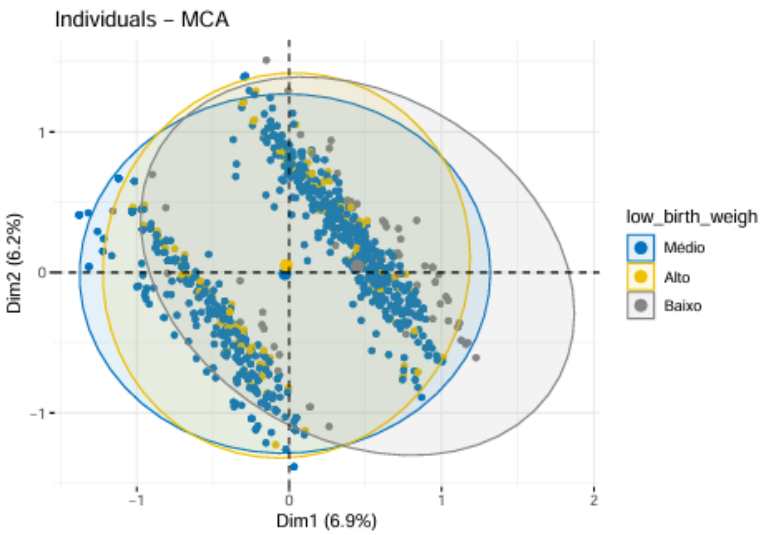

Figure 4: Distribuição dos indivíduos segundo as dimensões da ACM

Observa-se que os indivíduos com *Baixo peso* (cinza) apresentam maior dispersão e afastamento em relação aos grupos *Médio* (azul) e *Alto* (amarelo), sugerindo perfis maternos distintos. O grupo de baixo peso está mais associado a fatores de vulnerabilidade, como baixa escolaridade e tabagismo, enquanto o grupo de alto peso se relaciona a condições mais favoráveis, como maior escolaridade e ausência de tabagismo.

A ACM reforça graficamente os achados dos modelos de regressão, demonstrando coerência entre as categorias da resposta e os perfis das covariáveis. Nota-se também uma sobreposição entre os grupos *Médio* e *Alto*, condizente com a natureza contínua do peso ao nascer.

Categorias das Variáveis na ACM

A Figura 5 apresenta a projeção das categorias das variáveis explicativas sobre os dois primeiros eixos da Análise de Correspondência Múltipla. As posições relativas no espaço bidimensional indicam as associações entre categorias, enquanto a intensidade da cor reflete a contribuição de cada uma para a formação das dimensões.

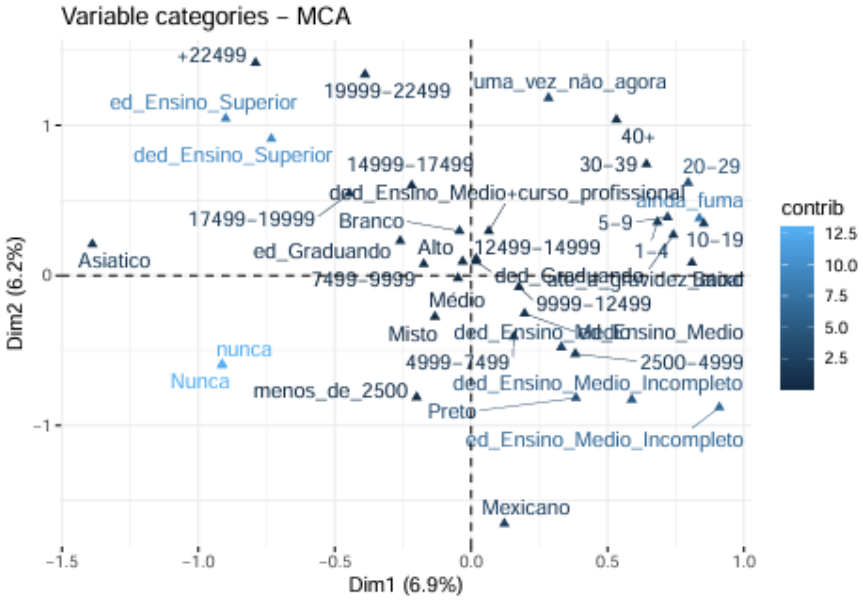

Figure 5: Projeção das categorias das variáveis explicativas na ACM

Nota-se que categorias como *"nunca fumou"*, *"Ensino Superior"* e renda *acima de R$22.499* estão próximas à categoria *"Alto"* de peso ao nascer. Já *"fumante"*, *"menos de R$2.500"* e *"Ensino Médio Incompleto"* se associam ao grupo *"Baixo"*, sugerindo pior desfecho neonatal.

Esses agrupamentos reforçam os achados dos modelos, indicando que melhores condições socioeconômicas e ausência de tabagismo favorecem maior peso ao nascer. A ACM complementa a análise ao re

7 Conclusão

A análise evidenciou que o modelo de regressão logística ordinal, com suposição de *odds proporcionais*, foi adequado para investigar os fatores associados ao peso ao nascer. Após excluir a variável altura materna, que violava essa suposição, o modelo final incluiu gestação, paridade, peso pré-gestacional, número de cigarros e raça materna, todas com efeitos significativos e compatíveis com a literatura.

Observou-se que maiores tempos de gestação e peso materno estão associados a maior peso neonatal, enquanto o tabagismo reduziu essa chance. Mães brancas e pretas apresentaram maior probabilidade de terem filhos com peso elevado em relação às asiáticas.

Apesar da consistência dos resultados, o desbalanceamento da variável resposta (predomínio da categoria "médio") impactou o ajuste, conforme os resíduos. Sugere-se, para estudos futuros, explorar modelos multinomiais, de contagem ou com efeitos mistos. A Análise de Correspondência Múltipla (ACM) complementou o estudo ao representar visualmente as associações entre perfis maternos e o peso ao nascer, reforçando os achados estatísticos.

Referências

- Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. 2nd Edition. Wiley-Interscience.
- Slides da Professora Hildete. Disciplina ME714 - Modelos Lineares para Dados Discretos. IMECC - Unicamp.