UNIVERSIDADE FEDERAL DO CEARÁ
PROGRAMAS DE PÓS-GRADUAÇÃO EM ECONOMIA - CAEN
MESTRADO ACADÊMICO EM ECONOMIA

LUIZ EGUIBERTO LOPES RODRIGUES FILHO

ANÁLISE DOS DETERMINANTES SOCIOECONÔMICOS DO DIABETES AUTORREFERIDO NO BRASIL: UMA ABORDAGEM ECONOMÉTRICA

Fortaleza/CE
2026

LUIZ EGUIBERTO LOPES RODRIGUES FILHO

ANÁLISE DOS DETERMINANTES SOCIOECONÔMICOS DO DIABETES AUTORREFERIDO NO BRASIL: UMA ABORDAGEM ECONOMÉTRICA

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia do Programa de Pós-Graduação em Economia - CAEN - da Universidade Federal do Ceará, como requisito parcial para a obtenção do título de Mestre em Economia.

Orientador: Prof. Dr. João Mário Santos de França

Fortaleza/CE

2026

Página reservada para ficha catalográfica.

Utilize a ferramenta *online* Catalog! para elaborar a ficha catalográfica de seu trabalho acadêmico, gerando-a em arquivo PDF, disponível para download e/ou impressão. (http://www.fichacatalografica.ufc.br/)

LUIZ EGUIBERTO LOPES RODRIGUES FILHO

ANÁLISE DOS DETERMINANTES SOCIOECONÔMICOS DO DIABETES AUTORREFERIDO NO BRASIL: UMA ABORDAGEM ECONOMÉTRICA

Dissertação apresentado ao Curso de Mestrado Acadêmico em Economia do Programa de Pós-Graduação em Economia - CAEN - da Universidade Federal do Ceará, como requisito parcial para a obtenção do título de Mestre em Economia.

Aprovado em 24/02/2026

BANCA EXAMINADORA

Prof. Dr. João Mário Santos de França (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dra. Guaracyane Lima Campelo
Universidade Federal do Ceará (UFC)

Dra. Natália Cecília de França
Controladoria e Ouvidoria Geral do Estado do Ceará (CGE)

À minha companheira, Júlia.

AGRADECIMENTOS

A Yahweh, o Eterno, que mesmo sendo universal cuida de mim com uma individualidade ímpar.

À minha companheira, Júlia, que em tudo é minha cúmplice e que acredita em mim mais do que eu mesmo.

À minha mãe, Jacinta, que desde a primeira vez que me colocou nos braços tem sido uma força da natureza, veementemente imparável, e que tem agido sempre para o meu bem.

Ao meu pai, Bebeto, que por meio de suas histórias me permite extrair diversos ensinamentos sobre a vida e que me tem sido uma torre segura.

Aos meus irmãos, Israel e Gabriel, que são as pessoas que mais partilham experiências comigo e que são uma constante na minha vida.

Aos meus tios, Gontran e Filomena que me possibilitaram ir além, por me hospedarem em sua casa.

Aos meus cunhados, Pedro e Kerolainy que são aquisições excepcionais para minha família e que tornaram minha vida mais leve.

Ao professor João Mário, que me guiou neste trabalho e que me deu a honra de ser seu orientando.

Aos amigos Gustavo e Júnior, que tornaram minha jornada acadêmica mais divertida e que foram verdadeiros professores no que diz respeito à obsessão pela excelência.

E, por fim, mas não menos importante, ao CAEN, que exigiu de mim mais do que eu acreditava poder me doar e que me mostrou que o esforço traz resultados recompensadores, além de me dar a oportunidade de aprender com professores excepcionais, trazendo à luz saberes antes invisíveis.

A todos vocês, sou grato!

“O conhecimento não é apenas
para ser compreendido, mas para
ser usado em favor da melhoria
das condições de vida.”
Amartya Sen

RESUMO

Este estudo analisa os determinantes socioeconômicos associados ao diagnóstico de diabetes mellitus autorreferido no Brasil, utilizando os microdados da Pesquisa Nacional de Saúde (PNS) de 2019. Parte-se da hipótese de que o diabetes não se distribui aleatoriamente na população, mas segue gradientes sistemáticos ao longo da idade e da posição socioeconômica. A estratégia empírica combina a estimação de um modelo logit binário, o cálculo do Índice de Concentração corrigido de Erreygers e técnicas de decomposição, incluindo a decomposição não linear de Oaxaca–Blinder. Os resultados indicam forte concentração do diabetes entre indivíduos mais velhos e um gradiente socioeconômico mais moderado quando ordenado pela renda. A idade, a hipertensão e a obesidade estão fortemente associadas à maior probabilidade de diagnóstico, enquanto níveis mais elevados de escolaridade exercem efeito protetivo relevante, mesmo após o controle por renda e hábitos de vida. As análises de decomposição indicam que as diferenças educacionais na prevalência do diabetes não podem ser plenamente atribuídas às características observáveis incluídas no modelo, embora a decomposição não permita identificar com precisão estatística a contribuição relativa dos componentes. De forma semelhante, no caso racial, embora existam diferenças na prevalência do diabetes entre grupos, a decomposição sugere que essas disparidades estão associadas, em parte, à distribuição desigual de características socioeconômicas e de acesso aos serviços de saúde, devendo tais resultados ser interpretados com cautela. Em conjunto, os resultados reforçam a interpretação do diabetes como um desfecho moldado por processos sociais acumulados ao longo do ciclo de vida. Conclui-se que o enfrentamento do diabetes no Brasil demanda políticas intersetoriais que combinem o fortalecimento do acesso à saúde com o combate às desigualdades estruturais de educação e renda.

Palavras-chave: Diabetes, Determinantes Socioeconômicos da Saúde, Modelo Logit, Índice de Concentração, Decomposição Blinder-Oaxaca.

ABSTRACT

This study analyzes the socioeconomic determinants associated with self-reported diabetes mellitus in Brazil, using microdata from the 2019 National Health Survey (Pesquisa Nacional de Saúde – PNS). It is based on the hypothesis that diabetes is not randomly distributed in the population, but follows systematic gradients across age and socioeconomic position. The empirical strategy combines the estimation of a binary logit model, the calculation of the Erreygers-corrected Concentration Index, and decomposition techniques, including the nonlinear Oaxaca–Blinder decomposition. The results indicate a strong concentration of diabetes among older individuals and a more moderate socioeconomic gradient when ranked by income. Age, hypertension, and obesity are strongly associated with a higher probability of diagnosis, while higher levels of education exhibit a relevant protective effect, even after controlling for income and lifestyle factors. The decomposition analyses suggest that educational differences in diabetes prevalence cannot be fully attributed to observable characteristics included in the model, although the decomposition does not allow precise statistical identification of the relative contribution of explained and unexplained components. Similarly, in the case of racial differences, while disparities in diabetes prevalence are observed across groups, the decomposition suggests that these differences are partly associated with the unequal distribution of socioeconomic characteristics and access to health services, and should therefore be interpreted with caution. Overall, the findings reinforce the interpretation of diabetes as an outcome shaped by social processes accumulated over the life course. The study concludes that effectively addressing diabetes in Brazil requires intersectoral public policies that combine improved access to healthcare with broader efforts to reduce structural inequalities in education and income.

Keywords: Diabetes; Socioeconomic inequalities; Health economics, Logit Model, Concentration Index, Blinder–Oaxaca decomposition.

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

PNS	Pesquisa Nacional de Saúde
IBGE	Instituto Brasileiro de Geografia e Estatística
MICE	Imputação Múltipla por Equações Encadeadas
MAR	*Missing at Random*
EPV	Eventos por Parâmeros
VIF	*Variance Inflation Factor*
ROC	Característica de Operação do Receptor
AUC	Área sob a Curva ROC
LRT	Teste da Razão de Verossimilhança
CI	Índice de Concentração
EI	Índice de Concentração de Erreygers
IC	Intervalo de Confiança
DP	Desvio Padrão
IMC	Índice de Massa Corporal
AOR	Razões de Chances Ajustadas
AME	Efeitos Marginais Médios
SE	Erro-padrão
DCNTs	Doenças Crônicas Não Transmissíveis
OMS	World Health Organization (Organização Mundial da Saúde)

SUMÁRIO

1. Introdução

As doenças crônicas não transmissíveis (DCNTs) constituem atualmente o principal desafio de saúde pública em escala global. Responsáveis por cerca de 70% das mortes no mundo, essas doenças, que incluem enfermidades cardiovasculares, câncer, doenças respiratórias crônicas e diabetes, afetam populações de todas as regiões e faixas etárias, impondo impactos expressivos sobre os sistemas de saúde, a produtividade econômica e o bem-estar social (World Health Organization – OMS, 2022; Bennett et al., 2018). Isso porque diferentemente das doenças transmissíveis, as DCNTs caracterizam-se por evolução prolongada, demandando acompanhamento contínuo e gerando custos acumulados ao longo do ciclo de vida (OMS, 2022).

Apesar de sua presença global, a carga das DCNTs não se distribui de forma homogênea. Segundo Bennett et al. (2018), em *NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4*, as evidências indicam que países de baixa e média renda concentram a maior parte das mortes prematuras associadas a essas doenças, bem como apresentam maiores dificuldades em atingir as metas internacionais de redução da mortalidade evitável. Ademais, o relatório *Invisible numbers The true extent of noncommunicable diseases and what to do about them* de atribuição da OMS (2022) salienta que mesmo dentro de um mesmo país, observa-se que indivíduos inseridos em contextos socioeconômicos mais desfavoráveis tendem a apresentar maior exposição a fatores de risco e piores condições de saúde, reforçando a existência de um gradiente socioeconômico na distribuição das DCNTs.

Complementarmente, o relatório da OMS (2022) enfatiza que, entre as DCNTs, o Diabetes Mellitus ocupa posição de destaque. Trata-se de uma doença de elevada prevalência, progressiva e frequentemente silenciosa em seus estágios iniciais, mas associada a complicações severas no longo prazo, como doenças cardiovasculares, insuficiência renal e comprometimentos funcionais. Adicionalmente, estimativas recentes, como a do relatório *Diabetes* redigido pela OMS, indicam que a grande maioria dos casos – aproximadamente 95% – de diabetes em adultos está associada ao tipo 2, fortemente relacionado a fatores comportamentais e condições de vida (OMS, 2024).

Nesse sentido, o debate contemporâneo em saúde pública tem avançado na compreensão de que comportamentos frequentemente associados ao diabetes, como sedentarismo, dietas inadequadas e excesso de peso, são moldados por fatores estruturais. Ou seja, o ambiente social, econômico e físico no qual os indivíduos estão inseridos pode facilitar ou restringir escolhas saudáveis, tornando determinadas trajetórias de saúde mais prováveis do que outras (OMS, 2022). Assim, renda, escolaridade, condições de moradia, hábitos e acesso a serviços públicos tornam-se elementos centrais para compreender a distribuição do diabetes mellitus na população.

Somado a isso, além de seus efeitos sobre a saúde individual, o diabetes impõe custos econômicos relevantes. Por seu caráter crônico, a doença está associada a gastos contínuos com tratamento, monitoramento e manejo de complicações, bem como a perdas de produtividade decorrentes de incapacidades e mortalidade prematura. As estimativas indicam que os custos indiretos associados às DCNTs, relacionados à perda de produtividade, superam amplamente os custos diretos com serviços de saúde, afetando de forma mais intensa países e grupos populacionais marcados por maior vulnerabilidade socioeconômica (OMS, 2022).

Gráfico 01 – DCNTs Prejudicam Economia

Fonte: OMS

No Brasil, esse quadro assume particular relevância. O país combina uma elevada carga de DCNTs com profundas desigualdades socioeconômicas e regionais, criando um cenário no qual o risco de desenvolver diabetes e suas consequências não se distribui de maneira uniforme. Assim, compreender como essa doença se organiza ao longo dos estratos socioeconômicos e quais fatores contribuem para sua concentração em determinados grupos é fundamental para o debate sobre equidade em saúde e para o desenho de políticas públicas mais eficazes (Bennett et al., 2018; OMS, 2022).

Assim, o diferencial deste estudo reside na aplicação conjunta do Índice de Concentração de Erreygers e da decomposição de Oaxaca–Blinder aos dados de saúde no Brasil, enquanto a literatura nacional tem se concentrado predominantemente em modelos de regressão. O trabalho integra modelos econométricos às técnicas de decomposição, permitindo não apenas identificar fatores associados à doença, mas também quantificar a contribuição relativa desses fatores para as disparidades observadas entre grupos populacionais.

É nesse contexto que se insere o presente trabalho. Esta dissertação tem como objetivo realizar uma análise dos determinantes socioeconômicos do diabetes autorreferido no Brasil, utilizando dados da Pesquisa Nacional de Saúde de 2019. A partir da estimação de modelos econométricos e do uso de medidas de desigualdade em saúde, busca-se investigar como a doença se distribui ao longo da renda e da idade, e também decompor essa distribuição de modo a identificar a contribuição relativa de diferentes características observáveis. Deste modo, ao adotar essa abordagem, o estudo não pretende estabelecer relações causais estritas, mas oferecer uma leitura analítica sobre os mecanismos socioeconômicos associados ao diabetes autorreferido no contexto brasileiro.

Além desta introdução, a dissertação está organizada em quatro capítulos. O segundo apresenta o referencial teórico, abordando os determinantes socioeconômicos da saúde e os principais instrumentos de mensuração e decomposição das desigualdades. O terceiro descreve a base de dados e a estratégia metodológica adotada. O quarto expõe e discute os resultados empíricos, incluindo o modelo logit, o Índice de Concentração de Erreygers e a decomposição de Oaxaca–Blinder não linear. Por fim, apresentam-se as conclusões do estudo.

2. Referencial Teórico

2.1 – Determinantes Socioeconômicos da Saúde

A literatura em economia da saúde e saúde pública tem demonstrado de forma consistente que os desfechos de saúde não se distribuem aleatoriamente na população, mas seguem um gradiente socioeconômico sistemático, no qual indivíduos em posições sociais mais desfavoráveis apresentam, em média, piores condições de saúde. Essa regularidade empírica, amplamente documentada em diferentes contextos institucionais e epidemiológicos, constitui o ponto de partida para análises que buscam compreender desigualdades em saúde a partir de fatores socioeconômicos.

Um marco fundamental nessa literatura é o relatório *Fair Society, Healthy Lives*, conduzido no Reino Unido e coordenado por Marmot (2012), que consolida evidências de que as desigualdades em saúde refletem diferenças persistentes nas condições materiais, educacionais e sociais ao longo do ciclo de vida. O relatório, que analisa tendências observadas principalmente entre os anos 1990 e o final dos anos 2000, argumenta que o gradiente social da saúde não se limita à comparação entre extremos de renda ou escolaridade, mas se manifesta de forma contínua ao longo de toda a distribuição socioeconômica. Assim, cada nível adicional de desvantagem social está associado, em média, a piores resultados em termos de morbidade, mortalidade e bem-estar.

Essa perspectiva desloca o foco analítico de explicações estritamente biomédicas para uma abordagem mais ampla, na qual fatores como escolaridade, renda, ocupação e ambiente social atuam como determinantes fundamentais da saúde. A escolaridade, por exemplo, influencia o acesso à informação, a capacidade de compreender recomendações médicas e a adoção de comportamentos preventivos. A renda, por sua vez, afeta o acesso a bens e serviços de saúde, a qualidade da alimentação, as condições de moradia e a exposição a ambientes mais ou menos saudáveis. Esses mecanismos operam de forma cumulativa, reforçando desigualdades ao longo do tempo.

Além de documentar a existência de um gradiente socioeconômico contínuo, o relatório *Fair Society, Healthy Lives* enfatiza que desigualdades em saúde tendem a persistir mesmo em contextos de expansão do acesso aos serviços de saúde.

Segundo Marmot (2012), políticas focadas exclusivamente na ampliação da cobertura assistencial são insuficientes para eliminar disparidades, uma vez que estas têm origem em determinantes sociais mais amplos, como condições de trabalho, ambiente social, nível educacional e exposição acumulada a situações de estresse ao longo do ciclo de vida. Assim, ainda que o acesso formal aos serviços seja ampliado, diferenças associadas à posição socioeconômica continuam a se refletir nos resultados em saúde.

Essa interpretação é reforçada por evidências empíricas sintetizadas na literatura sobre desigualdades socioeconômicas em saúde, conforme discutido por Safieddine et al. (2023). Esse estudo, conduzido na Alemanha, explora três janelas entre 2005–2007, 2010–2012 e 2015–2017, e mostra, com base na estimação de probabilidades preditas de grupos de comorbidades e do número de comorbidades em indivíduos com diabetes — por meio de regressões logísticas e regressão ordinal, com termos de interação para verificar mudanças temporais nas disparidades por status socioeconômico e estratificação das análises por subgrupos — que diversos indicadores de saúde, incluindo medidas autorreferidas e desfechos clinicamente observáveis, apresentam padrões sistemáticos de concentração ao longo da hierarquia socioeconômica. Em particular, os autores destacam que a magnitude e o sinal dessas desigualdades variam conforme o contexto institucional e o estágio da transição epidemiológica, mas a presença de um gradiente social permanece como característica recorrente.

A partir dessas contribuições, tanto conceituais quanto empíricas, a saúde pode ser compreendida como um resultado fortemente condicionado por fatores socioeconômicos, e não apenas por características individuais ou biológicas. Esse entendimento fornece a base analítica para abordagens que vão além da comparação de médias populacionais e buscam examinar como os desfechos de saúde se distribuem ao longo do gradiente socioeconômico. De modo que essa perspectiva é particularmente relevante para o estudo de doenças crônicas não transmissíveis, como o diabetes mellitus, cuja ocorrência pode resultar da interação entre fatores biológicos, comportamentais e sociais acumulados ao longo do ciclo de vida.

2.2 – Diabetes Mellitus sob a Perspectiva Socioeconômica

O diabetes mellitus tem sido amplamente analisado na literatura como uma doença crônica de natureza multifatorial, cuja ocorrência pode resultar da interação entre fatores biológicos, comportamentais e socioeconômicos. Diferentemente de explicações estritamente clínicas, abordagens socioeconômicas enfatizam que a probabilidade de desenvolver a doença pode estar associada à posição dos indivíduos na estrutura social, refletindo diferenças no acesso a recursos, informação e condições de vida ao longo do ciclo de vida.

No contexto brasileiro, Flor e Campos (2017) analisam os determinantes socioeconômicos do diabetes autorreferido utilizando dados da Pesquisa Nacional de Saúde (PNS) de 2013 e modelo logístico multivariado. Os autores encontram associação robusta entre diabetes e idade, bem como com fatores comportamentais e clínicos, como obesidade e hipertensão. Em relação aos determinantes socioeconômicos, a escolaridade emerge como variável central, com indivíduos de menor nível educacional apresentando maior probabilidade de relatar diagnóstico de diabetes, mesmo após o controle por outras características individuais. Por outro lado, a renda não se mostra estatisticamente significativa em todas as especificações, o que os autores discutem à luz de possíveis vieses de diagnóstico e de sobrevivência, além de desigualdades no acesso aos serviços de saúde. Essa discussão é particularmente relevante ao indicar que a ausência de significância estatística para renda não implica inexistência de desigualdades, mas pode refletir limitações associadas à mensuração do diagnóstico autorreferido.

Ademias, evidências em países de renda média reforçam essa interpretação. Kundu et al. (2022) investigam a prevalência de diabetes, hipertensão e comorbidades em adultos de Bangladesh entre 2017–2018, utilizando modelos logísticos e medidas de desigualdade socioeconômica, como Curvas de Concentração e Índices de Concentração. De modo que os resultados indicam que o diabetes apresenta um padrão de associação significativo com variáveis socioeconômicas e comportamentais, como escolaridade, condição econômica e estilo de vida. Os autores mostram ainda que o gradiente socioeconômico da doença persiste mesmo após o controle por fatores individuais, sugerindo que comportamentos de risco não esgotam os mecanismos que conectam posição socioeconômica e diabetes. Essa evidência reforça a adequação do uso de modelos probabilísticos para captar a

relação entre características observáveis e a ocorrência do diabetes em contextos socioeconômicos heterogêneos.

Evidências mais recentes para o contexto brasileiro reforçam esse padrão. Kluthcovsky e Beraldo (2024), ao analisarem a evolução da prevalência de diabetes autorreferido nas capitais brasileiras e no Distrito Federal entre 2010 e 2021, com base em dados do Vigitel, mostram que a ocorrência da doença apresenta associação sistemática com características socioeconômicas, especialmente escolaridade e condições de vida, com maior prevalência entre indivíduos em posições sociais mais desfavoráveis. De forma complementar, Garces et al. (2023), por meio de um estudo ecológico de abrangência nacional que analisa os 5.570 municípios brasileiros no período de 2010 a 2020, identificam, a partir de uma abordagem espacial e temporal, que regiões com piores condições socioeconômicas apresentam maiores taxas de mortalidade associadas ao diabetes. Em conjunto, esses resultados reforçam a presença de um gradiente socioeconômico persistente no país, alinhado às evidências internacionais.

Uma perspectiva complementar é fornecida por Hosseinpoor et al. (2012), que analisam a prevalência de doenças crônicas não transmissíveis — incluindo o diabetes — em países de baixa e média renda, com base em dados do World Health Survey (2002–2004). O estudo utiliza Curvas de Concentração e Índices de Concentração para avaliar a distribuição dessas doenças ao longo do gradiente socioeconômico, considerando variáveis como índice de riqueza domiciliar, escolaridade, idade e sexo. Os autores documentam que o sinal e a magnitude das desigualdades associadas ao diabetes variam substancialmente entre países, refletindo diferenças institucionais, estágios da transição epidemiológica e padrões de acesso ao diagnóstico. Ainda assim, o estudo mostra que o diabetes raramente se distribui de forma neutra ao longo da hierarquia socioeconômica, reforçando a relevância de análises empíricas que considerem explicitamente a posição relativa dos indivíduos.

Em conjunto, esses trabalhos sustentam a interpretação do diabetes autorreferido como um desfecho adequado para análises econométricas baseadas em modelos binários, como o logit, e para a investigação de desigualdades socioeconômicas em saúde. Em acréscimo, a literatura também indica que variáveis como idade, escolaridade, condições socioeconômicas e fatores comportamentais

desempenham papéis distintos, mas inter-relacionados, na determinação da probabilidade de diagnóstico da doença. Além disso, os resultados divergentes observados em diferentes contextos reforçam a importância de análises específicas a cada país, bem como a necessidade de cautela na interpretação dos coeficientes estimados, especialmente quando o desfecho é autorreferido.

Essa abordagem justifica a estratégia empírica adotada neste trabalho, que combina a modelagem probabilística da ocorrência do diabetes com a análise das desigualdades socioeconômicas associadas ao desfecho. Ao dialogar com evidências nacionais e internacionais, a literatura oferece suporte tanto para a escolha do modelo econométrico quanto para a seleção das covariáveis, ao mesmo tempo em que reconhece as limitações inerentes à mensuração do diagnóstico e à heterogeneidade dos contextos analisados.

2.3 – Desigualdades em Saúde e Medidas de Gradiente Socioeconômico

A análise das desigualdades socioeconômicas em saúde requer instrumentos que sejam capazes de captar não apenas diferenças médias entre grupos, mas também como os desfechos de saúde se distribuem ao longo da hierarquia socioeconômica. Nesse contexto, a literatura de economia da saúde tem enfatizado o uso de medidas baseadas em rankings de posição socioeconômica, que permitem avaliar a existência, a direção e a magnitude de gradientes sociais associados a diferentes indicadores de saúde.

Uma referência central nessa literatura é o manual *Analyzing Health Equity Using Household Survey Data* desenvolvido por O'Donnell et al. (2007), sob a égide do Banco Mundial, que sistematiza métodos amplamente utilizados para mensurar desigualdades em saúde a partir de dados de pesquisas domiciliares. No manual, os autores apresentam a Curva de Concentração e o Índice de Concentração como ferramentas fundamentais para avaliar se um determinado desfecho de saúde se encontra desproporcionalmente concentrado entre indivíduos de menor ou maior nível socioeconômico. Diferentemente de medidas baseadas em comparações binárias entre grupos extremos, essas abordagens exploram toda a distribuição da variável de ordenação, fornecendo uma representação mais completa do gradiente socioeconômico.

O Índice de Concentração, conforme definido na obra supracitada, resume numericamente a informação contida na Curva de Concentração, assumindo valores positivos quando o desfecho está concentrado entre os indivíduos mais favorecidos e valores negativos quando a concentração ocorre entre os mais desfavorecidos. Essa propriedade torna o índice particularmente útil para análises comparativas entre populações, períodos ou desfechos distintos. No entanto, os autores também ressaltam que a interpretação do índice deve considerar a natureza da variável de interesse, especialmente quando se trata de indicadores de saúde limitados ou binários, como a prevalência de doenças crônicas.

Nesse ponto, a contribuição metodológica de *Correcting the Concentration Index* (Erreygers, 2009) é fundamental. Nele o autor demonstra que o Índice de Concentração tradicional apresenta limitações importantes quando aplicado a variáveis binárias, uma vez que seus limites dependem da média do desfecho, dificultando comparações entre populações com diferentes níveis de prevalência. Para contornar esse problema, o autor propõe uma versão corrigida do índice, conhecida como Índice de Concentração de Erreygers, que satisfaz propriedades normativas desejáveis, como simetria e invariância a transformações lineares do desfecho. Essa correção torna o índice particularmente adequado para a análise de desigualdades em saúde quando o desfecho é dicotômico, como o diagnóstico de diabetes.

A aplicabilidade empírica dessas medidas é ilustrada em estudos que utilizam a decomposição do Índice de Concentração para investigar os fatores associados às desigualdades observadas. Sharma et al. (2022), em *Decomposing socioeconomic inequality in blood pressure and blood glucose testing*, analisam desigualdades socioeconômicas no acesso a testes de pressão arterial e glicemia em quatro distritos do estado de Kerala, na Índia, com base em um estudo transversal domiciliar realizado entre julho e outubro de 2019. A partir de uma amostra de 6.383 indivíduos com 30 anos ou mais, os autores empregam estatísticas descritivas, índices de concentração de Erreygers e sua decomposição para estimar a contribuição relativa de diferentes covariáveis para a desigualdade observada. Os resultados mostram que o acesso aos testes apresenta desigualdades socioeconômicas relevantes, cuja magnitude varia entre os distritos, e que fatores socioeconômicos e demográficos desempenham papel importante na explicação desse gradiente. Embora o desfecho analisado não seja o

diabetes em si, o estudo oferece uma referência metodológica clara ao demonstrar como a decomposição do Índice de Concentração pode ser aplicada para identificar os mecanismos estatísticos associados às desigualdades em saúde.

Corroborando com isso, Kumar et al. (2025), em *Decomposing socioeconomic inequality in lean diabetes among middle-aged adults and elderly in India*, analisam a desigualdade socioeconômica do diabetes entre adultos de meia-idade e idosos na Índia, utilizando os microdados da primeira onda do *Longitudinal Ageing Study in India* (LASI), com amostra analítica de 58.824 indivíduos. O estudo combina diagnóstico autorreferido de diabetes com o índice de massa corporal para identificar casos de diabetes magro, empregando estatísticas descritivas, regressão logística multivariada, o Índice de Concentração de Erreygers e sua decomposição. Os resultados indicam desigualdade pró-pobre na distribuição desse subtipo de diabetes, com maior concentração entre os estratos socioeconômicos mais desfavorecidos. A decomposição evidencia que a condição econômica domiciliar responde pela maior parcela da desigualdade observada, seguida pelo local de residência e pela escolaridade. Esses achados reforçam a utilidade da decomposição do Índice de Concentração para identificar os fatores associados ao gradiente socioeconômico de desfechos em saúde, inclusive em contextos nos quais a distribuição da doença assume padrões distintos segundo a posição social.

Em conjunto, esses trabalhos fornecem a base conceitual e metodológica para a adoção de medidas de desigualdade sensíveis à posição socioeconômica relativa dos indivíduos. A utilização do Índice de Concentração, e, em particular, de sua versão corrigida proposta por Erreygers, permite avaliar de forma consistente se a prevalência do diabetes pode estar associada a um gradiente socioeconômico e em que direção esse gradiente se manifesta. Além disso, a possibilidade de decompor o índice amplia o escopo da análise, ao permitir investigar quais características observáveis contribuem para a desigualdade medida, em consonância com a literatura de economia da saúde aplicada.

2.4 – Evidências e Estratégias de Decomposição das Desigualdades

A mensuração das desigualdades socioeconômicas em saúde, embora informativa, não é suficiente para compreender os mecanismos associados à sua

formação. Para avançar além da identificação do gradiente, a literatura propõe o uso de técnicas de decomposição, que permitem avaliar em que medida as desigualdades observadas podem ser associadas à distribuição dos determinantes socioeconômicos ou às diferenças na forma como esses determinantes se relacionam com o desfecho de interesse. Nesse contexto, as decomposições assumem um papel central ao conectar resultados empíricos a interpretações analíticas mais estruturadas.

No caso de desfechos binários, como o diagnóstico de diabetes mellitus, a aplicação direta das decomposições tradicionais de Oaxaca–Blinder, originalmente desenvolvidas para modelos lineares, não é apropriada. Em *An Extension of the Blinder–Oaxaca Decomposition to Logit and Probit Models* (Fairlie, 2005) propõe uma extensão dessa metodologia para modelos não lineares, como logit e probit, permitindo decompor diferenças médias em probabilidades previstas entre grupos populacionais. A abordagem desenvolvida por Fairlie baseia-se na comparação contrafactual das distribuições das covariáveis, mantendo fixos os coeficientes estimados e explorando a substituição contrafactual das distribuições das covariáveis entre os grupos, o que possibilita quantificar a parcela das diferenças observadas atribuível às características observáveis dos grupos.

Assim, a principal contribuição deste método reside na sua capacidade de adaptar a lógica da decomposição de Oaxaca–Blinder a contextos em que o desfecho é dicotômico, preservando a interpretação econômica dos resultados. Ao invés de decompor diferenças em médias lineares, a metodologia avalia como a substituição da distribuição das covariáveis de um grupo pela de outro afeta a média das probabilidades estimadas. Essa característica torna a abordagem particularmente adequada para o estudo de desigualdades em saúde, nas quais muitos desfechos relevantes são naturalmente modelados como variáveis binárias.

Complementarmente, *Decomposing Differences in the First Moment* (Yun, 2004) desenvolve uma formulação geral para a decomposição de diferenças no primeiro momento da distribuição, aplicável tanto a modelos lineares quanto não lineares. Yun propõe um procedimento sistemático para atribuir pesos às contribuições individuais das covariáveis, assegurando que a soma das contribuições reproduza exatamente a diferença total observada entre os grupos. Essa formalização é particularmente útil para organizar a decomposição detalhada dos efeitos de

composição, evitando problemas de ordenação e facilitando a interpretação dos resultados.

No âmbito das análises em saúde, estratégias de decomposição têm sido amplamente utilizadas para investigar disparidades associadas a características como escolaridade, renda, raça e gênero. Conforme sistematizado em *Analyzing Health Equity Using Household Survey Data* (O'Donnell et al., 2007), essas abordagens permitem avaliar em que medida as desigualdades observadas em desfechos de saúde podem ser associadas à distribuição das características observáveis ao longo do gradiente socioeconômico. Embora os resultados empíricos variem conforme o contexto e o desfecho analisado, a literatura aponta que parte relevante das desigualdades pode ser atribuída a essas características, enquanto uma fração residual permanece associada a fatores não observados ou a diferenças estruturais mais profundas. Essa interpretação reforça o uso das decomposições como instrumentos analíticos e descritivos, sem atribuir a elas uma leitura causal estrita.

Adicionalmente, evidências empíricas recentes reforçam a aplicabilidade dessas técnicas no estudo das desigualdades em saúde. Tabrizi et al. (2023), em *Socioeconomic inequality in hypertension and its determinants in people over 60 years in Fasa, southern Iran: a Blinder–Oaxaca decomposition*, analisam disparidades socioeconômicas na prevalência de hipertensão em indivíduos com 60 anos ou mais no sul do Irã, com base em dados de inquérito populacional coletados entre 2014 e 2016. Os autores utilizam uma decomposição do tipo Blinder–Oaxaca para investigar diferenças na probabilidade de ocorrência da condição entre grupos populacionais, mostrando que parcela significativa das desigualdades observadas é explicada pela distribuição desigual de características como escolaridade, renda e condições de vida, enquanto uma fração relevante permanece associada a componentes não explicados, indicando a presença de mecanismos estruturais não capturados pelas variáveis observáveis. Esses resultados evidenciam a adequação das técnicas de decomposição para a análise de desigualdades em saúde, especialmente em contextos nos quais desfechos binários e fatores socioeconômicos interagem de forma complexa.

Deste modo, no presente estudo, a adoção da decomposição de Oaxaca-Blinder com a contribuição de Fairlie, aliada à formalização proposta por Yun, para caso não lineares, permite examinar diferenças na prevalência do diabetes entre

grupos definidos por escolaridade e raça, identificando a contribuição relativa das características observáveis para essas disparidades. Essa estratégia é complementar à decomposição do Índice de Concentração, pois possibilita analisar desigualdades sob uma ótica alternativa, baseada em comparações diretas entre grupos. Em conjunto, essas abordagens ampliam a compreensão dos padrões de desigualdade socioeconômica associados ao diabetes, ao articular medidas de gradiente populacional com análises detalhadas dos mecanismos subjacentes às diferenças observadas.

Em síntese, o referencial teórico e empírico apresentado neste capítulo fornece a base conceitual e metodológica para a análise das desigualdades socioeconômicas associadas ao diabetes mellitus autorreferido. A literatura discutida evidencia que os desfechos de saúde seguem um gradiente social persistente e que o diabetes constitui um desfecho multifatorial, sensível a fatores socioeconômicos, demográficos e comportamentais. Ademais, foram discutidas as principais medidas de mensuração e decomposição das desigualdades em saúde, justificando a adoção do Índice de Concentração — em especial sua versão corrigida — e das decomposições não lineares. Com base nesse arcabouço, o capítulo seguinte apresenta a estratégia empírica e os procedimentos metodológicos adotados neste estudo.

3. Metodologia

3.1 – Base de Dados

A base de dados utilizada neste trabalho é a da Pesquisa Nacional de Saúde (PNS) de 2019, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em parceria com o Ministério da Saúde.

A PNS constitui o principal inquérito domiciliar voltado à avaliação das condições de saúde, estilo de vida e acesso aos serviços de saúde da população brasileira. Seu objetivo é “coletar informações sobre o desempenho do sistema nacional de saúde no que se refere ao acesso e uso dos serviços disponíveis e à continuidade dos cuidados, bem como sobre as condições de saúde da população, a vigilância de doenças crônicas não transmissíveis e os fatores de risco a elas associados.” (IBGE, [s.d.]).

A pesquisa possui caráter amostral probabilístico e representatividade nacional, sendo composta por três questionários complementares: o Domiciliar, o de Todos os Moradores e o Individual. Essa estrutura permite captar informações tanto em nível coletivo quanto individual, acerca das condições de saúde, hábitos e percepções do entrevistado selecionado.

Assim, a escolha da PNS 2019 justifica-se pela sua abrangência temática e qualidade metodológica, que possibilitam investigar, sob uma ótica socioeconômica, os fatores associados à probabilidade de ocorrência de diabetes entre indivíduos. Trata-se de uma base que combina, com elevado rigor técnico, dados sobre renda, hábitos de vida, autopercepção de saúde e comorbidades, o que a torna especialmente adequada à análise proposta.

Ademais, destaca-se que a edição de 2019 é a mais recente disponível e que seus microdados são de acesso público, disponibilizados pelo IBGE em formato de arquivo texto de largura fixa.[1]

A partir desses arquivos, elaborou-se um procedimento de extração e modelagem destinado a integrar as informações dos questionários, selecionar as variáveis de interesse e preparar a base analítica final, utilizada nas etapas de estimação econométrica subsequentes.

Ressalta-se que considerando o desenho amostral complexo da PNS, as estimações realizadas neste estudo incorporam os pesos amostrais individuais

[1] https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=PNS/2019/Microdados/Dados

fornecidos pelo IBGE, de modo a garantir a representatividade dos resultados para a população brasileira. Isso pois, a utilização dos pesos corrige a probabilidade desigual de seleção dos indivíduos e ajusta possíveis distorções decorrentes do plano amostral, permitindo que as estimativas reflitam adequadamente a distribuição populacional. Assim, tanto as estimações do modelo econométrico quanto os cálculos dos índices de concentração e das decomposições realizadas são conduzidos com base em estatísticas ponderadas.

3.2 – Modelagem da Base de Dados

A etapa de modelagem, assim como as demais etapas do trabalho, foi inteiramente realizada no software R, e se consistiu em sucessivas filtragens e exclusões destinadas a refinar o conjunto de observações e adequar a amostra às necessidades do estudo. O objetivo central foi obter uma base de dados composta apenas por entrevistas válidas, realizadas com indivíduos para os quais há informações diretas e consistentes sobre características socioeconômicas, hábitos de vida e condições de saúde.

Em primeiro lugar, foram mantidos apenas os registros com V0015 = 1, o que assegura entrevistas efetivamente concluídas. Essa filtragem inicial elimina observações referentes a questionários interrompidos ou inconsistentes, garantindo a validade das respostas utilizadas nas análises subsequentes.

Na sequência, restringiu-se a amostra aos indivíduos que tiveram medidas antropométricas aferidas, isto é, com V0025B = 1. Tal etapa foi fundamental, pois variáveis como peso e altura, indispensáveis para o cálculo de sobrepeso e obesidade via Índice de Massa Corporal (IMC), são derivadas dessas aferições. A ausência desses dados inviabilizaria a correta mensuração de fatores associados ao risco de diabetes.

Após essa etapa, manteve-se na amostra apenas os casos em que M001 = 1, ou seja, aqueles em que o próprio morador selecionado respondeu ao questionário individual. Essa escolha decorre do fato de que várias informações cruciais (como percepção de saúde, hábitos alimentares, prática de atividade física e histórico de doenças) são de caráter autodeclarado, não devendo, portanto, ser respondidas por terceiros.

Ademais, com o intuito de evitar distorções nos resultados, foram excluídas as gestantes, identificadas pela variável P005 = 1. Pois, a gravidez acarreta variações

fisiológicas temporárias, sobretudo no peso e nos hábitos alimentares, que poderiam comprometer a consistência estatística das inferências sobre o risco de diabetes.

Também foram removidas da base as observações em que M00203 = 2, correspondentes a pessoas consideradas não aptas a responder. Essa decisão visa assegurar a coerência metodológica do estudo, uma vez que, nesses casos, as respostas foram fornecidas por outra pessoa do domicílio. Além disso, o número de indivíduos nessa condição é ínfimo, de modo que sua exclusão não afeta a representatividade amostral.

Por fim, realizou-se o tratamento dos valores ausentes. Nos casos em que a ausência de informação correspondia a menos de 2,15% do total da amostra, aplicou-se o método de *listwise deletion*, que consiste na exclusão completa das observações com dados faltantes. Essa opção foi adotada por se tratar de um percentual residual, cuja remoção não compromete a robustez da amostra, permitindo, ao mesmo tempo, preservar a consistência das variáveis utilizadas.

O detalhamento das variáveis selecionadas e tratadas — com seus respectivos códigos, descrições e transformações — encontra-se no Apêndice A, onde também são apresentadas as recodificações e agrupamentos utilizados na construção das variáveis analíticas.

3.3 – Tratamento de Dados Faltantes e Imputação Múltipla

Após a etapa de filtragem e exclusão das observações inconsistentes, focalizou-se no tratamento dos valores faltantes remanescentes na base de dados. A adequada abordagem desse problema é essencial, uma vez que a omissão ou o tratamento inadequado de dados ausentes pode comprometer a validade estatística das estimativas e introduzir viés nos resultados.

Inicialmente, como já descrito, aplicou-se o método de *listwise deletion* apenas às variáveis cujas ausências correspondiam a menos de 2,15% do total da amostra. Esse procedimento foi considerado suficiente para casos residuais, em que a exclusão de observações não acarreta perda significativa de informação. Entretanto, para as variáveis com proporções de ausência mais relevantes, adotou-se uma abordagem mais robusta, baseada na Imputação Múltipla por Equações Encadeadas (*Multiple Imputation by Chained Equations* – MICE).

O método MICE baseia-se em uma estratégia iterativa de imputação condicional, também conhecida como *Fully Conditional Specification*. Segundo van

Buuren (2018), nessa abordagem, cada variável com valores ausentes é imputada por meio de um modelo estatístico condicional às demais variáveis disponíveis, respeitando o tipo de dado envolvido — modelos lineares para variáveis contínuas, logísticos para binárias e logito multinomiais para variáveis categóricas com mais de duas categorias. Assim, a cada iteração, as estimativas das variáveis imputadas são atualizadas sucessivamente até que a distribuição conjunta dos dados imputados atinja estabilidade.

Em termos formais, o MICE pode ser interpretado como uma aproximação iterativa à inferência bayesiana plena descrita por Rubin (1987) em *Multiple Imputation for Nonresponse in Surveys*, na qual as imputações múltiplas são amostras provenientes da distribuição posterior dos dados completos $P(X_j^{mis} \mid X^{(obs)})$. Ou seja, a cada iteração, as estimativas são atualizadas de modo a refletir a incerteza associada à ausência dos dados, o que assegura a validade estatística das inferências subsequentes.

De forma sintética, se $X = (X_1, X_2, \dots, X_p)$ representa o conjunto de variáveis e $X_j^{(mis)}$ denota os valores ausentes da variável j, o MICE estima sucessivamente:

$$P(X_j^{(mis)} \mid X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p),$$

gerando imputações condicionais para cada variável até atingir convergência. Esse processo é repetido m vezes, resultando em m bases completas, cada uma contendo estimativas ligeiramente diferentes dos valores faltantes.

Segundo o autor supracitado, a principal vantagem da imputação múltipla é que ela preserva a variabilidade inerente à incerteza dos dados ausentes, ao contrário da imputação simples, que tende a subestimar a variância dos estimadores. Dessa forma, o MICE oferece estimativas consistentes e intervalos de confiança mais realistas, sendo especialmente adequado para inquéritos amostrais complexos, como a PNS, em que o padrão de ausência pode depender de múltiplos fatores observáveis.

No presente estudo, a aplicação do MICE teve por objetivo lidar com as ausências observadas nas variáveis relevantes à modelagem econométrica, assegurando que a base final preservasse o máximo possível de informação. Ademais, ressalta-se que antes da imputação, realizou-se um diagnóstico do mecanismo de ausência, por meio de regressões da variável indicadora de ausência ($R = 1$ se faltante, $R = 0$ caso contrário) sobre o conjunto de covariáveis disponíveis. Assim, a significância estatística dessas covariáveis e a ausência de evidência de que

a falta estivesse relacionada aos próprios valores ausentes permitiram considerar os dados compatíveis com o mecanismo *Missing at Random* (MAR), segundo o qual a probabilidade de ausência depende apenas de características observáveis, como idade, escolaridade ou renda, e não do valor faltante em si. Dessa forma, as imputações múltiplas realizadas pelo MICE basearam-se em uma suposição empiricamente plausível, garantindo estimativas não viesadas e estatisticamente consistentes. (Rubin, 1987; van Buuren, 2018).

Ora, as variáveis empregadas como auxiliares nos modelos condicionais de imputação foram agrupadas de acordo com cinco dimensões principais: 1. demográficas, abrangendo idade, sexo e cor/raça; 2. socioeconômicas, incluindo o logaritmo da renda domiciliar per capita, escolaridade, estado civil e localidade do domicílio (urbano ou rural); 3. comportamentais, compostas por hábito de fumar, consumo de bebidas alcoólicas, tempo diário de exposição à televisão e prática de atividade física; 4. de saúde, englobando hipertensão e obesidade; e 5. de acesso a serviços, representadas pela posse de plano de saúde.

Levando isso em consideração, neste trabalho, apenas três variáveis foram submetidas à imputação: (i) *esc_serie* (D00901), (ii) *esc_concluiu* (D014) e (iii) a variável dependente *diabetes* (Q03001). Suas taxas de ausência observadas foram respectivamente de aproximadamente 6%, 17% e 7%.

Para a escolaridade, adotou-se uma estratégia em duas etapas que respeita a ordem causal entre as informações: primeiro, imputou-se *esc_serie* sem utilizar *esc_concluiu* como preditor; em seguida, imputou-se *esc_concluiu*, incluindo *esc_serie* como variável explicativa, garantindo coerência interna entre série cursada e conclusão declarada.

Quanto à variável dependente *diabetes*, optou-se por incluí-la no processo de imputação, mesmo não sendo prática universal, em razão da fração de ausência observada (≈7%) e da natureza do mecanismo de falta previamente diagnosticado como MAR. Nessa situação, a exclusão dos casos incompletos poderia gerar viés de seleção, uma vez que a probabilidade de ausência se mostrou associada a características observáveis, como idade, sexo, escolaridade e condição urbano/rural. Assim, a imputação dessa variável foi conduzida por meio de regressão logística condicional às covariáveis do modelo e às variáveis auxiliares, de modo a preservar a estrutura de dependência entre a variável explicada e seus determinantes socioeconômicos. Essa escolha metodológica é sustentada pela literatura (Donders

et al., 2006; von Hippel, 2007), que recomenda imputar o desfecho quando o mecanismo de ausência não é completamente aleatório, a fim de evitar distorções nas relações de interesse e assegurar consistência estatística entre o modelo de imputação e o modelo substantivo.

Assim, o processo de imputação gerou 30 bases completas e independentes, resultantes de cadeias distintas do algoritmo MICE. O número de imputações foi definido com base na recomendação de Rubin (1987), segundo a qual o número de bancos deve ser, no mínimo, equivalente à fração percentual de dados ausentes, além de considerações práticas sobre a estabilidade das estimativas. O uso de 30 imputações garantiu a convergência das distribuições condicionais e reduziu o erro de Monte Carlo associado ao processo iterativo, assegurando maior precisão nas variâncias combinadas.

Após a etapa de imputação, verificou-se a estabilidade das distribuições das variáveis imputadas ao longo das bases geradas. Para cada variável imputada, foram calculadas as médias e variâncias dentro e entre as imputações, observando-se baixa variabilidade entre bases e consistência das médias imputadas em relação às observadas originalmente, o que indica convergência adequada do algoritmo e plausibilidade dos valores gerados.

Em seguida, as estimativas e variâncias associadas a cada conjunto de imputações foram então combinadas segundo as Regras de Rubin, incorporando tanto a variabilidade dentro das imputações quanto a variabilidade entre elas. Formalmente, seja $\hat{Q}_j$ o estimador obtido na j-ésima base imputada ($j = 1, 2, \dots, m$) e U_j sua variância associada, define-se a média dos estimadores como

$$\bar{Q} = \frac{1}{m} * \sum_{j=1}^{m} \hat{Q}_j,$$

a variância intra-imputação como

$$\bar{U} = \frac{1}{m} * \sum_{j=1}^{m} U_j,$$

e a variância entre imputações como

$$B = \frac{1}{m - 1} * \sum_{j=1}^{m} (\hat{Q}_j - \bar{Q})^2.$$

A variância total combinada é obtida por

$$T = \bar{U} + (1 + \frac{1}{m}) * B,$$

de onde derivam os erros-padrão e intervalos de confiança ajustados à incerteza da imputação múltipla. Deste modo, essa combinação final permite que as estimativas reflitam não apenas a variabilidade amostral, mas também a incerteza inerente ao processo de imputação, produzindo resultados não viesados, consistentes e estatisticamente eficientes (Rubin, 1987; van Buuren, 2018).

Por fim, o processo de imputação múltipla resultou na consolidação de uma base de dados imputada e representativa, sendo aplicado exclusivamente às variáveis necessárias à estimação dos modelos econométricos subsequentes, em consonância com os objetivos analíticos do estudo. Deste modo, as estimações foram realizadas com base nas 30 bases completas geradas, cujos resultados foram combinados segundo as Regras de Rubin (Rubin, 1987), incorporando simultaneamente a variabilidade intra e inter-imputação. Esse procedimento assegura estimativas consistentes e intervalos de confiança que refletem adequadamente a incerteza associada aos dados faltantes, de modo que todos os resultados apresentados neste e nos capítulos seguintes decorrem desse processo de pooling, conforme as melhores práticas metodológicas.

3.4 – Diagnósticos da Amostra e Qualidade dos Dados

Antes das estimações econométricas, realizaram-se testes e diagnósticos para avaliar a qualidade da base de dados e a adequação da amostra às exigências do modelo. Essa etapa teve como objetivo conferir se as variáveis selecionadas apresentavam variabilidade suficiente, ausência de colinearidade excessiva e estrutura amostral compatível com a análise proposta.

A amostra final contém 6.325 observações. Considerando o conjunto de bases imputadas, a prevalência média não ponderada de diabetes foi de aproximadamente 8,4%, com baixa variabilidade entre imputações. Ao incorporar os pesos amostrais da PNS, a prevalência estimada na população foi de cerca de 7,9%, valor próximo ao observado na amostra analítica. Esse resultado sugere que o processo de imputação e os critérios de seleção não introduziram distorções relevantes na distribuição do desfecho.

Considerando as 16 variáveis explicativas a serem incluídas na regressão, obteve-se uma razão de 32,94 eventos por parâmetro (EPV), valor superior ao mínimo

de 10 eventos recomendado pela literatura (Peduzzi et al., 1996; Vittinghoff e McCulloch, 2007). Essa proporção assegura estimativas estáveis, intervalos de confiança confiáveis e risco mínimo de sobreajuste (*overfitting*).

As variáveis categóricas, por sua vez, apresentaram frequências elevadas em todas as categorias, com contagens superiores a 1.000 observações. Além disso, a análise de células cruzadas entre cada variável explicativa e a variável explicada revelou ausência de *sparsity*, uma vez que nenhuma combinação apresentou menos de 90 observações. Tal resultado indica adequada variabilidade interna e garante suporte estatístico suficiente para todas as *dummies* utilizadas no modelo.

A verificação de multicolinearidade entre as variáveis explicativas (realizados previamente à estimação do modelo, de modo a garantir a viabilidade da especificação) mostrou correlações máximas de |r| = 0,35, sem qualquer sinal de redundância preocupante. Os valores do *Variance Inflation Factor* (VIF) variaram entre 1,0 e 1,53, muito abaixo dos limites críticos geralmente aceitos (5 para atenção e 10 para severa colinearidade). Esses resultados indicam que as regressoras são estatisticamente independentes o suficiente para serem incluídas simultaneamente, assegurando a estabilidade e a identificabilidade do modelo.

As variáveis contínuas, idade e logaritmo da renda domiciliar per capita, também apresentaram comportamento adequado. A idade média dos respondentes foi de 47,1 anos (DP = 17,3), com mediana de 46 anos e valores compreendidos entre 15 e 104. Já o logaritmo da renda per capita apresentou média de 6,78 (DP = 0,98) e mediana de 6,91, com dispersão compatível com o padrão esperado para a renda na amostra analisada. Além disso, foram identificados apenas dois outliers para idade e 263 para renda (cerca de 4% da amostra), valores plausíveis e não influentes. Diante disso, não houve necessidade de truncamento ou winsorização, uma vez que esses casos representam variação legítima dentro do fenômeno estudado.

Por fim, os diagnósticos realizados indicam que a amostra analítica apresenta consistência interna e perfil compatível com o esperado para a população brasileira, considerando a estrutura da PNS 2019. Adicionalmente, verificou-se que as variáveis que mais diferenciam indivíduos com e sem diagnóstico de diabetes são idade, hipertensão e obesidade, seguidas por renda e hábitos sedentários, resultados que se mostram empiricamente plausíveis. Não foram identificados indícios relevantes de viés estrutural decorrente dos filtros aplicados ou do processo de imputação,

sugerindo que os modelos econométricos capturam diferenças sistemáticas entre os indivíduos, e não distorções introduzidas no processo de construção da base.

3.5 – Seleção e Descrição das Variáveis

Após o tratamento dos dados e a verificação da qualidade amostral, foram definidas as variáveis utilizadas na estimação dos modelos econométricos. A seleção foi realizada com base em critérios empíricos e de plausibilidade teórica, buscando representar, de forma abrangente, os fatores demográficos, econômicos, de acesso à saúde, de estilo de vida e de condição clínica que podem influenciar a probabilidade de ocorrência do diabetes.

O modelo empírico proposto é formalizado da seguinte forma:

$$diabetes_i = f(idade\_c, idade2\_c, homem, negro, urbano, ln\_renda\_pc, plano\_saude, esc\_fund2, esc\_medio, esc\_superior, ativ\_fis3m, tv\_3h, fumou\_dia, saude\_ok, hipertensao, obesidade)$$

onde a variável dependente (*diabetes*) é binária, assumindo valor 1 quando o indivíduo declara diagnóstico médico de diabetes e 0 caso contrário.

Complementando, as variáveis estão descritas no Quadro 01, que apresenta o nome e uma breve descrição de cada variável incluída no modelo.

Quadro 01 – Variáveis do Modelo.

Variável	Descrição
idade_c	idade em anos centralizada em torno da média
idade2_c	idade em anos ao quadrado centralizado em torno da média
homem	1, se a pessoa é do gênero masculino; 0, caso contrário
negro	1, se a pessoa é preta ou parda; 0, caso contrário
urbano	1, se domicílio localizado na zona urbana; 0, caso contrário
ln_renda_pc	Logaritmo natural da renda domiciliar per capita em reais
plano_saude	1, se a pessoa tem plano de saúde; 0, caso contrário
esc_fund2	1, se a pessoa tem ensino fundamental completo ou médio incompleto; 0, caso contrário
esc_medio	1, se a pessoa tem ensino médio completo ou superior incompleto; 0, caso contrário
esc_superior	1, se a pessoa tem ensino superior completo; 0, caso contrário

ativ_fis3m	1, se, nos últimos 3 meses, a pessoa praticou algum tipo de atividade física; 0, caso contrário
tv_3h	1, se a pessoa assiste 3 horas ou mais de televisão por dia; 0, caso contrário
fumou_dia	1, se a pessoa fuma ou fumou diariamente em algum momento da vida; 0, caso contrário
saude_ok	1, se a pessoa se autoavalia com saúde razoável ou melhor; 0, caso contrário
hipertensao	1, se algum médico já deu o diagnóstico de hipertensão arterial para a pessoa; 0, caso contrário
obesidade	1, se a pessoa tem IMC acima de 30; 0, caso contrário
diabetes	1, se algum médico já deu o diagnóstico de diabetes para a pessoa; 0, caso contrário

Fonte: dados elaborados pelo autor a partir da PNS 2019

Observa-se que a variável que informa a idade foi incluída de forma linear e quadrática para capturar possíveis efeitos não lineares sobre a probabilidade de ocorrência da doença e ambas as variáveis foram centralizadas em torno de sua média, pois este é um procedimento que reduz a colinearidade entre os termos e facilita a interpretação dos coeficientes no modelo logit.

Ademais, as variáveis de escolaridade foram transformadas em *dummies* (*esc_fund2, esc_medio e esc_superior*), tomando como categoria de referência os indivíduos com ensino fundamental I completo ou menos. Já as outras covariáveis representam fatores demográficos (*homem, negros, urbano*), econômicos (*ln_renda_pc*), de acesso à saúde (*plano_saude*), comportamentais (*ativ_fis3m, tv_3h, fumou_dia*) e clínicos (*hipertensão, saúde_ok, obesidade*).

Deste modo, o conjunto final de variáveis foi definido de forma a assegurar equilíbrio entre relevância teórica, disponibilidade de dados e parcimônia do modelo, de modo que os coeficientes estimados possam refletir adequadamente os principais determinantes socioeconômicos e comportamentais associados ao diabetes mellitus.

3.6 – Modelo Econométrico Logit

Com o intuito de investigar os determinantes associados à probabilidade de um indivíduo apresentar diagnóstico médico de diabetes, estimou-se um modelo logit binário. Essa escolha metodológica se justifica pelo fato de a variável dependente ser

binária: assume valor 1 se o indivíduo declarou diagnóstico médico de diabetes e 0 caso contrário. Segundo Gujarati (2011), o modelo logit é apropriado nesse contexto por evitar limitações do modelo de probabilidade linear, como a possibilidade de predições fora do intervalo [0,1], heterocedasticidade dos erros e inconsistência dos estimadores.

Conforme Greene (2019), o modelo logit pode ser derivado a partir de uma variável latente y_i^*, representando a propensão não observável de o indivíduo ser diagnosticado com diabetes, da forma:

$$y_i^* = \beta_0 + \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i$$

em que $\boldsymbol{x_i}$ é o vetor de covariáveis observáveis, $\boldsymbol{\beta}$ é o vetor de parâmetros a estimar, e ε_i é um termo de erro logisticamente distribuído com média zero e variância constante. A variável observável y_i está relacionada a y_i^* por:

$$y_i = \begin{cases} 1, se\ y_i^* > 0 \\ 0, caso\ contrário \end{cases}$$

Em complemento, dado que ε_i segue uma distribuição logística padrão, a probabilidade condicional de ocorrência do evento é:

$$P(y_i = 1|x_i) = \Lambda(\beta_0 + \boldsymbol{x_i}\boldsymbol{\beta}) = \frac{1}{1 + e^{-(\beta_0 + x_i\beta)}},$$

garantindo que $0 < P(y_i = 1) < 1$ para todos os valores de $\boldsymbol{x_i}\boldsymbol{\beta}$.

Assim, o modelo logit utilizado neste trabalho é aplicado para estimar a probabilidade de um indivíduo relatar diagnóstico de diabetes, a partir das seguintes variáveis explicativas:

$$P_i = P(diabetes_i = 1) = \Lambda(\beta_0 + \beta_1 idade\_c_i + \beta_2 idade2\_c_i + \beta_3 homem_i + \beta_4 negro_i + \beta_5 urbano_i + \beta_6 ln\, renda_i + \beta_7 plano\_saude_i + \beta_8 esc\_fund2_i + \beta_9 esc\_medio_i + \beta_{10} esc\_superior_i + \beta_{11} ativ\_fis3m_i + \beta_{12} tv\_3h_i + \beta_{13} fumou\_dia_i + \beta_{14} saude\_ok_i + \beta_{15} hipertensao_i + \beta_{16} obesidade_i).$$

E a estimação dos parâmetros $\boldsymbol{\beta}$ foi realizada por máxima verossimilhança. Sob a hipótese de independência condicional das observações, a função log-verossimilhança é dada por:

$$\log \boldsymbol{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i log \Lambda(\boldsymbol{x_i}\boldsymbol{\beta}) + (1 - y_i) \log\left(1 - \Lambda(\boldsymbol{x_i}\boldsymbol{\beta})\right) \right],$$

maximizada numericamente por meio de algoritmos iterativos.

Prosseguindo, a interpretação dos coeficientes estimados se dá em termos do logaritmo da razão de chances (log-odds):

$$\log\left(\frac{P(y_{\boldsymbol{i}}=1)}{1-P(y_i=1)}\right)=\beta_0+\boldsymbol{x}_i\boldsymbol{\beta}.$$

Para efeitos mais substantivos, pode-se considerar os efeitos marginais. Para variáveis contínuas, o efeito marginal é:

$$\frac{\partial P(y_i=1|\boldsymbol{x_i})}{\partial x_{ij}}=\lambda(\boldsymbol{x_i}\boldsymbol{\beta})*\beta_j,$$

em que $\lambda(z)=\frac{e^{-z}}{(1+e^{-z})^2}$ é a função densidade da distribuição logística. Já para variáveis binárias, o efeito marginal é expresso por:

$$\Delta_j=\Lambda(\beta_0+\beta_j+\sum_{k\neq\boldsymbol{j}}\beta_k x_{ik})-\Lambda(\beta_0+\sum_{k\neq\boldsymbol{j}}\beta_k x_{ik}),$$

representando a variação na probabilidade estimada de diabetes quando x_j muda de 0 para 1, mantendo as demais variáveis constantes.

Em resumo, o modelo logit oferece uma estrutura teórica robusta e empiricamente adequada para analisar a influência de características individuais sobre a probabilidade de ocorrência de diabetes, respeitando as propriedades estatísticas exigidas para estimação e inferência válidas.

Por fim, após a estimação, o modelo foi submetido a uma série de testes diagnósticos com o intuito de avaliar sua significância estatística, a robustez dos coeficientes estimados e a qualidade do ajuste. Nesse sentido, os procedimentos adotados foram:

1. Teste de Wald global, para avaliar a significância conjunta dos coeficientes;
2. Pseudo-R² de McFadden, como medida do poder explicativo relativo;
3. Área sob a curva ROC (AUC), como indicador da capacidade discriminatória do modelo;
4. Fator de inflação da variância (VIF), para verificar eventual multicolinearidade entre os regressores.

Ressalta-se que esses testes são recomendados pela literatura para garantir a validade das inferências em modelos logit (Greene, 2019; Hosmer, Lemeshow e Sturdivant, 2013) e seus resultados serão apresentados no capítulo seguinte.

Dessa forma, a especificação e estimação do modelo logit atendem aos requisitos teóricos e práticos do problema em questão, oferecendo uma base estatística apropriada para investigar os determinantes individuais do diagnóstico de diabetes autorreferido.

3.7 – Índice de Concentração Segundo Erreygers e Decomposição

Após a estimação do modelo econométrico e de seus respectivos testes de ajuste, com o objetivo de avaliar a desigualdade na distribuição do diagnóstico de diabetes segundo marcadores socioeconômicos e etários, adotou-se o Índice de Concentração de Erreygers (EI). Esse índice é especialmente adequado à análise de variáveis binárias, como é o caso da variável dependente deste estudo, e supera as limitações do índice de concentração tradicional (CI) ao satisfazer simultaneamente propriedades desejáveis como a simetria, o princípio de transferência de Pigou-Dalton, a independência de nível e a invariância a transformações lineares (Erreygers, 2009).

Enquanto o CI tradicional varia com a média do desfecho e perde significado em variáveis limitadas, o EI corrige essa dependência e assegura que o índice assuma valores entre –1 e 1, independentemente da prevalência do evento. Além disso, quando o desfecho é binário (como no caso de presença ou ausência de diabetes), o EI mantém a simetria, ou seja, calcular a desigualdade em relação ao não evento gera o mesmo valor em módulo, com sinal oposto (Kjellsson e Gerdtham, 2013).

Formalmente, para uma variável y com média $\bar{y}$, mínimo y_{min}, máximo y_{max} e ordenação $R \in [0,1]$, o EI é definido como:

$$EI = \frac{4\bar{y}}{y_{max} - y_{min}} * Cov(y, R).$$

No caso de variáveis binárias normalizadas entre 0 e 1 (ou seja, $y_{min} = 0 \; e \; y_{max} = 1$), essa expressão simplifica-se para:

$$EI = 8 * Cov(y, R)$$

ou de forma equivalente, $EI = 4\bar{y} * CI$, onde CI é o índice de concentração tradicional (O'Donnell et al., 2007).

Nesse contexto, o EI foi calculado a partir de duas ordenações distintas: a idade e o logaritmo da renda domiciliar per capita. A escolha desses rankings visa capturar desigualdades segundo diferentes dimensões — biológica e socioeconômica, respectivamente.

Além do cálculo do EI, também se realizou sua decomposição para quantificar as contribuições de cada covariável explicativa na desigualdade observada no desfecho. Segundo O'Donnell et al. (2007), essa técnica parte da estimação de um modelo econométrico para o desfecho de interesse — neste caso, um modelo logit binário — e utiliza a decomposição da covariância entre $y \; e \; R$ com base na

sensibilidade de cada y a cada variável explicativa e na desigualdade da própria covariável.

Para modelos lineares da forma $y_i = \alpha + \sum_k \beta_k x_{ki} + \varepsilon_i$, o índice de concentração pode ser decomposto como:

$$CI_y = \sum_k \left(\frac{\beta_k \bar{x}_k}{\bar{y}}\right) CI_{xk} + \frac{GC_\varepsilon}{\bar{y}}$$

onde CI_{xk} representa o índice de concentração da covariável $x_k, \bar{x}_k$, sua média, e GC_ε o índice de concentração generalizado do termo de erro.

No caso de variáveis binárias e do EI, aplica-se uma transformação simples sobre essa equação. Como $ECI = 4\bar{y} * CI$, tem se:

$$EI_y = \sum_k 4\beta_k \bar{x}_k * CI_{xk} + 4GC_\varepsilon.$$

Essa expressão permite decompor o EI em parcelas atribuíveis a cada covariável explicativa do modelo logit. A contribuição de uma variável x_k depende do seu coeficiente estimado no modelo, da sua média e do grau de desigualdade (CI) em relação ao ranking utilizado. Assim, covariáveis associadas positivamente ao desfecho e concentradas em grupos de maior status aumentam a desigualdade total, por outro lado, covariáveis com sinais opostos ou distribuídas de forma mais equitativa podem atenuá-la (Erreygers, 2009; O'Donnell et al., 2007).

Vale destacar que a soma das contribuições das covariáveis explica a maior parte da desigualdade medida pelo EI; a parcela remanescente corresponde ao termo residual $4GC_\varepsilon$, que representa a desigualdade associada a fatores não incluídos no modelo ou ao erro aleatório.

Em suma, a decomposição do Índice de Concentração de Erreygers oferece uma ferramenta valiosa para compreender quais fatores contribuem para a concentração do diagnóstico de diabetes entre grupos sociais distintos, e com que intensidade.

3.8 – Decomposição de Oaxaca–Blinder Não Linear

Ademais, com o intuito de investigar em que medida as diferenças observadas na ocorrência de diabetes podem ser atribuídas a fatores observáveis versus diferenças nos seus efeitos marginais, foi empregada a técnica de decomposição de Oaxaca–Blinder para modelos não lineares, conforme proposta por Yun (2004) e operacionalizada nos moldes de Fairlie (2005). Essa abordagem é especialmente

apropriada quando a variável dependente é binária, como é o caso do presente estudo.

A formulação tradicional da decomposição de Oaxaca–Blinder (Blinder, 1973; Oaxaca, 1973) é amplamente utilizada para desagregar diferenças médias em desfechos contínuos entre dois grupos. No entanto, sua aplicação direta a modelos com variável dependente binária é inadequada, uma vez que a relação entre os regressores e a probabilidade do evento é não linear. Como argumenta Fairlie (2005) em *An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models*, em modelos como o logit, a expectativa da função de probabilidade não é igual à função avaliada na média dos regressores devido à não linearidade (violação da igualdade de Jensen). Assim, métodos adaptados são necessários para assegurar validade teórica e interpretativa da decomposição.

Neste trabalho, a decomposição foi aplicada em dois contextos de desigualdade social:

1. Entre indivíduos com alta escolaridade (ensino superior completo) e aqueles com baixa escolaridade (menos que superior);
2. Entre indivíduos que se autodeclaram negros (pretos ou pardos) e os que se declaram não negros (brancos).

Considerando esses pares de comparação, estima-se a diferença nas probabilidades médias de ocorrência de diabetes com base em modelos logit ajustados separadamente para cada grupo.

Seja $F(\cdot)$ a função logística acumulada, $\hat{\beta}^A$ *e* $\hat{\beta}^B$ os vetores de coeficientes estimados para os grupos A (alta escolaridade ou não negros) e B (baixa escolaridade ou negros). A diferença média prevista é decomposta como:

$$\bar{P}_A - \bar{P}_B = \underbrace{\left[E_{x|A}\left\{F\left(X\hat{\beta}^B\right)\right\} - E_{x|B}\left\{F\left(X\hat{\beta}^B\right)\right\}\right]}_{Componente\ de\ características\ (endowment)} + \underbrace{\left[E_{x|A}\left\{F\left(X\hat{\beta}^A\right)\right\} - E_{x|A}\left\{F\left(X\hat{\beta}^B\right)\right\}\right]}_{Componente\ do\ coeficiente}.$$

O primeiro termo reflete o quanto da diferença se deve às características observáveis (idade, renda, hábitos etc.), mantendo constantes os coeficientes do grupo B — ou seja, trata-se do componente “explicado”. O segundo termo, por sua vez, corresponde ao componente “não explicado”, captando quanto da diferença se deve a distintas sensibilidades ou retornos dos grupos às mesmas características.

Ademais, para assegurar invariância à ordem das variáveis e evitar a instabilidade típica dos métodos sequenciais, utilizou-se a formulação proposta por Yun (2004) em *Decomposing differences in the First Moment*, que aplica ponderações

apropriadas na desagregação dos efeitos. Essa abordagem oferece maior robustez e consistência, mantendo a lógica da decomposição tradicional e permitindo atribuir o hiato total em probabilidades entre os grupos a dois blocos fundamentais: características e coeficientes.

Nesse sentido, a implementação computacional seguiu as etapas padrão da decomposição de Fairlie e Yun: estimação de modelos logit separados, construção de contrafactuais e cálculo dos componentes explicado e não explicado.

Em continuidade, seguindo a prática recomendada por Oaxaca e Ransom (1994) em *On Discrimination and the Decomposition of Wage Differentials*, adotou-se como grupo de referência aquele com baixa escolaridade ou identificação como negro, de forma que o componente explicado representa a diferença esperada caso ambos os grupos tivessem a mesma distribuição de características, mas com os retornos (coeficientes) do grupo de referência.

Em resumo, a decomposição de Oaxaca–Blinder não linear, portanto, fornece um instrumento valioso para distinguir em que medida as disparidades na ocorrência de diabetes entre grupos sociais decorrem de desigualdades nos perfis observáveis ou de diferenças nos retornos associados a esses fatores. Sua aplicação ao modelo logit estimado permite uma análise estrutural das desigualdades por escolaridade e raça/cor, compatível com a natureza binária do desfecho e com os pressupostos estatísticos da abordagem adotada.

Por fim, esclarece-se que a seção seguinte apresentará os resultados empíricos derivados das estratégias aqui descritas, abrangendo a análise descritiva, a estimação do modelo logit, o cálculo e a decomposição do Índice de Concentração de Erreygers e a decomposição de Oaxaca–Blinder não linear. De modo que todos os achados são interpretados à luz da fundamentação metodológica apresentada neste capítulo.

4. Resultado

4.1 – Estatística Descritiva

Esta seção apresenta as estatísticas descritivas da amostra final utilizada nas análises empíricas. Ressalta-se que os diagnósticos relativos à qualidade da amostra, à adequação do tamanho amostral, à verificação de colinearidade e à consistência das variáveis foram previamente conduzidos e discutidos na Seção 3.4 do capítulo metodológico. Assim, o foco desta seção recai exclusivamente sobre a descrição dos principais padrões observados nos dados.

Destaca-se que tanto as estatísticas aqui apresentadas como o as estimações do modelo econométrico, dos índices de concentração e das decomposições realizadas incorporam os pesos amostrais da PNS, bem como a estrutura do desenho amostral complexo, considerando unidades primárias de amostragem (UPAs) e estratos. Dessa forma, os resultados devem ser interpretados como representativos da população brasileira, e não apenas da amostra analítica utilizada no estudo.

Dando seguimento, a análise descritiva tem início pela variável central deste estudo: o diagnóstico autorreferido de diabetes. Deste modo, a apresentação da prevalência do desfecho permite situar quanto à frequência da condição na amostra analisada e fornece a base empírica sobre a qual se desenvolvem as análises subsequentes.

O Quadro 02 apresenta a prevalência média estimada de diabetes autorreferido. Observa-se que 7,9% dos indivíduos da amostra relataram diagnóstico médico da doença, com intervalo de confiança de 95% entre 7,7% e 8,0%. Em termos práticos, esse resultado indica que o diabetes está presente em uma parcela considerável da população estudada, o que reforça a pertinência da investigação de seus determinantes individuais sob diferentes dimensões — demográfica, socioeconômica, comportamental e clínica.

Quadro 02 – Prevalência de Diabetes

Prevalência de Diabetes	
% de Diabetes	**IC 95%**
7,9%	7,7% - 8,0%

Fonte: dados da pesquisa

A Tabela 01 apresenta as estatísticas descritivas das variáveis utilizadas na análise, comparando indivíduos com e sem diagnóstico autorreferido de diabetes. As estimativas correspondem a médias e proporções ponderadas, obtidas a partir do

conjunto de dados com imputação múltipla, sendo os resultados agregados conforme as regras de Rubin. Os testes de diferença entre os grupos foram realizados por meio de regressões simples ponderadas, permitindo a obtenção de p-valores consistentes com a estratégia empírica adotada ao longo do trabalho.

Tabela 01 – Estatísticas Descritivas

Estatísticas Descritivas por Diagnóstico de Diabetes				
variavel	no_diabetic	diabetic	p_value	sig
idade	42.2274	58.2083	0.0000	***
homem	0.4726	0.4089	0.0082	***
negro	0.5784	0.5392	0.0954	*
urbano	0.8648	0.8851	0.2119	
ln_renda_pc	6.8672	7.0959	0.0000	***
plano_saude	0.2480	0.2859	0.0652	*
esc_fund2	0.2250	0.1694	0.0068	***
esc_medio	0.3381	0.2312	0.0000	***
esc_superior	0.1359	0.0748	0.0001	***
ativ_fis3m	0.4360	0.4014	0.1459	
tv_3h	0.1868	0.2848	0.0000	***
fumou_dia	0.3458	0.3751	0.1965	
saude_ok	0.9520	0.8040	0.0000	***
hipertensao	0.1963	0.6061	0.0000	***
obesidade	0.2128	0.3580	0.0000	***

Fonte: dados elaborados pelo autor a partir da PNS 2019

Os resultados evidenciam diferenças estatisticamente significativas entre indivíduos diabéticos e não diabéticos em diversas dimensões. Em termos demográficos, observa-se que indivíduos com diabetes apresentam idade média substancialmente mais elevada (58,2 anos) em comparação aos não diabéticos (42,2 anos), diferença estatisticamente significativa ao nível de 1%. Além disso, a proporção de homens é menor entre os diabéticos, enquanto a variável de raça/cor apresenta diferença marginalmente significativa, com menor proporção de indivíduos negros entre aqueles com diagnóstico da doença.

No que se refere aos fatores socioeconômicos, indivíduos diabéticos apresentam, em média, maior renda domiciliar per capita, diferença estatisticamente

significativa, e maior proporção de posse de plano de saúde, ainda que com significância marginal. Em contrapartida, observa-se um claro gradiente educacional, com menor proporção de indivíduos com níveis mais elevados de escolaridade entre os diabéticos, resultado consistente com a literatura sobre determinantes sociais da saúde.

Entre os comportamentos relacionados ao estilo de vida, destaca-se a maior proporção de indivíduos que assistem televisão por três horas ou mais por dia entre os diabéticos, diferença estatisticamente significativa. Por outro lado, não se observam diferenças estatisticamente significativas na prática de atividade física recente ou no hábito de fumar.

Por fim, as diferenças mais expressivas concentram-se nas condições de saúde. Indivíduos com diabetes apresentam maior prevalência de hipertensão (60,6% contra 19,6%) e obesidade (35,8% contra 21,3%), além de menor proporção de autopercepção de saúde positiva. Essas diferenças são estatisticamente significativas e refletem a forte associação entre diabetes e outras condições crônicas, padrão amplamente documentado na literatura (Flor & Campos, 2017; Safieddine et al., 2023).

Em conjunto, os resultados descritivos fornecem evidências iniciais de que o diabetes está associado a fatores etários, condições clínicas e características socioeconômicas, antecipando padrões que serão aprofundados nas análises econométricas subsequentes.

4.2 – Resultados do Modelo Logit

Doravante serão apresentados os resultados do modelo logit estimado para a probabilidade de diagnóstico autorreferido de diabetes, bem como os principais indicadores de qualidade do ajuste e consistência estatística do modelo.

A Tabela 02 apresenta os coeficientes estimados do modelo logit, seus erros-padrão, razões de chances ajustadas (Adjusted Odds Ratios – AOR), intervalos de confiança de 95% e efeitos marginais médios (AME). Os resultados indicam que idade, condições de saúde e escolaridade exercem papel central na probabilidade de diagnóstico de diabetes, mesmo após o controle por características demográficas, socioeconômicas e comportamentais.

A idade apresenta efeito estatisticamente significativo e não linear sobre a probabilidade de diabetes. O coeficiente positivo do termo linear indica que a

probabilidade da doença aumenta com a idade. Embora o termo quadrático apresente sinal negativo, seu coeficiente não é estatisticamente significativo, sugerindo evidência limitada de não linearidade no formato da relação. Ainda assim, o padrão geral é coerente com a dinâmica do risco ao longo do ciclo de vida e reforça a importância do controle etário na análise. Resultado semelhante é documentado por Flor e Campos (2017), que encontram forte associação entre idade e diagnóstico de diabetes na população brasileira, bem como por Kundu et al. (2022), que observam aumento da prevalência da doença em faixas etárias mais elevadas. Em termos conceituais, esse padrão também é consistente com a perspectiva do ciclo de vida discutida por Marmot (2012), segundo a qual riscos à saúde tendem a se acumular ao longo do envelhecimento.

Entre as condições de saúde, a hipertensão arterial destaca-se como o principal fator associado ao aumento na probabilidade de diabetes. Indivíduos hipertensos apresentam uma razão de chances aproximadamente 2,53 vezes maior de relatar diagnóstico da doença em comparação aos não hipertensos, mantendo-se constantes as demais variáveis. A obesidade também apresenta associação positiva e marginalmente significativa, com aumento relevante na probabilidade estimada do desfecho. Esses resultados são consistentes com a literatura empírica, que documenta forte associação entre diabetes e outras condições metabólicas. Flor e Campos (2017) identificam hipertensão e obesidade como importantes fatores associados ao diagnóstico de diabetes na população brasileira, enquanto Kundu et al. (2022) mostram que a presença de hipertensão e outras comorbidades está significativamente relacionada à prevalência da doença. De forma complementar, Safieddine et al. (2023) evidenciam a elevada incidência de comorbidades entre indivíduos com diabetes, reforçando a interrelação entre essas condições no perfil epidemiológico da doença.

A autopercepção de saúde razoável ou boa está associada a uma redução significativa na probabilidade de diagnóstico de diabetes, sugerindo coerência entre a avaliação subjetiva de saúde e a presença de condições crônicas diagnosticadas.

No que se refere aos fatores socioeconômicos, observa-se um gradiente educacional claro. Em relação à categoria de referência (ensino fundamental I completo ou menos), indivíduos com ensino médio completo apresentam efeito negativo e marginalmente significativo, enquanto aqueles com ensino superior completo apresentam probabilidades significativamente menores de diagnóstico de

diabetes. O efeito associado ao ensino superior é particularmente expressivo, refletindo um importante diferencial educacional mesmo após o controle por renda, hábitos e condições de saúde. Esse resultado é consistente com evidências empíricas da literatura. Flor e Campos (2017) encontram maior prevalência de diabetes entre indivíduos com menor escolaridade no Brasil, enquanto Kundu et al. (2022) também identificam associação entre nível educacional e prevalência da doença. Evidências mais recentes para o contexto brasileiro, como em Kluthcovsky e Beraldo (2024), reforçam esse padrão ao apontar a persistência de diferenciais socioeconômicos na prevalência de diabetes autorreferido. Em perspectiva comparada, Hosseinpoor et al. (2012) mostram que doenças crônicas não transmissíveis, incluindo o diabetes, tendem a apresentar distribuição desigual ao longo da hierarquia socioeconômica, evidenciando a presença de gradientes sociais em saúde.

Por sua vez, a renda domiciliar per capita apresenta coeficiente positivo e marginalmente significativo, sugerindo que indivíduos com maior renda possuem maior probabilidade de diagnóstico da doença. Esse resultado pode refletir, ao menos em parte, diferenças no acesso ao diagnóstico e à utilização de serviços de saúde, indicando possível viés de detecção, conforme sugerido por Flor e Campos (2017).

A posse de plano de saúde privado está associada a maior probabilidade de diagnóstico de diabetes, embora o coeficiente não seja estatisticamente significativo ao nível convencional, o que sugere cautela na interpretação desse efeito. Ainda assim, o sinal estimado é consistente com a hipótese de que o acesso a serviços de saúde influencia a probabilidade de diagnóstico.

Entre os hábitos comportamentais, assistir televisão por três horas ou mais por dia apresenta associação positiva com a probabilidade de diabetes, embora sem significância estatística. De forma semelhante, a prática de atividade física não apresenta efeito estatisticamente significativo após o controle pelas demais covariáveis. Por outro lado, o tabagismo apresenta coeficiente negativo e marginalmente significativo, resultado que deve ser interpretado com cautela, podendo refletir efeitos de seleção ou outras características não observadas.

Tabela 02 – Modelo Logit

Modelo Logit para Diabetes (PNS 2019)

Coeficientes (β), odds ratios ajustados e efeitos marginais médios

Variável	β (IC95%)	SE	AOR (IC95%)	AME	Sig.
(Intercept)	-3.7181 (-5.5196 ↔ -1.9165)	0.9192	0.0243 (0.0040 ↔ 0.1471)		
idade (centr.)	0.0435 (0.0213 ↔ 0.0657)	0.0113	1.0445 (1.0216 ↔ 1.0679)	0.0027	***
idade² (centr.)	-0.0005 (-0.0011 ↔ 0.0002)	0.0003	0.9995 (0.9989 ↔ 1.0002)	-0.0000	
homem	0.0318 (-0.3666 ↔ 0.4303)	0.2033	1.0323 (0.6931 ↔ 1.5377)	0.0020	
negro	-0.0029 (-0.3728 ↔ 0.3671)	0.1887	0.9971 (0.6888 ↔ 1.4435)	-0.0002	
urbano	0.1185 (-0.2100 ↔ 0.4471)	0.1676	1.1258 (0.8106 ↔ 1.5638)	0.0075	
renda pc (ln)	0.2518 (-0.0307 ↔ 0.5343)	0.1441	1.2863 (0.9698 ↔ 1.7062)	0.0159	*
plano de saúde	0.1849 (-0.3410 ↔ 0.7109)	0.2683	1.2031 (0.7110 ↔ 2.0358)	0.0117	
ens. fund. II	-0.2023 (-0.7373 ↔ 0.3327)	0.2729	0.8169 (0.4784 ↔ 1.3947)	-0.0128	
ens. médio	-0.4241 (-0.8583 ↔ 0.0100)	0.2215	0.6543 (0.4239 ↔ 1.0100)	-0.0267	*
ens. superior	-1.0839 (-1.8843 ↔ -0.2835)	0.4084	0.3383 (0.1519 ↔ 0.7532)	-0.0684	***
ativ. fís. ú3m	0.2741 (-0.1168 ↔ 0.6651)	0.1995	1.3154 (0.8898 ↔ 1.9446)	0.0173	
tv ≥ 3h/dia	0.2748 (-0.0812 ↔ 0.6307)	0.1816	1.3162 (0.9221 ↔ 1.8789)	0.0173	
fumou diariamente	-0.3077 (-0.6462 ↔ 0.0309)	0.1727	0.7352 (0.5240 ↔ 1.0314)	-0.0194	*
saúde ok (autoav.)	-1.0749 (-1.8032 ↔ -0.3466)	0.3716	0.3413 (0.1648 ↔ 0.7071)	-0.0678	***
hipertensão	0.9277 (0.5650 ↔ 1.2904)	0.1851	2.5286 (1.7594 ↔ 3.6342)	0.0585	***
obesidade	0.4269 (-0.0066 ↔ 0.8603)	0.2212	1.5324 (0.9934 ↔ 2.3640)	0.0269	*

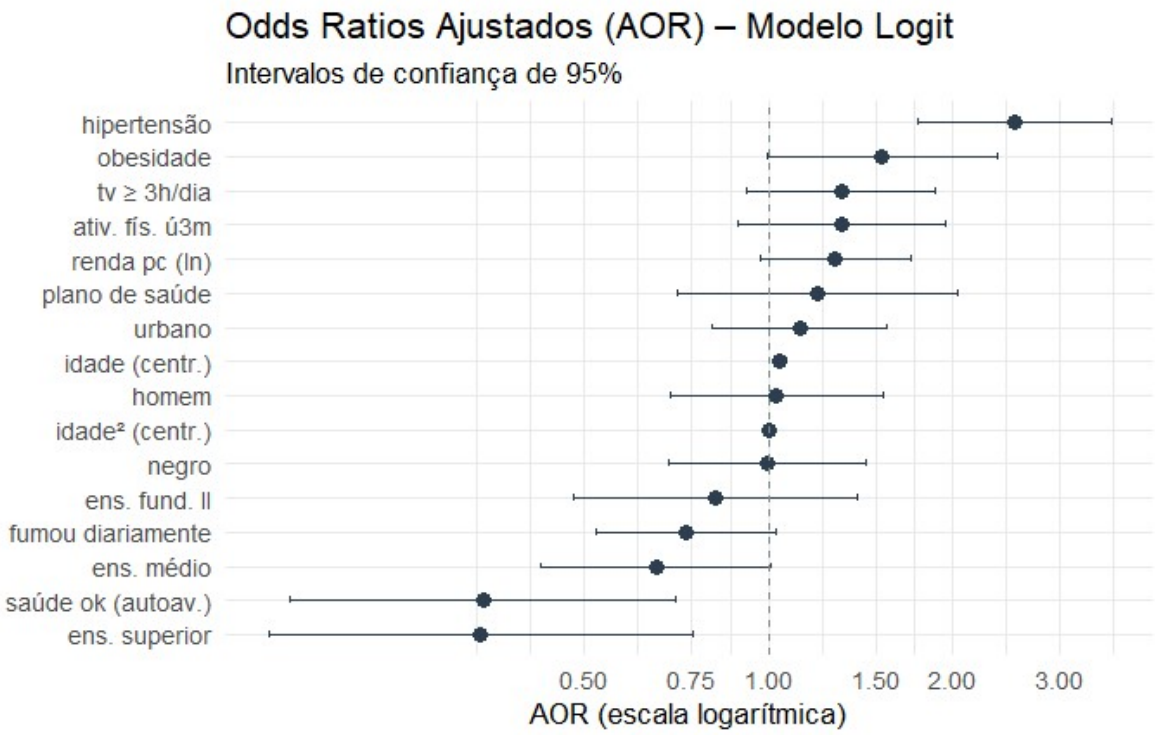
N = 6325 | Pseudo R^2 (McFadden) = 0.182 | AUC = 0.788 | Wald global: F = 14.03, p <0.001 | Asteriscos: *** p<0.01, ** p<0.05, * p<0.10.

Fonte: dados elaborados pelo autor a partir da PNS 2019

Em complemento, o Gráfico 02 apresenta o *forest plot* das razões de chances ajustadas (AOR) em escala logarítmica. O gráfico permite visualizar os efeitos estimados e seus intervalos de confiança, destacando a magnitude das associações mais relevantes, como hipertensão, obesidade, escolaridade e hábitos sedentários. Observa-se que as variáveis associadas a condições clínicas apresentam os maiores desvios em relação à linha de neutralidade (AOR = 1).

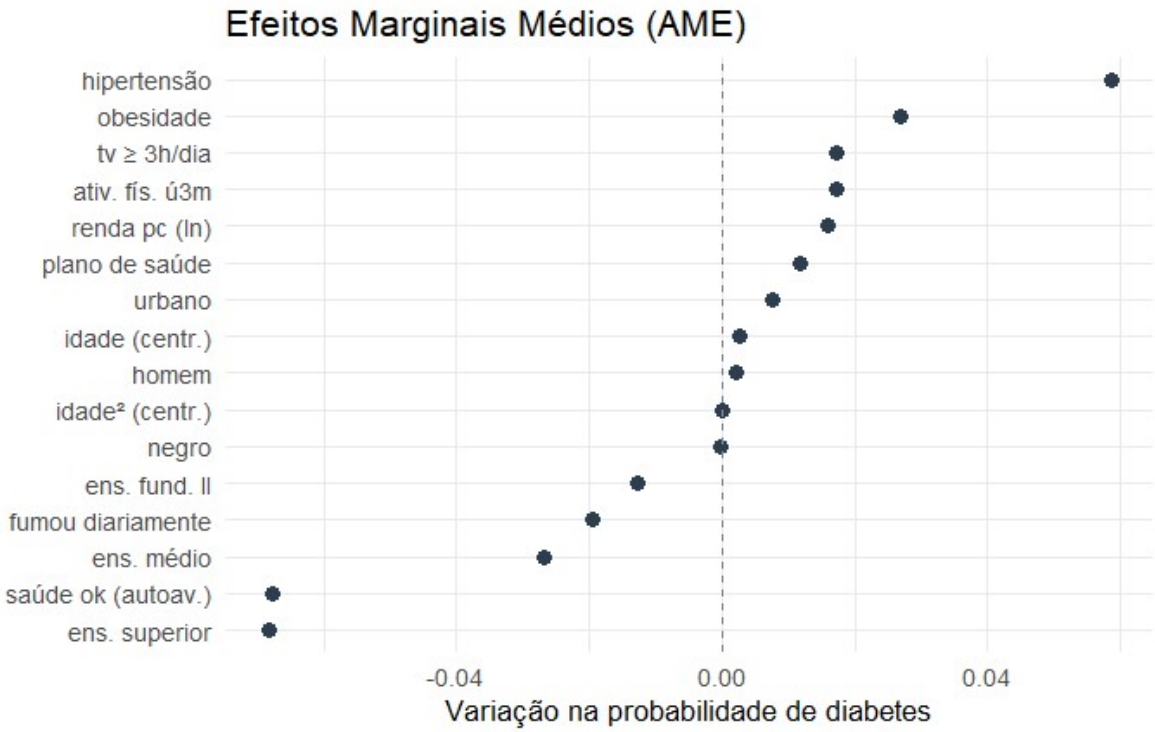
O Gráfico 03 complementa essa análise ao apresentar os efeitos marginais médios (AME), expressos como variação absoluta na probabilidade de diagnóstico de diabetes. Esse gráfico evidencia que, em termos de impacto absoluto, hipertensão e obesidade são os fatores que mais alteram a probabilidade prevista do desfecho, seguidos por escolaridade e hábitos sedentários.

Gráfico 02 – Forest Plot



Fonte: dados elaborados pelo autor a partir da PNS 2019

Gráfico 03 – Efeitos Marginais Médios



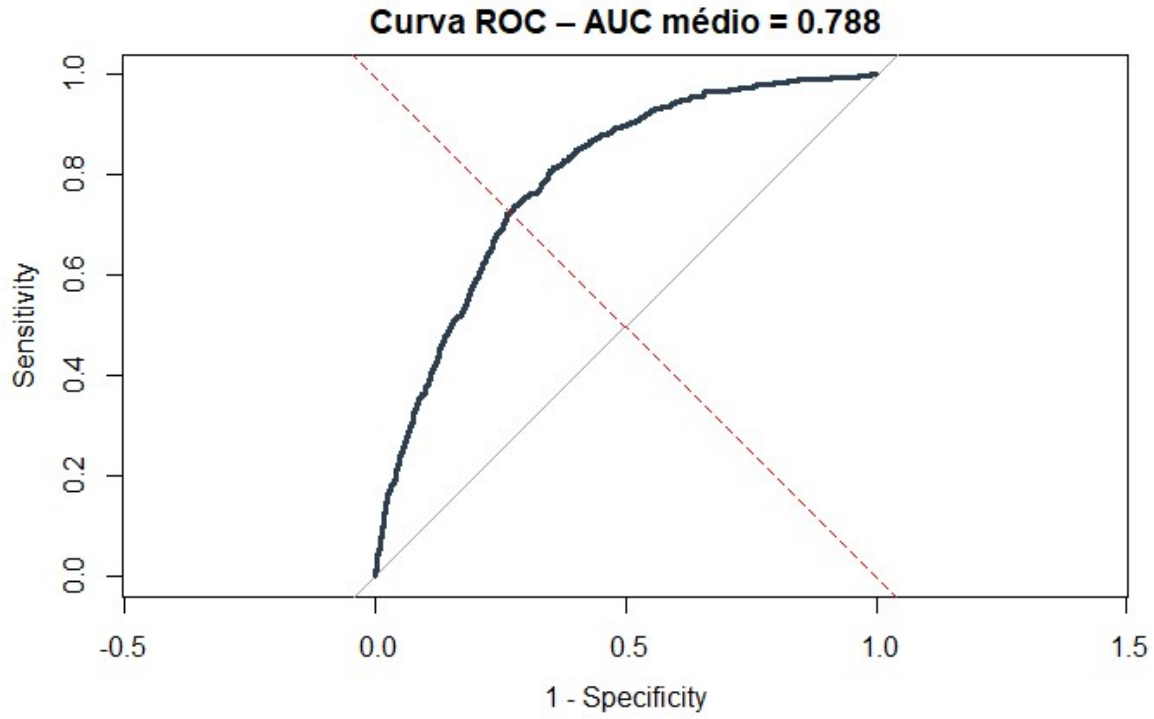
Fonte: dados elaborados pelo autor a partir da PNS 2019

Ademais, o modelo apresenta bom desempenho estatístico e adequada capacidade discriminatória. O teste global de significância indica que o conjunto dos coeficientes é estatisticamente diferente de zero ($p < 0,001$). O pseudo-R^2 de

McFadden é igual a 0,182 e, embora valores desse indicador sejam geralmente inferiores aos observados em modelos lineares, a literatura sobre modelos logit indica que valores dessa magnitude já podem ser considerados satisfatórios em aplicações com dados individuais (McFadden, 1978).

Já a área sob a curva ROC (AUC) média é de 0,788, conforme apresentado no Gráfico 04, o que indica boa capacidade do modelo em discriminar corretamente indivíduos com e sem diagnóstico de diabetes. Esse resultado sugere que o conjunto de variáveis explicativas utilizado apresenta desempenho satisfatório na distinção entre os grupos, capturando de forma consistente fatores relevantes associados ao desfecho analisado.

Gráfico 04 – Área sob a Curva



Fonte: dados elaborados pelo autor a partir da PNS 2019

Adicionalmente, a análise de multicolinearidade não revelou problemas relevantes. Os valores do fator de inflação da variância (VIF) permaneceram baixos para todas as variáveis explicativas, com valores máximos próximos a 3, indicando ausência de colinearidade excessiva e garantindo estabilidade das estimativas.

Tabela 03 – VIF

term	VIF_mean	VIF_sd	VIF_max
<chr>	<dbl>	<dbl>	<dbl>
idade_c	2.945192	0.0593242024	3.057491
idade2_c	2.369757	0.0634087804	2.492717
ln_renda_pc	1.584745	0.0042492857	1.594942
esc_superior	1.370539	0.0091489620	1.389429
esc_medio	1.359108	0.0059029013	1.371967
plano_saude	1.343766	0.0027202391	1.350992
hipertensao	1.178954	0.0042622902	1.188910
esc_fund2	1.169930	0.0051493930	1.180154
urbano	1.150036	0.0017046817	1.153575
ativ_fis3m	1.105118	0.0016949063	1.108222
negro	1.099419	0.0007774524	1.100940
fumou_dia	1.071400	0.0010936191	1.074182
homem	1.071356	0.0009141368	1.072464
saude_ok	1.069713	0.0006778507	1.071182
obesidade	1.068492	0.0009106704	1.071413
tv_3h	1.039915	0.0005407000	1.040986

Fonte: dados elaborados pelo autor a partir da PNS 2019

Em conjunto, os resultados do modelo logit indicam que a probabilidade de diagnóstico autorreferido de diabetes está fortemente associada a fatores etários e a condições clínicas pré-existentes, além de apresentar gradientes claros segundo escolaridade e alguns aspectos relacionados ao estilo de vida. O bom desempenho estatístico e a ausência de evidências de problemas relevantes de especificação conferem robustez às estimativas obtidas, fornecendo uma base consistente para as análises de desigualdade e decomposição apresentadas nas seções seguintes.

4.3 – Índice de Concentração Segundo Erreygers e Decomposição

A partir dos resultados obtidos, aqui se irá analisar a desigualdade na distribuição do diagnóstico autorreferido de diabetes a partir do Índice de Concentração de Erreygers (EI), considerando duas ordenações distintas: idade e renda domiciliar per capita. Em seguida, procede-se à decomposição do índice com o objetivo de identificar a contribuição relativa das covariáveis para a desigualdade observada.

Quando a ordenação é realizada pela idade, o índice de concentração de Erreygers assume valor positivo (EI = 0,1507), indicando que o diagnóstico de diabetes está concentrado entre indivíduos de maior idade. Esse resultado é consistente com o perfil etário da doença e com evidências da literatura que apontam o envelhecimento como um dos principais fatores associados ao aumento da incidência de doenças crônicas ao longo do ciclo de vida (Marmot, 2012; Flor & Campos, 2017).

O termo residual apresenta magnitude elevada (0,1207), o que é esperado nesse contexto, uma vez que a própria variável de ordenação (idade) capta parcela substantiva da variação do fenômeno analisado. Esse padrão é recorrente em aplicações do índice de concentração em saúde, nas quais a variável de

ranqueamento possui forte correlação com o desfecho (Hosseinpoor et al., 2012), limitando a capacidade explicativa dos fatores observáveis incluídos na decomposição.

Tabela 04 - Decomposição do EI em Idade

Decomposição do EI (diabetes) ao longo da idade

EI total = 0.1507 | Resíduo = 0.1207

Variável	média(x)	CI(x|idade)	AME	Contrib.	% do EI
Idade (anos)	43.4833	0.2290	0.0019	0.0192	12.7286
Hipertensão	0.2285	0.4983	0.0585	0.0067	4.4230
Renda (ln)	6.8852	0.0157	0.0159	0.0017	1.1381
Autoavaliação boa	0.9403	-0.0214	-0.0678	0.0014	0.9077
Fumou diariamente	0.3481	0.1556	-0.0194	-0.0011	-0.6978
Ensino médio	0.3297	-0.1152	-0.0267	0.0010	0.6743
Atividade física (3m)	0.4333	-0.1311	0.0173	-0.0010	-0.6520
Ens. fundamental II	0.2206	-0.2565	-0.0128	0.0007	0.4791
Obesidade	0.2242	0.1082	0.0269	0.0007	0.4336
TV ≥ 3h/dia	0.1945	0.1361	0.0173	0.0005	0.3046
Plano de saúde	0.2510	0.0724	0.0117	0.0002	0.1406
Ensino superior	0.1311	-0.0071	-0.0684	0.0001	0.0425
Homem	0.4676	-0.0181	0.0020	0.0000	-0.0113
Pessoa negra	0.5753	-0.0502	-0.0002	0.0000	0.0035
Urbano	0.8664	0.0004	0.0075	0.0000	0.0017

Fonte: dados elaborados pelo autor a partir da PNS 2019

Quando a ordenação é feita pela renda domiciliar per capita, o índice de concentração de Erreygers apresenta menor magnitude (EI = 0,0484), mantendo sinal positivo, o que indica que o diagnóstico de diabetes se encontra relativamente mais concentrado entre indivíduos de maior renda. O valor reduzido do índice sugere que o gradiente socioeconômico do diabetes, quando medido pela renda, é mais moderado do que o gradiente etário, ainda que presente. Esse padrão é compatível com a literatura sobre desigualdades em saúde, que aponta a existência de gradientes socioeconômicos na distribuição de doenças crônicas e destaca o papel da renda e de outros determinantes sociais na explicação dessas disparidades (Marmot, 2012;

Hosseinpoor et al., 2012; Flor & Campos, 2017). Evidências recentes utilizando a decomposição do Índice de Concentração, como em Kumar et al. (2025), também indicam que fatores socioeconômicos, especialmente condição econômica, escolaridade e acesso a serviços, desempenham papel central na explicação do gradiente observado em desfechos relacionados ao diabetes, reforçando a adequação dessa abordagem para a análise das desigualdades em saúde.

A decomposição do índice evidencia que a própria renda exerce contribuição relevante para o nível de concentração observado, respondendo por aproximadamente 16,9% do EI total. Adicionalmente, a idade também apresenta papel expressivo, contribuindo com cerca de 8,4%, o que reforça a interação entre fatores etários e socioeconômicos na determinação da distribuição do diabetes. Por outro lado, variáveis como ensino superior apresentam contribuição negativa (−10,2%), indicando efeito atenuador sobre a concentração da doença entre os indivíduos de maior renda.

De modo geral, os resultados sugerem que o gradiente de renda no diabetes resulta da combinação de fatores diretamente associados à posição socioeconômica, como a própria renda e o acesso a serviços de saúde, e de características correlacionadas, como idade e escolaridade, que influenciam simultaneamente o risco da doença e a posição no ranking de renda.

Tabela 05 - Decomposição do EI em Renda

Decomposição do EI (diabetes) ao longo da renda

EI total = 0.0484 | Resíduo = 0.0380

Variável	média(x)	CI(x|renda)	AME	Contrib.	% do EI
Renda (ln)	6.8852	0.0748	0.0159	0.0082	16.8906
Ensino superior	0.1311	0.5513	-0.0684	-0.0049	-10.2082
Idade (anos)	43.4833	0.0486	0.0019	0.0041	8.4046
Plano de saúde	0.2510	0.4650	0.0117	0.0014	2.8113
Hipertensão	0.2285	0.0841	0.0585	0.0011	2.3238
Atividade física (3m)	0.4333	0.1299	0.0173	0.0010	2.0097
Autoavaliação boa	0.9403	0.0137	-0.0678	-0.0009	-1.8003
Urbano	0.8664	0.0569	0.0075	0.0004	0.7614
Ensino médio	0.3297	0.0312	-0.0267	-0.0003	-0.5685

Ens. fundamental II	0.2206	-0.0937	-0.0128	0.0003	0.5446
Fumou diariamente	0.3481	-0.0265	-0.0194	0.0002	0.3699
TV ≥ 3h/dia	0.1945	-0.0341	0.0173	-0.0001	-0.2378
Obesidade	0.2242	0.0139	0.0269	0.0001	0.1730
Homem	0.4676	0.0360	0.0020	0.0000	0.0702
Pessoa negra	0.5753	-0.1566	-0.0002	0.0000	0.0343

Fonte: dados elaborados pelo autor a partir da PNS 2019

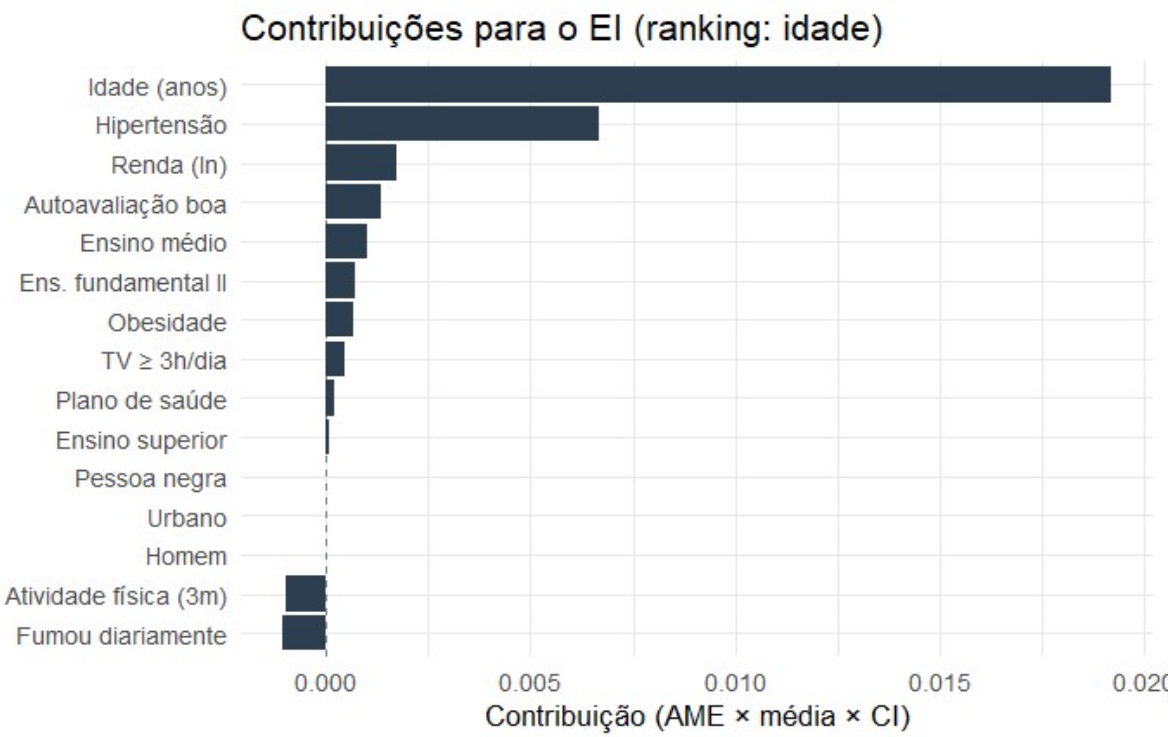
A decomposição do EI com ordenação pela idade permite identificar os fatores que contribuem para a desigualdade etária do diabetes. Conforme apresentado na Tabela 04, a própria idade responde pela maior parcela da desigualdade, contribuindo com aproximadamente 12,7% do EI total. Esse resultado é esperado, dado que a idade é simultaneamente o critério de ordenação e um determinante central do risco de diabetes.

Entre as covariáveis incluídas no modelo, a hipertensão arterial destaca-se como o principal mediador observável da desigualdade etária, respondendo por cerca de 4,4% do EI. Esse achado indica que a maior prevalência de hipertensão entre indivíduos mais velhos contribui de forma relevante para a concentração do diabetes ao longo do ciclo de vida.

As variáveis de escolaridade apresentam contribuições positivas, porém de menor magnitude. O ensino médio e o ensino superior contribuem modestamente para a desigualdade etária, refletindo a distribuição historicamente menos favorável desses níveis educacionais entre as faixas etárias mais avançadas. Outras covariáveis, como renda, autopercepção de saúde, hábitos comportamentais e características demográficas, apresentam contribuições reduzidas, algumas inclusive com sinal negativo, indicando efeitos atenuadores sobre a concentração do diabetes entre os mais velhos.

O Gráfico 05 de contribuições reforça visualmente esses resultados, evidenciando a dominância da idade e da hipertensão na explicação da desigualdade etária do diabetes.

Gráfico 05 – Contribuições para o EI (Idade)



Fonte: dados elaborados pelo autor a partir da PNS 2019

O elevado componente residual indica que parcela expressiva da desigualdade etária do diabetes decorre de fatores não observados ou de determinantes biológicos e históricos não plenamente capturados pelo conjunto de covariáveis incluído no modelo, resultado compatível com a natureza do ranking adotado.

A decomposição do EI com ordenação pela renda revela um padrão distinto. Conforme a Tabela 05, a própria renda domiciliar per capita constitui a principal contribuição positiva para a desigualdade socioeconômica do diabetes, respondendo por aproximadamente 16,9% do EI total. Esse resultado indica que a concentração do diabetes entre indivíduos de maior renda está diretamente associada à posição no ranking socioeconômico.

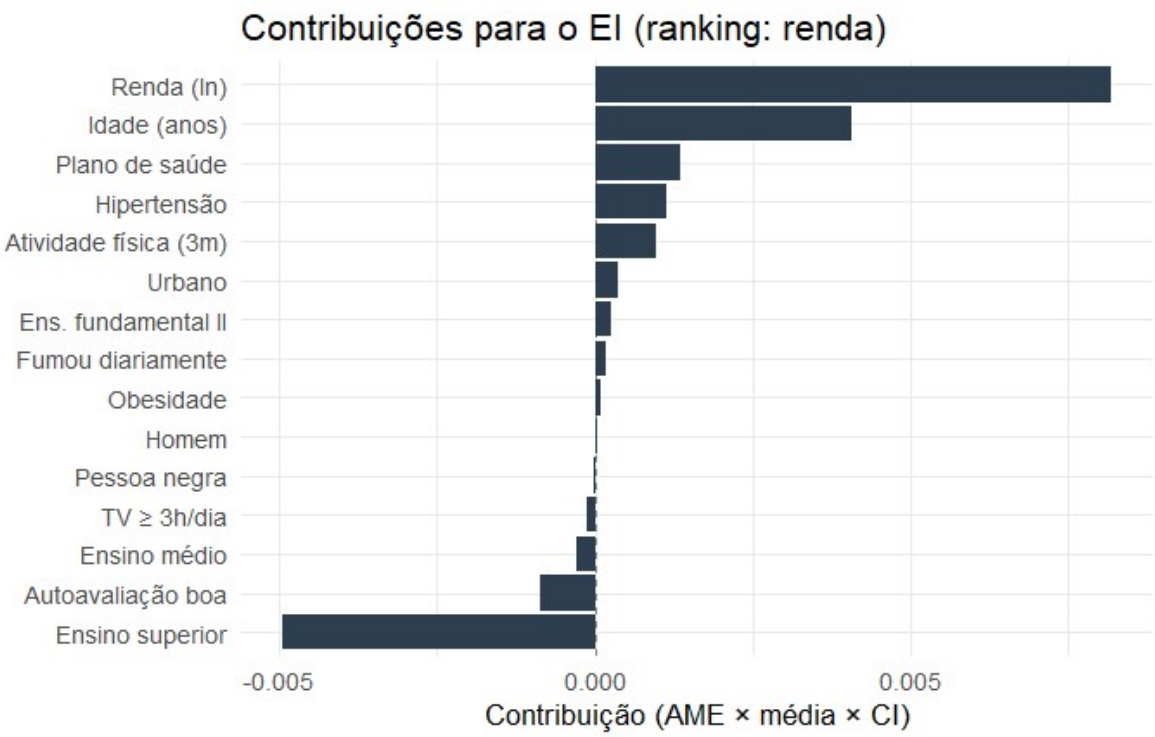
A idade também desempenha papel relevante, contribuindo com cerca de 8,4% do índice, sugerindo que diferenças etárias entre os estratos de renda ajudam a explicar parte da concentração observada. Adicionalmente, variáveis como plano de saúde (2,8%) e hipertensão (2,3%) apresentam contribuições positivas, refletindo sua maior incidência ou diagnóstico entre indivíduos de renda mais elevada.

Em contraste, o ensino superior exerce contribuição negativa expressiva (−10,2%), atuando no sentido de reduzir a desigualdade socioeconômica do diabetes. Esse resultado indica que, embora indivíduos com maior renda apresentem maior

prevalência diagnosticada, a maior escolaridade atua como fator protetivo, atenuando o gradiente socioeconômico da doença.

O Gráfico 06 ilustra essas contribuições, destacando o papel central da renda e da idade na ampliação do EI, bem como o efeito compensatório da escolaridade elevada.

Gráfico 06 – Contribuições para o EI (Renda)



Fonte: dados elaborados pelo autor a partir da PNS 2019

Assim como no caso da ordenação etária, observa-se a presença de um componente residual não desprezível, indicando que parte da desigualdade socioeconômica do diabetes decorre de fatores não incluídos no modelo ou de mecanismos mais complexos de interação entre características individuais.

Em síntese, os resultados do Índice de Concentração de Erreygers indicam que o diabetes autorreferido apresenta forte gradiente etário e um gradiente socioeconômico mais moderado quando ordenado pela renda. A decomposição do índice evidencia que a desigualdade etária está fortemente associada à própria idade e a condições clínicas relacionadas ao envelhecimento, enquanto a desigualdade socioeconômica reflete a interação entre fatores como renda, idade, acesso a serviços de saúde e escolaridade. Destaca-se ainda que a variável de ordenação exerce papel relevante na explicação do índice, em parte devido à sua associação mecânica com

o ranking utilizado. Esses achados fornecem a base empírica para a análise estrutural das diferenças entre grupos sociais apresentada na seção seguinte.

4.4 – Decomposição de Oaxaca–Blinder Não Linear

Dando seguimento, esta seção apresenta os resultados da decomposição de Oaxaca–Blinder não linear aplicada às diferenças na prevalência de diabetes autorreferido entre grupos sociais selecionados. Conforme descrito no capítulo metodológico, a análise tem por objetivo decompor o diferencial médio observado em dois componentes: (i) o componente explicado, associado às diferenças na composição das características observáveis, e (ii) o componente não explicado, associado às diferenças nos retornos dessas características, captando heterogeneidades estruturais não atribuíveis à composição dos grupos.

A decomposição foi realizada em dois contextos: baixa versus alta escolaridade e negros versus não negros, utilizando modelos logit estimados separadamente para cada grupo e combinados segundo as Regras de Rubin.

O Quadro 03 apresenta os resultados da decomposição da diferença de prevalência de diabetes entre indivíduos com baixa escolaridade (menos que ensino superior completo) e aqueles com alta escolaridade (ensino superior completo).

Observa-se uma diferença média de prevalência de −3,9 pontos percentuais, indicando que indivíduos com alta escolaridade apresentam menor probabilidade de diagnóstico de diabetes em comparação aos de baixa escolaridade. Essa diferença é estatisticamente significativa, conforme evidenciado pelo intervalo de confiança que não inclui o zero. Ademais, esse resultado também é consistente com a literatura sobre determinantes sociais da saúde, que destaca a importância da educação na formação de gradientes de saúde ao longo do ciclo de vida (Marmot, 2012; Hosseinpoor et al., 2012).

No que se refere à decomposição, o componente explicado apresenta magnitude de −1,38 ponto percentual, correspondendo a aproximadamente 35,5% da diferença total. Contudo, esse efeito não é estatisticamente significativo, uma vez que o intervalo de confiança inclui o zero. De forma semelhante, o componente não explicado, de −2,51 pontos percentuais (cerca de 64,5% do total), também não apresenta significância estatística.

Quadro 03 – Decomposição Oaxaca-Blinder - Escolaridade

Decomposição de Oaxaca–Blinder Não Linear: Baixa vs Alta Escolaridade				
Componente	**Estimativa**	**SE**	**CI_low**	**CI_high**
Diferença de prevalência (baixa - alta)	-0.0389	0.0017	-0.0422	-0.0355
Explicado (composição de X)	-0.0138	0.0227	-0.0583	0.0308
Não explicado (coeficientes/retornos)	-0.0251	0.0238	-0.0716	0.0215

Fonte: dados elaborados pelo autor a partir da PNS 2019

Esses resultados sugerem que, embora exista um diferencial educacional significativo na prevalência de diabetes, não é possível atribuir, com precisão estatística, a decomposição desse diferencial entre os componentes explicado e não explicado. Evidências empíricas baseadas em decomposições do tipo Oaxaca–Blinder aplicadas a desfechos em saúde, como em Tabrizi et al. (2025), também indicam que, embora parte relevante das desigualdades possa ser associada à distribuição de características socioeconômicas, uma fração não desprezível permanece vinculada a componentes não explicados, refletindo mecanismos estruturais não capturados pelas variáveis observáveis.

O Quadro 04 apresenta os resultados da decomposição da diferença de prevalência de diabetes autorreferido entre indivíduos negros (pretos e pardos) e não negros (brancos).

A diferença média de prevalência é de −1,16 ponto percentual, indicando menor prevalência diagnosticada entre indivíduos negros em relação aos não negros. Essa diferença é estatisticamente significativa, conforme evidenciado pelo intervalo de confiança que não inclui o zero.

No que se refere à decomposição, o componente explicado apresenta magnitude de −1,29 ponto percentual, superando ligeiramente a diferença total observada. Entretanto, esse componente não é estatisticamente significativo, dado que seu intervalo de confiança inclui o zero. De forma análoga, o componente não explicado, de 0,12 ponto percentual, também não apresenta significância estatística.

Dessa forma, embora as magnitudes estimadas sugiram que as diferenças na composição das características observáveis — como idade, escolaridade, renda e condições de saúde — possam desempenhar papel relevante no diferencial de prevalência entre negros e não negros, a ausência de significância estatística impede afirmar com precisão a contribuição relativa dos componentes explicado e não explicado.

Ainda assim, o sinal e a ordem de magnitude dos resultados permanecem compatíveis com a literatura sobre determinantes sociais da saúde, que destaca o papel das condições socioeconômicas na produção de desigualdades em saúde entre grupos populacionais (Marmot, 2012; Hosseinpoor et al., 2012), devendo-se interpretar tais evidências com cautela diante da incerteza associada às estimativas.

Quadro 04 – Decomposição Oaxaca-Blinder - Cor

Decomposição de Oaxaca–Blinder Não Linear: Negros vs Não Negros				
Componente	**Estimativa**	**SE**	**CI_low**	**CI_high**
Diferença de prevalência (negros - não negros)	-0.0116	0.0013	-0.0142	-0.0090
Explicado (composição de X)	-0.0129	0.0219	-0.0588	0.0301
Não explicado (coeficientes/retornos)	0.0012	0.0222	-0.0423	0.0448

Fonte: dados elaborados pelo autor a partir da PNS 2019

Em síntese, os resultados da decomposição de Oaxaca–Blinder não linear revelam padrões distintos de desigualdade na prevalência de diabetes segundo escolaridade e raça/cor. No caso educacional, observa-se um diferencial expressivo e estatisticamente significativo na prevalência da doença, indicando menor probabilidade de diagnóstico entre indivíduos com maior escolaridade. Contudo, a decomposição desse diferencial não permite identificar, com precisão estatística, a contribuição relativa dos componentes explicado e não explicado, uma vez que ambos apresentam elevada incerteza. De forma semelhante, no caso racial, embora o diferencial de prevalência seja menor e estatisticamente significativo, os componentes da decomposição também não se mostram estatisticamente distintos de zero. Assim, apesar das magnitudes sugerirem possíveis contribuições das características observáveis, não é possível afirmar com segurança se o diferencial decorre predominantemente de efeitos de composição ou de diferenças nos retornos. Esse padrão de resultados reflete, em parte, o maior rigor metodológico adotado na análise, que incorpora a ponderação amostral e o pooling das múltiplas imputações, implicando aumento da variância das estimativas. Como consequência, as inferências sobre a decomposição tornam-se mais conservadoras, ainda que o diferencial total permaneça bem identificado.

Esses achados complementam as evidências obtidas a partir do modelo logit e da análise de desigualdade via Índice de Concentração de Erreygers, oferecendo uma visão integrada dos mecanismos associados às disparidades sociais na ocorrência do diagnostico autorreferido de diabetes.

5. Conclusão

Os resultados deste estudo sugerem que o diabetes mellitus autorreferido no Brasil não pode ser compreendido apenas como um evento clínico isolado, mas como o resultado de processos que se acumulam ao longo do tempo e se distribuem de forma desigual entre os indivíduos. A combinação entre idade, condições de saúde pré-existentes e fatores socioeconômicos revela que o risco da doença pode ser moldado por trajetórias de vida marcadas por diferentes exposições a fatores de risco, oportunidades e restrições.

O papel central da idade na probabilidade de ocorrência do diabetes reforça essa leitura. O risco da doença tende a aumentar ao longo do ciclo de vida, não apenas pelo envelhecimento em si, mas pela maior probabilidade de convivência prolongada com condições como hipertensão e obesidade. De modo que essas condições funcionam como elos intermediários entre o tempo vivido e o diagnóstico do diabetes, ajudando a explicar por que a doença se concentra de forma tão clara entre indivíduos mais velhos. Nesse sentido, o gradiente etário observado reflete menos um efeito pontual da idade e mais o acúmulo de riscos ao longo da vida.

Já quando se observa a dimensão socioeconômica, a escolaridade emerge como um fator particularmente relevante. Indivíduos com maior nível educacional apresentam menor probabilidade de diagnóstico de diabetes mesmo quando se consideram diferenças de renda, hábitos de vida e acesso a serviços de saúde. Esse resultado sugere que a escolaridade atua como um recurso que influencia decisões, comportamentos e a forma como os indivíduos se relacionam com o sistema de saúde. Mais do que um indicador de status econômico, a educação parece refletir capacidades acumuladas ao longo da vida, como acesso à informação, compreensão de orientações médicas e maior autonomia para adotar comportamentos preventivos.

Adicionalmente, a análise das desigualdades reforça essa interpretação, pois, embora o diabetes apresente apenas um gradiente moderado quando ordenado pela renda, a decomposição do Índice de Concentração indica que a escolaridade contribui para atenuar essa desigualdade. Em outras palavras, mesmo em um contexto em que o diagnóstico tende a se concentrar entre indivíduos de maior renda, níveis educacionais mais elevados estão associados à redução dessa concentração, apontando para a relevância de fatores estruturais de longo prazo na explicação das desigualdades em saúde.

Em complemento, as diferenças observadas entre grupos educacionais tornam essa leitura ainda mais clara, ainda que a decomposição de Oaxaca–Blinder não permita identificar, com precisão estatística, a contribuição relativa dos componentes explicado e não explicado. Ainda assim, as magnitudes estimadas sugerem que parte do diferencial de prevalência de diabetes entre indivíduos com alta e baixa escolaridade pode estar associada a mecanismos não plenamente capturados pelas variáveis observáveis, relacionados a padrões cumulativos de exposição ao risco, condições de trabalho, acesso contínuo à prevenção e cuidados de saúde.

Por outro lado, no caso das diferenças raciais, os resultados apontam para uma dinâmica distinta. Embora exista um diferencial de prevalência entre negros e não negros, a decomposição não permite afirmar com precisão estatística a contribuição dos componentes explicado e não explicado. Ainda assim, as evidências sugerem que as diferenças observadas estão associadas, em grande medida, à distribuição desigual de características socioeconômicas e de acesso a serviços de saúde, indicando que a desigualdade racial no diagnóstico da doença pode operar, sobretudo, por meio de condições sociais e econômicas desiguais.

Em conjunto, os achados deste estudo indicam que o diabetes autorreferido reflete desigualdades que se constroem ao longo do tempo e se manifestam de forma diferenciada entre grupos sociais. A evidência empírica sugere que fatores estruturais, como educação e condições de vida acumuladas, desempenham papel relevante na distribuição da doença, reforçando a necessidade de interpretações que considerem o caráter histórico e social das desigualdades em saúde.

Assim, os resultados apresentados neste estudo oferecem subsídios relevantes para a reflexão sobre políticas públicas voltadas à prevenção e ao enfrentamento do diabetes mellitus no Brasil, ainda que não permitam inferências causais estritas. A principal contribuição nesse sentido está em evidenciar que a distribuição da doença reflete processos sociais e econômicos de longo prazo, que não podem ser plenamente enfrentados por intervenções focadas apenas no tratamento clínico ou na ampliação pontual do acesso aos serviços de saúde.

O forte gradiente etário observado indica que estratégias de prevenção tendem a ser mais eficazes quando pensadas ao longo do ciclo de vida. A concentração do diabetes em idades mais avançadas não decorre apenas do envelhecimento biológico, mas do acúmulo de exposições a fatores de risco, como hipertensão, obesidade e hábitos sedentários. Nesse contexto, políticas que promovam a

prevenção precoce de doenças crônicas e o acompanhamento contínuo da saúde ao longo da vida podem contribuir para reduzir a incidência futura do diabetes, mitigando seus efeitos tanto individuais quanto coletivos.

Em contrapartida, a centralidade da escolaridade nos resultados sugere que políticas educacionais exercem papel indireto, porém estrutural, sobre os desfechos de saúde. A evidência de que indivíduos mais escolarizados apresentam menor probabilidade de diagnóstico de diabetes, mesmo após o controle por renda e acesso a serviços, aponta para a educação como um fator que amplia a capacidade de prevenção, de compreensão das informações de saúde e de interação com o sistema de cuidados. Assim, políticas que elevem o nível educacional da população e reduzam desigualdades educacionais podem produzir efeitos persistentes sobre a saúde, ainda que esses efeitos se manifestem apenas no médio e longo prazos.

Os achados relacionados ao acesso aos serviços de saúde também merecem atenção, visto que a associação positiva entre posse de plano de saúde e diagnóstico de diabetes sugere que parte das desigualdades observadas pode refletir diferenças no acesso ao diagnóstico, e não necessariamente na incidência da doença. Esse resultado destaca a importância de fortalecer a atenção primária e as estratégias de rastreamento no sistema público, de modo a reduzir desigualdades no diagnóstico precoce e evitar que grupos socialmente mais vulneráveis permaneçam subdiagnosticados.

No que se refere às desigualdades raciais, os resultados sugerem que as diferenças na prevalência do diabetes estão associadas, em parte, à distribuição desigual de características socioeconômicas e de acesso a serviços. Ainda que a decomposição não permita identificar com precisão estatística a contribuição dos diferentes componentes, as evidências apontam para a relevância de fatores estruturais na compreensão dessas disparidades. Nesse sentido, intervenções voltadas à redução das desigualdades de renda, escolaridade e acesso a serviços tendem a ter efeitos indiretos relevantes sobre as desigualdades raciais em saúde.

Em conjunto, os resultados deste estudo sugerem que políticas públicas voltadas ao enfrentamento do diabetes tendem a ser mais efetivas quando articuladas a estratégias mais amplas de redução das desigualdades sociais. Ao evidenciar os canais pelos quais fatores socioeconômicos se relacionam com a distribuição da doença, a análise empírica oferece elementos que podem orientar o desenho de políticas de prevenção e promoção da saúde, respeitando os limites da evidência

observacional, mas reforçando a importância de abordagens estruturais e de longo prazo.

Ademais, este trabalho contribui para a literatura ao oferecer uma análise integrada dos determinantes socioeconômicos do diabetes autorreferido no Brasil a partir de diferentes abordagens empíricas complementares. Enquanto grande parte dos estudos nacionais se concentra na identificação de associações por meio de modelos de regressão, esta dissertação combina a estimação de modelos logit com a análise de desigualdades via Índice de Concentração de Erreygers e técnicas de decomposição, permitindo não apenas identificar fatores associados ao diagnóstico da doença, mas também analisar como esses fatores se relacionam com as disparidades observadas entre grupos populacionais. Ao articular essas abordagens, o estudo amplia a compreensão dos mecanismos por meio dos quais características demográficas, socioeconômicas e de saúde se associam à distribuição do diabetes na população brasileira.

Em contrapartida, apesar das contribuições analíticas e empíricas apresentadas, este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. A principal delas diz respeito à natureza observacional e transversal dos dados utilizados. A Pesquisa Nacional de Saúde de 2019 oferece um retrato abrangente da população brasileira em um determinado momento do tempo, mas não permite acompanhar trajetórias individuais nem estabelecer relações causais estritas entre os determinantes socioeconômicos e a ocorrência do diagnóstico de diabetes mellitus.

Outra limitação relevante refere-se à utilização do diagnóstico autorreferido de diabetes como variável dependente. Embora essa medida seja amplamente empregada na literatura e apresente boa validade para análises populacionais, ela pode refletir desigualdades no acesso ao diagnóstico médico, especialmente em contextos marcados por heterogeneidade socioeconômica. Nesse sentido, parte das associações observadas pode estar relacionada a diferenças na probabilidade de diagnóstico, e não necessariamente à incidência da doença em si. Complementando, cabe destacar que a variável dependente utilizada se refere ao diagnóstico autorreferido de diabetes, sem distinção entre os diferentes tipos da doença. Embora a maior parte dos casos em adultos esteja associada ao diabetes mellitus tipo 2 — que corresponde a aproximadamente 95% dos casos globais (WHO, 2023) —, a impossibilidade de distinguir entre os tipos constitui uma limitação da análise.

Adicionalmente, apesar do uso de técnicas robustas de tratamento de dados faltantes, como a imputação múltipla, é possível que permaneçam fontes de incerteza associadas a informações não observadas ou imperfeitamente mensuradas. Variáveis como histórico familiar, qualidade da alimentação e condições de trabalho não estão plenamente capturadas nos microdados da PNS, mas podem desempenhar papel relevante na determinação do risco de diabetes ao longo do ciclo de vida.

Essas limitações, contudo, não invalidam os achados do estudo, mas delimitam seu alcance interpretativo. Ao contrário, elas apontam caminhos naturais para pesquisas futuras. Estudos baseados em dados longitudinais poderiam explorar de forma mais direta os mecanismos dinâmicos associados ao desenvolvimento do diabetes, permitindo avaliar a acumulação de riscos ao longo do tempo. Da mesma forma, a integração de informações clínicas ou biomarcadores objetivos poderia contribuir para distinguir de maneira mais precisa entre desigualdades no diagnóstico e desigualdades na ocorrência efetiva da doença.

Por fim, investigações futuras podem aprofundar a análise das desigualdades educacionais e raciais em saúde, explorando interações mais complexas entre fatores socioeconômicos, contextos institucionais e trajetórias individuais. Ao avançar nessas direções, a literatura poderá contribuir para uma compreensão mais refinada dos determinantes do diabetes e para o desenho de políticas públicas mais eficazes e sensíveis às desigualdades sociais.

Referências

BENNETT, J. E. et al. **NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4**. The Lancet, v. 392, n. 10152, p. 1072–1088, set. 2018. Disponível em: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31992-5/fulltext. Acesso em: 21 abr. 2025

BLINDER, Alan S. **Wage Discrimination: Reduced Form and Structural Estimates**. The Journal of Human Resources, v. 8, n. 4, p. 436, 1973.

DONDERS, A. Rogier T. *et al.* **Review: A gentle introduction to imputation of missing values**. Journal of Clinical Epidemiology, v. 59, n. 10, p. 1087–1091, out. 2006.

ERREYGERS, Guido. **Correcting the Concentration Index**. Journal of Health Economics, v. 28, n. 2, p. 504–515, mar. 2009.

FAIRLIE, Robert W. **An extension of the Blinder-Oaxaca decomposition technique to logit and probit models**. Journal of Economic and Social Measurement, v. 30, n. 4, p. 305–316, 1 nov. 2005.

FLOR, Luisa Sorio; CAMPOS, Monica Rodrigues. **Prevalência de diabetes mellitus e fatores associados na população adulta brasileira: evidências de um inquérito de base populacional**. Revista Brasileira de Epidemiologia, v. 20, n. 1, p. 16–29, mar. 2017.

GARCES, Thiago Santos et al. **Relação entre indicadores de desenvolvimento social e mortalidade por Diabetes Mellitus no Brasil: análise espacial e temporal**. *Revista Latino-Americana de Enfermagem*, v. 31, p. e3971, 2023.

GREENE, W. H. **Econometric Analysis**. 5ª. ed. Upper Saddle River, N.J.: Prentice Hall, 2003. p. 162–338

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica**. 5. ed. Porto Alegre: AMGH, 2011. 924 p. 534-609

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied Logistic Regression**. *[S.l.]*: Wiley, 2013.

HOSSEINPOOR, Ahmad Reza *et al.* **Socioeconomic inequality in the prevalence of noncommunicable diseases in low- and middle-income countries: Results from the World Health Survey**. BMC Public Health, v. 12, n. 1, p. 474, 22 dez. 2012.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Pesquisa Nacional de Saúde: o que é**. [s.d.]. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=&t=o-que-e. Acesso em: 12 jan. 2026.

KJELLSSON, Gustav; GERDTHAM, Ulf-G. **On correcting the concentration index for binary variables**. Journal of Health Economics, v. 32, n. 3, p. 659–670, maio 2013.

KLUTHCOVSKY, Ana Claudia Garabeli Cavalli; BERALDO, Maria Luiza Julinhaque**. Análise de tendência e diferenciais socioeconômicos na prevalência de diabetes autorreferido em um inquérito de base populacional**. *Cadernos de Educación y Desarrollo*, v. 16, n. 6, p. e4537, 2024.

KUMAR, A. et al. **Decomposing socioeconomic inequality in lean diabetes among middle-aged adults and elderly in India**. , 20 fev. 2025.

KUNDU, Satyajit *et al.* **Diabetes, Hypertension, and Comorbidity among Bangladeshi Adults: Associated Factors and Socio-Economic Inequalities**. Journal of Cardiovascular Development and Disease, v. 10, n. 1, p. 7, 23 dez. 2022.

MCFADDEN, Daniel. **Quantitative methods for analysing travel behaviour of individuals: some recent developments**. In: HENSHER, David A.; STOPHER, Peter R. (ed.). Behavioural travel modelling. London: Croom Helm, 1978. p. 279–318.

MARMOT, M.; BELL, R. **Fair society, healthy lives. Public Health**, v. 126, p. S4–S10, set. 2012.

OAXACA, Ronald. **Male-Female Wage Differentials in Urban Labor Markets**. International Economic Review, v. 14, n. 3, p. 693, out. 1973.

OAXACA, R. L.; RANSOM, M. R. **On discrimination and the decomposition of wage differentials**. Journal of Econometrics, v. 61, n. 1, p. 5–21, mar. 1994.

O'DONNELL, Owen *et al.* **Analyzing Health Equity Using Household Survey Data**. *[S.l.]*: The World Bank, 2007.

PEDUZZI, Peter *et al.* **A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology**, v. 49, n. 12, p. 1373–1379, dez. 1996.

RUBIN, Donald B. **Multiple Imputation for Nonresponse in Surveys**. *[S.l.]*: Wiley, 1987.

SAFIEDDINE, Batoul *et al.* **Socioeconomic inequalities in type 2 diabetes comorbidities in different population subgroups: trend analyses using German health insurance data**. Scientific Reports, v. 13, n. 1, p. 10855, 5 jul. 2023.

SHARMA, Santosh Kumar *et al.* **Decomposing socioeconomic inequality in blood pressure and blood glucose testing: evidence from four districts in Kerala, India**. International Journal for Equity in Health, v. 21, n. 1, p. 128, 9 set. 2022.

TABRIZI, Reza et al. **Socioeconomic inequality in hypertension and its determinants in people over 60 years in Fasa, southern Iran: a Blinder–Oaxaca decomposition**. *BMC Public Health*, v. 25, n. 1, p. 274, 2025.

VAN BUUREN, Stef. **Flexible Imputation of Missing Data**, Second Edition. Second edition. | Boca Raton, Florida: CRC Press, [2019] |: Chapman and Hall/CRC, 2018.

VITTINGHOFF, E.; MCCULLOCH, C. E. **Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression**. American Journal of Epidemiology, v. 165, n. 6, p. 710–718, 12 jan. 2007.

VON HIPPEL, Paul T. 4. **Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data**. Sociological Methodology, v. 37, n. 1, p. 83–117, 1 ago. 2007.

WORLD HEALTH ORGANIZATION. **Diabetes**. Geneva: WHO, 2024. Disponível em: https://www.who.int/news-room/fact-sheets/detail/diabetes. Acesso em: 21 out. 2025.

WORLD HEALTH ORGANIZATION. **Invisible numbers: the true extent of noncommunicable diseases and what to do about them.** Geneva: WHO, 2022. 42 p. Disponível em: https://www.who.int/publications/i/item/9789240057661. Acesso em: 21 abr. 2025

YUN, Myeong-Su. **Decomposing differences in the first moment**. Economics Letters, v. 82, n. 2, p. 275–280, fev. 2004.

Apêndice A – Tratamento e Construção da Base de Dados

O objetivo deste apêndice é apresentar, de forma sistemática e detalhada, os procedimentos adotados no tratamento da base de dados utilizada nesta pesquisa. Busca-se, com isso, garantir clareza metodológica, transparência nas decisões empíricas e reprodutibilidade dos resultados, permitindo que outros pesquisadores compreendam exatamente como a amostra final e as variáveis analíticas foram construídas a partir dos microdados originais.

A seguir, descrevem-se, passo a passo, os filtros aplicados, as recodificações realizadas, as exclusões adotadas e os métodos utilizados para lidar com valores ausentes e construção das variáveis explicativas e da variável de interesse.

A1 – Filtros Iniciais da Amostra

A base foi inicialmente restrita às entrevistas efetivamente realizadas, mantendo-se apenas as observações com V0015 = 1. Em seguida, foram selecionados exclusivamente os indivíduos com medidas antropométricas aferidas, identificados por V0025B = 1.

Foram mantidas apenas as respostas fornecidas pelo morador selecionado (M001 = 1). Adicionalmente, indivíduos gestantes (P005 = 1) e pessoas consideradas não aptas a responder (M00203 = 2) foram excluídos da amostra.

A2 – Construção e Recodificação das Variáveis de Escolaridade

A escolaridade foi tratada de forma cuidadosa devido à sua relevância socioeconômica e à presença de valores ausentes. A variável original D00901, que informa a etapa de ensino mais alta frequentada, foi combinada com D00301, que indica a etapa de ensino cursada no momento da entrevista, nos casos em que D00901 estava em branco. Dessa combinação resultou a variável consolidada D00901', representando a etapa máxima de ensino do indivíduo.

Posteriormente, os códigos originais foram agregados em cinco grandes grupos: (i) ensino fundamental I ou menos, (ii) ensino fundamental II, (iii) ensino médio, (iv) ensino superior – graduação, e (v) pós-graduação. Essa classificação respeita a equivalência entre sistemas educacionais antigos e atuais e facilita a interpretação analítica.

Devido à presença de valores ausentes relevantes, aplicou-se imputação múltipla por equações encadeadas (MICE) às variáveis de escolaridade.

Primeiramente, foi imputada a variável ordinal referente à etapa de ensino (esc_serie), sem utilizar como preditor a variável de conclusão. Em seguida, foi imputada a variável binária esc_concluiu (D014), incorporando a etapa de ensino como variável explicativa, respeitando a lógica temporal e causal entre nível cursado e conclusão.

Após a imputação, utilizou-se a informação de conclusão para regredir o indivíduo em um nível educacional quando a etapa não havia sido concluída, exceto nos casos em que o indivíduo já se encontrava no último nível. Em seguida, as categorias de graduação e pós-graduação foram agrupadas em uma única categoria denominada grad_mais. A variável final de escolaridade foi desdobrada em três dummies — Ensino Fundamental II, Ensino Médio e Graduação ou mais — adotando-se Ensino Fundamental I ou menos como categoria de referência nos modelos logit.

A3 – Construção das Variáveis de Saúde

A variável de hipertensão arterial foi construída a partir de Q00201 e Q00202, excluindo-se os casos em que o diagnóstico ocorreu apenas durante a gravidez. Procedimento análogo foi adotado para a variável de diabetes mellitus, combinando Q03001 e Q03002, resultando na variável Q03001'.

A variável Q03001' apresentou aproximadamente 7% de valores ausentes. A análise dos padrões de missing indicou que a ausência estava associada a características observáveis, como idade, sexo, condição urbano/rural e escolaridade. Diante disso, optou-se por imputar essa variável por meio de MICE, incorporando-a ao sistema de equações de imputação.

A autoavaliação de saúde (N001) foi recodificada em uma variável binária, distinguindo indivíduos com avaliação satisfatória (muito boa, boa ou regular) daqueles com avaliação insatisfatória (ruim ou muito ruim).

A4 – Construção de Variáveis Antropométricas e Comportamentais

A partir das variáveis de peso (P00104) e altura (P00404), foi calculado o Índice de Massa Corporal (IMC). Com base nesse indicador, construiu-se a variável binária OB001, que identifica indivíduos obesos.

O tempo médio diário assistindo televisão (P04501) foi recodificado em uma variável binária que indica se o indivíduo assiste três ou mais horas de TV por dia.

O consumo de bebida alcoólica (P027) foi transformado em uma variável binária que identifica indivíduos que consomem álcool mais de uma vez ao mês.

O tabagismo foi tratado a partir das variáveis P050 e P052, resultando na variável P050', que indica se o indivíduo fuma ou já fumou diariamente.

A5 – Variáveis Socioeconômicas Adicionais

A renda domiciliar per capita (VDF003) foi transformada por meio do logaritmo natural, originando a variável VDF003'.

O estado civil (C011) foi recodificado em uma variável binária que indica se o indivíduo é casado ou não.

A variável de cor/raça (C009) foi tratada excluindo-se as categorias amarela e indígena. Em seguida, construiu-se a variável binária Negros, na qual indivíduos pretos e pardos foram classificados como 1 e indivíduos brancos como 0.

A6 – Tratamento de Valores Ausentes e Transformações Finais

Nos casos em que a proporção de valores ausentes em determinadas variáveis era inferior a 2,15% do total da amostra, aplicou-se remoção por lista completa (listwise deletion).

Para captar possíveis efeitos não lineares da idade, foi incluído o termo idade ao quadrado. A idade foi previamente centralizada em torno da média da amostra, e tanto o termo linear centralizado quanto seu quadrado foram utilizados nas estimações.

Esses procedimentos resultaram na base final empregada nas análises empíricas deste trabalho.