UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

Trabalho de Conclusão de Curso

Cálculo dos anos de vida ajustados pela produtividade (PALY) para estimar o custo do Diabetes na população brasileira

RODOLFO ARNALDO MONTECINOS DE ALMEIDA

Porto Alegre
11 de abril de 2023

Rodolfo Arnaldo Montecinos de Almeida

Cálculo dos anos de vida ajustados pela produtividade (PALY) para estimar o custo do Diabetes na população brasileira

Trabalho de Conclusão submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Orientador(a): Prof[a]. Dr[a] Paula Andreghetto Bracco

Porto Alegre

Abril de 2023

Instituto de Matemática e Estatística
Departamento de Estatística

Modelo de estimação da perda da produtividade causado por Diabetes na população brasileira

Rodolfo Arnaldo Montecinos de Almeida

Banca examinadora:

Prof. Dr. Rodrigo Citton Padilha dos Reis
UFRGS

Porto Alegre
Abril de 2023

AGRADECIMENTOS

Gostaria de expressar minha sincera gratidão à minha família, em especial à minha esposa Fernanda, por todo o carinho, compreensão e paciência que me concederam durante este curso. Sem o apoio e a ajuda de vocês, tenho certeza de que teria sido muito mais desgastante e difícil. À minha filha Manuela, por todos os momentos de distração que me proporcionou e que me permitiram esquecer todas as preocupações.

Aos “Guris”, por todo suporte durante o curso, principalmente durante a pandemia. Sem as nossas conversas e mensagens lembrando dos prazos, acredito que não estaria apresentando o TCC hoje.

Agradeço especialmente à minha orientadora Profª. Paula Bracco, que aceitou o desafio de me orientar, mesmo não me conhecendo. Muito obrigado por toda ajuda, conversas e ensinamentos. Sem a sua orientação, paciência e empenho, certamente não teria sido possível concluir este trabalho.

Por fim, gostaria de agradecer a todos aqueles que “prejudicaram” o andamento do curso e deste trabalho. Acreditem ou não, sem vocês, a vida seria muito mais chata.

“Lasciate ogne speranza, voi ch'intrate”

“Deixai toda esperança, Vós que entrais”

Dante Alighieri – Divina Comédia

RESUMO

O diabetes mellitus é uma das doenças mundialmente mais prevalentes em adultos e está entre as principais causas de perda de anos de vida saudável, o que se agrava com o acelerado envelhecimento populacional no Brasil. Este estudo visa estimar os valores de presenteísmo e absenteísmo devido ao diabetes a partir dos dados da Pesquisa Nacional de Saúde (PNS) de 2019, utilizando-se um modelo linear considerando o plano de amostragem complexa. Combinando esses dados com a prevalência do diabetes e taxa de mortalidade obtidos em outros estudos, estimou-se a diferença entre o PALY (anos de vida ajustados pela produtividade) dos indivíduos com e sem diabetes. Calculou-se então a consequente redução no produto interno bruto (PIB). A soma do presenteísmo e absenteísmo resultou em uma redução de 16% na produtividade de indivíduos com diabetes quando comparado com aqueles sem diabetes. Estimamos de forma conservadora que na coorte de brasileiros de 2019 com idade entre 20 e 60 ou 65 anos, o diabetes será o responsável por uma perda no PIB de aproximadamente 113 bilhões de dólares, em um cenário com redução anual do PALY de 3% e sem aumento anual do PIB de 1.3%. Essa perda representa em torno de 7% do PIB brasileiro de 2021. Com o uso dos dados provenientes da PNS conseguiu-se, de forma inédita, construir um modelo para estimação dos anos de vida produtivos perdidos devido ao diabetes na população brasileira, enriquecendo as estimativas da carga da doença no Brasil e consequentemente nos países de baixa e média renda.

Palavras-chave: PNS, PALY, modelos lineares, amostragem complexa, Diabetes.

ABSTRACT

Diabetes mellitus is one of the most prevalent diseases worldwide in adults and is among the leading causes of loss of healthy years of life, which is exacerbated by the accelerated population aging in Brazil. This study aims to estimate the values of presenteeism and absenteeism due to diabetes from the data of the National Health Survey (PNS) of 2019, using a linear model considering the complex sampling plan. Combining these data with the prevalence of diabetes and mortality rate obtained in other studies, we estimated the difference between PALY (years of life adjusted for productivity) of individuals with and without diabetes. Then the reduction in GDP was calculated. The sum of presenteeism and absenteeism resulted in a 16% reduction in the productivity of individuals with diabetes compared to those without diabetes. It was conservatively estimated that in the cohort of Brazilians in 2019 aged between 20 and 60 or 65 years, diabetes would be responsible for a loss in GDP of approximately 113 billion dollars, in a scenario with an annual reduction of PALY of 3% and without an annual increase in GDP of 1.3%. This loss represents around 7% of Brazil's GDP in 2021. Using data from the PNS, we were able to develop, for the first time, a model for estimating productive life years lost due to diabetes in the Brazilian population, enriching estimates of the disease burden in Brazil and consequently in low- and middle-income countries.

Keywords: PNS, PALY, linear models, complex, Diabetes.

LISTA DE FIGURAS

Figura 1: Exemplo de amostragem por conglomerado (OCHOA, 2015a). 7
Figura 2: Exemplo de amostragem estratificada (OCHOA, 2015b) 8
Figura 3: P-P plot de resíduos (RODRIGUES, 2022).. 16
Figura 4: Gráfico de resíduos versus valores ajustados (RODRIGUES, 2022).. 16
Figura 5: Resumo do cálculo do PALY. 28
Figura 6: Criação da variável Labor. 31
Figura 7: Diagrama com a criação da variável "Dias". 32
Figura 8: Diagrama de criação da variável Limitpct. 36
Figura 9: Histograma dos resíduos para o modelo de presenteísmo para o sexo feminino. 44
Figura 10: a) Gráfico dos resíduos pelo valor estimado; b) resíduos padronizados para cada elemento. 44

LISTA DE TABELAS

Tabela 1: Forma abreviada da tabela de vida para dos Estado Unidos da América - EUA (DAY; REYNOLDS; KUSH, 2015) 25
Tabela 2: Prevalência do diabetes 41
Tabela 3:Distribuição da variável resposta Days. 42
Tabela 4:Resultado do absenteísmo para o sexo feminino. 42
Tabela 5:Distribuição da variável resposta Days. 43
Tabela 6: Resultado do absenteísmo para o sexo masculino. 43
Tabela 7: Distribuição da variável Limitpct para o sexo feminino. 44
Tabela 8: Resultado do presenteísmo para o sexo feminino. 44
Tabela 9: Distribuição da variável Limitpct para o sexo masculino. 45
Tabela 10: Resultado do presenteísmo para o sexo masculino. 45
Tabela 11: P_{index} para ambos os sexos, dos indivíduos com e sem diabetes. 46
Tabela 12: Anos de vida vividos. 47
Tabela 13: Cálculo do PIB utilizando o PALY sem desconto e com aumento anual do PIB de 1,3%, para o sexo feminino. 49
Tabela 14: Cálculo da redução do PIB utilizando o PALY com desconto anual de 3% e com aumento anual do PIB de 1,3%, para o sexo feminino. 50
Tabela 15: Cálculo do PIB utilizando o PALY sem desconto e com aumento anual do PIB de 1,3%, para o sexo masculino 50
Tabela 16: Cálculo do PIB utilizando o PALY com desconto anual de 3% e com aumento anual do PIB de 1,3%, para o sexo masculino. 51
Tabela 17: Cálculo do PIB utilizando o PALY sem desconto sem aumento anual do PIB de 1,3%, para o sexo feminino. 51
Tabela 18: Cálculo do PIB utilizando o PALY com desconto anual de 3% sem aumento anual do PIB de 1,3%, para o sexo feminino. 52
Tabela 19: Cálculo do PIB utilizando o PALY com desconto sem aumento do PIB de 1,3%, para o sexo masculino. 52
Tabela 20: Cálculo do PIB utilizando o PALY com desconto sem aumento do PIB de 1,3%, para o sexo masculino. 53

Tabela 21: Resumo dos PIB's para todos os casos discutidos. 53

Sumário

1 Introdução 1
2 Revisão Bibliográfica 3
2.1 Diabetes Mellitus 3
2.1.1 Diabetes Tipo I – DM1 3
2.1.2 Diabetes Tipo 2 – DM2 4
2.1.3 Diabetes Gestacional - DMG 4
2.1.4 Complicações do Diabetes 4
2.2 Carga do diabetes no Brasil 5
2.3 Bases de dados disponíveis 6
2.3.1 Pesquisa Nacional de Saúde – PNS 6
2.3.2 Estudo Longitudinal de Saúde do Adulto – ELSA-Brasil 6
2.3.3 Instituto Brasileiro de Geografia e Estatística - IBGE 7
2.4 Amostragem complexa 7
2.4.1 Amostragem por Conglomerados 7
2.4.2 Amostragem Estratificada 8
2.4.3 Amostragem em três estágios 9
2.5 Modelos de Regressão Linear 13
2.5.1 Regressão Linear Múltipla 13
2.5.2 Interações 14
2.5.3 Pressupostos 14
2.5.4 Estimação dos Parâmetros do modelo 14
2.5.5 Análise de Variância 15
2.5.6 Análise de Resíduos 16
2.6 Regressão Linear em Amostras Complexas 20
2.6.1 Diagnósticos da Regressão 21

2.7 Efeitos Marginais 21
2.7.1 Modelo de regressão linear: 22
2.8 Pacote Survey 22
2.8.1 Svydesing 23
2.8.2 Survey-weighted generalised linear models – svyglm 23
2.8.3 Svypredmeans 23
2.8.4 Diagnóstico do Modelo 24
2.9 Tabela de Vida 24
2.9.1 Tipos de tabela de vida 26
2.10 Taxa de Mortalidade 27
2.11 PALY – Anos de Vida Ajustado pela Produtividade 28
3 Métodos 30
3.1 Coleta dos dados 30
3.2 Prevalência do Diabetes e Mortalidade 30
3.3 Índice de Produtividade 31
3.3.1 Absenteísmo 32
3.3.2 Presenteísmo 36
3.3.3 Cálculo do Índice de Produtividade 38
3.3.4 PALY - Anos de vida Ajustados pela Produtividade 39
4 Resultados e discussão 41
4.1 Absenteísmo 42
4.1.1 Indivíduos do Sexo Feminino 42
4.1.2 Indivíduo do Sexo Masculino 43
4.2 Presenteísmo 44
4.2.1 Indivíduos do Sexo Feminino 44
4.2.2 Indivíduos do Sexo Maculino 45

4.3 Análise dos resíduos 46
4.5 Custo do Diabetes 49
5 Discussão 55
6 Conclusões 58
7 Trabalhos futuros 59
8 Referências 60

1 Introdução

Segundo o *International Diabetes Federation (*IDF*),* em 2021 a América do Sul e central tinha aproximadamente 33 milhões de pessoas entre 20 e 79 anos diagnosticadas com diabetes, resultando em uma prevalência de 9.5%. O Brasil é o país com o maior número de pessoas diagnosticadas na região - 15.7 milhões de pessoas, representando uma prevalência de 10.5% (INTERNATIONAL DIABETES FEDERATION, 2021). Esse valor é similar ao estimado pela pesquisa de vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico (VIGITEL) do Ministério da Súde, onde o percentual de adultos (>18 anos) brasileiros que referiram diagnóstico médico de diabetes em 2021 foi de 9.1% (MINISTÉRIO DA SAÚDE, SECRETARIA DE VIGILÂNCIA EM SAÚDE, DEPARTAMENTO DE, 2021). Considerando que em 2017 a prevalência estimada do diabetes no Brasil foi de 8.1%, as estimativas de 2021 indicam um crescimento de 12% em quatro anos e realçam o fato de que o Brasil vem sofrendo um aumento considerável na carga do diabetes (BRACCO, 2019).

Usualmente a carga de diabetes no Brasil é caracterizada pelas estimativas de prevalência inferida através dos dados do VIGITEL ou através da Pesquisa Nacional de Saúde (PNS) (IBGE, 2020; MINISTÉRIO DA SAÚDE, SECRETARIA DE VIGILÂNCIA EM SAÚDE, DEPARTAMENTO DE, 2021), e pela mortalidade observada em certificados de óbitos, os quais apresentam um série de limitações na identificação do diabetes (SAYDAH et al., 2004).

Nos últimos anos, estudos têm sido realizados para a utilização de banco de dados nacionais como PNS (Pesquisa Nacional de Saúde), SIM (Sistema de Informação de Mortalidade) (“Sistema de Informação sobre Mortalidade – SIM”, 2022) e SIH (Sistema de Informação de Hospitalização) (MINISTÉRIO DA SAÚDE; SECRETARIA DE ATENÇÃO À SAÚDE; DEPARTAMENTO DE REGULAÇÃO, AVALIAÇÃO E CONTROLE, 2006) , além de dados provenientes de estudos longitudinais com abrangência nacional, como é o caso do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil), para obter inferências mais complexas, como a proporção de mortalidade nacional devido ao diabetes (BRACCO et al., 2020) , o risco ao longo da vida de desenvolver diabetes e os anos de vida perdidos devido ao diabetes (BRACCO et al., 2021). Ainda há poucos estudos de representatividade nacional para estimativas do custo econômico do diabetes na população brasileira. Em 2016 foi estimado em aproximadamente 2,15 bilhões de dólares o custo direto e indireto das hospitalizações decorrentes do diabetes no brasil (PEREDA et al., 2022). Há, no

entanto, uma metodologia diferente que vêm sendo utilizada para estimar o custo indireto do diabetes a partir da perda de produtividade devido a doença, representada pela diferença de PALYs (Productivity-Adjusted Life-Years) (MAGLIANO et al., 2018) entre os indivíduos com e sem diabetes. Essa estimativa tem como vantagem considerar todos os trabalhadores diagnosticados com diabetes e não apenas os indivíduos hospitalizados.

Para a estimação da diferença de PALYs consideramos que uma doença crônica como o diabetes pode levar a perda de produtividade de duas formas:

- dias de trabalho perdido devido à doença - absenteísmo,
- redução da produtividade no trabalho devido à doença - presenteísmo

Com esses dois elementos e os anos de vida populacional, consegue-se encontrar a diferença de PALY entre aqueles com e sem diabetes.

Para a obtenção dessas duas métricas: absenteísmo e presenteísmo, os bancos de dados nacionais podem ser utilizados de uma forma análoga ao que foi realizado para obter a proporção de mortalidade nacional devido ao diabetes e os anos de vida perdidos devido ao diabetes. Porém um estudo dos questionários se faz necessário para selecionar os dados pertinentes, além de adaptações de modelos e métodos a fim de estimar os componentes do índice de produtividade.

O objetivo desse trabalho é estimar o PALY utilizando os bancos de dados nacionais (PNS e IBGE), para estimar o absenteísmo e o presenteísmo, além de estimar o PALY para os indivíduos com e sem diabetes e estimar o custo da produtividade devido à doença na população brasileira.

2 Revisão Bibliográfica

2.1 Diabetes Mellitus

De acordo com a Sociedade Brasileira de Diabetes:

> *Diabetes mellitus* (DM) não é uma doença, mas um grupo heterogêneo de distúrbios metabólicos que apresenta em comum a hiperglicemia, resultada de defeitos na ação da insulina, na secreção de insulina ou ambas (SBD, 2015).

A insulina é um hormônio essencial produzido pelo pâncreas. Ela é responsável por converter a glicose em energia ou armazená-la nas células, e pela síntese de proteínas e gordura. O indicador clínico da diabete é a falta de insulina, ou quando as células não respondem a ela, o que leva a hiperglicemia (altos níveis de glicose no sangue) (INTERNATIONAL DIABETES FEDERATION, 2021).

Se a falta de insulina no sangue não for tratada por um longo período, diversos problemas podem ocorrer, desde complicações que deixam sequelas até problemas que causam risco de vida, como doenças cardiovasculares, danos aos nervos, danos aos rins, amputações de membros inferiores, perda parcial da visão e até cegueira (INTERNATIONAL DIABETES FEDERATION, 2021).

A DM normalmente é classificada em diabetes tipo 1 (DM1), diabetes tipo 2 (DM2), diabetes gestacional (DMG), além de outros tipos de diabetes mais específicos. Outras classificações têm sido propostas, incluindo classificação em subtipos de DM levando em conta características clínicas (RODACKI et al., 2022). A classificação da DM permite o tratamento adequado e a definição de estratégias de rastreamento de comorbidades e complicações crônicas (SILLER et al., 2020).

2.1.1 Diabetes Tipo I – DM1

Esse tipo de diabetes é caracterizado por um processo autoimune no qual o sistema imunológico ataca as células produtoras de insulina do pâncreas. Com isso o corpo produz pouca ou não produz insulina. Essa condição pode ser desenvolvida em qualquer idade, no entanto ocorre mais frequentemente em crianças e em adultos jovens. Diabetes tipo I é uma das doenças crônicas mais comuns na infância (SUN et al., 2022).

Além de necessitar a injeção diária de insulina, viver com DM1 é um desafio até para os familiares, mesmo tendo acesso às injeções diárias, monitoramento de glicose e médicos especializados. Isso porque, além das complicações causadas pela hipoglicemia e da cetoacidose diabética, um controle mal feito pode levar ao crescimento deficiente e ao início precoce de complicações circulatórias(INTERNATIONAL DIABETES FEDERATION, 2021).

2.1.2 Diabetes Tipo 2 – DM2

A DM2 é o tipo mais comum, correspondendo a 90% dos casos no mundo, e será o foco deste trabalho. Quando começa a haver resistência à insulina, inicialmente ocorre um aumento na produção de insulina para manter a homeostase da glicose, no entanto, com o passar do tempo ocorre a diminuição da produção da insulina. Ela é normalmente observada em pessoas com 45 anos ou mais, porém os casos em crianças, adolescentes e adultos jovens vem crescendo devido ao aumento dos casos de obesidade, falta de atividade física e dieta desbalanceada (GOYAL; JIALAL, 2022).

Por apresentar sintomas mais brandos que a DM1, o diagnóstico é normalmente mais tardio, quando já existem evidências de complicações. Porém mudanças de comportamento, como melhora na dieta, aumento da atividade física e o uso de medicamentos hipoglicêmicos podem auxiliar no tratamento, não sendo necessário o uso contínuo de insulina na maioria das vezes (BRACCO, 2019).

2.1.3 Diabetes Gestacional - DMG

De acordo com a World Health Organization (WHO) e a Federação Internacional de Ginecologia e Obstetrícia, a hiperglicemia durante a gravides é classificada como diabetes gestacional (HOD et al., 2015; WORLD HEALTH ORGANIZATION, 2013). Mulheres com essa condição tendem a ter um aumento no risco de eventos adversos da gravidez, além de, junto com o filho, apresentar maior risco de desenvolver diabetes tipo 2 ao longo da vida (INTERNATIONAL DIABETES FEDERATION, 2021).

2.1.4 Complicações do Diabetes

Como condição crônica o diabetes exige da pessoa o seguimento de um regime terapêutico de autocuidado diário, que é necessário para a manutenção do controle metabólico. A necessidade

de adaptar a rotina e a dieta somada a doença pode atuar negativamente na qualidade de vida das pessoas (RODRIGUES, 2017). Assim de forma geral as pessoas que têm diabetes apresentam qualidade de vida menor que pessoas sem a doença (ZULIAN et al., 2013).

As complicações de saúde devido ao diabetes podem afetar diversos órgãos, membros e funções corporais e são causas de incapacidade, diminuição da qualidade de vida e morte prematura (BRACCO, 2019). As complicações mais significativas relacionam-se às pancreáticas, vasculares, oculares e neurológicas, advindas de distúrbios metabólicos causados pela hiperglicemia. Segundo a WHO, a diabetes pode ter os seguintes efeitos no corpo (WORLD HEALTH ORGANIZATION, [s.d.]):

- Aumentar de 2 a 3 vezes mais risco de ataque cardíaco em adultos.
- Causar retinopatia, que é uma das principais causas de cegueira e ocorre devido ao dano acumulado nos vasos sanguíneos da retina.
- Causar neuropatia resultante da isquemia dos nervos devido a efeitos diretos da hiperglicemia sobre os neurônios e alterações metabólicas intracelulares que alteram a função dos nervos.
- Causar a redução do fluxo sanguíneo nos pés aumentando as chances de úlceras, infecções e a necessidade de amputação.
- Causar falha nos rins (nefropatias);
- Fazer com que a pessoa tenha sintomas mais graves para diversas infecções, como por exemplo COVID-19.

Todas essas complicações acabam por fazer com que as pessoas tenham de se ausentar do trabalho por um determinado tempo caso fiquem hospitalizadas, tenham que se aposentar de forma prematura, que não consigam render tanto quanto renderiam caso elas estivem com 100% das capacidades, ou que ocorra o óbito da mesma.

2.2 Carga do diabetes no Brasil

A carga representada por uma doença pode ser expressa não só pela prevalência e incidência, mas também a partir do risco de desenvolver a doença ao longo da vida, dados de mortalidade, morbidade (anos vividos com incapacidade), anos de vida perdido ajustados por incapacidade

(DALY) e perda, em anos de vida e expectativa de vida, decorrentes da doença (BRACCO, 2019).

Segundo Muzy et al. em 2021, a prevalência do diabetes no Brasil foi de 9,2%, pelo modelo multinomial, e a prevalência da PNS corrigida (autorreferida + alterada na hemoglobina glicosilada - HbA1c $\geq$ 6,5) foi de 9,4%, apresentando uma grande diferença de acordo com a região do Brasil: de 6,3% no Norte, 7,2% no Sul, 7,6% centro-oeste, 12,2% para o nordeste e 12,8% no sudeste. Há diferença também de acordo com o sexos sendo de 10,2% para o sexo feminino e 8,1% para o sexo masculino (MUZY et al., 2021).

2.3 Bases de dados disponíveis

Na área da saúde existem uma infinidade de bancos de dados disponíveis, e devido a complexidade do estudo, será necessário combinar as estatísticas de diferentes bancos de dados. Os dados utilizados neste trabalho foram obtidos da PNS, do Instituto Brasileiro de Geografia e Estatística (IBGE) e estimativas realizadas com dados ELSA-Brasil.

2.3.1 Pesquisa Nacional de Saúde – PNS

A pesquisa nacional de saúde foi realizada até o momento em 2013 e 2019. O banco de dados utilizado neste trabalho se refere a pesquisa realizada em 2019, com representatividade nacional brasileira com base domiciliar, e amostragem probabilística complexa. A população alvo da PNS foram pessoas com idade de 15 anos ou mais. Foram excluídos da pesquisa moradias "especiais" como bases militares, penitenciárias/colônias penais, orfanatos, asilos, regiões indígenas e quilombolas (STOPA et al., 2020).

Neste questionário foram feitas aproximadamente 650 perguntas, e entre elas há perguntas mais específicas relacionadas à diabetes, à participação dos indivíduos na força de trabalho e à presença de limitações na realização das atividades habituais (IBGE, 2020).

O plano amostral da PNS será discutido posteriormente no capítulo 2.7.3.

2.3.2 Estudo Longitudinal de Saúde do Adulto – ELSA-Brasil

O ELSA-Brasil é um estudo longitudinal multicêntrico com quase 15 anos de existência. Neste estudo participam seis instituições públicas de ensino superior de três regiões do Brasil (nordeste, sul e sudeste), onde ocorreram o recrutamento de participantes e onde são realizadas as

entrevistas, exames clínicos e coleta, processamento, estocagem e transferência de amostras biológicas (AQUINO et al., 2012; LOTUFO, 2013). O ELSA-Brasil conta com a participação voluntária de 15105 mil funcionários (ativos ou aposentados) dessas instituições, que no início do estudo tinham entre 35-74 anos. Além de acompanhamento telefônico anual, o estudo está atualmente realizando a sua quarta visita presencial de acompanhamentos dos participantes.

Este estudo tem como objetivo contribuir com informações relevantes com respeito ao desenvolvimento e progressão de doenças crônicas, em particular doenças cardiovasculares e diabetes. Os dados do ELSA-Brasil são únicos no sentido de permitirem o acompanhamento de pessoas com e sem diabetes até o óbito.

2.3.3 Instituto Brasileiro de Geografia e Estatística - IBGE

Pelo IBGE consegue-se informações gerais da população, como a taxa de mortalidade e a projeção populacional, estratificado por idade e sexo ("IBGE - Instituto Brasileiro de Geografia e Estatística", [s.d.]).

2.4 Amostragem complexa

Em amostragem aleatória simples, todos os elementos da população têm igual probabilidade de fazerem parte da amostra e consideramos a amostra independente e identicamente distribuída de uma distribuição não conhecida (LEOTTI, 2019). Quando o tamanho da amostra é suficientemente grande, a lei dos grandes números e o teorema central do limite justificam a maioria das análises paramétricas.

No entanto, quando nosso interesse é construir amostras a partir de populações imensas, caso dos bancos de dados apresentados no capítulo 2.3, é inviável listar todos os componentes da população para amostragem aleatória simples, assim, utiliza-se a estratégia de conglomerados ou estratos, para obter uma amostra que seja representativa da população.

2.4.1 Amostragem por Conglomerados

Esse tipo de amostragem explora a existência de grupos (clusters/conglomerados) na população. Espera-se que os conglomerados sejam homogêneos entre si e que internamente representem adequadamente a população em relação a característica desejada, ou seja, possuam a variabilidade da população (OCHOA, 2015a). Com essa condição, podemos selecionar apenas uma parte dos conglomerados para fazer parte da amostra, conforme pode ser visto da Figura 1.

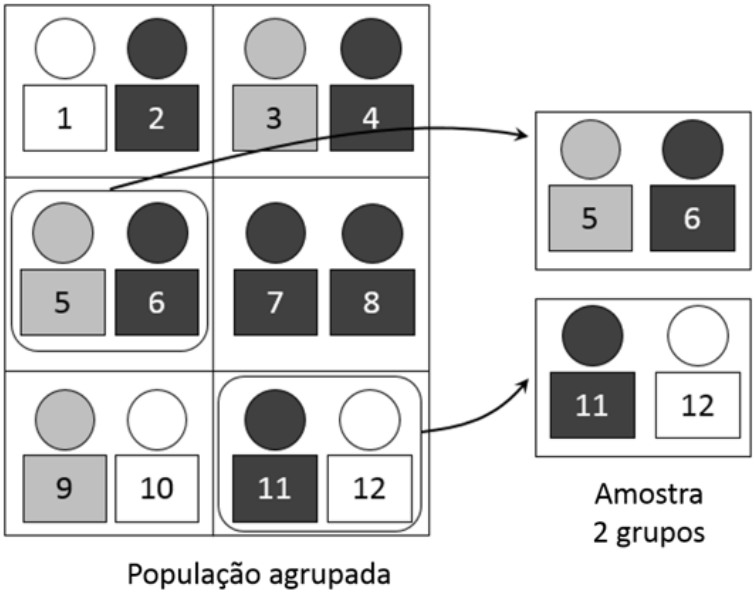

Figura 1: Exemplo de amostragem por conglomerado (OCHOA, 2015a).

Para realizar esse tipo de amostragem, primeiro deve-se definir os conglomerados, ou seja, identificar a característica que permite a divisão da população em grupos diferentes. Em pesquisas nacionais é usual a utilização de delimitadores geográficos, como os setores censitários (unidade territorial estabelecida para fins de controle cadastral), para construção dos conglomerados (STOPA et al., 2020). Para determinar o número de conglomerados que serão sorteados para fazer parte da amostra pode-se utilizar amostragem aleatória simples ou sistemática.

Apesar da amostragem por conglomerados apresentar a vantagem de ser mais simples selecionar um conglomerado do que realizar uma amostragem aleatória de todos os elementos da população, uma das desvantagens é o risco dos conglomerados não apresentarem homogeneidade entre eles.

2.4.2 Amostragem Estratificada

A amostragem estratificada consiste na divisão de uma população em grupos (chamados estratos) segundo alguma(s) característica(s) conhecida(s) na população em estudo. De cada estrato são selecionadas amostras (LEOTTI, 2019). Ou seja, divide-se a população em diferentes subgrupos, de maneira que um indivíduo pode fazer parte de apenas um estrato e após, seleciona-se indivíduos utilizando alguma técnica de amostragem, como a amostragem aleatória simples, ver Figura 2.

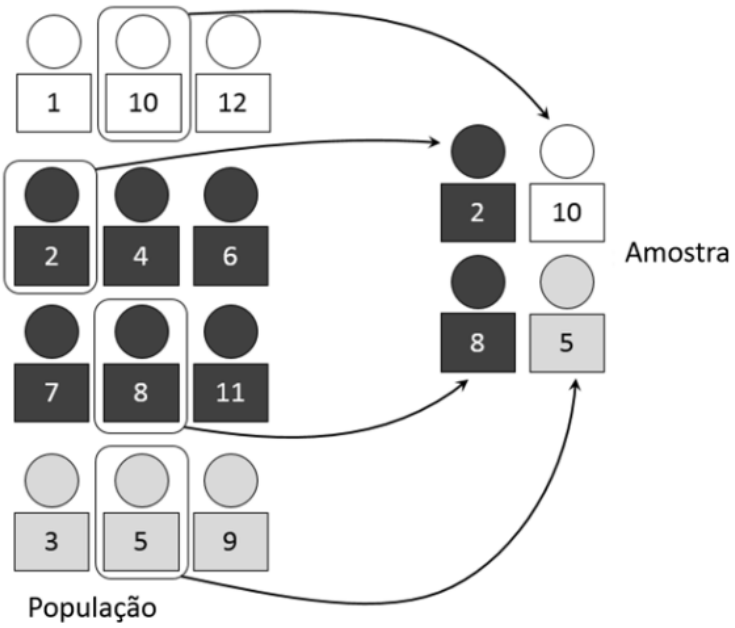

Figura 2: Exemplo de amostragem estratificada (OCHOA, 2015b)

As camadas ou estratos são grupos internamente homogêneos de elementos da população, que por sua vez, são heterogêneos entre si. Como por exemplo estudos dos mercados em uma cidade, pode-se construir três estratos: mercados pequenos, de tamanho médio e um terceiro estrato para os mercados grandes.

Assim, sendo os estratos homogêneos internamente e heterogêneos entre si, o uso da amostragem aleatória reduz o erro amostral melhorando a precisão dos resultados ao realizar um estudo sobre a amostra (OCHOA, 2015b).

2.4.3 Amostragem em três estágios

> A amostragem conglomerada em vários estágios é caracterizada por unidades populacionais arranjadas em grupos conforme uma hierarquia, com seleção de grupos nos vários níveis da hierarquia até chegar às unidades elementares (de referência) da pesquisa que serão investigadas (SILVA; BIANCHINI; DIAS, 2020).

Na amostragem conglomerada em três estágios, por exemplo, adota-se a seguinte terminologia, onde cada estágio da amostragem considera um tipo de unidade: Unidades Primárias de Amostragem – UPA’s; Unidades Secundárias de Amostragem – USA’s; e unidades elementares (SILVA; BIANCHINI; DIAS, 2020).

Com isso os três estágios para selecionar uma amostra de unidades elementares que serão investigadas podem ser definidos como (SILVA; BIANCHINI; DIAS, 2020):

- Estágio 1: selecionar uma amostra de UPA's.
- Estágio 2: selecionar uma amostra de USA's em cada uma das UPA's selecionadas no primeiro estágio.
- Estágio 3: selecionar uma amostra de unidades elementares em cada uma das USA's selecionadas no segundo estágio, que irão compor a amostra, *s* de unidades elementares a serem investigadas.

Assim de forma geral os pesos de uma pesquisa com amostragem complexa, mapeiam o caminho para uma representação não enviesada da população. Geralmente os pesos da análise final para cada elemento *i* considera o peso da seleção do elemento ($w_{sei,i}$), um fator de ajuste devido a não respostas ($w_{nr,i}$) e um fator de ajuste de pós-estratificação ($w_{ps,i}$) (BEST; WOLF, 2015).

$$w_{final,i} = w_{sei,i} \times w_{nr,i} \times w_{ps,i} \qquad \text{Eq. 2.4.1}$$

Plano Amostral PNS

Tendo como base esses três estágios, pode-se ter como exemplo o plano amostral realizado na PNS (FIOCRUZ, 2019; STOPA et al., 2020):

- Estágio 1: a seleção das unidades primárias de amostragem (setores censitários ou composição de setores) foi realizada por amostragem aleatória simples, mantendo a estratificação da Amostra Mestra da PNAD (Pesquisa Nacional de Amostra por Domicílios).
- Estágio 2: um número fixo de domicílios particulares permanentes foi selecionado aleatoriamente em cada UPA selecionada no primeiro estágio, a partir do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE).
- Estágio 3: dentro de cada domicílio da amostra, um morador (com 18 ou mais anos de idade, em 2013, e com 15 anos ou mais de idade, em 2019) foi selecionado com equiprobabilidade, a partir de uma lista de moradores elegíveis construída no momento da entrevista, para responder à entrevista individual.

Definição dos pesos

Em uma amostra construída por um plano complexo, os elementos da população não possuem mais a mesma probabilidade de participarem da amostra. Dessa forma, os elementos amostrais precisam incorporar a sua probabilidade de seleção durante as análises estatísticas. No caso da PNS, o peso do morador selecionado foi definido da seguinte forma (FREITAS, 2014, STOPA et al., 2020):

- Definiu-se a probabilidade de seleção do morador dentro do domicílio. O peso básico é dado por:

$$w_{hijk} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * O_{hij}.$$

- O peso com ajuste de não resposta do domicílio inteiro é dado por

$$w'_{hijk} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * \frac{n_{hi}^*}{n_h^{**}} * O_{hij}.$$

Onde:

- w_{hi}: é o peso básico da UPA *i* do estrato *h*.
- m_h: é o número da UPA's selecionadas no estrato *h* para Amostra Mestra.
- N_{hi}: é o número de domicílios particulares permanentes ocupados, ocupados sem entrevistas realizadas (equivalentes aos domicílios fechados) e vagos na UPA *i* do do estrato *h*.
- N_h é o número de domicílios particulares permanentes ocupados, ocupados sem entrevistas realizadas (equivalentes aos domicílios fechados) e vagos no estrato h , dados atualizados do CNEFE (Cadastro Nacional de Endereços para Fins Estatísticos) no momento da seleção das UPA's para Amostra Mestra.
- N_{hi}^*: é o número de domicílios particulares permanentes ocupados e fechados na UPA *i* do estrato *h*, dados do CNEFE, no momento da seleção dos domicílios;
- n_{hi}: número de domicílios selecionados na UPA *i* do estrato *h*,
- n_{hi}^*: número de domicílios selecionados com morador na UPA *i* do estrato *h*,

- n_h^{**}: número de domicílios selecionados com entrevista realizada na UPA *i* do estrato *h*,
- O_{hij}: é o número de moradores com 18 anos ou mais de idade no domicílio *j* na UPA *i* do estrato *h;*
- m_h^{PNS}: é o número de UPAs selecionadas no estrato h para a PNS

Após a seleção do morador houve perda de entrevista, foi realizada a correção dos pesos para compensar a não resposta. O ajuste foi realizado por sexo, pois houve diferença entre as perdas masculina e feminina (FREITAS, 2014).

$$w_{hijk}^{M} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^{*}}{n_{hi}} * \frac{n_{hi}^{*}}{n_h^{**}} * O_{hij} * \frac{\sum_j^{n_{hijk}^{***}} w\prime_{hijk} * \alpha_{hijk}^{M}}{\sum_i^{n_{hijk}^{***}} w\prime_{hijk} * \alpha_{hijk}^{M}}.$$

$$w_{hijk}^{F} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^{*}}{n_{hi}} * \frac{n_{hi}^{*}}{n_h^{**}} * O_{hij} * \frac{\sum_j^{n_{hijk}^{***}} w\prime_{hijk} * \alpha_{hijk}^{F}}{\sum_i^{n_{hijk}^{***}} w\prime_{hijk} * \alpha_{hijk}^{F}}.$$

- n_{hijk}^{***}: é o número de moradores selecionados com entrevista realizada na UPA *i* do estrato *h*;
- α_{hijk}^{M}: indica se o morador selecionado no domicílio j da UPA i do estrato h é do sexo masculino;
- α_{hijk}^{F}: indica se o morador selecionado no domicílio j da UPA i do estrato h é do sexo feminino;

Como é feita uma amostra aleatória simples de um morador dentro do domicílio, é natural que, por conta da aleatoriedade de seleção, os totais populacionais obtidos com os fatores de expansão do morador selecionado não sejam exatamente iguais aos totais populacionais obtidos com os fatores de expansão de domicílio (FREITAS, 2014).

No entanto, os moradores dos domicílios formam uma amostra maior que os moradores selecionados. Para igualar estas estimativas calibrou-se o peso do morador selecionado fazendo que os totais populacionais por sexo e classes de idade correspondessem aos totais obtidos com o peso do domicílio. As quatro classes de idade utilizadas foram de 18 a 24 anos, de 25 a 39 anos, de 40 a 59 anos e mais de 60 anos (FREITAS, 2014).

$$w_{hijk}^{M*} = w_{hijk}^{M} \frac{\hat{P}_{a,M,c}^{M}}{\hat{P}_{a,M,c}^{S}}.$$

Onde:

- $\hat{P}_{a,M,c}^{M}$: é a estimativa populacional obtida com os dados dos moradores dos domicílios da pesquisa para o nível geográfico *a* sexo M e classe de idade *c* ;

> Os pesos para os domicílios e todos os seus moradores, utilizados para a estimação das características investigadas para todos os moradores e para todos os idosos, foram definidos levando-se em conta o peso da UPA correspondente e ajustes para correção de não respostas e também para calibrar as estimativas com totais populacionais conhecidos de outras fontes (FREITAS, 2014).

A descrição dos outros pesos utilizados na PNS pode ser encontrado na Pesquisa Nacional de Saúde: Plano Amostral (FREITAS, 2014).

2.5 Modelos de Regressão Linear

A análise de regressão estuda a relação entre uma variável dependente e outras variáveis independentes. Esta relação é representada por um modelo matemático, ou seja, uma equação que associa a variável dependente (Y) com as variáveis independentes ($X_1, \dots, X_p$) (RODRIGUES, 2022).

Os modelos lineares podem ser simples, quando há apenas uma variável independente (X_1), ou múltiplos quando há *p* variáveis independentes ($X_1, \dots, X_p$).

Por se tratar de uma generalização do método, abordaremos o tema da regressão linear múltipla.

2.5.1 Regressão Linear Múltipla

O modelo de regressão linear múltipla com p variáveis dependentes pode ser definido conforme equação abaixo (SEBER; LEE, 2003).

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon_i, \qquad \text{Eq. 2.5.1}$$

em que,

$(X_1, \dots, X_p)$ é um vetor de variáveis independentes;

$\beta_0, \beta_1, \dots, \beta_p$ são os parâmetros do modelo;

Y_i é valor da variável resposta na observação i;

ε_i é o erro aleatório.

A interpretação para o parâmetro β_k do modelo é a mudança esperada na variável resposta, quando a variável X_k sofre um aumento unitário, enquanto todas as outras variáveis $X_j, j \neq k$ são mantidas constantes.

O parâmetro β_0 corresponde ao intercepto do plano de regressão, se existir a variável X_j = 0, então β_0 será a média deste ponto (RODRIGUES, 2022).

2.5.2 Interações

Caso considerarmos um modelo mais complexo, no qual existe interação entre as variáveis obtemos um modelo da forma (RODRIGUES, 2022):

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon_i$$

Se a interação $X_1 X_2$, for significativa e existir, o efeito de X_1 na resposta depende do nível X_2 e vice-versa.

2.5.3 Pressupostos

Os pressupostos para o modelo de regressão linear múltipla são (YANG; TU; CHEN, 2019):

a) Média dos erros nula, $E(\varepsilon_i) = 0, i = 1, \dots, n$;
b) Os erros são independentes;
c) Variância do erro é constante, $V(\varepsilon_i) = \sigma^2, i = 1, \dots, n$;
d) Os erros têm distribuição normal.

Destes pressupostos, concluímos que $\varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$ e consequentemente que Y tem distribuição normal com σ^2 e, para o caso de modelo definido na equação 2.5.1

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n.$$

2.5.4 Estimação dos Parâmetros do modelo

Como a relação entre X e Y é linear, têm-se a necessidade de estimar os valores de β. Para definir os parâmetros do modelo utiliza-se o método dos mínimos quadrados. Este método visa

minimizar a soma dos quadrados dos desvios e_i (SQE), para encontrar o vetor de estimadores $\underline{\hat{\beta}}$, com componentes $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n)$ (RODRIGUES, 2022).

$$SQE = \sum_{i=1}^{n} e_i^2 = \underline{e}'\underline{e} = (\underline{Y} - \underline{X}\underline{\beta})'(\underline{Y} - \underline{X}\underline{\beta})$$

$$= \underline{Y}'\,\underline{Y} - \underline{Y}'\,\underline{X}\underline{\beta} - \underline{\beta}'\underline{Y}\,\underline{X}' + \underline{\beta}'\underline{X}\,\underline{X}'\underline{\beta}$$

$$= \underline{Y}'\,\underline{Y} - 2\underline{\beta}'\underline{Y}\,\underline{X}' + \underline{\beta}'\underline{X}\,\underline{X}'\underline{\beta},$$

em que

$\underline{e} = [e_1 \ \vdots \ e_n\ ]$;

$\underline{e}' = matriz\ transposta\ de\ \underline{e}$.

Sendo válido o análogo para os outros termos.

Calculando as derivadas parciais em $\underline{\beta}$, igualando a zero e substituindo β por $\hat{\beta}$, obtemos:

$$\frac{\partial SQE}{\partial \underline{\beta}} = -2\underline{X}'\underline{Y} + 2\underline{X}'\underline{X}\underline{\beta} = 0$$

$$(\underline{X}'\underline{X})\underline{\hat{\beta}} = \underline{X}'\underline{Y}$$

$$\underline{\hat{\beta}} = \underline{X}'\underline{Y}\,(\underline{X}'\underline{X})^{-1}$$

Tendo encontrado o vetor de estimadores $\underline{\hat{\beta}}$, temos que:

$$\underline{\hat{Y}} = \underline{X}\underline{\hat{\beta}},$$

em que

$\underline{\hat{Y}}$: Vetor com valores preditos de Y calculados a partir dos $\underline{\hat{\beta}}$ estimados .

Com isso o vetor com os resíduos é:

$$\underline{e} = \underline{Y} - \underline{\hat{Y}}$$

2.5.5 Análise de Variância

Após encontrar os estimadores para os parâmetros é importante determinar se as variáveis independentes do modelo possuem poder de explicação, para isso faz-se o teste F.

Para esse teste decompõem-se a soma de quadrados total, SQT (variância da variável resposta), soma dos quadrados explicada, SQR (variação da variável resposta explicada pelo

modelo) e a soma dos quadrados dos resíduos, SQE (variação não explicada pelo modelo) (RODRIGUES, 2022).

$$F = \frac{\frac{SQR}{p}}{\frac{SQE}{n-p-1}} = \frac{QMR}{QME} \dot{\sim} F_{p,n-p-1}$$

Onde:

- p número de variáveis independentes
- n dimensão da amostra,

Neste teste testa-se duas hipóteses, a H_0 afirma que os parâmetros β são iguais e iguais a zero, e a hipótese alternativa que afirma a existência de ao menos um β diferente de zero.

$$\{H_0: \beta_0 = \beta_1 = \ldots = \beta_n = 0 \ ; \ H_1: \exists_j: \beta_j \neq 0, j = 1, \ldots, p$$

Assim, com relação a H_0, a estatística F segue uma distribuição F central com p e $n - (p + 1)$ graus de liberdade. Portanto se, $F_{obs} > F_{p,n-p-1}$ rejeita-se a hipótese H_0, concluindo que pelo menos uma das variáveis independentes contribui significativamente para o modelo (RODRIGUES, 2022).

Após essa definição, pode-se testar cada termo do modelo individualmente com um teste t.

2.5.6 Análise de Resíduos

Após as variáveis terem sido determinadas, e analisada a significância delas, diversos problemas ainda podem existir no modelo como (SOUZA, 2006):

• Presença de observações discrepantes;

• Inadequação das pressuposições para os erros aleatórios ou para as médias;

• Colinearidade entre as colunas da matriz do modelo;

• Forma funcional do modelo inadequada;

• Presença de observações influentes.

Para determinar a existência destes problemas, realiza-se a análise de resíduos para testar os pressupostos citados anteriormente no capítulo 2.4.3.

Diagnóstico de normalidade

Para diagnosticar se os resíduos têm uma distribuição normal, pode-se utilizar o gráfico P-P plot. Nesse gráfico, compara-se a distribuição de probabilidades dos valores observados, com os valores esperados (linha diagonal), conforme pode-se ver na figura 1 (RODRIGUES, 2022).

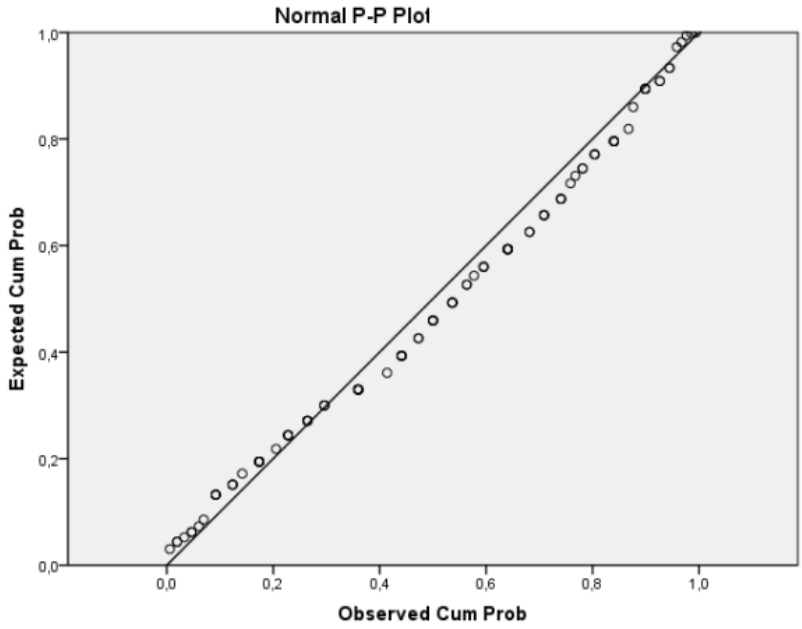

Figura 3: P-P plot de resíduos (RODRIGUES, 2022)..

Para determinar que os resíduos têm uma distribuição normal, as observações devem estar localizadas próximas da linha diagonal.

A normalidade dos resíduos também pode ser avaliada por um histograma estandardizado, no qual se avalia se há mudanças com relação à forma simétrica de uma distribuição normal. Pode ser feito também pelo teste Kolmogorov-Smirnov (K-S), ou pelo teste Shapiro Wilk (RODRIGUES, 2022)..

Diagnóstico de Homoscedasticidade

Para verificar se as variâncias estão constantes, pode-se criar um gráfico dos resíduos versus os valores preditos, mostrado na figura 2 (RODRIGUES, 2022)..

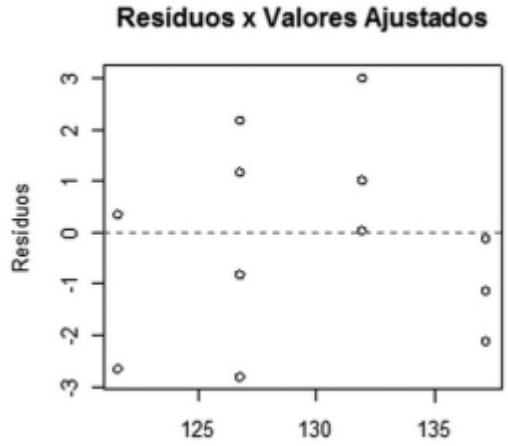

Figura 4: Gráfico de resíduos versus valores ajustados (RODRIGUES, 2022)..

Como os resíduos devem ter variância constante, eles devem estar igualmente distribuídos ao redor da reta y=0, sem nenhuma tendência ou comportamento. Um exemplo de presença de variância é a forma de "funil", no qual os pontos ficam próximos da linha y=0, mas começam a se afastar de acordo com que os valores ajustados (eixo x) aumenta (RODRIGUES, 2022).

Diagnóstico de Independência

Para diagnosticar a presença de autocorrelação, pode-se utilizar o teste de Durbin-Whatson (DW) (PESTANA; GAGEIRO, 2008).

Neste teste testa-se as duas hipóteses:

$\{H_0: Não\ existe\ autocorrelação\ dos\ resíduos\ \ H_1: Existe\ autocorrelação\ dos\ resíduos$

Para testar as hipóteses, utiliza-se a seguinte fórmula:

$$dw = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} {e_i}^2}$$

Esta medida mede a correlação entre cada resíduo e o resíduo correspondente à observação anterior.

A multicolineariedade testa se há correlação entre as variaáveis utilizadas, uma forma de de diagnosticar a multicolineariridade é o teste VIF – Fator de Inflação de Variância. Supondo que as variáveis estão centradas e padronizadas, tem-se que $R = (\underline{X}'\underline{X})^{-1}$ em que os elementos da diagonal dessa matriz são chamados de fatores de inflação de variância (VIF) e representam o incremento da variância devido à presença de multicolinearidade (MONTGOMERY; PECK; VINING, 2006). O VIF pode ser calculado da seguinte forma:

$$VIF = \frac{1}{1 - R_j^2}, \qquad j = 1,2, \dots, p$$

Em que: p é o número das variáveis preditoras; R_j^2 é o coeficiente de correlação múltipla, resultante da regressão de *Xj* nos outros *p-1* regressores. Com isso, consegue-se afirmar que quanto maior foi o VIF, maior é a correlação *Xj* com os outros *p-1* regressores. Assim pode-se considerar que um VIF acima de 10 a multicolinearidade pode estar influenciando as estimativas de mínimos quadrados (MONTGOMERY; PECK; VINING, 2021).

Diagnóstico de Outliers e Observações Influentes

Outliers são observações extremas, muito afastadas da maioria dos dados, deixando dúvidas com relação a validade da mesma. Caso esses *outliers* sejam influentes com relação ao modelo de regressão ajustado a inclusão ou não destes pontos modifica substancialmente o modelo e com isso os valores ajustados (PESTANA; GAGEIRO, 2008).

Uma medida que serve para diagnosticar os *outliers* é o *Leverage* (LEV). Um LEV elevado indica que essa observação se distancia do centro das observações influenciando o valor previsto. Para considerar o valor como elevado depende do *p* número de variáveis independentes e o *n* dimensão da amostra, conforme as equações abaixo (PESTANA; GAGEIRO, 2008).

$$LEV > \frac{3(p+1)}{n}\ , amostras\ de\ dimensão\ reduzidas$$

$$LEV > \frac{2(p+1)}{n}\ , amostras\ grandes$$

Após identificar se a observação é um outlier, é importante determinar se essa observação é influente . A definição de observação influente é: "a observação que individualmente ou em conjunto com as outras observações demonstram ter mais impacto do que as restantes no cálculo dos estimadores" (RODRIGUES, 2022).

Algumas das técnicas que podem ser utilizadas para determinar a influência são (RODRIGUES, 2022):

- SDFFIT

  Trata-se de uma medida estandardizada que mede a influência que a observação i tem sobre o seu valor ajustado, considera-se influente se:

$$|SDFFIT| > 2\sqrt{\frac{p+1}{n-p-1}}$$

- SDFBETA

  Também é uma medida estandarizada, porém corresponde à alteração nos coeficientes estimados, $\hat{\beta}_j$, *j=0,...,p*. A definição de influência é dada por:

$$|SDFBETA| > 1{,}96; amostras\ com\ dimensões\ reduzidas$$

$$|SDFBETA| > \frac{2}{\sqrt{n}}; amostras\ grandes$$

- Distância de Cook

A distância de Cook mede a influência da i-ésima observação sobre todos os *n* valores ajustados $\hat{y}_j$. Um valor elevado indica que o e_i é elevado. Assim a observação pode ser considerada influente quando:

$$COOK > \frac{4}{n-p-1}$$

Caso o valor da distância de Cook seja maior que 1 considera-se que a observação é extremamente influente.

2.6 Regressão Linear em Amostras Complexas

Conforme mencionado no capítulo 2.4, quando se trata de amostras complexas, as análises têm que considerar alguns ajustes como o peso para cada amostra devido ao estrato, conglomerado, etc. Considerando a equação 2.4.1, que mostra uma generalização para o peso de um elemento de amostra complexa, um exemplo de estimador para as estatísticas da população pode ser visto na equação 2.6.1 (BEST; WOLF, 2015).

$$\hat{B} = \frac{\sum_{i=1}^{n} w_{sei,i} * y_i x_i}{\sum_{i=1}^{n} w_{sei,i} * {x_i}^2} \qquad \text{Eq. 2.6.1}$$

Onde:

$w_{sei,i}$: peso da seleção do elemento *i*;

x_i : observação *i*;

y_i : valor da variável resposta na observação *i*;

$\hat{B}$: estimador da inclinação da reta.

Se considerarmos os pesos finais considerados em uma pesquisa com amostragem complexa de uma população finita, os estimadores dos parâmetros da regressão β são aqueles que minimizam a seguinte função para uma população finita de tamanho *N* (BEST; WOLF, 2015):

$$f(B) = \sum_{i=1}^{N} (y_i - x_i B)^2 \qquad \text{Eq. 2.6.2}$$

Considerando que a função definida na equação 2.6.2 como a soma dos quadrados dos resíduos (SQE_{pop}). Um estimador não enviesado incorporando os pesos utilizados da pesquisa pode ser escrito como (BEST; WOLF, 2015):

$$\widehat{SQE}_{pop} = \sum_{h}^{H} \sum_{\alpha}^{\alpha_h} \sum_{i}^{i_{h\alpha}} w_{h\alpha i} (y_{h\alpha i} - x_{h\alpha i} B)^2,$$

onde *h* é o índice relacionado ao estrato, α é relacionado ao conglomerado (ou à UPA) e o *i* é o i-ésimo elemento do cluster α.

Quando os pesos de uma pesquisa são fornecidos (como no caso da PNS), o estimador do parâmetro B, pode ser definido da seguinte forma:

$$\hat{B} = \underline{X}'\underline{W}\,\underline{Y}\,(\underline{X}'\underline{W}\,\underline{X})^{-1}$$

Onde, $\underline{W}$ é uma matriz diagonal n x n onde os valores finais do peso, $w_{h\alpha i}$, estão localizados na diagonal para cada caso na amostra.

2.6.1 Diagnósticos da Regressão

As análises de diagnóstico como por exemplo, gráficos de resíduos ajustados, p-p plots, podem ser usadas para identificar problemas com a estrutura do modelo em uma população grande, assim como violações das suposições de distribuição para a variável dependente ou valores discrepantes que podem ter grande alavancagem e/ou influência no ajuste de um modelo (BEST; WOLF, 2015).

Segundo Best et al., 2015, só recentemente pesquisadores começaram a examinar como a computação e a interpretação de outras ferramentas de diagnóstico comuns para modelos de regressão linear podem ser adaptadas para acomodar recursos de amostragem complexos (BEST; WOLF, 2015).

No entanto, ferramentas computacionais foram desenvolvidas nos últimos anos, como por exemplo o pacote *svydiags* (VALLIANT; VALLIANT, 2018), que têm funções que testam os diagnósticos em modelos de regressão linear. A descrição deste pacote será realizada no capítulo 3.8.4.

2.7 Efeitos Marginais

Segundo Onukwugha (2015) efeitos marginais mostram como a variavel dependente muda quando determinada variável independente sofre alguma mudança, e as outras covariáveis são mantidas constantes. Mais especificamente, esse efeito marginal de covariáveis contínuas pode ser definido como o valor numérico da derivada parcial da variável resposta com relação a uma pequena mudança na covariável. Ou em caso de uma covariável discreta, é medido como uma mudança incremental na variável resposta devido ao aumento na covariável discreta. Assim se temos uma regressão dada pela função $y = f(x;\beta)$, onde y é a variavel dependente, $x=(x_1,\ldots, x_k)$ é

um vetor de *k* variáveis independentes, e *b* é um vetor de parâmetros a serem estimados, o efeito marginal é (ONUKWUGHA; BERGTOLD; JAIN, 2015):

$$Efeito\ Marginal\text{: } \frac{\partial(x;\beta)}{\partial x_k} = g(x;\beta)$$

Por ser calculado diretamente da função obtida pela regressão, teremos diferentes efeitos marginais de acordo com o tipo de modelo (ONUKWUGHA; BERGTOLD; JAIN, 2015):

2.7.1 Modelo de regressão linear:

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$Efeito\ Marginal\text{: } \frac{\partial(x;\beta)}{\partial x_2} = \beta_2 \qquad \text{Eq. 2.7.1}$$

Como pode-se ver na equação 2.7.1, em modelos onde há apenas relações lineares das variáveis, o efeito marginal é constante. Porém, o cálculo do efeito marginal é igual, ainda que a regressão inclua transformações não lineares como potências, logarítmicos ou até interações entre as variáveis, que ela ainda será uma função linear dos coeficientes.

2.8 Pacote Survey

Para a realização de análises em amostras originárias de pesquisas por amostragem complexa, devem ser utilizados programas de análise estatística que possuam pacotes destinados a dados complexos, ou seja, que possuam um algoritmo capaz de considerar os efeitos da estratificação e da conglomeração na estimação dos indicadores e suas medidas de precisão (STOPA et al., 2020).

Para a realização de análises analíticas, como a construção de modelos envolvendo dados provenientes de amostragem complexas, existem diversos programas/pacotes estatísticos capazes de realizar essas análises (tais como MINITAB, R, SAS, SPSS, etc). Neste capítulo o foco será na linguagem R, mais especificamente no pacote *survey* (LUMLEY, 2004).

Conforme (LUMLEY, 2004), analisar uma amostra estratificada como se ela fosse uma amostra aleatória simples irá superestimar os erros padrões, analisar uma amostra por clusters (conglomerados) como se fosse uma amostra aleatória simples, normalmente irá subestimar os erros padrões, isso também ocorre ao analisar uma amostra com probabilidades desiguais como uma amostra aleatória simples.

O pacote survey tem muitas ferramentas para colaborar e tornar possível a análise com amostras complexas.

2.8.1 Svydesing

Uma das principais objetivos do pacote survey é vincular os metadados do delineamento (design) aos dados que serão analisados, para que os ajustes corretos possam ser realizados de forma confiável e automática. Isso é feito com as funções ***svydesign***, que cria objeto que contém o bancop de dados e informações adicionais, como estratos e pesos, conforme pode ser visto no exemplo:

pnsd <- svydesign(id=~UPA,strata=~V0024, weights=~V00291, data=pns)

Conforme o exemplo mostra o objeto ***pnsd*** contém as seguintes informações:

- ***id*** – define a unidade primária de amostragem (UPA);
- ***strata*** – a coluna que será utilizada para estratificar a população (ver capítulo 2.7.2);
- ***weights –*** indica a coluna que será utilizada para definir os pesos de cada elemento da população.
- ***data*** – banco de dados que será utilizado para definir o design.

2.8.2 Survey-weighted generalised linear models – svyglm

Com o design definido, pode-se utilizar o svyglm para gerar um ajuste nos dados com um modelo linear generalizado, como regressão linear, modelos da família binomial ou Poisson. Para gerar modelos dos últimos dois casos pode-se utilizar um argumento *family = quasibinomial()* e *family = quasipoisson()*. Exemplos de utilização desta função podem ser vistas os códigos a seguir.

Linear <- svyglm(Days ~ DM, design=pns)

Os códigos mostrado mostra a utilização da função ***svyglm***. Mostrando a criação de um modelo linear, de duas variáveis (variável dependente quantitativa *Days* (dias que faltou ao trabalho) e variável independente *DM*), utilizando um design definido conforme explicado na seção 2.8.1.

2.8.3 Svypredmeans

Dependendo do tipo a análise desejada, deve ser realizada a obtenção do resultado de previsão marginal pode ser necessária, como por exemplo quantificar a média de dias que

indivíduos com diabetes se ausentaram do trabalho, ajustado por características socioeconômicas (ONUKWUGHA; BERGTOLD; JAIN, 2015).

Para obter os dados das previsões marginais, o pacote survey apresenta a função *svypredmeans()*, como pode ser visto no código abaixo:

svypredmeans(model1,~DM)

Essa função têm alguns parâmetros que são obrigatórios como:

- Model – o modelo de regressão criado;
- DM – fórmula, especificando o fator para o qual se quer encontrar a marginal.

2.8.4 Diagnóstico do Modelo

Conforme descrito no capítulo 2.5, após a definição do modelo é necessário confirmar se o modelo respeita os pressupostos. Para calcular os diagnósticos, apesar das dificuldades da amostra ser complexa, a biblioteca *svydiags* realiza esses cálculos para modelos de regressão linear obtidos pelo através do pacote *survey* (VALLIANT; VALLIANT, 2018).

O diagnóstico de normalidade e homocedasticidade, pode ser realizado visualizando o histograma dos resíduos do modelo e pela função *svystdres* do pacote *svydiags* respectivamente.

Para definir a independência, e diagnosticar a multicolinearidade pode-se realizar a análise VIF (*Variance Inflation Factor* ou Fator de Inflação da Variância) utilizando a função *svyvif*.

A identificação de Outliers e observações influentes pode ser realizada pela função *svydffits*. Essa função calcula o DFFITS.

2.9 Tabela de Vida

Tabela de vida, também chamada de tabela de mortalidade é uma técnica estatística muito utilizada para apresentar a mortalidade de uma população de uma forma que seja possível responder a perguntas como: qual a probabilidade de uma pessoa com 30 anos viver até os 70?(NAMBOODIRI; SUCHINDRAN, 1987). Por possibilitar esse tipo de resposta a tabela de vida é muito utilizada na área de saúde pública, onde pode-se considerar o efeito que determinada doença tem na população, através do estudo de uma coorte. Um exemplo de tabela de vida pode ser vista na tabela 1.

Tabela 1: Forma abreviada da tabela de vida para dos Estado Unidos da América - EUA (DAY; REYNOLDS; KUSH, 2015)

Idade	I(x)	d(x)	q(x)	L(x)	T(x)	e(x)
0	100000	612	0,006	99694	7866328	78,7
1	99388	43	0,000	99367	7766634	78,1
2	99345	27	0,000	99332	7667268	77,2
3	99318	21	0,000	99308	7567936	76,2
4	99297	16	0,000	99289	7468629	75,2
5	99281	14	0,000	99274	7369340	74,2
...	...	...	...	...	...	...
20	98910	74	0,001	98873	5881863	59,5
30	98011	101	0,001	97961	4897086	50
40	96798	161	0,002	96718	3922589	40,5
50	94295	394	0,004	94098	2965247	31,4
60	88770	778	0,009	88381	2046832	23,1
70	78069	1526	0,020	77306	1206570	15,5
80	57188	2868	0,050	55754	519263	9,1
90	23619	3375	0,143	21932	108501	4,6
100	1968	692	0,352	1622	4611	2,3

A tabela apresentada mostra algumas informações que usualmente são utilizadas como:

- Idade: a idade no início de cada intervalo;
- I(x): o número de pessoas vivas no início da idade;
- d(x): quantidade de pessoas que morreram na idade;
- q(x): probabilidade de morte durante a idade;
- L(x): Anos de vida vividos na idade, considerando que as pessoas morreram no meio do ano;
- T(x): total de anos de vida vividos;
- e(x): expectativa de vida.

A expectativa de vida apresentada na tabela 1 mostra a média dos anos que as pessoas da coorte, em determinada idade, ainda tem para viver. É importante entender que a expectativa de vida não é uma previsão dos anos que qualquer indivíduo vai viver. Por esse motivo, ao nascer – idade 0, a expectativa de vida é de 78,7 anos, porém mais de 20% da população considerada na tabela 1 não vai viver até os 70 anos, no entanto mais de 20% das pessoas vão viver 90 anos ou mais, tendo uma expectativa de vida de 4,6 anos (na idade de 90 anos) e 2,3 anos (na idade de 100 anos) (DAY; REYNOLDS; KUSH, 2015).

Apesar da expectativa de vida ser a resposta mais utilizada em estudos de sobrevivência (DAY; REYNOLDS; KUSH, 2015), dependendo do contexto da análise, outras respostas podem ser mais importantes como por exemplo a probabilidade de morrer até uma determinada idade (REDDY; KAR, 2019), ou anos de vida perdidos devido a determinada doença (BRACCO, 2019).

2.9.1 Tipos de tabela de vida

Tabelas de vida podem ser nomeadas de duas formas (LAHIRI, 2018):

- Tabela de vida de coorte ou de geração;
- Tabela de vida convencional.

A tabela de vida de coorte, requer o acompanhamento de um grupo desde o nascimento, até a morte do último membro deste grupo, adquirindo então a taxa de mortalidade. Devido a necessidade de acompanhamento por toda a vida dessa coorte, esse tipo de tabela de vida acaba tendo uma dificuldade elevada de ser calculada/estimada/construída (LAHIRI, 2018).

Já as tabelas de vidas convencionais, que são mais comuns em análise demográfica, geram uma visão transversal da mortalidade em diferentes idades em dada população durante um curto espaço de tempo (10 anos ou menos). Ela gera o tempo de vida de um indivíduo de uma coorte hipotética (normalmente 100.000 indivíduos), assumindo que indivíduos em uma idade particular estarão submetidos à mesma taxa de mortalidade que foi observada na população durante o tempo de estudo. Ela acaba gerando o tempo de vida que um recém-nascido assumindo que a mesma sequência de mortalidade seja seguida durante a vida (LAHIRI, 2018).

Dentro destes dois tipos de tabela de vida, elas podem ter diferentes classificações de acordo com os fatores utilizados.

Tabelas unidecrementais

A tabela de vida mostrada na tabela 1 é considerada uma tabela de vida unidecremental, isto é, para a sua construção foi utilizada apenas um fator (mortalidade).

Tabelas multidecrementais

Tabelas de vida com mais de um fator, são chamadas de tabelas de vida multidecrementais. Nesse tipo de tabela, leva-se em consideração dois ou mais fatores que operam em conjunto. Como por exemplo, o tamanho da população de pessoas solteiras, ela sofre o efeito tanto da mortalidade,

quanto do primeiro casamento. Esses dois fatores em conjunto produzem uma tabela multidecremental de dois fatores (NAMBOODIRI; SUCHINDRAN, 1987).

Tabelas incremental-decremental

Em ambos os casos citados até o momento, tabelas uni e multidecrementais, não há aumento no grupo de pessoas em estudo, no entanto existem casos onde pode ocorrer tanto a redução quanto o aumento do grupo em questão. Um exemplo é o estudo de força de trabalho, no qual pode haver a redução por diferentes motivos (morte, aposentadoria, ausência temporária devido a doenças), porém no caso de haver a ausência temporária devido a doenças, acontece de haver o incremento no final do período de ausência. Esse tipo de tabela pode ser chamada de tabela incremental-decremental ou tabela multiestado.(NAMBOODIRI; SUCHINDRAN, 1987).

2.10 Taxa de Mortalidade

Apesar da taxa de mortalidade ser o fator mais utilizado para a criação de tabelas de vida de uma população com dados de registros nacionais, há uma dificuldade quando o objetivo é determinar a mortalidade de indivíduos com uma determinada doença em uma população na qual não há registros suficientes para essa inferência (BRACCO, 2019).

Em 2017, Jacobs et. al. (JACOBS et al., 2017) aplicou um método matemático para calcular a taxa de mortalidade de pessoas com e sem diabetes da Alemanha. Esse método é relevante especialmente quando não possuímos um registro nacional que permita inferir a taxa de mortalidade separadamente em indivíduos com e sem diabetes (BRACCO, 2019).

Os passos para realizar o método desenvolvido por Jacobs et. al. são (JACOBS et al., 2017):

1. Cálculo da Razão de Taxa de Mortalidade (RTM), para cada sexo s e idade i, entre indivíduos com (M_1) vs. sem diabetes (M_0) na população

$$RTM(i,s) = \frac{M_1(i,s)}{M_0(i,s)} \qquad \text{Eq. 2.10.1}$$

2. Com os dados da taxa de mortalidade da população brasileira (M_t), da prevalência de diabetes (P) e da RTM para cada sexo e idade, pode-se estimar as taxas de mortalidade de indivíduos com (M_1) vs. sem diabetes (M_0).

   Sabendo que:

$$M_t(s,i) = P(i,s) \times M_1(s,i) + (1 - P(s,i)) \times M_0(s,i) \qquad \text{Eq. 2.10.2}$$

Para isolar M_1 e M_0 através da Equação 2.10.1 e 2.10.2, seguem-se os passos abaixo:

$$M_t(s,i) = P(i,s) \times M_1(s,i) + M_0(s,i) - P(i,s) \times M_0(s,i)$$

$$M_t(s,i) = P(i,s) \times \big(M_1(s,i) - M_0(s,i)\big) + M_0(s,i)$$

$$M_t(s,i) = \left[\frac{P(i,s) \times \big(M_1(s,i) - M_0(s,i)\big) + M_0(s,i)}{M_0(s,i)}\right] \times M_0(s,i)$$

$$M_t(s,i) = \left[P(i,s) \times \left(\frac{M_1(s,i)}{M_0(s,i)} - 1\right) + 1\right] \times M_0(s,i) \qquad \text{Eq. 2.10.3}$$

Utilizando a equação 2.10.1 na equação 2.10.3:

$$M_t(s,i) = [P(i,s) \times (RTM(s,i) - 1) + 1] \times M_0(s,i) \qquad \text{Eq. 2.10.4}$$

Assim, a taxa de mortalidade para os indivíduos sem diabetes pode ser calculada da seguinte forma:

$$M_0(s,i) = \frac{M_t(s,i)}{[P(s,i) \times (RTM(s,i) - 1) + 1]} \qquad \text{Eq. 2.10.5}$$

Pela equação 4 e 5, temos que a taxa de mortalidade de indivíduos com diabetes é:

$$M_1(s,i) = RTM(s,i) \times M_0(s,i) \qquad \text{Eq. 2.10.6}$$

2.11 PALY – Anos de Vida Ajustado pela Produtividade

Segundo Magliano (MAGLIANO et al., 2018), a estimação do PALY ("productivity-adjusted life-years") ocorre por modelos de tabela de vida com ciclo de um ano. Os cálculos de estimação do PALY serão realizados separadamente para cada sexo *s*. Para cada idade *i* o cálculo se inicia a partir do número de indivíduos vivos estimado com diabetes na idade *i* (N_i).

Após, é necessário calcular a taxa de mortalidade de indivíduos com ($M_1(i)$) e sem diabetes ($M_0(i)$), ver equações 2.11.1 e 2.11.2. Com essas informações calcula-se (separadamente, considerando ou mortalidade com diabetes ou a mortalidade sem diabetes) o número de mortes (M_i) e consequente número de indivíduos ainda vivos ao final dessa idade e início da idade i + 1 (N_{i+1}).

Utilizando $M_1(i)$:

$$M_i = N_i M_1(i) \qquad \text{Eq. 2.11.1}$$

$$N_{i+1} = N_{i-1} - M_i \qquad \text{Eq. 2.11.2}$$

Com os valores de N_i e N_{i+1} calcula-se o total de anos de vida populacional estimados para determinada idade *i*.

A seguir, a partir do índice de produtividade da idade *i* (PALY index (i)), o qual é um índice que a variação vai de 0 (produtividade nula) a 1 (produtividade total) e é calculado como a soma do absenteísmo e do presenteísmo em cada idade *i*, estima-se os anos de vida vividos com produtividade ou ajustados pela produtividade ($PALY_i$):

$$PALY_i = Anos\ de\ vida\ _i \times\ Pindex_i \qquad \text{Eq. 2.11.3}$$

O PALY index é estimado como 1 para indivíduos sem diabetes. Para indivíduos com diabetes estimamos a redução percentual de dias trabalhados (absenteísmo) e redução percentual de produtividade devido à limitações de saúde (presenteísmo) observada quando comparada com os indivíduos sem diabetes. Um resumo do cálculo do PALY pode ser visto na figura 5.

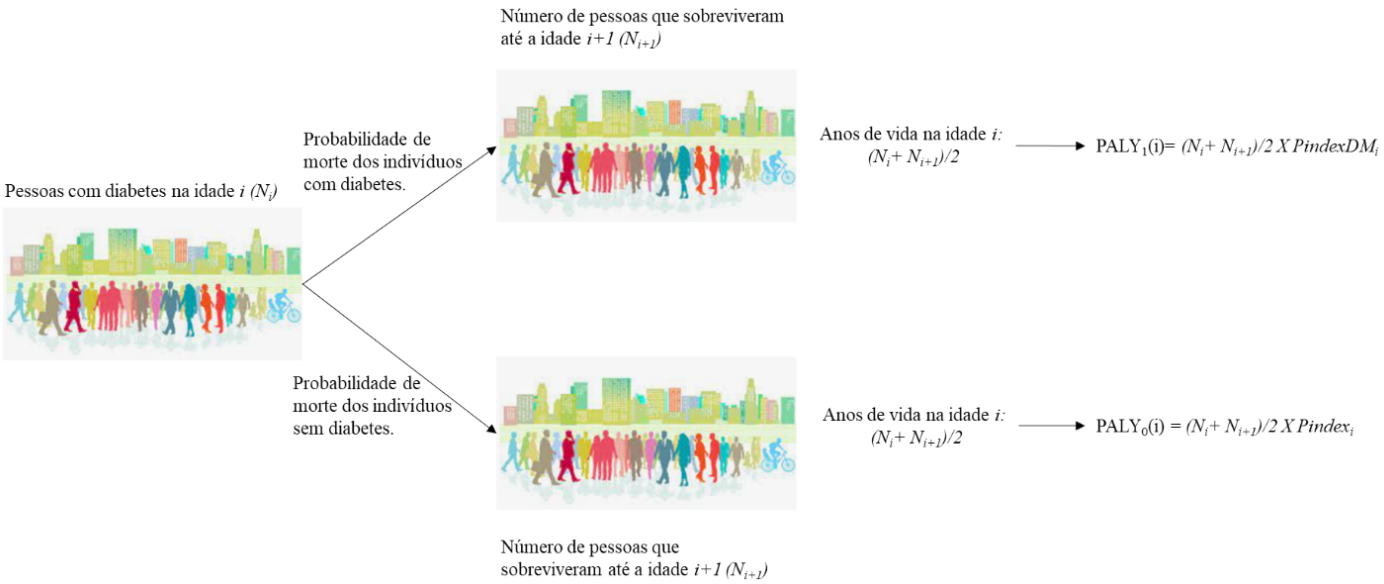

Figura 5: Resumo do cálculo do PALY.

Com a diferença entre o PALY para as população com diabetes ($PALY_1(i)$) e sem diabetes ($PALY_0(i)$) se consegue chegar no dado dos anos de vida produtivos perdidos (AVP) devido a presença do diabetes em cada idade *i*.

$$AVP(i) = PALY_0(i) - PALY_1(i) \qquad \text{Eq. 2.11.7}$$

3 Métodos

3.1 Obtenção dos arquivos dos dados

Após estudo das bases de dados disponíveis com representatividade nacional, nesta etapa do trabalho foram coletados dados de duas fontes: PNS e IBGE. Todos os bancos de dados estão disponíveis para download pelos endereços eletronicos: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=&t=microdados e https://www.ibge.gov.br/, respectivamente (FIOCRUZ, 2019; IBGE, 2020).

No caso da PNS, por se tratar de um banco de dados construídos para múltiplas utilizações, têm-se a necessidade de uma limpeza/filtro nos mesmos, seja selecionando as questões que são relevantes para cada um dos cálculos, seja transformando as variáveis para uma melhor compreensão, para isso a utilização do dicionario de variáveis é extremamente importante (podem ser obtidos no mesmo endereço citado anteriormente). Essas etapas serão discutidas nos próximos capítulos.

3.2 Prevalência do Diabetes e Mortalidade

A estimativa das taxas de mortalidade para indivíduos com e sem diabetes foram previamente calculadas (BRACCO, 2019; BRACCO et al., 2021) a partir dos dados da ELSA-Brasil e do IBGE.

Para o cálculo da prevalência do diabetes, foram selecionados indivíduos entre 20 e 65 anos do banco de dados PNS e com informação a respeito de diabetes auto referido pela variável Q03001, que diz respeito à pergunta "*Algum médico já lhe deu diagnóstico de diabetes?*". Para esse cálculo foram desconsideradas como sendo diabetes as pessoas que responderam de forma positiva a pergunta Q03002 – "Esse diabetes ocorreu durante apenas durante o período de gravidez?". Ou seja não consideramos no cálculo de prevalência casos de diabetes gestacional.

Como a resposta desta pergunta é dicotômica a prevalência da diabetes foi estimada por regressão logística por idade e sexo. A prevalência por idade foi então multiplicada pela projeção da população em 2019 fornecida pelo IBGE (IBGE, 2018) para a estimação do número absoluto de indivíduos com diabetes na população brasileira para cada idade e sexo.

A taxa de mortalidade para indivíduos com e sem diabetes foi calculada conforme descrito na seção 2.9. A mortalidade total da população por idade e sexo foi obtida pelo IBGE e a razão de taxa de mortalidade foi estimada com os dados do ELSA-Brasil (BRACCO, 2019).

3.3 Índice de Produtividade

O índice de produtividade é calculado como a soma da proporção de absenteísmo e de presenteísmo estimado e possui valores entre 0 (completamente improdutivo) e 1 (100% produtivo).

Para estimar ambos os parâmetros (absenteísmo e presenteísmo de pessoas com diabetes) utilizou-se os seguintes passos:

1. Banco de dados utilizado: PNS;
2. Seleção de variáveis;
3. Modificação/criação de variáveis;
4. Definição do design;
5. Criação do modelo;
6. Predição marginal;
7. Diagnóstico do modelo.

Como o interesse é a população de pessoas com diabetes, selecionou-se apenas os dados da PNS referente aos indivíduos e excluiu-se aqueles com *missings* na variável Q03001 (que foi renomeada como DM), que responde à pergunta: "*Algum médico já lhe deu o diagnóstico de diabetes?"*. Todas as análises deste trabalho foram realizadas utilizando a linguagem R versão 4.2.1, pelo RStudio V. 2022.12.0+353.

Para as demais seleções de observações para análise, não houve exclusão dos dados e sim foi utilizada a função *subset* do R. As análises deste trabalho foram realizadas considerando apenas os indivíduos presentes na força de trabalho e ocupados.

Para isso, foi necessário criar a variável referente a ocupação – *Labor*, cruzando-se duas outras variáveis: VDE001 (condição em relação a força de trabalho; 1 = sim ; 2 = não) e VDE002 (condição de ocupação nas últimas duas semanas ; 1 = sim ; 2 = não), conforme pode ser visto na Figura 6:

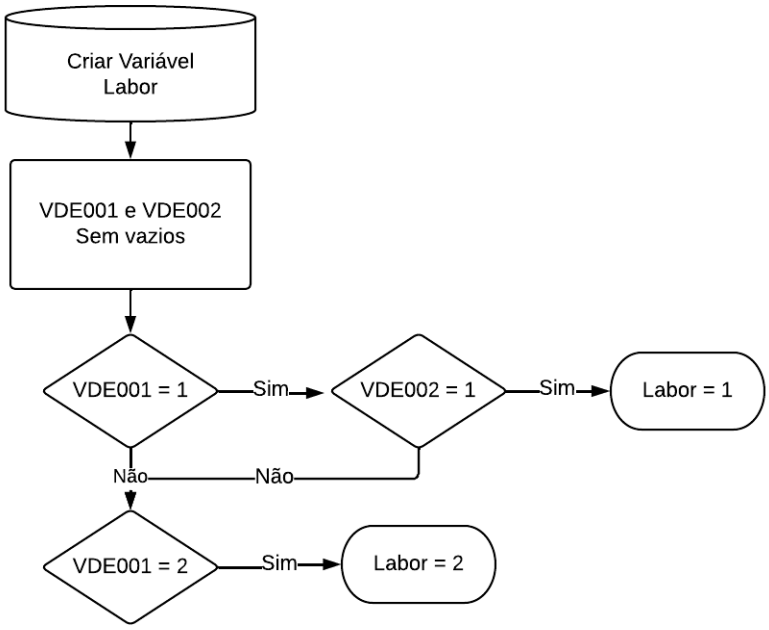

Figura 6: Criação da variável Labor.

Por ser de interesse apenas as pessoas que estão com trabalho, foram consideradas trabalhando (*Labor* = 1), aqueles que estão na força de trabalho (VDE001 = 1) e que estão ocupadas (VDE002 = 1) na semana da pesquisa, nos outros casos, considerou-se que o indivíduo está fora da força de trabalho (*Labor* = 2).

3.3.1 Absenteísmo

A metodologia que será descrita foi utilizada para estimar o absenteísmo para os indivíduos com e sem diabetes, separadamente para homens e mulheres. Para ambos os sexos se utilizou a mesma metodologia, com exceção do filtro de idade que será explicado mais adiante.

3.3.1.1 Seleção e criação de variáveis

Por se tratar de um termo que remete à ausência do trabalhador no posto de trabalho, precisou-se encontrar uma pergunta que indicasse esse fato, para isso utilizou-se duas variáveis J002 e J003, que perguntam respectivamente:

- *"Nas duas últimas semanas, ___ deixou de realizar quaisquer de suas atividades habituais (trabalhar, ir à escola, brincar, afazeres domésticos etc.) por motivo da própria saúde";*

- "*Nas duas últimas semanas, quantos dias _____ deixou de realizar suas atividades habituais (trabalhar, ir à escola, brincar, afazeres domésticos etc.), por motivo da própria saúde*".

Com essas duas perguntas, criou-se a variável resposta, *Dias* da seguinte forma:

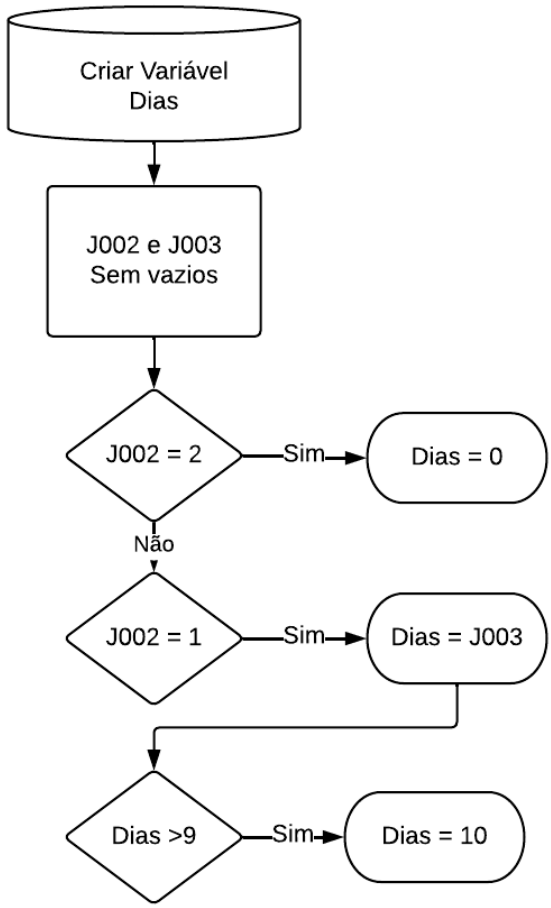

Figura 7: Diagrama com a criação da variável "Dias".

Para criar a variável *Dias*, primeiro selecionou-se apenas os dados com J002 e J003 sem *missings*, após verificou-se se a variável J002 era 1 (não deixou de realizar atividades), ou 2 (deixou de realizar atividades), então definindo o valor de *Dias*. No entanto definiu-se que o número máximo de dias úteis em duas semanas é 10, apesar de haver tipos de trabalho com 6, ou 7 dias úteis por semana, preferiu-se considerar o convencional de "segunda a sexta-feira", dessa forma atribuímos como 10 dias faltantes para todos que responderam 10, 11, 12 ou 14 dias.

Demais variáveis que serão consideradas no modelo:

- V00291 – Peso do morador selecionado com calibração;
- C008 – Idade do morador;
- C009 – Cor/raça;
- VDD004A – Nível de instrução mais elevado;
- VDF002 – Rendimento domiciliar.
- E01401 – Tipo do trabalho (doméstico, público, privado, próprio, militar)
- E01602 – Rendimento bruto mensal do trabalho

A variável *escolaridade* foi recategorizada da seguinte forma:

- VDD004A = 1 e 2, *escol* = "F";
- VDD004A = 3 e 4, *escol* = "M";
- VDD004A = 5, 6 e 7, *escol* = "S";

Onde F, M, S correspondem à: fundamental (incompleto ou completo), médio (incompleto ou completo) e superior (incompleto e completo) respectivamente.

A variável *raça*, também precisou ser recategorizada. Temos a variável C009, que diz respeito à raça. Para simplificar o estudo, e devido a pequena amostragem de indivíduos que não sejam brancos ou negros, simplificou-se a variável. Assim a *raça* é igual a "branca" quando C009 é 1 e "Outra" para outros valores de C009.

Definimos *raça* igual a "branca" quando C009 é 1 e "Outra" para outros valores de C009.

Foi criada também a variável IMC (índice de massa corporal), que consiste no quociente do peso do indivíduo (variável P00104) e a altura do mesmo (variável P00404) ao quadrado.

Em relação a variável E01401, foram excluídas as pessoas que eram trabalhadores não remunerados/estagiários/em treinamento.

Definição do design

Conforme descrito no capítulo 2.9, o pacote *Survey* do R realiza a modelagem de dados com amostras complexas. Assim, um dos componentes necessários para realizar essa análise é definir o design, utilizando a função *svydesign*.

Então definimos 3 parâmetros, o id (clusters), os estratos e os pesos:

pnsd <- svydesign(id=~UPA_PNS, strata=~V0024, weights=~V00291, data=pns)

- ID – será utilizado a váriavel UPA_PNS, que corresponde ao código da UPA;
- Estratos – V0024, que corresponde aos estratos utilizados na PNS;

- Peso – V00291, que é o peso do morador selecionado com correção de não entrevista com calibração pela projeção de população para morador selecionado.

Após a criação do objeto do design, foi realizado o filtro relacionado à idade. Como busca-se apenas as pessoas dentro da força de trabalho, filtrou-se para análise aqueles com idades entre 20 anos e 60 anos (caso das mulheres) e nos homens aqueles com idades entre 20 anos e 65 anos.

Criação do modelo

Para criar o modelo utilizou-se a função *svyglm* do pacote *Survey.* Foram obtidos modelos diferentes de acordo com o sexo. Para o sexo femino, o modelo final foi:

svyglm(Dias~C008 + as.factor(raça) + as.factor(escol) + VDF002 + IMC + as.factor(E01401) + E01602, design=subF) Eq. 3.3.1

Já para o sexo marculino, o modelo foi:

svyglm(Dias~C008 + as.factor(raça) + as.factor(escol) + VDF002 + IMC + as.factor(tipo) + E01602, design=subM) Eq. 3.3.2

A diferença entre os dois modelos ocorreu devido a colinearidade das variáveis para o sexo masculino. No sexo masculino a variável *E01401* foi agrupada da seguinte forma:

- E01401 =1 , tipo = ‘doméstico’,
- E01401 = 2 ou E1401 =4, tipo = ‘público ou militar’,
- E01401 = 3, tipo = ‘privado’
- E01401 = 6 ou E1401 = 6, tipo = ‘próprio’

A função *as.factor* foi utilizada para indicar que a variável deve ser considerada categórica no modelo.

Predição marginal

Nota-se que no modelo considerado anteriormente, não há a presença da variável que indica o diabetes - DM. Isso porque essa variável será considerada para encontrar a média da predição marginal do modelo para um fator, utilizando a função *svypredmeans*. Dessa forma, calculamos a média de dias de trabalhos perdidos (absenteísmo) entre os indivíduos com e sem diabetes. Se a média de dias perdidos foi 0.4, de um total de 10 dias, estimamos que o absenteísmo foi de 0.04 (ou 4%).

Diagnóstico do modelo

Para o diagnóstico dos resíduos, utilizou-se o modelo definido anteriormente, porém considerando no modelo a variável DM.

svyglm(Days~ DM + C008 + as.factor(raca) + as.factor(escol) + VDF002 + IMC + as.factor(tipo) + E01602, design=subM) Eq. 3.3.2

Com a função *svyhist* construiu-se o histograma para verificar a normalidade dos resíduos,

Para determinar a homocedasticidade utilizou-se a função *svystdres* para obtenção dos resíduos padronizados, considerando os mesmos estratos e ID definidos no design.

Para definir se há multicolinearidade utilizou-se a função *svyvif* do pacote *survey.* Se o resultado apresentar um valor elevado, há multicolinearidade.

A análise de outliers e observações influentes, utilizou-se a função *svydffits.* Após realizou-se o cálculo do modelo retirando-se as observações indicadas como outliers ou influentes. E comparou-se os resultados dos dois modelos, com as observações influentes e sem as observações influentes.

3.3.2 Presenteísmo

Assim como no absenteísmo, a estimativa do presenteísmo foi calculada para aqueles com e sem diabetes, separadamente para homens e mulheres. Para ambos os casos se utilizou a mesma metodologia, com exceção do filtro de idade que foi explicado no capítulo 3.3.1, que explica a metodologia do absenteísmo.

Seleção e criação de variáveis

Todas as variáveis selecionadas, criadas e/ou filtradas explicadas na seção 3.3.1.1 foram utilizadas também no cálculo do presenteísmo, com exceção da variável *Dias*.

Para o cálculo do presenteísmo entre indivíduos com e sem diabetes criou-se a variável resposta que indica a limitação dos indivíduos sendo 0 = sem limitações e 1 = limitação máxima (100%). Nomeamos essa variável de *Limitpct.* O valor da variável depende de diversas condições conforme a figura 8.

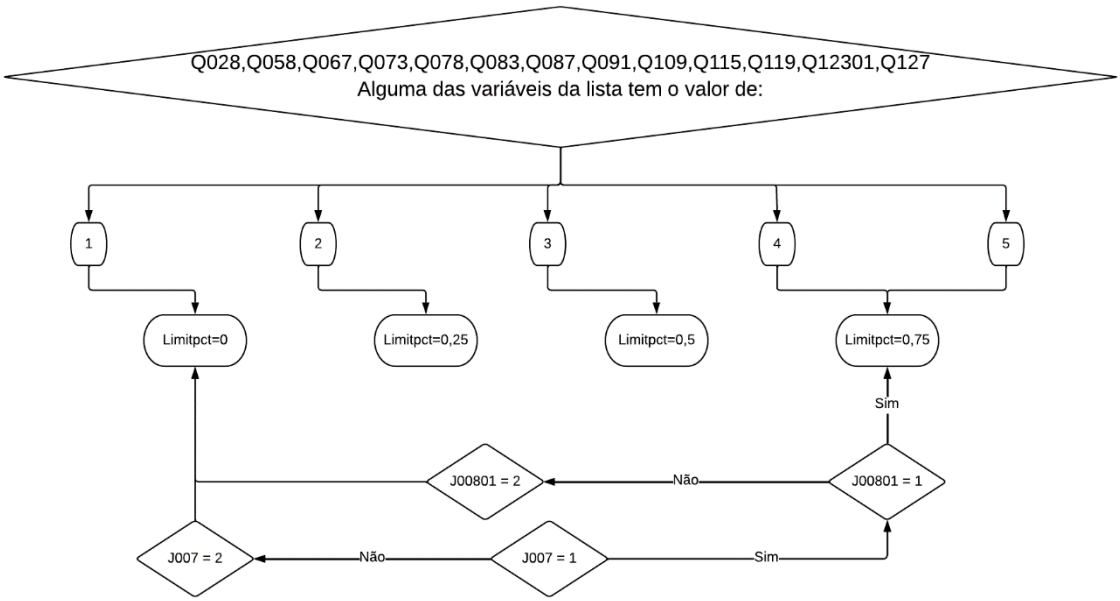

Figura 8: Diagrama de criação da variável *Limitpct*.

A figura 7 mostra que a variável *Limitpct* pode ter 4 valores, 0,0.25,0.5 e 0.75. Essa variável representa a limitação que as doenças crônicas podem causar de acordo com a resposta (que varia de 1 a 5) de qualquer doença da lista que aparecem na figura em questão. As doenças que aparecem na lista são: hipertensão, doença do coração, AVC, diabetes, asma, artrite, dores na coluna, DORT/LER, depressão e saúde mental, doença pulmonar, câncer, doença renal e outras. Considerou-se apenas 4 níveis para essa variável porque a frequência de respostas indicando limitação máxima (=1) foi muito baixa.

Outra condição definiu o *Limitpct* igual a 0.75. Se o participante não respondeu às perguntas específicas, mas respondeu de forma positiva duas perguntas mais genéricas: "*Algum médico já deu o diagnóstico de alguma doença crônica, física ou mental, ou doença de longa duração (de mais de 6 meses de duração)*?" e "*Alguma dessas doenças limita de alguma forma suas atividades habituais (trabalhar, ir à escola, brincar, afazeres domésticos, etc.)*".

O design definido para a estimativa do presenteísmo foi o mesmo utilizado para o absenteísmo.

Criação do modelo

Para estimar o presenteísmo, temos um modelo muito parecido com o utilizado para o absenteísmo, porém com a variável resposta diferente (aqui consideramos a variável *Limitpct*). Neste caso também têm-se modelos diferentes para homens e mulheres. Para as mulheres o modelo ficou:

svyglm(Limitpct ~C008 + as.factor(raça) + as.factor(escol) + VDF002 + IMC + as.factor(E01401) + E01602, design=subF) Eq. 3.3.1

Já para o sexo marculino, o modelo foi:

svyglm(Limitpct ~C008 + as.factor(raça) + as.factor(escol) + VDF002 + IMC + as.factor(tipo) + E01602, design=subM) Eq. 3.3.2

A predição marginal e os diagnósticos foram realizados de forma análoga ao utilizado para o absenteísmo. A partir da função *svypredmeans* estimamos a média do presenteísmo em termos percentuais para os indivíduos com e sem diabetes. Por exemplo, se a média estimada foi de 0.14, consideramos que o presenteísmo foi de 14%.

3.3.3 Cálculo do Índice de Produtividade

Com os valores de absenteísmo (transformado em percentual) e de presenteísmo (já obtido em percentual), o índice de produtividade (P_{index}) foi calculado da seguinte forma.

P_{index} = 1 – Absenteísmo - Presenteísmo

Calculou-se o P_{index}, para ambos os sexos de forma separada (P_{indexM} e P_{indexF}) e também para os casos com, e sem diabetes ($P_{indexMCDM}$ e $P_{indexMSDM}$). Calculou-se então 4 P_{index} diferentes:

- índice de produtividade masculino com diabetes;
- índice de produtividade masculino sem diabetes;
- índice de produtividade feminino com diabetes;
- índice de produtividade feminino sem diabetes;

Com esses dados calculou-se a redução percentual, já que considera-se o P_{index} sem diabetes como referência igual a 1.

$P_{indexMascDM} = 1 - \frac{P_{indexMascSemDM} - P_{indexMascComDM}}{P_{indexMascSemDM}}$, índice de produtividade masculino com diabetes.

A mesma variável foi calculada para indivíduos do sexo feminino.

3.3.4 PALY - Anos de vida Ajustados pela Produtividade

Com os dados calculados e obtidos criou-se uma tabela de vida, com os seguintes dados para cada sexo e para todas as idades entre 20 e 60 anos (mulhres) e entre 20 e 65 anos (homens):

- Sexo;
- Idade (20 – 60 mulheres, 20 – 65 homens) ;
- Taxa de mortalidade,
- Prevalência da diabetes;
- População;
- Total de indivíduos com diabetes: População x Prevalência da diabetes;
- Mortalidade dos indivíduos sem diabetes:
- Mortalidade dos indivíduos com diabetes
- P_{index} sem diabetes (igual a 1);
- $P_{indexMascDM}$ quando for masculino e $P_{indexFemDM}$ quando for feminino.
- PIB por trabalhador, estimativa fornecida pelo World Bank (WORLD BANK, [s.d.]), iniciando com a estimativa de 2019 para a idade *i* e considerando um aumento anual de 1.3% a cada idade *i+1*;

Com essas informações criou-se tabelas de vida separadas iniciando em diferentes idades a cada cinco anos (formando um ciclo). Assim criou-se uma tabela iniciando em 22 anos, uma iniciando em 27, assim por diante até criar a última para o sexo feminino iniciando em 57, e para o sexo masculino iniciando em 62. Em cada tabela foram calculadas para cada ciclo de idade *i*:

- Número de mortes no ciclo (mortalidade x população com diabetes) – calculado separadamente utilizando-se a mortalidade de indivíduos com e sem diabetes (dois cenários);
- Anos de vida da população: média entre o número de indivíduos no início do ciclo e número de indivíduos vivos ao final do ciclo para cada cenário;
- PALY: anos de vida da população multiplicado pelo PALYindex (separadamente para os cenários com e sem diabetes) – calculado sem desconto e considerando um desconto de 3% ao ano, de acordo com as recomendações da OMS. Esse desconto é considerado para diminuir o peso da estimativa do PALY para os anos seguintes;

- Anos de vida produtivos perdidos devido a presença do diabetes: diferença entre os PALY (com e sem diabetes);
- PIB perdido: é estimativa do custo de cada ano de produtividade perdida devido ao diabetes calculada por *PIB por trabalhador x Diferença entre os PALY*;

A figura 5, do (cap 2.10) ilustra os passos para o cálculo do PALY em cada ciclo.

Para o cálculo do desconto do PALY considerou-se uma redução de 3% ao ano, de acordo com o ciclo que o indivíduo está, conforme a equação abaixo (EDEJER et al., 2003):

$$PALY_{Desc} = \frac{PALY}{(1+0{,}03)^{ciclo-1}} \qquad \text{Eq. 3.3.5}$$

O desconto foi aplicado para o PALY calculado nos dois cenários: com e sem diabetes.

Os cálculos foram repetidos sem considerar o aumento anual do PIB para estimativas mais conservadoras.

4 Resultados e discussão

Conforme as etapas descritas no capítulo 3, a mortalidade e a taxa de mortalidade com e sem diabetes, foram obtidas anteriormente (BRACCO, 2019). Com relação à prevalência o resultado obtido pode ser visto na Tabela 2.

Tabela 2:Prevalência do diabetes.

Faixa Etária	Sexo	População	Prevalência do diabetes (%)	Pessoas com diabetes
20-24	Feminino	8575788	1,294122	110981,2
25-29	Feminino	8519370	1,837104	156509,7
30-34	Feminino	8708998	2,599203	226364,5
35-39	Feminino	8611601	3,668013	315874,6
40-44	Feminino	7854763	5,14435	404076,5
45-49	Feminino	6961901	7,195792	500963,9
50-54	Feminino	6500061	9,962178	647547,6
55-60	Feminino	6881146	14,03777	965959,7
20-24	Masculino	8744065	0,803416	70251,21
25-29	Masculino	8485534	1,27423	108125,2
30-34	Masculino	8549320	1,915663	163776,1
35-39	Masculino	8260741	2,823003	233201
40-44	Masculino	7400750	4,08001	301951,4
45-49	Masculino	6472175	5,635839	364761,4
50-54	Masculino	5965268	8,165136	487072,2
55-59	Masculino	5186871	11,6876	606220,5
60-65	Masculino	4936078	19,60691	967812,5

Pela Tabela 2, a prevalência do diabetes ela foi crescente de acordo com a idade, e também foi maior para as mulheres quando comparadas com os homens, para todas as idades (de 20 a 60 anos). A maior diferença encontrada foi para a idade de 20 anos, no qual houve uma diferença de 64% entre mulheres e homens (1,1% e 0,6%) e a menor diferença foi aos 60 anos com 13 % (16% e 14%).

Para a razão da taxa de mortalidade (estimada utilizando-se o ELSA-Brasil) reduziu com a idade e obteve-se valores entre 1,87 e 2,70 para o sexo feminino e 1,96 e 2,78 para o sexo masculino.

4.1 Absenteísmo

4.1.1 Indivíduos do Sexo Feminino

Após realizar a metodologia descrita no capítulo 3.3, no qual variáveis foram criadas, recategorizadas e filtradas, resultou em um banco de dados com 17990 mulheres, com idade entre 20 e 60 anos. Fazendo a ponderação pelos pesos calibrados da PNS a grande maioria 95%, não tem diabetes e 5%, tem diabetes.

A variável resposta criada para a estimativa do absenteísmo (*Days*) ficou distribuída da seguinte forma, ver tabela 3.

Tabela 3:Distribuição da variável resposta *Days*.

Days	0	1	2	3	4	5	6	7	8	9	10
Quantidade	16115	351	348	293	119	118	24	222	49	4	347
Quantidade Ponderada (%)	90,6	2,0	1,6	1,4	0,6	0,6	0,1	1,1	0,3	0,0	1,7

De acordo com a tabela tabela 3 tem-se que 90,6% dos indivíduos não apresentaram faltas ao trabalho nas últimas duas semanas devido à saúde, e apenas 1.7% apresentaram 10 ou mais dias de ausência nas últimas duas semanas devido à saúde.

Com a regressão linear estimou-se o absenteísmo dos indivíduos com e sem diabetes. O resultado pode ser visto na tabela 4.

Tabela 4:Resultado do absenteísmo para o sexo feminino.

	Média	Desvio padrão
Com Diabetes	0,769	0,003
Sem Diabetes	0,405	0,0164

Pela tabela 4, observamos um absenteísmo de 7,7% nos indivíduos com diabetes e 4% nos indivíduos sem diabetes. Pensando que um ano tem 240 dias úteis isso nos leva a estimar que a presença do diabetes ocasionaria aproximadamente 9 dias a mais de falta ao trabalho.

Os pressupostos de normalidade dos resíduos, a homocedasticidade, multicolinearidade, presença de outliers e observações influentes foram testados, validados, e serão discutidos posteriormente.

4.1.2 Indivíduo do Sexo Masculino

Após utilizar os filtros para os indivíduos do sexo masculino restou um banco de dados com 22674 homens, com idades entre 20 e 60 anos. Considerando os pesos obtidos pela PNS, 95,5% dos homens não têm diabetes, e 4,5% têm diabetes. Desses casos a varíavel resposta *Dias* para o sexo masculino ficou distribuida da seguinte forma:

Tabela 5:Distribuição da variável resposta *Diass*.

Dias	0	1	2	3	4	5	6	7	8	9	10
Quantidade	21339	209	248	186	79	85	21	145	30	2	330
Quantidade Ponderada (%)	94,8	0,9	0,8	0,7	0,3	0,3	0,1	0,6	0,1	0	1,5

De acordo com a tabela tabela 5 tem-se que 94,8% dos indivíduos não apresentaram faltas ao trabalho nas últimas duas semanas devido à saúde, e apenas 1.5% apresentaram 10 ou mais dias de ausência nas últimas duas semanas devido à saúde.

Observa-se valores aproximados nos percentuais encontrados para ambos os sexos (90,6% e 94% para *Days* = 0 e 1,7% e 1.5% para *Days* = 10).

Com a regressão linear estimou-se o absenteísmo dos indivíduos com e sem diabetes. O resultado pode ser visto na tabela tabela 6.

Tabela 6: Resultado do absenteísmo para o sexo masculino.

	Média	Desvio padrão
Com Diabetes	0,649	0,015
Sem Diabetes	0,253	0,126

Pela tabela 6, encontra-se que o absenteísmo para os indivíduos com diabetes é mais que o dobro do absenteísmo dos indivíduos sem diabetes, 6,5% e 2,5% respectivamente. Diferente do sexo feminino, a diferença de ausência ao trabalho foi de 6 dias entre os homens com e sem diabetes, 3 dias a menos que o sexo feminino..

Os pressupostos de normalidade dos resíduos, a homocedasticidade, multicolineariedade, presença de outliers e observações influentes foram testados, validados, e serão discutidos posteriormente.

4.2 Presenteísmo

4.2.1 Indivíduos do Sexo Feminino

Após realizar a metodologia descrita no capítulo 3.3, no qual variáveis foram criadas, recategorizadas e filtradas, resultou em um banco de dados com 17990 mulheres, com idade entre 20 e 60 anos. Desses casos a variável resposta *Limitpct* ficou distribuída da seguinte forma:

Tabela 7: Distribuição da variável *Limitpct* para o sexo feminino.

Limitpct	0	0,25	0,5	0,75
Quantidade	12130	2229	1402	2229
Quantidade Ponderada (%)	68.6%	12.2%	7.7%	11.5%

Nota-se que ponderando pelo peso, 68,6%, foram considerados sem limitação de acordo com as respostas no questionário e 11,5% foram consideradas com o nível de limitação 0,75.

Com a variável *Limitpct*, conseguiu-se criar o modelo e calcular a média da predição marginal do modelo para o fator diabetes, obtendo o presenteísmo para o sexo feminino, conforme pode ser visto na tabela 8.

Tabela 8: Resultado do presenteísmo para o sexo feminino.

	Média	Desvio padrão
Com Diabetes	0,244	0,003
Sem Diabetes	0,151	0,0164

O presenteísmo dos indivíduos do sexo feminino sem diabetes foi estimado em aproximadamente 15%, já quando considerou-se a presença de diabetes, houve um aumento relativo de 61%, indo para 24%.

Os pressupostos de normalidade dos resíduos, a homocedasticidade, multicolinearidade, presença de outliers e observações influentes foram testados, validados, e serão discutidos posteriormente.

4.2.2 Indivíduos do Sexo Maculino

Após utilizar os filtros para os indivíduos do sexo masculino temos um banco de dados com 22674 homens, com idades entre 20 e 60 anos. Desses casos a variável resposta *Limitpct* ficou distribuída da seguinte forma:

Tabela 9: Distribuição da variável *Limitpct* para o sexo masculino.

Limitpct	0	0,25	0,5	0,75
Quantidade	17394	2135	1180	1965
Quantidade Ponderada (%)	78%	9,2%	4,6%	8,2%

A distribuição de indivíduos do sexo masculino de acordo com a variavel resposta ficou diferente da encontrada para os individuos do sexo feminino. Utilizando-se os pesos para ponderar 78% dos indivíduos são classificados como sem limitação (*Limitpct = 0*), e 8,2% foram considerados com o máximo de limitação (*Limitpct = 0,75*).

Com a variável *Limitpct*, conseguiu-se criar o modelo e calcular a média da predição marginal do modelo para o fator diabetes, obtendo o presenteísmo para o sexo masculino, conforme pode ser visto na tabela 10.

Tabela 10: Resultado do presenteísmo para o sexo masculino.

	Média	Desvio padrão
Com Diabetes	0,202	0,003
Sem Diabetes	0,103	0,016

O presenteísmo dos indivíduos com diabetes (20%), foi aproximadamente o dobro do presenteísmo dos indivíduos sem diabetes (10%) .

Ao comparar o resultado do presenteismo masculino com o feminino, nota-se valores menores para o sexo masculino, tanto nos indivíduos com e sem diabetes, que provavelmente deve-se ao fato da prevalência ser maior para o sexo feminino.

Os pressupostos de normalidade dos resíduos, a homocedasticidade, multicolineariedade, presença de outliers e observações influentes foram testados, e serão discutidos posteriormente.

4.3 Análise dos resíduos

Os resíduos dos quatro modelos foram testados, de forma geral apresentaram resultados similares entre eles. Como todos modelos apresentaram resultados similares, apresentando a mesma peculiaridade na distribuição dos resíduos, gráficos parecidos dos resíduos padronizados, valores do teste VIF e resultado do teste DFFITS similares, será apresentado figuras apenas do resultado de um modelo. Para entender a distribuição dos resíduos utilizou-se o histograma. Os resíduos do modelo de presenteísmo para as mulheres pode ser visto na figura 9.

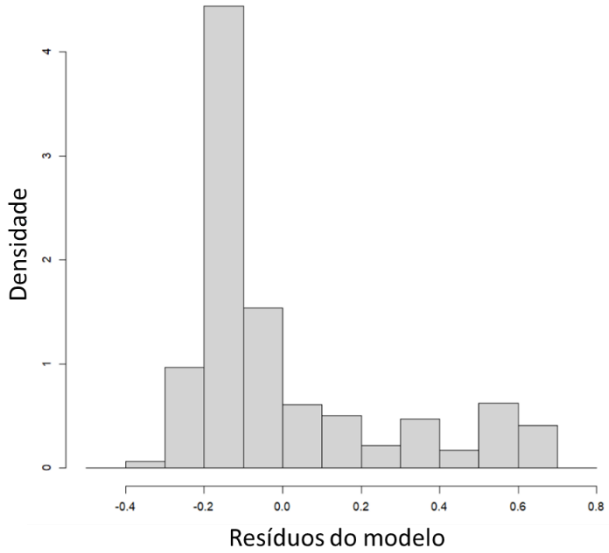

Figura 9: Histograma dos resíduos para o modelo de presenteísmo para o sexo feminino.

Nota-se pela figura 9, que há uma cauda nos resíduos positivos, apresentando uma densidade maior do que a esperada nos resíduos acima de 0,3, mostrando elementos nos resíduos até 0,6. A distribuição dos resíduos para os outros 3 modelos, também apresentaram a mesma peculiaridade na distribuição e em valores similares. Para melhor analisar os resíduos, fez-se dois gráficos com os resíduos apresentados na figura 10.

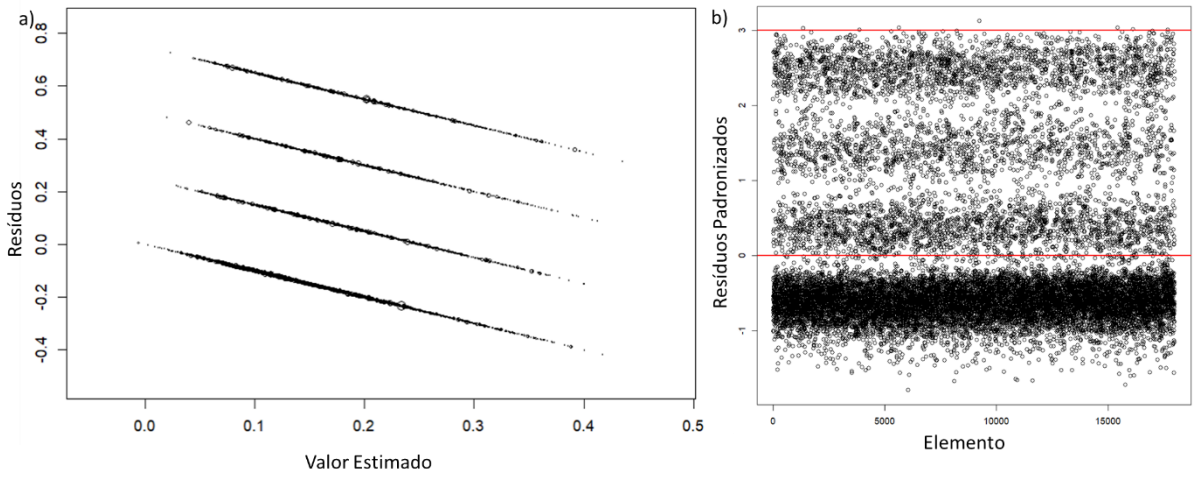

Figura 10: a) Gráfico dos resíduos pelo valor estimado; b) resíduos padronizados para cada elemento.

Conforme mostrado pelo histograma, na figura 10 temos 4 faixas de resíduos para cada valor observado. Pelos resíduos padronizados tem-se também uma distribuição maior dos resíduos para valores acima de 0. Apesar dessa distribuição não simétrica, apenas 1 elemento ficou com valor acima da linha de 3 desvios padrões.

A multi colinearidade também foi testada utilizando a metodologia VIF apresentada no capítulo 2.5.6, e para todos os modelos apresentou valores entre 1 e 2,7 em todas as covariáveis, indicando que não há correlação entre as variáveis consideradas nos modelos.

No teste da presença de outliers, os quatro modelos mostraram uma quantidade de valores considerados valores extremos próximos a 200. Esses indivíduos apresentaram pouca idade, alta renda, baixo IMC, e consequentemente apresentaram um valor baixo de predição, porém reportaram uma alta limitação ou um número elevado de dias ausentes do trabalho. Apesar do teste SDFFIT mostrar que há valores extremos, decidiu-se manter esses indivíduos no banco de dados, justamente porque esses indivíduos não podem ser considerados como um "problema" de medição, são pessoas que apresentaram peculiaridades e existem na população. A magnitude da diferença média do absenteísmo e presenteísmo estimado entre indivíduos com e sem diabetes ficou muito similar sem e com a exclusão dessas observações faltantes.

4.4 Cálculo do PALY

Com os resultados de absenteísmo e presenteísmo, calculou-se os P_{index} com e sem diabetes para ambos os sexos,

Tabela 11: P_{index} para ambos os sexos, dos indivíduos com e sem diabetes.

	Sexo	P_{index}	Pindex Diabetes (1-redução percentual)
Com Diabetes	Fem	0,68	0,84
Sem Diabetes	Fem	0,81	
Com Diabetes	Masc	0,73	0,84
Sem Diabetes	Masc	0,87	

Como esperado, conforme pode ser visto na tabela 11, os P_{index} para os indivíduos do sexo feminino foram menores que os respectivos P_{index} para os indivíduos do sexo masculino, devido aos maiores valores de presenteísmo e absenteísmo. A redução percentual entre o índice para os indivíduos com diabetes quando comparado com aqueles sem diabetes foi a mesma para os dois sexos. Dessa forma, consideramos o Pindex dos indivíduos sem diabetes como 1 e para os indivíduos com diabetes como 0,84 tanto para homens quanto para mulheres. O Pindex específico foi então multiplicado aos anos de vida da população com e sem diabetes para encontrarmos a redução dos nos anos de vida produtivos (PALY) devido ao diabetes.

Os anos de vida perdidos para ambos os sexos podem ser vistos na tabela 12.

Tabela 12: Anos de vida vividos.

Idade	Mortes com Diabetes	Anos de vida vividos com diabetes	Mortes sem Diabetes	Anos de vida vividos sem diabetes	Excesso de mortes	Aumento nas mortes (%)	Anos de vida Perdidos	Redução nos anos de vida vividos (%)
Mulheres								
20-24	3051	823044	1834	838169	1217	39,9	15125	1,8
25-29	4111	997294	2478	1015764	1633	39,7	18471	1,8
30-34	5719	1231856	3457	1254757	2263	39,6	22901	1,8
35-39	7555	1424483	4579	1450740	2976	39,4	26257	1,8
40-44	9555	1552015	5808	1579655	3747	39,2	27640	1,7
45-49	9332	1290458	5690	1311094	3642	39,0	20636	1,6
50-54	9213	1068469	5634	1081585	3579	38,9	13116	1,2
55-60	6028	625241	3694	629765	2334	38,7	4524	0,7
Total	**54565**	9227768	**33174**	**9385965**	**21391**	**39,2**	**158198**	**1,7**
Homens								
20-24	6076	510302	3176	566968	2901	47,7	56666	10,0

25-29	8810	704101	4618	773241	4192	47,6	69140	8,9
30-34	12619	928998	6627	1014223	5992	47,5	85225	8,4
35-39	17161	1136393	9022	1236534	8139	47,4	100141	8,1
40-44	21154	1239217	11124	1343709	10030	47,4	104492	7,8
45-49	22661	1158597	11903	1249377	10758	47,5	90780	7,3
50-54	26245	1160459	13739	1238927	12506	47,6	78469	6,3
55-60	24857	939079	12925	986043	11932	48,0	46964	4,8
60-64	15253	526063	7839	540817	7414	48,6	14753	2,7
Total	**154835**	**8467212**	**80973**	**9147067**	**73862**	**47,7**	**679855**	**7,4**

Pela tabela 12, pode-se ver que a presença de diabetes representou nos homens um maior aumento de morte que nas mulheres (47,7% e 39,2% respectivamente), assim como nos anos de vida populacionais perdidos: 7,4% para os homens e 1,7% para as mulheres com diabetes.

4.5 Custo do Diabetes

Com a redução do PALY calculada, calculou-se os custos do diabetes em diferentes faixas de idades, para ambos os sexos. Para mostrar esse resultado, utilizou-se a tabela de vida. Foram realizados 4 cenários diferentes, considerando-se um desconto anual no PALY de 3% e aumento anual do PIB de 1,3%, para ambos os sexos.

As tabelas 13 e 14, mostram os calculos da redução de PIB para o sexo feminino considerando-se dois casos, com desconto anual de 3% no PALY e sem desconto, utilizando-se o aumento anual do PIB.

Tabela 13: Cálculo do PIB utilizando o PALY sem desconto e com aumento anual do PIB de 1,3%, para o sexo feminino.

Idade	POP com DM	PALY DM	PALY S/ DM	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	110981	691357	838169	17,5	146812	6224860459	292312
25-29	156510	837727	1015764	17,5	178038	7734545308	261178
30-34	226365	1034759	1254757	17,5	219998	9786814969	228098
35-39	315875	1196566	1450740	17,5	254174	11572074857	192970
40-44	404077	1303693	1579655	17,5	275962	12851670723	155740
45-49	500964	1083985	1311094	17,3	227109	10813706353	116273
50-54	647548	897514	1081585	17,0	184071	8957744980	75009
55-60	965960	525203	629765	16,6	104562	4982579065	31647
TOTAL	**3328278**	**7751325**	**9385965**	**17,4**	**1634640**	**$75.140.106.155**	**$1.353.813**

Tabela 14: Cálculo da redução do PIB utilizando o PALY com desconto anual de 3% e com aumento anual do PIB de 1,3%, para o sexo feminino.

Idade	POP com DM	PALY DM desconto	PALY sem DM desconto	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	110981	423547	511120	17,1	87573	3575531837	165470
25-29	156510	544775	657903	17,2	113128	4779751247	159213
30-34	226365	715422	864609	17,3	149187	6507699715	149768
35-39	315875	880990	1065276	17,3	184286	8281772732	136518
40-44	404077	1023846	1238163	17,3	214317	9902653085	118765
45-49	500964	909609	1098857	17,2	189248	8974203319	95615
50-54	647548	806241	971055	17,0	164814	8007879085	66525
55-60	965960	502878	602912	16,6	100034	4982579065	31647
TOTAL	**3328278**	**5918218**	**7147438**	**17,2**	**1229220**	**$56.356.116.772**	**$923.832**

Para o sexo femino, o desconto anual de 3% do PALY, não resultou em uma diferença grande na redução percentual do PALY entre indivíduos com e sem diabetes (apenas 0,2% no total), no entanto a diferença no PIB perdido e no PIB percapto foi de aproximadamente 33% e 46% respectivamente.

As tabelas 15 e 16, mostram os calculos da redução de PIB para o sexo masculino considerando-se os mesmos cenários anteriores.

Tabela 15: Cálculo do PIB utilizando o PALY sem desconto e com aumento anual do PIB de 1,3%, para o sexo masculino

Idade	POP com DM	PALY DM	PALY S/ DM	Red. PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	70251	428654	566968	24,4	138314	6105574821	488997
25-29	108125	591444	773241	23,5	181796	8202587572	426440
30-34	163776	780358	1014223	23,1	233865	10777153760	371866
35-39	233201	954570	1236534	22,8	281964	13262793115	319598
40-44	301951	1040942	1343709	22,5	302767	14530253886	266118
45-49	364761	973221	1249377	22,1	276156	13517508032	210295
50-54	487072	974785	1238927	21,3	264142	10837792976	123382
55-60	606221	788826	986043	20,0	197217	10037058723	93211
60-64	967812	441893	540817	18,3	98924	5144333014	38099
TOTAL	**3303171**	**7112458**	**9147067**	**22,2**	**2034609**	**$95.343.590.620**	**$2.340.136**

Tabela 16: Cálculo do PIB utilizando o PALY com desconto anual de 3% e com aumento anual do PIB de 1,3%, para o sexo masculino.

Idade	POP com DM	PALY DM desconto	PALY sem DM desconto	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	70251	256300	330566	22,5	74266	3130273488	242753
25-29	108125	372796	477104	21,9	104307	4542411276	229391
30-34	163776	519777	663510	21,7	143734	6452489014	216957
35-39	233201	673049	859119	21,7	186070	8592045360	202416
40-44	301951	778157	993081	21,6	214924	10189007927	183046
45-49	364761	772598	983822	21,5	211223	10262769709	157137
50-54	487072	823353	1041324	20,9	217970	10837792976	123382
55-60	606221	710688	886416	19,8	175728	8929897730	82130
60-64	967812	423412	517921	18,2	94509	4913257138	36368
TOTAL	**3303171**	**5414631**	**6872122**	**21,2**	**1457491**	**$69.679.113.252**	**$1.474.711**

Diferente do sexo feminino, para o sexo masculino o desconto do PALY criou uma diferença maior na redução do PALY (1,0%). A diferença no PIB perdido e no PIB per capita foi de aproximadamente 35% e 58% respectivamente.

Os resultados do cálculo da redução do PIB sem considerar o ajuste de 1,3%. Será mostrado a seguir. Nas tabelas 17 e 18 vê-se os resultados para o sexo feminino.

Tabela 17: Cálculo do PIB utilizando o PALY sem desconto sem aumento anual do PIB de 1,3%, para o sexo feminino.

Idade	POP com DM	PALY DM	PALY S/ DM	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	110981	691357	838169	17,5	146812	5001026646	233613
25-29	156510	837727	1015764	17,5	178038	6448911808	216729
30-34	226365	1057593	1254757	15,7	197164	7567206688	175649
35-39	315875	1196566	1450740	17,5	254174	10303794716	171167
40-44	404077	1303693	1579655	17,5	275962	11782573161	142310
45-49	500964	1083985	1311094	17,3	227109	10186843797	109220
50-54	647548	897514	1081585	17,0	184071	8653620834	72288
55-60	965960	525203	629765	16,6	104562	4918704323	31239
TOTAL	**3328278**	**7774553**	**9385965**	**17,2**	**1611412**	**$66.722.139.038**	**$1.152.688**

Tabela 18: Cálculo do PIB utilizando o PALY com desconto anual de 3% sem aumento anual do PIB de 1,3%, para o sexo feminino.

Idade	POP com DM	PALY DM desconto	PALY sem DM desconto	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	110981	423547	511120	17,1	87573	2983106300	137485
25-29	156510	544775	657903	17,2	113128	4097759856	135970
30-34	226365	730679	864609	15,5	133929	5140244586	117882
35-39	315875	880990	1065276	17,3	184286	7470645065	122746
40-44	404077	1023846	1238163	17,3	214317	9150538038	109424
45-49	500964	909609	1098857	17,2	189248	8488621947	90208
50-54	647548	806241	971055	17,0	164814	7748311494	64224
55-60	965960	502878	602912	16,6	100034	4918704323	31239
TOTAL	**3328278**	**5933647**	**7147438**	**17,0**	**1213791**	**$51.152.748.410**	**$809.436**

Observando-se as tabelas 17 e 18, nota-se para o sexo femino, um resultado similar ao encontrado nas tabelas 13 e 14, isso ocorre porque o aumento do PIB ocorre de forma uniforme. Neste caso a diferença no PIB perdido e no PIB per capito foi de aproximadamente 30% e 42% respectivamente.

Nas tabelas 19 e 20 pode-se ver os resultados do cálculo da redução do PIB sem considerar o ajuste de 1,3%, para o sexo masculino.

Tabela 19: Cálculo do PIB utilizando o PALY com desconto sem aumento do PIB de 1,3%, para o sexo masculino.

Idade	POP com DM	PALY DM	PALY S/ DM	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	70251	428654	566968	24,4	138314	4711548096	373245
25-29	108125	591444	773241	23,5	181796	6585058857	189169
30-34	163776	780358	1014223	23,1	233865	8975796014	307131
35-39	233201	1006761	1236534	18,6	229773	9314597224	223819
40-44	301951	1040942	1343709	22,5	302767	12927040309	235350
45-49	364761	973221	1249377	22,1	276156	12386796711	191778
50-54	487072	974785	1238927	21,3	264142	10247328990	116250
55-60	606221	788826	986043	20,0	197217	9697254040	89803
60-64	967812	441893	540817	18,3	98924	5077602814	37599
TOTAL	**3303171**	**7165385**	**9147067**	**21,7**	**1981682**	**$82.341.336.622**	**$1.765.722**

Tabela 20: Cálculo do PIB utilizando o PALY com desconto sem aumento do PIB de 1,3%, para o sexo masculino.

Idade	POP com DM	PALY DM desconto	PALY sem DM desconto	Redução PALY (%)	Redução PALY	PIB perdido	PIB per capto com diabetes
20-24	70251	256300	330566	22,5	74266	2529798879	194300
25-29	108125	372796	477104	21,9	104307	3778238248	189169
30-34	163776	519777	663510	21,7	143734	5516534482	184089
35-39	233201	708550	859119	17,5	150569	6103823179	143341
40-44	301951	778157	993081	21,6	214924	9176480378	163952
45-49	364761	772598	983822	21,5	211223	9474288678	144416
50-54	487072	823353	1041324	20,9	217970	10247328990	116250
55-60	606221	710688	886416	19,8	175728	8640631988	79258
60-64	967812	423412	517921	18,2	94509	4851008684	35902
TOTAL	**3303171**	**5450453**	**6872122**	**20,7**	**1421669**	**$61.881.042.089**	**$1.251.617**

Análogo ao sexo feminino, para o sexo masculino (tabelas 19 e 20), um resultado similar ao encontrado nas tabelas 15 e 16. Com isso, mesmo com a redução do PALY, observa-se uma diferença grande na redução do PALY (1,0%), no entanto para este caso a diferença no PIB perdido e no PIB per capito foi de aproximadamente 32% e 41% respectivamente, uma diferença menor que a encontrada para o caso análogo das tabelas 15 e 16.

O resumo do PIB perdido total e o PIB per capito devido ao diabetes em todos os casos pode ser visto na tabela 21.

Tabela 21: Resumo dos PIB's para todos os casos discutidos.

Sexo	Desconto no PALY	Ajuste PIB	Redução PIB	Redução PIB per capto com diabetes
Fem	PALY SEM DESC	PIB COM AJUSTE	$75.140.106.155	$1.353.813
Fem	PALY DESC	PIB COM AJUSTE	$56.356.116.772	$923.832
Fem	PALY SEM DESC	PIB SEM AJUSTE	$66.722.139.038	$1.152.688
Fem	PALY DESC	PIB SEM AJUSTE	$51.152.748.410	$809.436
Masc	PALY SEM DESC	PIB COM AJUSTE	$95.343.590.620	$2.340.136
Masc	PALY DESC	PIB COM AJUSTE	$69.679.113.252	$1.474.711
Masc	PALY SEM DESC	PIB SEM AJUSTE	$82.341.336.622	$1.765.722

Masc	PALY DESC	PIB SEM AJUSTE	$61.881.042.089	$1.251.617

Nota-se que independente do ajuste do PIB, e desconto no PALY, os valores do PIB perdido são extremamente grandes, acima de U$50 bilhões. Para o caso do sexo feminino a redução estimada no PIB per capto devido ao diabetes foi de aproximadamente U$1 milhão , sendo entre U$1.3 milhões e U$809 mil, e para o sexo masculino observou-se uma redução maior do PIB per capto, próximo a U$1,5 milhões chegando a $2,3 milhões. Para o sexo masculino observou-se que a estimativa do PIB perdido chegou a 91 bilhões quando foi considerado o cenário de PALY sem desconto e com o ajuste anual no PIB.

5 Discussão

A utilização dos dados da PNS para estimar os índices de presenteísmo e absenteísmo resultou em uma redução de 16% na produtividade de indivíduos com diabetes quando comparado com aqueles sem diabetes. Em termos monetários, considerando a perda ocasionada também pelo excesso de mortalidade devido ao diabetes, estimamos de forma conservadora que na coorte de brasileiros de 2019 com idade entre 20 e 60 ou 65 anos, o diabetes será o responsável por uma perda no PIB de aproximadamente 113 bilhões de dólares, em um cenário com redução anual do PALY de 3% e sem aumento anual do PIB de 1.3%, essa perda representa em torno de 7% do PIB brasileiro de 2021.

A redução de 3% anual do PALY é uma recomendação da organização mundial da saúde para converter valores futuros de gastos ou efeitos de saúde aos seus valores presentes. Em outras palavras, essa prática define que os eventos futuros tenham um peso menor nas análises. Essa prática de desconto é controversa na literatura e por isso optamos por demonstrar os dois resultados nesse trabalho (EDEJER et al., 2003).

Quando não consideramos o desconto do PALY a perda no PIB chega a 149 bilhões se não considerarmos um ajuste anual no PIB de 1.3% e, considerando esse aumento, a perda chega a aproximadamente 170 bilhões de dólares.

Com relação ao absenteísmo,o valor estimado foi maior que o absenteísmo estimado pela *American Diabetes Associaton* - ADA para os Estados Unidos da América (EUA) para o ano de 2012, que foi de 3 dias por ano, representando aproximadamente 1,3% de redução em um ano (MAGLIANO et al., 2018).

Segundo (BOMMER et al., 2017) o absenteísmo para países desenvolvidos seria entre 2,1 e 4,3 dias (homens e mulheres respectivamente), o que corrobora o resultado encontrado por Magliano et al. (MAGLIANO et al., 2018). Para países não desenvolvidos (Boomer et al. considerou países da África e do sul da Ásia), mesmo não havendo evidências para estratificar por faixa etária e por sexo, foi estabelecido que os homens têm um absenteísmo devido ao diabetes de 1,9 dias, e as mulheres de 10,2 dias por ano, esse resultao ficou similar para o caso femino no qual encontramos 9 dias de ausência devido ao diabetes, já para o sexo masculino houve uma diferença maior, já que estimamos uma diferença de 6 dias para as pessaos com diabetes.

A estimação do presenteísmo ainda hoje é um desafio (LOHAUS; HABERMANN, 2019). No Brasil, especialmente a nível nacional, não há dados utilizando nenhum dos questionários

desenvolvidos até o momento para estimação do presenteísmo. Ao redor do mundo, análises de produtividade acabam se baseando em estimativas pré-estabelecidas como a de Goetzel et al, 2014 que estimou a partir da média de diversos questionários diferentes aplicados na população dos Estados Unidos, um presenteísmo de 11% devido ao diabetes (GOETZEL et al., 2004). No entanto, acreditamos que os dados da PNS sejam uma fonte valiosa de informação com dados da população brasileira e por isso estudamos uma adaptação desses dados para a estimação do presenteeismo, o qual ficou em aproximadamente de 10% para o sexo masculino e 9% para o sexo feminino, semelhante ao estimado por Goetlz, et al. Magliano et al. descreve que foi estimado pelo ADA uma redução na produtividade de 6,6% (MAGLIANO et al., 2018). Já Bommer et al. afirma que a queda de produtividade causada pelo presenteísmo para países desenvolvidos seria em torno de 1.6%, e para os países em desenvolvimento e subdesenvolvidos, aproximadamente 4% (BOMMER et al., 2017).

Pereda (PEREDA et al., 2022) foi um dos poucos trabalhos que se propôs a estimar o custo do diabetes no Brasil, observando um custo de US$ 2.15 bilhões para o ano de 2016, no entanto a análise foi baseada em dados de hospitalizações. Neste trabalho, foram considerados custos diretos (cerca de 30% do custo) e indiretos como: morte prematura, absenteísmo e aposentadoria precoce, sendo responsável por 70% do custo estimado. No entanto, não foram consideradas os custos causados por complicações crônicas do diabetes e por doenças relacionadas ao diabetes.

Outro estudo, que estimou o custo do diabetes em 2014 no Brasil, estimou um valor de US$15.67 bilhões (BAHIA et al., 2019). Destes US$6.89 bilhões (44%) foram causados por custo direto (custo médico), US$3.69 bilhões (23%) de custo direto, porém não médico, como custo devido a dieta e transporte para exames e consultas. Porém o custo indireto (causado pelo absenteísmo e aposentadoria precoce), foi estimado em US$3.69 bilhões (23%), destes US$1.8 bilhões foi causado pelo absenteísmo.

Apesar de ser esperados valores maiores para o custo do diabetes devido ao tempo (espera-se um aumento na prevalência do diabetes com o passar dos anos), as diferenças nos valores encontrados foi grande, isso deve-se, provavelmente, devido ao fato de considerarmos o presenteísmo como fator nos nossos cálculos (e não apenas absenteísmo), ao maior percentual de absenteísmo e presenteísmo nos indivíduos com diabetes quando em comparação com aqueles sem diabetes que observamos nos nossos dados, quando em comparação com os estimados para

população dos EUA usualmente utilizados nos demais estudos, e também porque consideramos o custo durante toda a vida da coorte de brasileiros de 2019 entre 20 e 60 ou 65 anos.

Apesar dos valores elevados, eles não se comparam com os valores obtidos por outros estudos, com metodologias similares, em outros países. No caso da China o estudo realizado por Hird (HIRD et al., 2019), estimou que cerca de 7.1% das pessoas em idade de trabalho (56.4 milhões de pessoas) têm diabetes, uma perda de anos de vida de 3.7%, com uma redução da PALY de 1.3 por pessoa, o que resultaria em uma perda de US$ 2.6 trilhões no PIB chinês devido ao diabetes. Para a Índia, assim como a China, também tem uma estimativa de mais de 50 milhões de pessoas com diabetes (54,4 milhões o que corresponde a 7.6%), tendo uma redução da PALY de 23.3%, o que representa um total de US$ 2.6 trilhões no PIB indiano (BANKER et al., 2021). Ainda, Boomer et al. estimou que o custo global do diabetes em 2015 foi de US$1.31 trilhões ou 1,8% do PIB mundial, e que é maior para países subdesenvolvidos e em desenvolvimento que para os países desenvolvidos (BOMMER et al., 2017).

6 Conclusões

Esse trabalho apresenta limitações, os dados da PNS precisaram ser adaptados para construirmos os desfechos de interesse: absenteísmo e presenteísmo. Como vimos pela análise dos resíduos, o modelo utilizado pode ser melhorado, talvez com o uso de um modelo linear generalizado considerando outra distribuição para os nossos desfechos, como por exemplo a distribuição Gama. Ainda, muitos estudos incorporam também os anos de vida perdidos devido a saída precoce dos indivíduos com diabetes da força de trabalho, o que não conseguimos estimar, apesar de existirem essas informações nos dados da PNS. Mesmo com essas ressalvas, esse trabalho demonstra a riqueza de dados da PNS e descreve de forma detalhada os procedimentos adequados para análise desses dados obtidos por amostragem complexa. Com o uso desses dados conseguimos de forma inédita construir um modelo para estimação dos anos de vida produtivos perdidos devido ao diabetes na população brasileira, enriquecendo as estimativas da carga da doença no Brasil e consequentemente nos países de baixa e média renda.

7 Trabalhos futuros

Estudar adaptações do modelo linear generalizado para melhorar o modelo utilizado na estimação do absenteísmo e presenteísmo.

Estender e incrementar o estudo com a perda de anos de vida devido a saída precoce da força de trabalho devido ao diabetes.

Utilizar dados mais atuais – quando disponíveis – pós pandemia de COVID19, para a realização dos cálculos, já que foram utilizados dados de 2019 (data do último PNS), para comparação com os resultados de 2019.

Referências

AQUINO, E. M. L. et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Objectives and Design. **American Journal of Epidemiology**, v. 175, n. 4, p. 315–324, 15 fev. 2012.

BAHIA, L. R. et al. Economic burden of diabetes in Brazil in 2014. **Diabetology & Metabolic Syndrome**, v. 11, n. 1, p. 54, dez. 2019.

BANKER, K. K. et al. The Impact of Diabetes on Productivity in India. **Diabetes Care**, v. 44, n. 12, p. 2714–2722, 1 dez. 2021.

BEST, H.; WOLF, C. (EDS.). **The SAGE handbook of regression analysis and causal inference**. Los Angeles [Calif.]: SAGE Reference, 2015.

BOMMER, C. et al. The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. **The Lancet Diabetes & Endocrinology**, v. 5, n. 6, p. 423–430, jun. 2017.

BRACCO, P. A. **Carga de Mortalidade do Diabetes, Risco de Desenvolver Diabetes ao Longo da Vida e Anos de Vida Perdidos Devido ao Diabetes na População Brasileira**. Porto Alegre: UFRGS, 2019.

BRACCO, P. A. et al. A nationwide analysis of the excess death attributable to diabetes in Brazil. **Journal of Global Health**, v. 10, n. 1, p. 010401, jun. 2020.

BRACCO, P. A. et al. Lifetime risk of developing diabetes and years of life lost among those with diabetes in Brazil. **Journal of Global Health**, v. 11, p. 04041, 3 jul. 2021.

DAY, S. M.; REYNOLDS, R. J.; KUSH, S. J. Extrapolating published survival curves to obtain evidence-based estimates of life expectancy in cerebral palsy. **Developmental Medicine & Child Neurology**, v. 57, n. 12, p. 1105–1118, dez. 2015.

EDEJER, T. T.-T. et al. **Making choices in health: WHO guide to cost-effectiveness analysis**. [s.l.] World Health Organization, 2003. v. 1

FIOCRUZ. **PNS - Bases de Dados**. Disponível em: <https://www.pns.icict.fiocruz.br/bases-de-dados/>. Acesso em: 5 mar. 2023.

FREITAS, MA. P. S. **Pesquisa Nacional de Saúde: Plano Amostral**. , 2014.

GOETZEL, R. Z. et al. Health, Absence, Disability, and Presenteeism Cost Estimates of Certain Physical and Mental Health Conditions Affecting U.S. Employers: **Journal of Occupational and Environmental Medicine**, v. 46, n. 4, p. 398–412, abr. 2004.

GOYAL, R.; JIALAL, I. Diabetes Mellitus Type 2. Em: **StatPearls**. Treasure Island (FL): StatPearls Publishing, 2022.

HIRD, T. R. et al. The impact of diabetes on productivity in China. **Diabetologia**, v. 62, n. 7, p. 1195–1203, jul. 2019.

HOD, M. et al. The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: A pragmatic guide for diagnosis, management, and care #. **International Journal of Gynecology & Obstetrics**, v. 131, p. S173–S211, out. 2015.

IBGE - Instituto Brasileiro de Geografia e Estatística. Disponível em: <https://www.ibge.gov.br/>.

IBGE - Instituto Brasileiro de Geografia e Estatística., **Projeções da População**. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?=&t=resultados>. Acesso em: 25 jul. 2022.

IBGE - Instituto Brasileiro de Geografia e Estatística., Pesquisa Nacional de Saúde : 2019 : informações sobre domicílios, acesso e utilização dos serviços de saúde : Brasil, grandes regiões e unidades da federação. p. 85, 2020.

INTERNATIONAL DIABETES FEDERATION. IDF Atlas 10th Edition. 2021.

JACOBS, E. et al. Burden of Mortality Attributable to Diagnosed Diabetes: A Nationwide Analysis Based on Claims Data From 65 Million People in Germany. **Diabetes Care**, v. 40, n. 12, p. 1703–1709, 1 dez. 2017.

LAHIRI, S. Survival Probabilities From 5-Year Cumulative Life Table Survival Ratios (Tx+ 5/Tx): Some Innovative Methodological Investigations. Em: **Handbook of statistics**. [s.l.] Elsevier, 2018. v. 39p. 481–542.

LEOTTI, V. **Notas de Aula - Semestre 2019/2**. , 16 set. 2019. . Acesso em: 5 mar. 2023

LOHAUS, D.; HABERMANN, W. Presenteeism: A review and research directions. **Human Resource Management Review**, v. 29, n. 1, p. 43–58, mar. 2019.

LOTUFO, P. A. Construcao do Estudo Longitudinal de Saude do Adulto (ELSA-Brasil). **Revista de Saúde Pública**, v. 47, n. suppl 2, p. 3–9, jun. 2013.

LUMLEY, T. Analysis of Complex Survey Samples. **Journal of Statistical Software**, v. 9, n. 8, 2004.

MAGLIANO, D. J. et al. The Productivity Burden of Diabetes at a Population Level. **Diabetes Care**, v. 41, n. 5, p. 979–984, 1 maio 2018.

MINISTÉRIO DA SAÚDE; SECRETARIA DE ATENÇÃO À SAÚDE; DEPARTAMENTO DE REGULAÇÃO, AVALIAÇÃO E CONTROLE. **Sistema de Informações Hospitalares do SUS – SIH/SUS**. Disponível em: <https://ces.ibge.gov.br/base-de-

dados/metadados/ministerio-da-saude/sistema-de-informacoes-hospitalares-do-sus-sih-sus.html>. Acesso em: 25 jul. 2022.

MINISTÉRIO DA SAÚDE, SECRETARIA DE VIGILÂNCIA EM SAÚDE. Vigitel Brasil 2021 : vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico : estimativas sobre frequência e distribuição sociodemográfica de fatores de risco e proteção para doenças crônicas nas capitais dos 26 estados brasileiros e no Distrito Federal em 2021. 2021.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [s.l.] John Wiley & Sons, 2021.

MUZY, J. et al. Prevalência de diabetes mellitus e suas complicações e caracterização das lacunas na atenção à saúde a partir da triangulação de pesquisas. **Cadernos de Saúde Pública**, v. 37, n. 5, p. e00076120, 2021.

NAMBOODIRI, N. K.; SUCHINDRAN, C. M. **Life table techniques and their applications**. Orlando [Fla.]: Academic Press, 1987.

OCHOA, C. **Amostra probabilística: Amostra por conglomerados**. Disponível em: <https://www.netquest.com/blog/br/blog/br/amostra-conglomerados>. Acesso em: 5 mar. 2023a.

OCHOA, C. **Amostragem probabilística: Amostra estratificada**. Disponível em: <https://www.netquest.com/blog/br/blog/br/amostragem-probabilistica-amostra-estratificada>. Acesso em: 5 mar. 2023b.

ONUKWUGHA, E.; BERGTOLD, J.; JAIN, R. A Primer on Marginal Effects—Part I: Theory and Formulae. **PharmacoEconomics**, v. 33, n. 1, p. 25–30, jan. 2015.

PEREDA, P. et al. Direct and Indirect Costs of Diabetes in Brazil in 2016. **Annals of Global Health**, v. 88, n. 1, p. 14, 3 mar. 2022.

PESTANA, M. H.; GAGEIRO, J. N. Análise de dados para ciências sociais: a complementaridade do SPSS. 2008.

REDDY, M.; KAR, S. Unconditional probability of dying and age-specific mortality rate because of major non-communicable diseases in India: Time trends from 2001 to 2013. **Journal of Postgraduate Medicine**, v. 65, n. 1, p. 11, 2019.

RODACKI, M. et al. Classificação do diabetes. Em: BERTOLUCI, M. C. et al. (Eds.). **Diretriz Oficial da Sociedade Brasileira de Diabetes**. 2022. ed. [s.l.] Conectando Pessoas, 2022.

RODRIGUES, J. B. **Diabetes Mellitus Tipo 2: Percepção da Qualidade de Vida**. [s.l.] UFMA, 2017.

RODRIGUES, S. C. A. **Modelo de Regressão Linear e suas Aplicações**. Covilhã: Universidade da Beira do Interior, out. 2022.

SAYDAH, S. H. et al. Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. **Annals of epidemiology**, v. 14, n. 7, p. 507–516, 2004.

SBD, S. B. DE D. **Diretrizes da Sociedade Brasileira de Diabetes: 2014-2015**. São Paulo: AC Farmacêutica, 2015.

SEBER, G. A. F.; LEE, A. J. **Linear regression analysis**. 2nd ed ed. Hoboken, N.J: Wiley-Interscience, 2003.

SILLER, A. F. et al. Challenges in the diagnosis of diabetes type in pediatrics. **Pediatric Diabetes**, v. 21, n. 7, p. 1064–1073, nov. 2020.

SILVA, P. L. DO N.; BIANCHINI, Z. M.; DIAS, A. J. R. **Amostragem: Teoria e Prática Usando R**. [s.l: s.n.].

Sistema de Informação sobre Mortalidade – SIM. Disponível em: <https://opendatasus.saude.gov.br/dataset/sim-2020-2021>. Acesso em: 16 jul. 2022.

SOUZA, É. C. DE. **Análise de influência local no modelo de regressão logística**. Mestrado em Estatística e Experimentação Agronômica—Piracicaba: Universidade de São Paulo, 9 fev. 2006.

STOPA, S. R. et al. Pesquisa Nacional de Saúde 2019: histórico, métodos e perspectivas. **Epidemiologia e Serviços de Saúde**, v. 29, n. 5, p. e2020315, 2020.

SUN, H. et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. **Diabetes Research and Clinical Practice**, v. 183, p. 109119, jan. 2022.

VALLIANT, R.; VALLIANT, M. R. **Package 'svydiags'**. [s.l: s.n.].

WORLD BANK. **GDP per person employed - Brazil**. Disponível em: <https://data.worldbank.org/indicator/SL.GDP.PCAP.EM.KD?end=2021&locations=BR&start=1991&view=chart>. Acesso em: 24 set. 2022.

WORLD HEALTH ORGANIZATION. **Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy**. [s.l: s.n.].

WORLD HEALTH ORGANIZATION. **Diabetes**. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/diabetes>.

YANG, K.; TU, J.; CHEN, T. Homoscedasticity: an overlooked critical assumption for linear regression. **General Psychiatry**, v. 32, n. 5, p. e100148, out. 2019.

ZULIAN, L. R. et al. Qualidade de vida de pacientes com diabetes utilizando o instrumento Diabetes 39 (D-39). **Revista Gaúcha de Enfermagem**, v. 34, n. 3, p. 138–146, set. 2013.