Marcson de Azevedo Araújo

Avaliação da pobreza na Região Metropolitana do Rio de Janeiro e o efeito da formalidade entre os ocupados

Niterói - RJ, Brasil

31 de Janeiro de 2022

Universidade Federal Fluminense

Marcson de Azevedo Araújo

Avaliação da pobreza na Região Metropolitana do Rio de Janeiro e o efeito da formalidade entre os ocupados

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dr. Rafael Santos Erbisti

Co-Orientador(a): Profa. Dra. Carolina Botelho Marinho da C. Hecksher

Niterói - RJ, Brasil

31 de Janeiro de 2022

Universidade Federal Fluminense

Marcson de Azevedo Araújo

Avaliação da pobreza na Região Metropolitana do Rio de Janeiro e o efeito da formalidade entre os ocupados

Monografia de Projeto Final de Graduação sob o título *"Avaliação da pobreza na Região Metropolitana do Rio de Janeiro e o efeito da formalidade entre os ocupados"*, defendida por Marcson de Azevedo Araújo e aprovada em 31 de Janeiro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Rafael Santos Erbisti
Departamento de Estatística – UFF

Profa. Dra. Carolina Botelho Marinho da C. Hecksher
SCN Lab/Mackenzie e DOXA/IESP/UERJ

Profa. Dra. Márcia Marques de Carvalho
Departamento de Estatística – UFF

Dr. Marcos Dantas Hecksher
Pesquisador – IPEA

Niterói, 31 de Janeiro de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

A658a Araújo, Marcson de Azevedo
Avaliação da pobreza na região metropolitana do rio de janeiro e o efeito da formalidade entre os ocupados / Marcson de Azevedo Araújo ; Rafael Santos Erbisti, orientador ; Carolina Botelho Marinho da C. Hecksher, coorientadora. Niterói, 2022.
74 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Inferência bayesiana. 2. Mercado de trabalho formal. 3. Pobreza. 4. Regressão logística. 5. Produção intelectual. I. Erbisti, Rafael Santos, orientador. II. Hecksher, Carolina Botelho Marinho da C., coorientadora. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O presente trabalho busca observar fatores de indivíduos e de seus domicílios associados a condição de pobreza monetária na Região Metropolitana do Rio de Janeiro no último trimestre de 2012, 2016 e 2020. São observados domicílios com alguma pessoa ocupada na PNAD Contínua, realizada pelo IBGE. O principal teste do estudo está relacionado ao mercado de trabalho, que visa estimar uma redução na chance de ser considerado pobre uma vez que o domicílio conte com alguém ocupado dentro do mercado de trabalho formal. É definido um recorte monetário de pobreza, a renda do trabalho mensal domiciliar per capita de elegibilidade do Bolsa Família. Para estimar o efeito de cada fator relacionado aos indivíduos pobres ou não, é utilizada uma Regressão Logística que aproxima a distribuição a posteriori destes fatores para identificar o comportamento de cada efeito. É encontrada redução da chance de ser pobre em um domicílio com ocupados no mercado formal em relação a domicílios com informais e o efeito é maior que o da relação de escolaridade do responsável na redução da chance de ser pobre.

Palavras-chave: Inferência bayesiana, Mercado de trabalho formal, Pobreza, Regressão logística.

Agradecimentos

Agradeço a toda minha família, que sempre me apoiou. Com o esforço incessável da minha esposa para me incentivar e meus pais, irmão e sogros que estão sempre disponíveis tecendo uma grande rede de apoio.

À todos os professores e pesquisadores que contribuíram para a minha formação, especialmente aos meus orientadores, Rafael e Carolina que mesmo nesse momento de pandemia sempre estiveram disponíveis para me orientar da melhor forma e fazer desse trabalho possível.

À todos os meus amigos, que me incentivaram e me ajudaram a passar por essa etapa.

Aos meus supervisores do estágio, que sempre buscaram me passar os conhecimentos técnicos e também sobre o ambiente de trabalho, o que me trouxe aos indicadores sociais e ter o grande prazer de trabalhar na área de pesquisa.

E à Universidade Federal Fluminense, pela infraestrutura e ensino transmitido.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O problema da pobreza impõe desafios à sociedade e aos governos. No final do século XIX, com o advento da segunda Revolução Industrial e expansão da globalização, o capitalismo começou a entrar na sua fase monopolista. Nesta fase, com o aumento da produção de mercadorias em níveis globais ocorreu o aumento do excedente econômico que, não obrigatoriamente atende às reais necessidades da população, corroborando para uma desigualdade de renda na sociedade.

A história mostra que medir pobreza de renda envolve riscos e escolhas. Próximo da virada do século para os anos 1900 foram realizadas as primeiras publicações realizadas na Inglaterra, em Booth (1892) foi identificada a importância da regionalização, encontrando diferentes valores para ser considerado pobre entre bairros londrinos. O maior impedimento para a produção deste conteúdo sempre foi um recorte que seja considerado objetivo, uma abordagem que mudou o curso da classificação de pobreza, foi vista pela primeira vez em Rowntree (1901). Neste estudo, foi considerado pobre quem não gerava renda semanal igual ou acima do valor para adquirir uma cesta de alimentos (medida em valor calórico) para um adulto manter o peso.

No Brasil, é em 1990 que a produção acadêmica sobre pobreza ganhou relevância, principalmente após o Plano Real, que abriu portas para que a questão seja amplamente discutida. Neste ponto, era cada vez mais perceptível que muitos brasileiros não tinham como atender às necessidades básicas dentro de casa e então, a redistribuição de riqueza começa a ser pensada como um contorno que ampliaria os direitos sociais e assim permitiria uma melhora direta na renda da família para ajudar na sobrevivência, permitindo geração de capital humano e que em algum momento, finalmente, não seria mais preciso a família participar do programa.

Dada esta discussão, houve a atualização de programas de transferência de renda que ocorriam para um número reduzido de pessoas e após o governo federal construir o próprio programa Bolsa Escola em 1997, houve um processo de consolidação de programas que

culminou no Programa Bolsa Família em 2003, como visto em Rocha (2019).

Apesar da efetividade dos programas criados para auxiliar a redução da pobreza, o índice de Gini de $0,539$ para o Brasil, em 2018, segundo o cálculo do Banco Mundial, evidencia o contraste social dentro do território nacional, pois em termos de distribuição de renda nosso país está entre os mais desiguais (posição de 156 entre 164 países). (IBGE (2020)).

Tendo em vista o baixo acúmulo de renda em grande parte da população, conhecer os fatores e quantificar os efeitos no aumento da pobreza é fundamental para identificar quais precisam ser priorizados para atender às necessidades da população. Através da identificação desses fatores, é possível adotar medidas de prevenção em busca de um maior suporte para melhorar a qualidade de vida dos pobres.

Ao considerar que a decisão sobre consumo é uma tarefa cotidiana entre os mais pobres, entende-se que eles são mais frágeis a uma variação de renda, o que resulta na importância da manutenção do trabalho para este grupo. Portanto um Mercado de Trabalho volátil com altos percentuais de contratações e de demissões, indica muita variação na renda dos mais pobres no decorrer dos meses.

Neste contexto, é observado o efeito do trabalho formal na pobreza de renda na Região Metropolitana(RM) do Rio de Janeiro com os dados da amostra da pesquisa nacional por amostra de domicílios contínua aplicados em um modelo de regressão logística, os tratamentos dos dados assim como a execução e validação dos modelos foi realizada no software R.

Como motivação geral para esse estudo, é observada a Agenda de Objetivos de Desenvolvimento Sustentável(ODS) de 2030 da ONU[1], cujo objetivo número um é erradicar a pobreza. Portanto, identificar características populacionais que auxiliem a geração de políticas públicas para redução da pobreza é relevante para o desenvolvimento social e econômico do Brasil.

1.1 Objetivos

O presente trabalho estima o comportamento da relação de duas categorias de um determinado fator quanto a condição de pobreza na RM e mede o efeito da formalidade com a relação entre as chances de um indivíduo que more com ocupados formais ser con-

[1]Implementado pelo PNUD, Programa das Nações Unidas para o Desenvolvimento.

siderado pobre comparada às chances de indivíduos que morem com ocupados informais estejam abaixo da linha da pobreza.

Objetivos específicos:

- Medir o efeito da formalidade na condição de pobreza de indivíduos que residem em domicílios com pelo menos um dos moradores ocupados no mercado de trabalho formal;
- Avaliar os efeitos de características gerais de indivíduos quanto sua condição socio-econômica;
- Analisar as diferenças entre indivíduos que residem na capital do estado do Rio de Janeiro e fora da capital pela condição de pobreza; e
- Analisar as diferenças entre características dos pobres na Região Metropolitana do Rio de Janeiro nos anos de 2012, 2016 e 2020.

1.2 Organização

O trabalho encontra-se dividido em cinco capítulos, incluindo este introdutório. O próximo capítulo apresenta uma breve revisão da literatura. No Capítulo 3, são definidos os materiais do estudo com uma análise dos dados selecionados e apresenta, em seguida, os aspectos metodológicos utilizados. No Capítulo 4 são mostrados os resultados dos modelos estimados. Por último, no Capítulo 5 serão discutidas as conclusões do trabalho.

2 Revisão de Literatura

A classificação de pobreza já era citada no final do século XIX. Booth (1892) realizou uma pesquisa que criou uma métrica absoluta, as Linhas de Booth, para classificar pobreza entre os londrinos dentro de regiões da cidade. Rowntree (1901), buscou identificar o mínimo necessário para não ser pobre na Cidade de York no Reino Unido, utilizando o custo do consumo calórico para não perder peso, definindo aqueles que não tinham renda para atingir este custo como pobres. Segundo Soares (2009) "(...) este método tem sido especialmente popular na América Latina com menção ainda mais honrosa no Brasil. As linhas de pobreza da CEPAL para a América Latina são, há três décadas, todas calculadas mediante o uso do método calórico indireto."

Na década de 1960, Orshansky (1968) escreveu "Shape of Poverty"[1], uma avaliação da população pobre não institucionalizada dos Estados Unidos, segundo os dados da SSA, (Administração de Seguridade Social) no período correspondente entre 1959 e 1966. A autora desagrega em diversos fatores as famílias norte-americanas para entender os resultados de políticas públicas da época. Na Europa, desenvolvido por Praag e Kapteyn (1980) , buscou-se entender a pobreza através da interpretação dos próprios membros da família, que se julgariam pobres por não conseguir fechar a conta de consumo e receita de sua família.

No Brasil, após o Plano Real, houve um expressivo aumento da produção textual relacionada à pobreza. Segundo Rocha (1996), o declínio da renda per capita entre 1980 e 1992, fez com que os níveis elevados de pobreza absoluta e de desigualdade se tornassem temas centrais em todos os fóruns.

Na literatura, há diversos trabalhos com informações sobre composição e distribuição de renda, agregando muito valor ao tema, pois trata de conceitos de famílias e composição de renda. Segundo Barros e Mendonça (1995), a família tem grande influência sobre todas as decisões com respeito à alocação do tempo de seus membros, desempenhando, portanto,

[1] traduzido como O Formato da Pobreza

um papel central na definição da estratégia de geração de renda e de investimentos em capital humano de seus membros.

Já Kanso (2004), visava encontrar a influência que determinadas características têm sobre a família estar acima ou abaixo da linha de pobreza nas Regiões Metropolitanas do Rio de Janeiro e Recife nos anos de 1970, 1980 e 1991. Borges (2005), avaliou o perfil da população pobre em comparação com a população não pobre da Região Metropolitana de Salvador em 1997 e 2003.

De acordo com Cobo, Athias e Mattos (2014), a representação de linhas de pobreza monetária "absolutas"tem uma distinção entre a pobreza extrema ou indigência que focam no mínimo para alimentação e a pobreza ou necessário para se viver dignamente, onde são mensuradas outras necessidades como habitação, transporte, higiene, etc.

De acordo com a Agência do Senado[2], existe um destaque para o Brasil como a 2^a maior concentração de renda entre mais de 180 países, classificado pela concentração de renda dos 1% mais ricos. E para os brasileiros 10% mais ricos a participação na renda total do país é de $41,9\%$. Ou seja, os outros 90% da população conseguem menos do que 60% da renda total. A publicação é baseada no estudo que pode ser visto em PNUD (2019).

A discussão feita por especialistas nas últimas duas décadas corrobora a relevância da análise que será desenvolvida neste trabalho.

[2]Publicação, Programa das Nações Unidas para o Desenvolvimento.

3 Materiais e Métodos

Neste capítulo serão analisados os dados e detalhados os métodos aplicados da estimação a ser realizada. Dentro da primeira seção o foco estará na origem da pesquisa aplicada no trabalho. Na Seção 3.2 são detalhados os recortes da população de acordo com o nível de pobreza, para construir a variável resposta do modelo. Na Seção 3.3, é identificada a população alvo e investiga-se o perfil dos envolvidos de forma a comparar duas dimensões, os habitantes da Capital com os da Região Metropolitana fora da capital e os classificados como pobres com os não pobres. E, a partir da Seção 3.4 são desenvolvidas etapas referentes aos modelos estimados.

3.1 Base de dados

Nesta seção é discutida a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), onde será contextualizada a pesquisa, expostas algumas características da amostra e descritas as variáveis avaliadas neste trabalho.

Em 1967, foi elaborada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) a Pesquisa Nacional por Amostra de Domicílios (PNAD), com coletas trimestrais nos primeiros anos e, a partir de 1971, a pesquisa foi realizada anualmente no último trimestre. A pesquisa visava com uso das pesquisas domiciliares, investigar características socioeconômicas e indicadores da evolução da força de trabalho do país para atender a demanda por informações estatísticas nos períodos intercensitários. A pesquisa foi encerrada após a divulgação dos dados de 2015.

De acordo com Suliano e Carvalho (2017), a implantação da Pesquisa Mensal de Emprego (PME), desde de 1980, aperfeiçoou as informações voltadas para a avaliação conjuntural do mercado de trabalho. A partir de 2007, o IBGE reformulou suas pesquisas por amostra de domicílios através da implantação gradual do Sistema Integrado de Pesquisas Domiciliares (SIPD), do qual a PNAD Contínua é um dos pilares básicos, substituindo a PNAD e a PME.

A PNAD Contínua foi instituída no ano de 2012 pelo IBGE, sendo realizada através de uma amostra de domicílios, com o objetivo de produzir informações básicas para o estudo do desenvolvimento socioeconômico do país. A população alvo da PNAD Contínua é constituída por pessoas moradoras em domicílios particulares permanentes ocupados. Sua abrangência geográfica é todo o território nacional, dividido nos setores censitários da Base Operacional Geográfica do Censo de 2010 e está distribuída na seguinte desagregação geográfica: Brasil, Grandes Regiões, Unidades da Federação, Regiões Metropolitanas que contêm Municípios das Capitais, Região Integrada de Desenvolvimento - RIDE e Municípios das Capitais.

O plano amostral escolhido foi um desenho estratificado, com a seleção de cada Unidade Primária de Amostragem (UPA) com probabilidade proporcional a uma medida de tamanho. A opção por este plano se justifica pelo fato das metodologias de todas as pesquisas domiciliares do IBGE incluírem planos amostrais que empregam algum tipo de estratificação das UPAs e seleção destas unidades com probabilidade proporcional ao tamanho, conforme descrito em Bianchini e Albieri (2002).

Para auxiliar na qualidade dos resultados, a pesquisa tem um esquema de rotação da amostra. Dessa forma, o domicílio é entrevistado um mês e fica fora da amostra por dois meses seguidos, sendo esta sequência repetida cinco vezes. Cada uma destas cinco entrevistas é contada como uma visita com diferentes questionários aplicados, apresentando, além das características gerais dos moradores, tópicos suplementares distribuídos entre as visitas, como habitação, outras formas de trabalho, entre outros.

A pesquisa foi reformulada em 2015 revisando os módulos para o aprimoramento da captação de alguns quesitos, entre eles, migração, fecundidade e trabalho infantil.

Será utilizado neste trabalho a coleta do último resultado trimestral disponível, o quarto trimestre de 2020. Assim como, para comparação, as pesquisas do quarto trimestre de 2012 e 2016.

A seguir, são especificadas as variáveis da base utilizadas no trabalho.

3.1.1 Variáveis

As características da população captadas como as mais importantes dentro da literatura obtidas da PNAD Contínua do quarto trimestre de todos os períodos estudados. Este conjunto de fatores são considerados para investigação da base de dados e para construção do modelo, fatores que são atribuídos como variáveis de indivíduos e de seus domicílios,

sendo citadas abaixo:

- Tipo de área: se o domicílio se encontra na Capital ou Região Metropolitana (excluindo a capital).
- Condição no Domicílio: pessoa responsável pelo domicílio, cônjuge ou companheiro(a) de sexo diferente, cônjuge ou companheiro(a) do mesmo sexo, filho(a) do responsável e do cônjuge, filho(a) somente do responsável, enteado(a), genro ou nora, pai, mãe, padrasto ou madrasta, sogro(a), neto(a), bisneto(a), irmão ou irmã, avô ou avó, outro parente, agregado(a) - não parente que não compartilha despesas, convivente - não parente que compartilha despesas, pensionista, empregado(a) doméstico(a), parente do(a) empregado(a) doméstico(a).
- Sexo: sexo do morador, classificado como "Masculino" e "Feminino".
- Idade: idade do morador na data de referência em anos. Varia de 0 a 130 anos.
- Cor ou raça: aponta a cor ou raça da pessoa. Esta variável está dividida em: branca, preta, amarela, parda, indígena e ignorado.
- Nível de instrução: nível de instrução mais elevado alcançado (para pessoas com 5 anos ou mais de idade) padronizado para o Ensino Fundamental com duração de 9 anos. Dividido em "Sem instrução e menos de 1 ano de estudo", "Fundamental incompleto ou equivalente", "Fundamental completo ou equivalente", "Médio incompleto ou equivalente", "Médio completo ou equivalente", "Superior incompleto ou equivalente" e "Superior completo".
- Condição de ocupação: condição de ocupação na semana de referência. os respondentes optam por pessoas ocupadas ou pessoas desocupadas.
- Posição na ocupação: posição na ocupação e categoria do emprego do trabalho principal na semana de referência para pessoas de 14 anos ou mais de idade ocupadas. Divididos em empregado no setor privado **com** carteira de trabalho assinada, empregado no setor privado **sem** carteira de trabalho assinada, trabalhador doméstico **com** carteira de trabalho assinada, trabalhador doméstico **sem** carteira de trabalho assinada, empregado no setor público **com** carteira de trabalho assinada, empregado no setor público **sem** carteira de trabalho assinada, militar e servidor estatutário, empregador, conta-própria e trabalhador auxiliar familiar.

- Contribuinte: contribuição para instituto de previdência em qualquer trabalho da semana de referência para pessoas de 14 anos ou mais de idade. Dividido em contribuinte e não contribuinte.

- Carteira assinada: se no trabalho tinha carteira assinada, as respostas são sim e não.

- Faixa de horas: faixa das horas efetivamente trabalhadas na semana de referência no trabalho principal para pessoas de 14 anos ou mais de idade, as respostas são até 14 horas, 15 a 39 horas, 40 a 44 horas, 45 a 48 horas, 49 horas ou mais.

- Renda: rendimento mensal efetivo de todos os trabalhos para pessoas de 14 anos ou mais de idade (apenas para pessoas que receberam em dinheiro, produtos ou mercadorias em qualquer trabalho). Valor em reais.

As **variáveis derivadas** são definidas a partir das variáveis selecionadas nesta seção e foram construídas de acordo com a literatura que permitiu identificar pontos importantes para o estudo.

Antes de definir as variáveis derivadas, é necessário classificar como será usada a formalidade neste trabalho que é dada pela variável de Posição na ocupação e de Contribuinte, os **formais** são: empregado no setor privado **com** carteira de trabalho assinada ou trabalhador doméstico **com** carteira de trabalho assinada ou empregado no setor público **com** carteira de trabalho assinada ou militar e servidor estatutário ou empregador e contribuinte ou conta-própria e contribuinte. E os **informais** são: empregado no setor privado **sem** carteira de trabalho assinada ou trabalhador doméstico **sem** carteira de trabalho assinada ou empregado no setor público **sem** carteira de trabalho assinada ou trabalhador familiar auxiliar ou empregador e não contribuinte ou conta-própria e não contribuinte.

Com estas é possível identificar novas características dos indivíduos e de seus domicílios, as variáveis derivadas são as citadas abaixo:

- Chave do domicílio: é a união de valores de identificação da PNAD Contínua, são os números da Unidade Primária de Amostragem (UPA), do estrato e do identificador do domicílio gerando um código único por domicílio com 18 caracteres.

- Quantidade no domicílio: número da soma de pessoas por cada chave do domicílio.

- Grupos de idade: derivada da idade de cada indivíduo, os valores são de 0 a 14 anos, 15 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 64 anos e 65 anos ou mais.

- Cor: identificado pela variável de Cor ou Raça, classifica-se cada indivíduo como brancos ou da agregação pretos ou pardos. Existe um filtro na população, definido na Seção 3.3, logo os valores identificáveis são estas duas categorias somente.

- Escolaridade: uma revisão da variável de Nível de instrução, que foi classificada como sem instrução e menos de 1 ano de estudo, fundamental incompleto, fundamental completo e médio incompleto, médio completo e superior incompleto e por último superior completo ou mais.

- Escolaridade do responsável: é encontrado o responsável pelo domicílio e a escolaridade de acordo com a variável derivada anterior, e então replica-se a escolaridade do responsável para todos do domicílio.

- Renda do trabalho domiciliar per capita: rendimento comum para todo o domicílio resultante da razão entre a soma da variável renda de trabalho de todos os integrantes do domicílio e a quantidade de pessoas no domicílio. Valor em reais.

- Pobres: classificação dada para cada indivíduo de acordo com o valor da variável da renda do trabalho domiciliar per capita estar abaixo da linha de pobreza ou não.

- Ocupação no domicílio: considera a variável Condição de ocupação, com ocupados do tipo formal e informal em cada domicílio, classificando todos os indivíduos por uma informação que será referente ao domicílio, que são: 1, ocupado formal, quando há somente ocupado(s) no mercado formal. 2, ocupado informal, quando há somente ocupado(s) fora da formalidade. E, 3, ocupado formal e informal, quando os ocupados no domicílio estão divididos dentro e fora do mercado de trabalho formal.

3.2 Linhas de pobreza

Neste trabalho a pobreza de renda é adotada para classificar os moradores de cada domicílio pelo resultado efetivo, ou seja, considera os valores de renda efetiva do trabalho da semana de referência contida no período coletado. Com isso, ao identificar o valor da renda de trabalho domiciliar per capita e dada uma linha de pobreza é possível classificar se cada indivíduo é pobre ou não pobre. Nesta seção busca-se entender os recortes monetários para definir tal Linha de Pobreza (LP).

De acordo com IBGE (2020), o conceito de pobreza é definido como privações de diferentes tipos e é cada vez mais visto como um fenômeno multidimensional, o que estaria ligado ao aumento da disponibilidade de dados e à influência de novos escritos na área como Athias e Oliveira (2016) e Alkire e Foster (2008). Ao mesmo tempo, permanecem mais recorrentes as análises unidimensionais e monetárias (pela renda ou pelo consumo) por razões de preferência ou disponibilidade de dados.

É necessário evidenciar que grande parte dos trabalhos encontrados utilizam pobreza monetária por grupo familiar, o que é uma limitação da pesquisa escolhida para este trabalho, uma vez que a PNAD Contínua permite identificar somente os moradores por grupos domiciliares. Esta diferença entre o objeto de análise pode ser dada pelo questionário/objetivo da pesquisa definido pelo pesquisador antes de realizá-la. Para mitigar esses efeitos foram removidos alguns componentes domiciliares, ao considerar uma interpretação de composição familiar e de composição de uma renda familiar.

Logo, para avaliar possíveis relações de renda entre família e domicílio foi discutido no Apêndice 1 uma comparação da distribuição de renda familiar per capita e renda domiciliar per capita. Utilizando as pesquisas anuais da PNAD de 2001 e 2009 e as pesquisas anuais da PNAD Contínua de 2016 e 2019.

No contexto da classificação da pobreza, existem diversas linhas de pobreza adotadas em todo o mundo. As linhas relativas são usadas normalmente em países com uma sociedade próspera e baixa desigualdade de renda, como a dos países nórdicos, onde a pobreza não é mais um conceito relevante, então é chamada de "pobreza" a renda dos 20% mais pobres. E, de outra forma, linhas de pobreza absolutas são valores calculados que dividem a população em pobres e não pobres, possibilitando a definição de um indicador de incidência de pobreza. Essa configuração é explorada principalmente em países muito desiguais economicamente para identificar a parte da população com problemas relacionados à fome e falta de qualidade de vida.(Rocha (1996)). Essa é a forma escolhida para calcular a variável resposta do modelo proposto neste trabalho.

Num aspecto internacional, a União Europeia (via Eurostat) e a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) assumem linhas de pobreza relativas, fixadas em 50% ou 60% da renda média nacional. Outra medida relevante é o recorte da linha de pobreza extrema internacional, construída a partir dos 15 países mais pobres, estabelecida como um indicador global de pobreza e calculada pelo Banco Mundial. O valor atual é de $1,90$ dólares por dia de renda ou consumo per capita em PPC, revisada em 2011. Mesmo se calculada a partir dos países mais pobres, essa linha tem muita

relevância no nível mundial, pois o relatório global de acompanhamento da Agenda para o Desenvolvimento Sustentável 2030 estimou que ainda havia 767 milhões de pessoas na pobreza extrema em 2013 (Nações Unidas, 2017). Esta é a linha de pobreza usada para o ODS número 1 da ONU[1].

No Brasil não há linha oficial de pobreza, mas há diversas linhas conhecidas como administrativas, utilizadas pelas políticas governamentais. Existem as linhas do Programa Brasil sem Miséria - PBSM – R$ 89, 00 (pobreza extrema) e R$ 178, 00 (pobreza) – e a linha do Benefício de Prestação Continuada (BPC) definida como o rendimento domiciliar per capita abaixo de 1/4 de salário mínimo. Essas linhas podem ser definidas por lei - como o BPC, na Lei n. 8.742, de 07.12.1993, atendendo ao princípio constitucional de as pessoas viverem e envelhecerem com dignidade - ou por decisões administrativas que são utilizadas por políticas governamentais como a do Programa Bolsa Família (PBF) que atenderá às famílias em situação de pobreza e de extrema pobreza, caracterizadas pela renda familiar mensal per capita de até R$ 178,00 (cento e setenta e oito reais) e R$ 89,00 (oitenta e nove reais), respectivamente.

Em relação ao Brasil, o IBGE considera em várias publicações o recorte para pobreza e extrema pobreza as linhas do Banco mundial de US$5.5 e US$1.9 por dia per capita em PPC 2001, respectivamente. A linha de US$5.5 é indicada para países de renda média alta e a linha de US$1.9 para comparação com a linha de extrema pobreza internacional.

Após uma busca na literatura, pela motivação de ser uma linha bem conhecida e permitir aqueles que estão abaixo dela receber um apoio assistencial do governo, é decidido para este trabalho o valor do PBF para ponto de corte na distribuição ordenada da renda do trabalho domiciliar per capita (RDPC). Com isso, para o presente estudo a classificação de pobre é dada para os residentes de domicílios onde a RDPC está abaixo de R$ 140 em 2012, R$ 170 em 2016 e R$ 178 em 2020. [2]

3.3 Área de estudo

A região delimitada para o estudo é a Região Metropolitana do Rio de Janeiro, constituída na década de 1970. A região passou por mudanças de Leis e Decretos para o desmembramento de municípios que resultou na sua formação atual com 22 municípios definida pelo Governo do Estado e adotada pelo IBGE como Região Administrativa In-

[1] Parte do Relatório de 2020 - Objetivo 1

[2] Os valores de linha de pobreza utilizados no trabalho serão corrigidos para valores correntes de acordo com a atualização do decreto : Decreto 5209/2004 - Artigo 18

tegrada de Desenvolvimento do Rio de Janeiro[3], com os municípios de: Belford Roxo, Cachoeiras de Macacu, Duque de Caxias, Guapimirim, Itaboraí, Itaguaí, Japeri, Magé, Maricá, Mesquita, Nilópolis, Niterói, Nova Iguaçu, Paracambi, Petrópolis, Queimados, Rio Bonito, Rio de Janeiro, São Gonçalo, São João de Meriti, Seropédica e Tanguá. A região Metropolitana do Rio de Janeiro é uma concentração de municípios em torno da capital, como visto na parte em roxo da Figura 1.

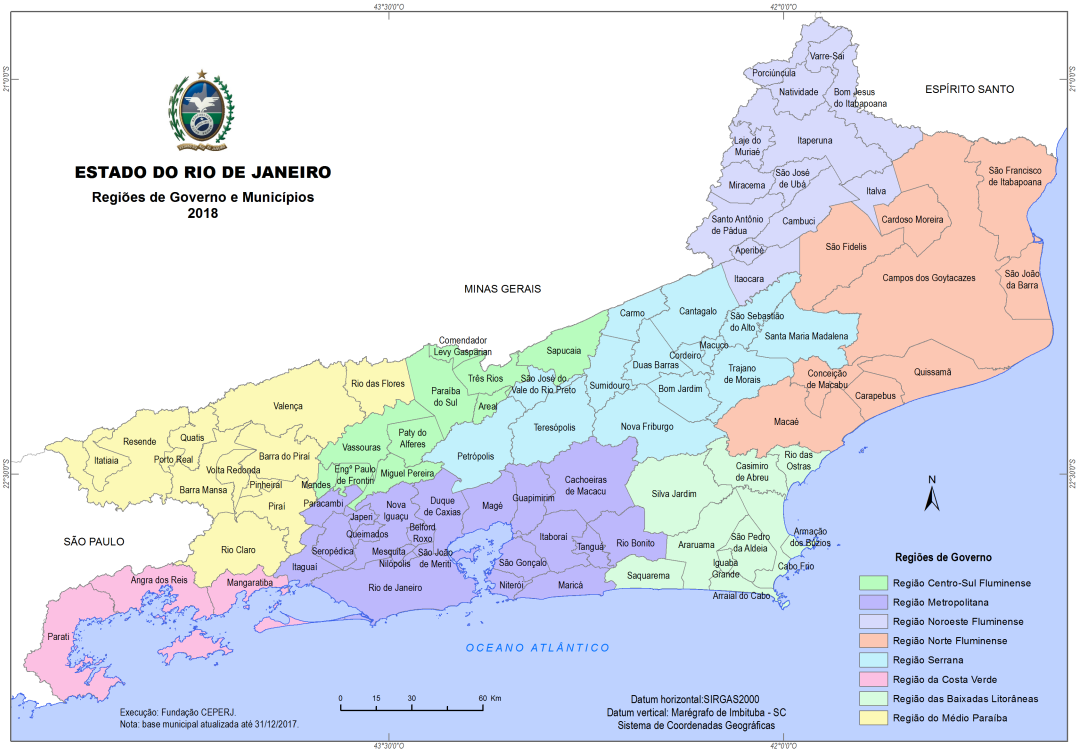

Figura 1: Mapa do Estado do Rio de Janeiro - Divisão por regiões - Fundação CEPERJ

De acordo com os dados da pesquisas utilizadas, o quarto trimestre da PNAD Contínua dos anos de 2012, 2016 e 2020, são encontradas $12,2$ milhões de pessoas nas estimativas para a Região Metropolitana do Rio de Janeiro em 2012, em 2016 é visto que a estimativa é de $12,5$ milhões e para 2020, de $12,7$ milhões de moradores.

A **população final** para composição das análises e dos modelos é selecionada a partir dos moradores de domicílios da Região Metropolitana do Rio de Janeiro. Para essa seleção, são removidos os domicílios onde nenhum morador indicou na condição de ocupação como pessoa ocupada. São removidos também os domicílios com moradores indígenas ou amarelos, pois é uma categoria de cor ou raça com presença muito pequena, mesmo agrupada, e para os domicílios remanescentes, excluímos alguns moradores de acordo

[3]IBGE: estrutura territorial. Acesso em Abril de 2021

com a variável de condição no domicílio, os removidos são: agregado(a) - Não parente que não compartilha despesas -, pensionista, empregado(a) doméstico(a) e parente do(a) empregado(a) doméstico(a). Concluindo com a base de dados do quarto trimestre de 2012 com 23.119 respostas (86% de toda a amostra da RM). Para 2016 com 21.052 observações (82%) e para 2020, 10.443 (73%).

Agora, com a população do trabalho definida, a informação de Região Metropolitana (RM) passa a ter significado associado a variável tipo de área, onde se define que RM representa os morados dos municípios da Região Metropolitana exceto os da capital, logo, Capital indicará os moradores somente do município do Rio de Janeiro.

3.3.1 Análise preliminar da população alvo

Nesta sessão, será explorado o perfil da população final adotada. Examinar as análises descritivas é de extrema relevância para a construção de um modelo, assim, é realizada uma análise introdutória das variáveis da base para investigar as comparações entre os anos de 2012, 2016 e 2020. Também são avaliados grupos de escolaridade do responsável, de comportamentos do mercado de trabalho comparados entre as mudanças de área e de condição de pobreza.

Tabela 1: Incidência de pobreza e percentual de pessoas em domicílios com ao menos um ocupado formal por ano e área

Ano	Área	Pobres (%)	Formais (%)
2012	RM	6,22	33,65
	Capital	4,50	38,91
2016	RM	3,91	32,41
	Capital	2,66	40,04
2020	RM	5,10	29,89
	Capital	3,80	35,58

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Para os diferentes anos e tipos de área definidas pelo estudo e mostradas na Tabela 1 há uma redução da incidência de pobreza no período de 2012 para 2016, porém no ciclo de 2016 para 2020, a taxa sobe 30% para a RM e aumenta 43% para a Capital.

O percentual de pessoas em domicílios com ao menos uma pessoa ocupada no mercado de trabalho formal varia entre 30% e 40% para todos os períodos e tipos de área, com destaque para a Capital em 2016, pois é o maior percentual de domicílios com formais e o menor valor da taxa de pobres.

A análise descritiva indica algumas mudanças nas pesquisas na incidência de pobreza entre grupos de idade comparado a população geral, para crianças e adolescentes de 0 a 13 anos da RM, a variação foi maior que 4 p.p., de $11,5\%$ em 2012 para $7,3\%$ em 2020.

E para os idosos, pessoas acima da 65 anos, a incidência de pobreza não mostra variação no período, neste caso, é mantida a taxa em torno de 5% para RM e de 3% para a Capital nos períodos investigados.

A Composição de cor ou raça mostra a tendência de redução da pobreza em ambas as categorias no ciclo $12 - 16$ em 30% e aumento da pobreza de 50% no ciclo $16 - 20$. Para os pobres em 2020 na RM, tem-se $1,9\%$ dos brancos e $3,2\%$ dos pretos ou pardos, agora, na Capital em 2020, brancos pobres são $1,6\%$ e pretos ou pardos pobres são $2,2\%$. Destaca-se a distancia de pobres entre brancos e pretos ou pardos na RM que é maior do que na Capital em todos os períodos.

Para o recorte por sexo, é observado que existem mais mulheres na população para todos os períodos e recortes regionais. Somada a dimensão de renda, a maior disparidade está na Capital em 2016 com 57% de mulheres entre os pobres.

Para o tema de educação, será analisado o comportamento das pessoas em domicílios dado o nível de escolaridade do responsável pelo domicílio.

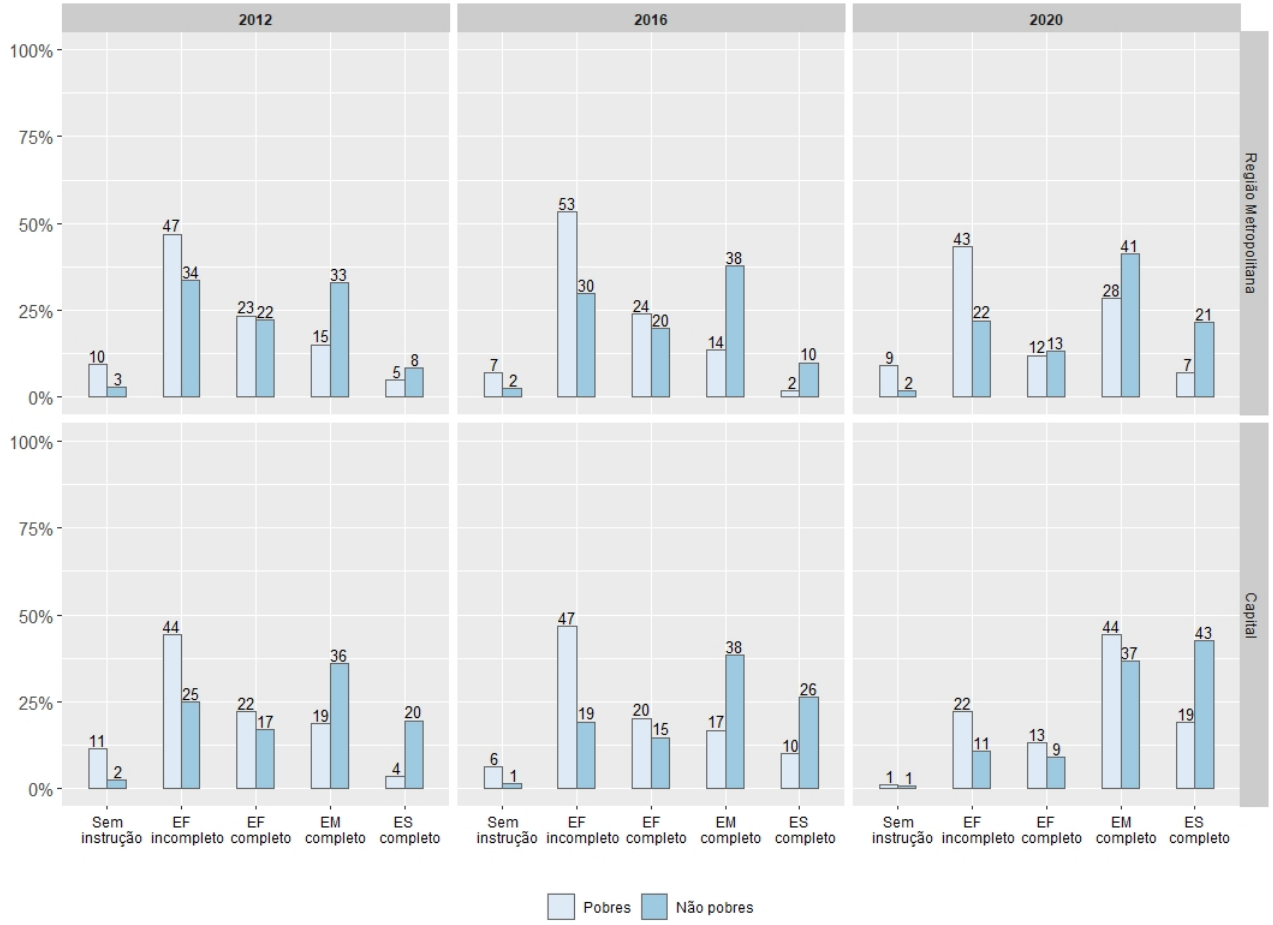

Figura 2: Gráfico da distribuição de pessoas dado o nível de escolaridade do responsável pelo domicílio segundo o ano e tipo de área

Quanto a Figura 2, é observada uma evolução no nível de escolaridade dos responsáveis na Capital ao longo das pesquisas para pobres e não pobres. Segundo a Tabela 3, referente a Capital, $18,8\%$ das pessoas em domicílios pobres em 2012 têm o responsável com o ensino médio completo, este número para 2020 é de $44,3\%$. De acordo com a Tabela 2, para a RM também existe tal avanço, ainda que não tão expressivo. Aprofundando um pouco a análise, pela Figura 2, com o decorrer das pesquisas é verificada uma redução de pessoas em domicílios com o responsável pelo domicílio com o ensino fundamental incompleto ou menos, principalmente na Capital, para pobres ou não. Por exemplo, com o olhar na Tabela 3, somadas as duas primeiras categorias de escolaridade do responsável para pobres na Capital representa $55,6\%$ em 2012 e $23,3\%$ em 2020.

Tabela 2: Nível de escolaridade do responsável pelo domicílio - Região Metropolitana

Ano	Categoria	Pobres			Não Pobres	
		Absoluto	Percentual		Absoluto	Perce
2012						
	Sem Instrução	29.205	9,5	%	135.398	2
	EF incompleto	144.006	46,9	%	1.558.759	33
	EF completo	71.880	23,4	%	1.026.629	22
	EM completo	46.449	15,1	%	1.519.564	32
	ES completo	15.379	5,0	%	387.553	8
2016						
	Sem Instrução	13.014	7,1	%	112.375	2
	EF incompleto	98.120	53,5	%	1.347.915	29
	EF completo	43.894	23,9	%	893.231	19
	EM completo	25.023	13,6	%	1.708.160	37
	ES completo	3.483	1,9	%	450.961	10
2020						
	Sem Instrução	19.274	9,0	%	78.422	2
	EF incompleto	92.739	43,4	%	876.246	22
	EF completo	25.797	12,1	%	522.239	13
	EM completo	60.851	28,5	%	1.640.247	41
	ES completo	14.880	7,0	%	852.298	21

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Tabela 3: Nível de escolaridade do responsável pelo domicílio - Capital

Ano	Categoria	Pobres		Não Pobres	
		Absoluto	Percentual	Absoluto	Perce
2012					
	Sem Instrução	28.150	11,4 %	129.127	2
	EF incompleto	109.568	44,2 %	1.315.559	25
	EF completo	54.920	22,1 %	897.724	17
	EM completo	46.582	18,8 %	1.890.892	35
	ES completo	8.793	3,5 %	1.030.897	19
2016					
	Sem Instrução	9.216	6,2 %	77.805	1
	EF incompleto	69.121	46,7 %	1.039.394	19
	EF completo	29.984	20,3 %	795.795	14
	EM completo	24.852	16,8 %	2.073.234	38
	ES completo	14.851	10,0 %	1.426.259	26
2020					
	Sem Instrução	2.214	1,1 %	33.598	0
	EF incompleto	43.453	22,2 %	538.027	10
	EF completo	25.821	13,2 %	452.307	9
	EM completo	86.525	44,3 %	1.811.094	36
	ES completo	37.366	19,1 %	2.110.068	42

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Em seguida, é explorado o tema de mercado de trabalho, observa-se a composição dos ocupados entre as pequisas e os recortes de área e condição de renda que estão sendo estudados.

Dada a Figura 3 é visto que o percentual de ocupados entre os não pobres está próximo de 50% para todas as dimensões. Quando observado entre os pobres, existe um aumento com a evolução de tempo entre as pesquisas, porém os maiores valores, em 2020, são de 33% para a RM e Capital.

Pelas Tabelas 4 e 5 é observado que a evolução dita anteriormente, para o percentual de ocupados entre os pobres, é realizada de formas diferentes dado a região, pois na

Capital, o aumento do percentual de ocupados entre os pobres só é identificado em 2020, para a RM, esta evolução é gradual.

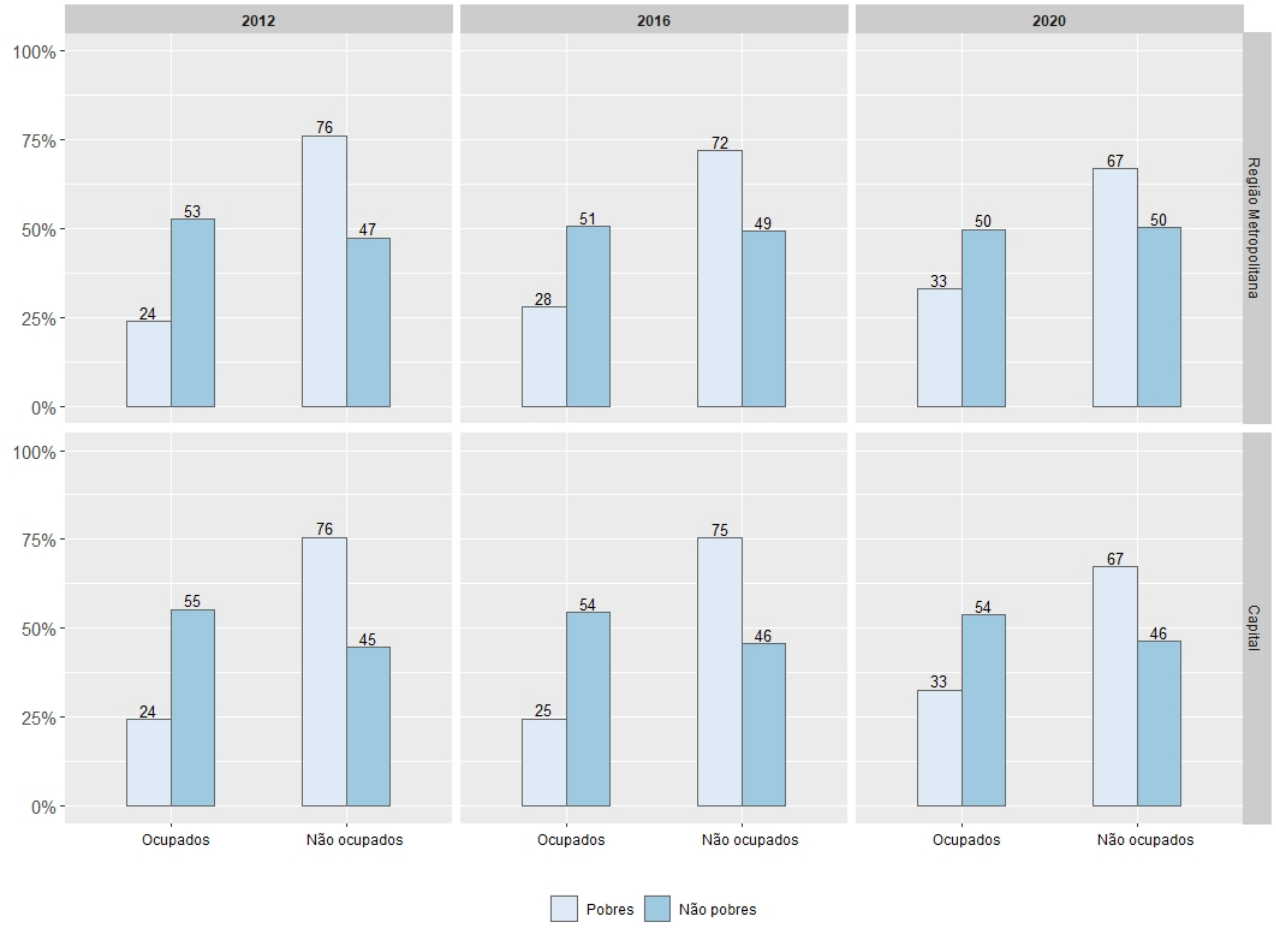

Figura 3: Gráfico da distribuição de pessoas dado a condição de ocupação segundo o ano e tipo de área

Tabela 4: Condição de ocupação - Região Metropolitana

Ano	Categoria	Pobres			Não Pobres	
		Absoluto	Percentual		Absoluto	Perce
2012						
	Ocupados	73.038	23,8	%	2.433.059	52
	Não ocupados	233.881	76,2	%	2.194.843	47
2016						
	Ocupados	51.405	28,0	%	2.289.756	50
	Não ocupados	132.129	72,0	%	2.222.886	49
2020						
	Ocupados	70.691	33,1	%	1.976.131	49
	Não ocupados	142.851	66,9	%	1.993.322	50

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Tabela 5: Condição de ocupação - Capital

Ano	Categoria	Pobres			Não Pobres	
		Absoluto	Percentual		Absoluto	Perce
2012						
	Ocupados	60.512	24,4	%	2.910.929	55
	Não ocupados	187.502	75,6	%	2.353.270	44
2016						
	Ocupados	36.272	24,5	%	2.942.961	54
	Não ocupados	111.752	75,5	%	2.469.526	45
2020						
	Ocupados	63.738	32,6	%	2.655.336	53
	Não ocupados	131.640	67,4	%	2.289.756	46

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Após uma avaliação inicial dos dados de ocupação, foi possível observar na posição na ocupação de **conta-própria** grande parte dos ocupados pobres. Os dados são analisados

pelo percentual calculado com o recorte no denominador de ano, área e condição de pobreza para toda a população, o que permite comparação com as Tabelas 4 e 5.

Entre os $32,6\%$ dos ocupados pobres em 2020 na Capital, $20,5\%$ estão na posição de conta-própria. Para a RM, dos $33,1\%$ dos ocupados, 22% se encontram na posição de conta-própria.

Os cruzamentos mostram também que entre os pobres, o percentual de ocupados na posição de conta-própria aumentou entre os períodos estudados para ambas as áreas. Para a Capital por exemplo, em 2012, de 100% dos pobres ocupados (60.512), 36% declararam a posição de conta-própria. Posteriormente em 2020, a razão entre os 40.147 na conta-própria e os ocupados(63.738 ou $32,6\%$ pela Tabela 5) indica que 63% dos pobres ocupados na Capital em 2020 estariam na posição de conta-própria.

Para concluir as análises descritivas, é vista a seguir a avaliação da variável derivada da condição de ocupação no domicílio.

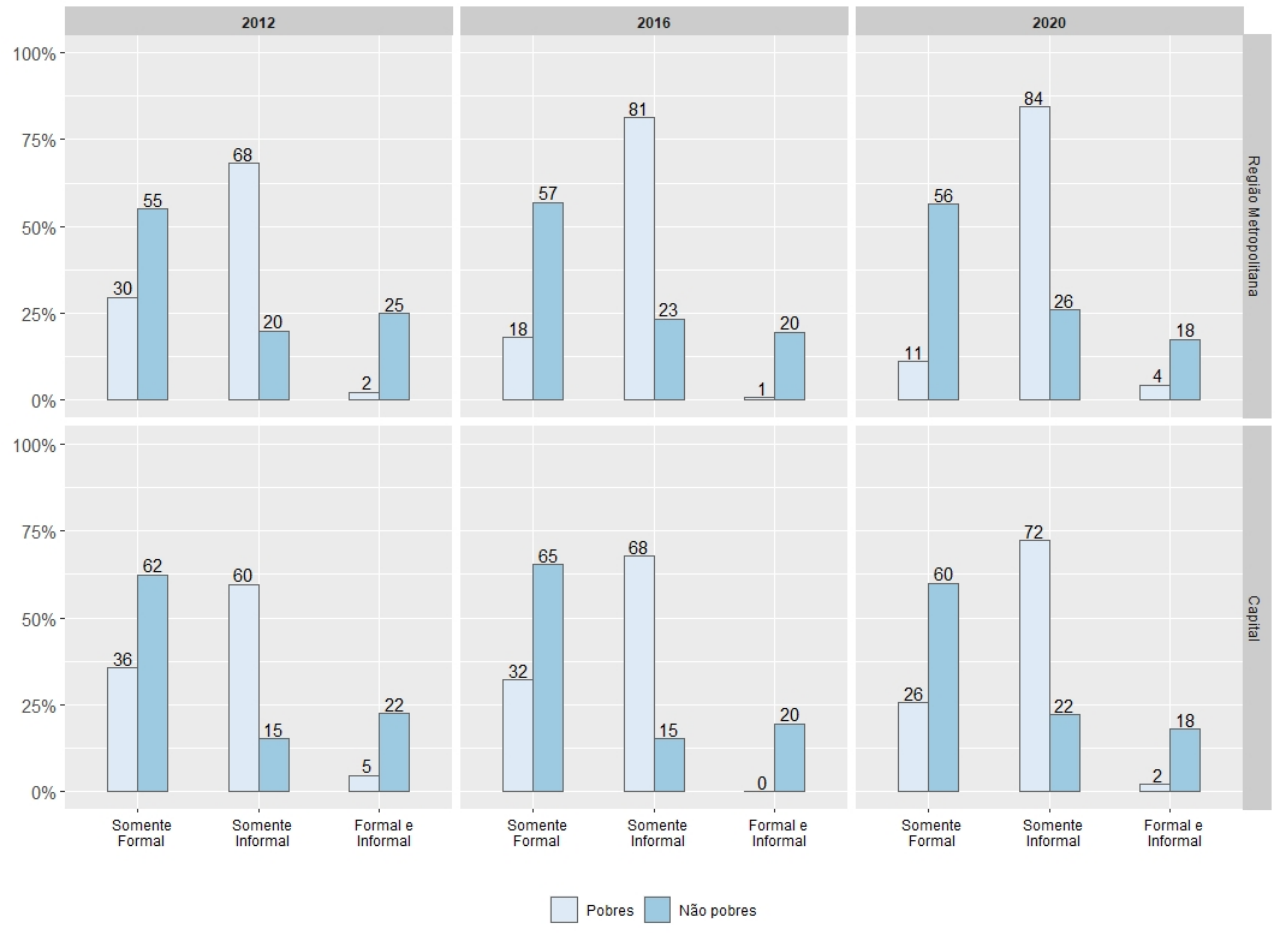

Figura 4: Gráfico da distribuição de pessoas dado a condição de ocupação no domicílio segundo o ano e tipo de área

O passar dos anos na Figura 4, mostrou que apesar do percentual de não pobres em domicílios com somente formal se manteve estável, houve uma redução dos pobres entre os moradores de domicílios com somente formal. Para a Capital, entre 2012 e 2020, abaixou 10p.p. e para RM reduziu 19p.p..

De acordo com as Tabelas 6 e 7, a categoria de ocupado informal é presente em pelo menos 60% dos pobres em todos os recortes de ano e área do estudo. Na RM, em 2016 e 2020, o percentual de pobres da categoria de ocupado informal passa de 80%.

Um detalhe importante para o restante do trabalho é que para os dados do ano de 2016, a amostra da categoria ocupado formal e informal para os pobres representa $0,024\%$ das observações, com 0 observações na Capital, o que dificulta a estimação e avaliação do resultado do fator associado a esta categoria no modelo para o ano de 2016.

Tabela 6: Condição de ocupação - Região Metropolitana

Ano	Categoria	Pobres			Não Pobre
		Absoluto	Percentual		Absoluto
2012					
	Ocupado formal	90.879	29,6	%	2.542.488
	Ocupado Informal	208.906	68,1	%	925.619
	Ocupado formal e informal	7.135	2,3	%	1.159.796
2016					
	Ocupado formal	33.108	18,0	%	2.571.276
	Ocupado Informal	148.978	81,2	%	1.054.219
	Ocupado formal e informal	1.449	0,8	%	887.147
2020					
	Ocupado formal	24.012	11,2	%	2.237.857
	Ocupado Informal	179.965	84,3	%	1.034.060
	Ocupado formal e informal	9.565	4,5	%	697.536

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Tabela 7: Condição de ocupação no Domicílio - Capital

Ano	Categoria	Pobres		Não Pobre
		Absoluto	Percentual	Absoluto
2012				
	Ocupado formal	88.536	35,7 %	3.282.692
	Ocupado Informal	147.789	59,6 %	802.912
	Ocupado formal e informal	11.689	4,7 %	1.178.594
2016				
	Ocupado formal	47.691	32,2 %	3.533.275
	Ocupado Informal	100.333	67,8 %	820.337
	Ocupado formal e informal	0	0,0 %	1.058.875
2020				
	Ocupado formal	50.184	25,7 %	2.964.154
	Ocupado Informal	141.276	72,3 %	1.090.032
	Ocupado formal e informal	3.919	2,0 %	890.906

Fonte: PNAD Contínua do quarto trimestre de 2012-2016-2020.

Finalizada as análises descritivas, segue-se o Capítulo, onde são definidos na Seção 3.4, os Modelos Lineares Generalizados e suas componentes, assim como explicadas as características singulares do modelo proposto. Na Seção 3.5 é desenvolvida a forma de estimação do modelo, na Seção 3.6 são apresentados métodos para validar as simulações realizadas e a Seção 3.7 apresenta a estimação aplicada ao modelo proposto.

3.4 Modelos Lineares Generalizados

A principal ferramenta deste trabalho é um Modelo Linear Generalizado que permitirá encontrar o efeito de fatores independentes que caracterizam o comportamento de uma variável resposta de interesse. Para utilizá-lo são necessárias outras técnicas para verificar os seus pressupostos e uma vez estimado cada modelo, validar suas hipóteses.

Um modelo estatístico é um modelo matemático que contém dois componentes, o determinístico e o aleatório, estes componentes são utilizados para expressar características de uma relação estatística entre duas partes, a dependente e a independente. Como resul-

tado encontra-se o valor estimado em média de uma variável resposta, seja $\mathbf{Y}$ a variável resposta, dada a ocorrência de uma ou mais variáveis independentes descritas por $\mathbf{X}$, é encontrado o valor estimado de $\mathbf{Y}$. Dessa forma, os valores esperados da variável dependente estão associados aos valores observados da variável ou das variáveis independentes. Esta relação é definida como, $\mathbf{E}[\mathbf{Y}|\ \mathbf{X}{=}\mathbf{x}]$, *lê se, a esperança de* $\boldsymbol{Y}$ *dada a ocorrência de* $\boldsymbol{X}$.

Os Modelos Lineares Generalizados são demonstrações de um conjunto de métodos e para todos os casos existem três componentes, uma distribuição de probabilidade que pertence a família exponencial para a variável resposta, um conjunto de variáveis explicativas (estrutura linear) e uma função de ligação, conforme definidos por Nelder e Wedderburn (1972) e após atualizações de literatura, trazidos aqui de acordo com Dobson e Barnett (2018).

Antes da discussão sobre os elementos de um MLG, é necessário definir a família exponencial de distribuições. Esta classificação é dada as funções de probabilidade de variáveis aleatórias discretas e contínuas quando a função pode ser escrita da forma da equação (3.1). Com a variável y, o parâmetro escalar θ e as funções reais conhecidas a, b, c e d.

$$f(y;\theta) = \exp\left[a(y)b\left(\theta\right) + c\left(\theta\right) + d\left(y\right)\right] \tag{3.1}$$

Dessa forma, de acordo com os autores, é definida a componente aleatória associada a variável resposta. Sejam os valores $y_1, y_2, \ldots, y_n$ as realizações das variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$ associadas à família exponencial, de forma que estas são independentes e identicamente distribuídas (iid) de uma distribuição de probabilidade com parâmetro θ_i, para $i = 1, \ldots, n$ pertencente à família exponencial. Assim, é reescrita a função de probabilidade para uma y_i na forma da equação (3.2).

$$f(y_i;\theta_i) = \exp\left[a(y_i)b\left(\theta_i\right) + c\left(\theta_i\right) + d\left(y_i\right)\right] \tag{3.2}$$

Como a distribuição de todos os Y_is são a mesma, então segundo Dobson e Barnett (2018), os índices de a, b, c e d não são necessários e a função de probabilidade conjunta (ou função de densidade de probabilidade conjunta) de $Y_1, Y_2, \ldots, Y_n$ associada aos parâmetros $\theta_1, \ldots, \theta_n$ é:

$$\begin{aligned} f(y_i, \ldots, y_n; \theta_1, \ldots, \theta_n) &= \prod_{i=1}^{n} \exp\left[a\left(y_i\right) b\left(\theta_i\right) + c\left(\theta_i\right) + d\left(y_i\right)\right] \\ &= \exp\left[\sum_{i=1}^{n} a\left(y_i\right) b\left(\theta_i\right) + \sum_{i=1}^{n} c\left(\theta_i\right) + \sum_{i=1}^{n} d\left(y_i\right)\right] \end{aligned} \tag{3.3}$$

Prosseguindo, a estrutura linear é composta por variáveis explicativas $X_1, X_2, \ldots, X_{p-1}$ ($p \leq n$) e por um vetor de parâmetros desconhecidos $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})$ que produzem um preditor linear η. O número p é a quantidade de parâmetros a serem estimados e representam os efeitos das $p-1$ covariáveis e do conjunto na variável resposta.

Assim, suponha que o valor esperado para cada valor i da variável dependente seja, $E\left[Y_i\right] = \mu_i$, onde μ_i é uma função de θ_i, ou seja, o valor esperado da variável dependente está associado ao(s) parâmetro(s) da mesma. Para os Modelos Lineares Generalizados existe uma transformação de μ_i, o preditor linear η_i, tal que

$$\eta_i = g\left(\mu_i\right) = \mathbf{x_i}\boldsymbol{\beta} \tag{3.4}$$

Onde, $g(.)$ é uma função monótona e diferenciável em relação a μ_i chamada de **função de ligação**. O vetor $\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \ldots, x_{ip-1})$ com dimensões $1 \times p$ está associado as variáveis explicativas, onde cada $\mathbf{x_i}$ representa a i-ésima linha da matriz de covariáveis do modelo. E $\boldsymbol{\beta}$, é o vetor de parâmetros $p \times 1$. Para a estimação do modelo, o interesse do trabalho é no vetor $\boldsymbol{\beta}$.

Assim, estão definidas as componentes de um MLG. Para mais informações, veja Cordeiro e Demétrio (2013), onde são apresentadas noções gerais desses modelos, algumas de suas extensões e aplicações.

3.4.1 Modelo de Regressão Logística

Entre os MLGs, um dos métodos utilizados para estimar valores associados à variável resposta com escala binária é o modelo de regressão logística, onde se busca estimar a probabilidade de uma variável dependente com a distribuição de Bernoulli dada a composição do preditor linear. Entre as opções de transformação conhecidas, a função de ligação **logito** é caracterizada como a forma linear de estimar probabilidade, pois é capaz de transformar valores no conjunto dos reais em valores dentro do intervalo $[0, 1]$. Esta transformação está entre as de mais fácil entendimento e interpretação. A outra escolha

comum para a transformação é a função **probito**, que não será explorada neste trabalho. O modelo com a transformação **logito** é construído da seguinte forma.

Após definir uma variável aleatória Y, como

$$Y_i \sim Ber(\pi_i) \tag{3.5}$$

ou seja, para os valores de y_i, que são 0 ou 1, a função de probabilidade é

$$f(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{(1-y_i)}$$

e reescrevendo-a é possível identificar que a distribuição de Bernoulli pertence à família exponencial de distribuições,

$$f(y_i; \pi_i) = \exp\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right]$$

Como já introduzido, o objetivo é medir π_i, a probabilidade de sucesso da variável resposta considerando as informações de cada observação de $\mathbf{x}_i$, o vetor de variáveis explicativas. Como a $\mathrm{E}(Y_i) = \pi_i$, as probabilidades π_i são representadas da forma

$$g(\pi_i) = \mathbf{x}_i\boldsymbol{\beta} \tag{3.6}$$

onde $\boldsymbol{\beta}$ é o vetor de parâmetros e $g(.)$ é a função de ligação.

Para garantir que π está restrito ao intervalo $[0, 1]$, este, é definido por uma função de distribuição acumulada, onde

$$\pi = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) = \int_{-\infty}^{t} f(s)\, ds \tag{3.7}$$

com $f(s) \geq 0$ e $\int_{-\infty}^{\infty} f(s)\, ds = 1$. $f(s)$ é chamada de **distribuição de tolerância**.

Distribuição que retorna a função de ligação do modelo de regressão logística, e quando substituído na equação (3.4) é encontrada a função de transformação logito.

$$\eta = g(\pi) = logito(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X\beta}$$
$$\pi = \frac{exp(\boldsymbol{X\beta})}{1 + exp(\boldsymbol{X\beta})} \tag{3.8}$$

Com as componentes do modelo definidas, a seguir, deve-se estimar o vetor de coeficientes $\boldsymbol{\beta}$ que indica o efeito em cada variável independente do modelo. Este processo será definido na Seção 3.5 onde será introduzida a teoria bayesiana e esclarecida a construção dos resultados de cada coeficiente. Uma vez com o vetor $\boldsymbol{\beta}$ estimado e validado pelo processo de inferência é encontrada a relação dos efeitos das covariáveis. E como os valores para cada $\boldsymbol{\beta_j}$ são amostrados, os resultados são indicadores da distribuição destes coeficientes, como pode ser visto na Tabela 8.

Tabela 8: Exemplo do resultado da estimação do modelo de regressão logística

Coeficiente	Média a posteriori	Desvio Padrão a posteriori	Razão de chances	IC (2.5% 97.5%)		Probabilidade de Significância
β_0	M(β_0)	DP(β_0)	-	-	-	p_{β_0}
β_1	M(β_1)	DP(β_1)	exp(M(β_1))	$q_{2.5}\beta_1$	$q_{97.5}\beta_1$	p_{β_1}
β_2	M(β_2)	DP(β_2)	exp(M(β_2))	$q_{2.5}\beta_2$	$q_{97.5}\beta_2$	p_{β_2}

A aplicação dos quantis (colunas 5 e 6) são para a construção do intervalo de credibilidade das razões de chances e a probabilidade de significância (coluna 7) indica o percentual do coeficiente estimado ser significativo.

A medida de razão de chances (coluna 4) são o grande resultado da regressão logística, a função de ligação definida na equação (3.8) tem a interpretação natural como o *log da razão de chances*, por isto, no exemplo da Tabela 8 este valor é representado pelo exponencial dos estimadores pontuais da média dos coeficientes. São apresentados mais detalhes na sequência.

A regressão logística pressupõe que o logaritmo da chance é linearmente relacionado com as variáveis explicativas. O que é uma das grandes vantagens desta modelagem, a possibilidade de interpretação direta dos coeficientes como medidas de associação, para compreender essa ferramenta de forma mais aplicada é destacada nesta seção a construção

em um caso específico, explorada no Capítulo 3 em Hosmer e Lemeshow (2000).

Considere inicialmente um modelo de regressão logística linear simples em que $\pi(x)$, a probabilidade de "sucesso"dado o valor x de uma variável explicativa qualquer, é definida tal que

$$log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{3.9}$$

em que β_0 e β_1 são parâmetros desconhecidos. Esse modelo poderia, por exemplo, ser aplicado para analisar a associação entre uma determinada doença e a ocorrência ou não de um fator particular.

Seriam então amostrados, independentemente, um grupo de n_1 indivíduos com presença do fator $(x = 1)$ e n_2 indivíduos com ausência do fator $(x = 0)$ e $\pi(x)$ seria a probabilidade de desenvolvimento da doença após um certo período fixo. Dessa forma, a chance de desenvolvimento da doença para um indivíduo com presença do fator fica dada por

$$\frac{\pi(1)}{1-\pi(1)} = \exp(\beta_0 + \beta_1 \times 1) = \exp(\beta_0 + \beta_1) \tag{3.10}$$

enquanto que a chance de desenvolvimento da doença para um indivíduo com ausência do fator é simplesmente dada por

$$\frac{\pi(0)}{1-\pi(0)} = \exp(\beta_0 + \beta_1 \times 0) = \exp(\beta_0) \tag{3.11}$$

Logo, a razão de chances fica dada por

$$OR = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \tag{3.12}$$

dependendo apenas do parâmetro β_1.

A razão de chances (*odds ratio* - OR), indica a chance de "sucesso"dado a ocorrência de um fator associado a um coeficiente. No modelo deste trabalho são observados algumas características de fatores associados aos indivíduos e a OR nos ajudará a interpretar a influência percentual de uma característica que indique a redução nas chances de ser considerado pobre.

Essa ocorrência pode ser dada diretamente pela presença ou não deste fator, como no

caso exemplificado acima, ou medida em relação a um nível base escolhido, forma aplicada para fatores com 3 ou mais níveis, onde será definido um nível base.

Para maior compreensão dos métodos estatísticos apresentados nesta seção, ver Capítulo 7 de Dobson e Barnett (2018). Com a definição das características do tipo de modelo que será trabalhado, na próxima seção, é definido o modelo proposto.

3.4.2 Modelo proposto

O modelo proposto deve ser capaz de estimar as chances de cada pessoa da população alvo ser pobre de acordo com as suas informações socioeconômicas, logo, é escolhido um MLG com a variável resposta seguindo uma distribuição de Bernoulli, dentro destes casos, é utilizada a regressão logística apresentada na Seção 3.4.1.

Neste trabalho serão estimados três modelos, referente as bases de dados estudadas para os anos de 2012, 2016 e 2020. Ao construir a estrutura do modelo é determinado também a informação dos fatores associados ao intercepto, com o efeito medido por β_0, neste caso será dado por moradores da Região metropolitana, do sexo masculino, na categoria de cor ou raça de brancos, em um domicílio com o responsável Sem instrução ou com menos de um ano na escola e em um domicílio com o(s) ocupado(s) fora do mercado de trabalho formal. E os coeficientes β_j, com $j = 1, \ldots, 9$, para medir os efeitos de cada fator em relação ao descrito no intercepto de forma individual, ou seja, o efeito de β_j será relacionado a categoria base deste fator encontrado no intercepto β_0. Por exemplo, para as medidas de escolaridade do responsável serão de comparações com indivíduos em domicílios com o responsável sem instrução.

Esse modelo, para cada indivíduo i resulta na seguinte estrutura:

$$\begin{aligned} Y_i \sim \text{Bern}\,(\pi_i) \quad & \mathit{logito}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{X_i\beta} \\ \mathit{logito}\,(\pi_i) &= \beta_0 + \beta_1 \times area_i + \beta_2 \times sexo_i + \beta_3 \times cor_i \\ &\quad + \beta_4 \times eresp2_i + \beta_5 \times eresp3_i + \beta_6 \times eresp4_i + \beta_7 \times eresp5_i \\ &\quad + \beta_8 \times ocupacaoSF_i + \beta_9 \times ocupacaoFI_i \end{aligned} \tag{3.13}$$

As variáveis explicativas usadas foram descritas na Seção 3.1.1 e a nomenclatura utilizada está representada na Tabela 9.

Tabela 9: Descrição das variáveis e as categorias utilizadas nos modelos

Nomenclatura	Variável	Categoria	Categoria base
area	Tipo de área	Capital	Região Metropolitana
sexo	Sexo	Feminino	Masculino
cor	Cor	Pretos ou pardos	Brancos
eresp2	Escolaridade do Responsável	Fundamental Incompleto	Sem Instrução
eresp3		Fundamental Completo	
eresp4		Médio Completo	
eresp5		Superior Completo	
ocupacaoSF	Ocupação no Domicílio	Somente Formal	Somente Informal
ocupacaoFI		Formal e Informal	

A seguir serão definidas as técnicas para realizar a estimação do modelo proposto, reunindo cada etapa do processo de estimação de um MLG com a inferência bayesiana.

3.5 Estimação Bayesiana

O processo de inferência consiste em reduzir os dados em estimativas, permitindo uma interpretação para o comportamento dos parâmetros que são desconhecidos.

O objetivo de usar a inferência dentro do conceito da modelagem estatística é estimar um conjunto de valores que representam a influência de cada fator numa variável dependente, onde cada número neste conjunto é referente as variáveis independentes, logo o tamanho deste conjunto é dado pelo número de fatores escolhidos para a construção do modelo. Neste caso, os valores são os coeficientes associados aos fatores socioeconômicos escolhidos para representar a pobreza, a variável dependente. E o conjunto, é o vetor de parâmetros do modelo, denominado por $\boldsymbol{\theta}$ e a tarefa da inferência é possibilitar a melhor estimação desse parâmetro.

Para estimar os parâmetros do MLG proposto de forma bayesiana, é visto nesta seção uma introdução para a teoria.

Para Paulino et al. (2018), as inferências bayesianas são baseadas em probabilidades subjetivas ou credibilidades a posteriori associadas com diferentes valores de um parâmetro de interesse $\boldsymbol{\theta}$ e condicionadas pelo particular valor de y observado.

Agora, definindo, seja Y uma variável aleatória definida em um espaço amostral Ω. Suponha que haja interesse em uma característica populacional desconhecida e relacionada a essa variável. Considere $p(Y = y \mid \boldsymbol{\theta})$ como a possível distribuição dessa variável condicionada a um vetor paramétrico $\boldsymbol{\theta}$ que representa essa característica, e que seja possível analisá-la com base nessa distribuição. Posto isto, é preciso inferir sobre esse vetor paramétrico, ou seja, fazer afirmações sobre $\boldsymbol{\theta}$. Para isso, pode-se utilizar um conjunto de dados.

Sob a abordagem bayesiana, publicada por Bayes (1763), o vetor paramétrico $\boldsymbol{\theta}$ é considerado uma variável aleatória. Logo, é possível incorporar uma distribuição de probabilidade a este vetor. Com isso, é elaborado outro atributo de uma inferência bayesiana, uma convicção inicial sobre $\boldsymbol{\theta}$, anterior à amostragem dos dados. Denote por $p(\boldsymbol{\theta})$ a distribuição a priori de $\boldsymbol{\theta}$ que representa probabilisticamente essa convicção inicial. A inferência sob $\boldsymbol{\theta}$ é realizada com base na distribuição a posteriori, denotada por $p(\boldsymbol{\theta} \mid Y = y)$, que é obtida através da distribuição da variável, $p(Y = y \mid \boldsymbol{\theta})$, e da distribuição a priori, $p(\boldsymbol{\theta})$, que resulta na equação (3.14), dada pelo Teorema de Bayes.

$$p(\boldsymbol{\theta} \mid Y = y) = \frac{p(Y = y \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(Y = y)} \tag{3.14}$$

A distribuição $p(Y = y)$ é chamada de distribuição marginal de Y e pode ser obtida combinando a distribuição $p(Y = y \mid \boldsymbol{\theta})$ com a distribuição $p(\boldsymbol{\theta})$. Se Y for uma variável aleatória contínua, então

$$p(Y = y) = \int_{\forall \boldsymbol{\theta}} p(Y = y \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_{\boldsymbol{\theta}}\left[p(Y = y \mid \boldsymbol{\theta})\right] \tag{3.15}$$

Caso Y seja uma variável discreta, basta considerar um somatório ao invés da integral. Seja $c^{-1} = p(Y = y)$, então a distribuição a posteriori dada na equação (3.14) pode ser reescrita como

$$p(\boldsymbol{\theta} \mid Y = y) = c \times p(Y = y \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{3.16}$$

Note que a constante c não depende de $\boldsymbol{\theta}$. Por isso, sob a inferência bayesiana, é comum utilizar a ideia de proporcionalidade e reescrever a equação (3.14) da seguinte forma:

$$p\left(\boldsymbol{\theta} \mid Y = y\right) \propto p\left(Y = y \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) \tag{3.17}$$

Desse modo, a expressão matemática costuma ser simplificada e torna-se mais fácil identificar o núcleo de uma distribuição conhecida para a posteriori. A constante c pode ser calculada recorrendo ao fato de que a integral (ou o somatório) de $p\left(\boldsymbol{\theta} \mid Y = y\right)$ com respeito a $\boldsymbol{\theta}$ tem que ser igual a 1.

Quando o vetor paramétrico $\boldsymbol{\theta}$ for desconhecido, ao calcular a distribuição $p\left(Y = y \mid \boldsymbol{\theta}\right)$ para um valor observado y da variável aleatória Y, obtém-se uma função que depende de $\boldsymbol{\theta}$. Essa função que é chamada de função de verossimilhança e passa a ser denotada por $l(\mathbf{y}; \boldsymbol{\theta})$. Essa expressão quando aplicada a diferentes valores de $\boldsymbol{\theta}$ informa quais valores parecem ser mais verossímeis. Logo, a distribuição a posteriori é encontrada de forma proporcional seguindo

$$p\left(\boldsymbol{\theta} \mid Y = y\right) \propto l\left(\mathbf{y}; \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) \tag{3.18}$$

Dessa forma é possível aplicar o Teorema definido na equação (3.14), relacionando a distribuição a posteriori com a verossimilhança e a priori. Inferências sumárias podem ser obtidas na forma de valores esperados a posteriori para escolhas apropriadas de $h(.)$, como

$$E\left[h(\boldsymbol{\theta} \mid y)\right] = \int h(\boldsymbol{\theta}) p\left(\boldsymbol{\theta} \mid Y = y\right) d\boldsymbol{\theta}$$

Assim, no caso contínuo, a operação de integração desempenha um papel fundamental em estatística bayesiana. Contudo, raramente é possível obter expressões explícitas para os integrais envolvidos. Consequentemente, as dificuldades encontradas na resolução dos integrais necessários para fazer inferências bayesianas serviram, durante longo tempo, como impedimento à aplicação generalizada desta metodologia. (Paulino et al. (2018))

Em Migon, Gamerman e Louzada (2014) há maiores detalhes sobre inferência bayesiana. Na Subseção 3.5.1 há uma discussão sobre como definir a distribuição a priori e na Subseção 3.5.2 é apresentado um método iterativo que pode ser utilizado para avaliar a distribuição a posteriori quando sua forma é desconhecida, impedindo a estimação de forma analítica.

3.5.1 Propostas para a distribuição a priori

A distribuição a priori $p(\boldsymbol{\theta})$ deve representar toda a crença probabilística sobre o parâmetro de interesse $\boldsymbol{\theta}$. Um especialista pode ter muito conhecimento prévio sobre o parâmetro desconhecido, tornando mais simples a tarefa de especificar uma distribuição a priori. Porém, ele também pode ter pouco conhecimento sobre $\boldsymbol{\theta}$ e saber especificar apenas a média e a variância, por exemplo. Ou, ainda, pode não haver qualquer informação sobre $\boldsymbol{\theta}$ antes do experimento. Quando há informação sobre a distribuição a priori, pode-se incorporar o conhecimento da média e da variância, se houver, através dos parâmetros dessa priori, chamados de hiperparâmetros. Caso não haja informação alguma, basta atribuir uma variância grande o suficiente para essa distribuição.

3.5.2 Métodos iterativos para a posteriori

Em razão da composição da distribuição a posteriori, definida na equação (3.18), em muitos casos, não é possível encontrar sua distribuição analiticamente. Com isso, uma solução para estimar os parâmetros da posteriori são os métodos iterativos que são apresentados nesta seção, construída para conceituar alguns deste métodos em etapas para uma compreensão mais dinâmica do que será aplicado neste trabalho. A construção completa para as simulações de cadeias de Markov, além das etapas a seguir descritas, podem ser vistas em Gelman et al. (2013).

Monte Carlo

Um processo para amostrar uma sequencia de valores aleatórios de uma distribuição proposta, **D**, com parâmetros fixos em todo o processo.

Cadeias de Markov

Uma cadeia de Markov de primeira ordem, ou processo de Markov de primeira ordem, é uma sequência de valores amostrados de variáveis aleatórias na forma $\theta_0, \theta_1, \ldots$ com espaço finito ou infinito enumerável, tal que a distribuição proposta para amostrar um novo valor é condicionada de θ_t, parâmetro desta distribuição. Isto é, o valor amostrado a cada iteração t é definido como o parâmetro da distribuição proposta para amostrar o valor $t + 1$, de forma que a probabilidade de uma cadeia assumir um certo valor futuro

depende apenas do seu estado atual, assim,

$$p\left(\theta_t \in A \mid \theta_0, \ldots, \theta_{t-1}\right) = p\left(\theta_t \in A \mid \theta_{t-1}\right) \tag{3.19}$$

para qualquer subconjunto A.

3.5.3 Monte carlo via cadeias de markov

Os métodos de Monte Carlo via Cadeias de Markov (MCMC) consistem em uma classe de algoritmos para amostrar valores de uma distribuição de probabilidade de interesse usando cadeias de Markov, sendo uma alternativa aos métodos não iterativos nos problemas em que as soluções analíticas são inviáveis ou complexas. De acordo com Gamerman e Lopes (2006), para que sejam usados os métodos MCMC, a cadeia deve ser **homogênea** (as probabilidades de transição de um estado para outro são invariantes), **irredutível** (cada estado pode ser atingido a partir de qualquer outro em um número finito de iterações) e **aperiódica** (não haja estados absorventes).

Na inferência bayesiana, os métodos de MCMC são muito utilizados para obter uma amostra da distribuição a posteriori de $\boldsymbol{\theta}$, permitindo assim inferir sobre o vetor paramétrico desconhecido. Com base na amostra obtida pelo MCMC, pode-se calcular as estimativas amostrais da distribuição de interesse.

Os métodos de MCMC mais utilizados na inferência bayesiana são o amostrador de Gibbs e o algoritmo de Metropolis-Hastings. As etapas seguintes fazem uma revisão sobre cada um desses algoritmos aplicados ao contexto da inferência bayesiana.

Amostrador de Gibbs

O Amostrador de Gibbs, proposto por Geman e Geman (1984) e introduzido por Gelfand e Smith (1990), é uma cadeia de Markov na qual não há um método de aceitação-rejeição, ou seja, a cadeia sempre irá para um novo valor.

Considere que o interesse esteja em amostrar um vetor ou matriz $\boldsymbol{\theta}$ da distribuição a posteriori $p\left(\boldsymbol{\theta} \mid Y = y\right)$, sendo y um conjunto de dados observados. Suponha que esse conjunto $\boldsymbol{\theta}$ seja particionado em d componentes e que cada componente possa ser um escalar ou um vetor ou mesmo uma matriz. Por simplicidade, considere que sejam d esca-

lares. O amostrador de Gibbs requer a obtenção das distribuições condicionais completas a posteriori, ou seja, das distribuições $p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i-1}, \ldots, \theta_d)$. As transições de um estado para o outro são feitas através dessas distribuições.

O amostrador de Gibbs é definido pelo esquema a seguir:

1. Inicialize o contador de iterações da cadeia $t = 0$;
2. Especifique valores iniciais $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_d^{(0)}\right)'$;
3. Obtenha um novo valor de $\boldsymbol{\theta}^{(t)}$ a partir de $\boldsymbol{\theta}^{(t-1)}$ através da geração sucessiva dos valores

$$\begin{aligned}
\theta_1^{(t)} &\sim p\left(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)}, \mathbf{y}\right) \\
\theta_2^{(t)} &\sim p\left(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)}, \mathbf{y}\right) \\
&\vdots \\
\theta_d^{(t)} &\sim p\left(\theta_d \mid \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{d-1}^{(t)}, \mathbf{y}\right)
\end{aligned}$$

4. Incremente o contador de t para $t + 1$ e retorne ao passo 2 até obter convergência.

Após a convergência, os valores resultantes formam uma amostra de $p(\boldsymbol{\theta} \mid Y = y)$.

Algoritmo de Metropolis-Hastings

Os algoritmos de Metropolis-Hastings, introduzido por Metropolis et al. (1953) e estendido por Hastings (1970) para o caso mais geral, tem por objetivo simular uma distribuição de probabilidade desconhecida. Este algoritmo nos garante a convergência para uma certa distribuição, chamada de distribuição de equilíbrio, que, na inferência bayesiana, pode ser a distribuição a posteriori. A grande diferença deste algoritmo é a etapa de validação do vetor de parâmetros $\boldsymbol{\theta}$, ou seja, insere uma regra de decisão após cada iteração t de substituir ou não $\boldsymbol{\theta}^{t-1}$ por $\boldsymbol{\theta}^{t}$ para ser usado para estimar os valores de $\boldsymbol{\theta}^{t+1}$.

Considere que o interesse esteja em amostrar a distribuição $p(\boldsymbol{\theta} \mid Y = y)$. Para isto, o algoritmo de Metropolis-Hastings consiste nos seguintes passos:

1. Inicialize o contador de iterações $t = 0$, especifique um valor inicial para o parâmetro denotando-o por $\boldsymbol{\theta}^0$;

2. Incremente o contador de t para $t + 1$;

3. Gere um valor $\boldsymbol{\theta}^p$ de uma distribuição auxiliar conhecida $q\left(\boldsymbol{\theta}^p \mid \boldsymbol{\theta}^{t-1}\right)$;

4. Calcule a seguinte probabilidade de aceitação

$$\alpha\left(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^p\right) = \min\left(1, \frac{p\left(\boldsymbol{\theta}^p \mid y\right) q\left(\boldsymbol{\theta}^{t-1} \mid \boldsymbol{\theta}^p\right)}{p\left(\boldsymbol{\theta}^{t-1} \mid y\right) q\left(\boldsymbol{\theta}^p \mid \boldsymbol{\theta}^{t-1}\right)}\right);$$

5. Gere $u \sim U(0, 1)$;

6. Se $u \leq \alpha$ então aceite o novo valor fazendo $\boldsymbol{\theta}^t = \boldsymbol{\theta}^p$ e incremente o contador *cont* para *cont* +1, caso contrário, rejeite o valor gerado fazendo $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$ e *cont* não se altera.

7. Volte ao passo 2 até obter convergência.

Para mais informações sobre os algoritmos explorados ou outros algoritmos relacionados, buscar pela Parte III em Gelman et al. (2013). Assim é concluída a subseção de métodos iterativos, a seguir, é vista a forma de validação a ser aplicada no modelo proposto.

3.6 Análise de convergência

Nesta seção é mostrado como serão validadas as simulações da posteriori resultante do processo de amostragem para cada parâmetro desconhecido do modelo. Métodos mais extensos e com maiores discussões sobre validação da estimação por MCMC de modelos bayesianos podem ser vistos em Gelman et al. (2013).

Para essa validação, além de Gilks, Richardson e Spiegelhalter (1995), são visitados os capítulos de diagnóstico de métodos MCMC em Paulino et al. (2018). A partir das cadeias geradas, são identificados os valores sugeridos de tamanho de amostra e do período de aquecimento (Raftery e Lewis (1992)). Com amostras mais robustas e um fator de dependência aceitável, são realizados outros testes para uma análise mais completa da convergência das cadeias. É verificada a convergência individual de cada coeficiente com o teste de Geweke (Geweke (1992)), verificada a estacionariedade das cadeias (Heidelberger

e Welch (1983)) e calculado o fator de redução de escala para indicar convergência em todas as cadeias.(Gelman e Rubin (1992)).

Para a construção das amostras piloto e das atualizações, é considerada uma interpretação com muitas dimensões, um número de iterações suficientemente grande, um período de aquecimento, o espaçamento entre as iterações, o ponto de partida para cada parâmetro e resultados como autocorrelação entre as iterações ou uma convergência lenta, além dos testes citados anteriormente, ou seja, após um processo de avaliação individual para cada modelo estimado, são definidas as cadeias efetivas. As cadeias efetivas contém os valores da amostra da distribuição a posteriori.

O diagnóstico da estimação de um MLG bayesiano com métodos MCMC sugerido por Gelman e Rubin (1992) busca verificar se uma distribuição alvo foi encontrada através das variâncias de sequências múltiplas das cadeias, para isso, é necessário ter mais de uma (1) cadeia estimada. As variâncias dentro da sequência e entre as sequências compõem o método que calcula o indicador de *fator de redução de escala* $\hat{R}$. Valores de $\hat{R} \approx 1$ são um indício de que cada uma das sequências simuladas se aproxima da distribuição alvo.

O método de Geweke (1992), é baseado em técnicas de séries temporais para constatar convergência nas sequências amostradas, através de um teste para estacionariedade. Para isto, ele realiza um teste de comparação de médias entre a média do início e a média do fim da sequência com uma estatística que tem distribuição normal padrão, o valor resultante é um quantil desta estatística, q_z, com ela é possível encontrar a probabilidade associada ao q_z para determinar se há ou não indicação de convergência.

Para encontrar o número de iterações total, N e do período de aquecimento, M o método proposto por Raftery e Lewis (1992) parte de condições previamente especificadas, que tem a suposição de estimar um quantil q da função a posteriori de uma parâmetro, com uma tolerância r e uma probabilidade, s, de estar dentro desta tolerância. Essa estimação resulta também no valor mínimo para uma amostra piloto, N_{min}, além de N e M. Que integram o *fator de dependência*, $I = (M+N)/N_{min}$. Valores elevados deste fator (> 5) podem indicar valores iniciais influentes, correlação elevada entre os coeficientes ou uma cadeia com fraca mistura no suporte da distribuição a posteriori.

Heidelberger e Welch (1983) propuseram uma estatística de teste, baseada no teste estatístico de Cramér-Von Mises (Darling (1957)), para testar a hipótese nula de que a cadeia de markov simulada provém de uma distribuição estacionária. O método busca um teste de estacionariedade que se repete, removendo um período inicial de 10% da sequência original a cada repetição, caso seja rejeitada a hipótese nula de estacionariedade.

Existe uma condição de parada, caso chegue a metade da sequência, 50%, dessa forma, o resultado do teste é que a sequência daquele parâmetro falhou no teste de estacionariedade. De outra forma, ao não se rejeitar a hipótese de estacionariedade em alguma etapa, é feita uma comparação com a sequência restante, usando a média amostral, o desvio padrão amostral e um nível de significância α pré-definido. O valor de α comum é 5% e este foi usado em todos os testes.

3.7 Estimação aplicada no modelo proposto

Nesta parte, o modelo proposto será estimado para definir as cadeias simuladas dos parâmetros da posteriori. Em Gilks, Richardson e Spiegelhalter (1995) é discutido o processo de implementação de amostragem por MCMC. Este livro é uma referência para as estruturas das técnicas aplicadas, porém, estas não são determinadas totalmente por ele, mas também por estudos atuais citados mais à frente.

O processo, de forma resumida, consiste em, primeiro é definido que cada coeficiente β_j é uma variável aleatória, junto a isso são definidas as distribuições a priori. Em sequência é determinada a verossimilhança da distribuição da variável resposta, e então, é definido o formato da posteriori e a forma de estimação de $\boldsymbol{\beta}$.

Para o início das etapas, conforme Acquah (2013), serão utilizadas prioris não informativas visando abranger mais possibilidades, logo para o nosso conhecimento inicial do conjunto de parâmetros desconhecidos, o vetor $\boldsymbol{\beta}$, de tamanho $p\times 1$, com $j = 0, 1, \ldots, p-1$, é escolhido para cada valor $\boldsymbol{\beta}_j$ uma distribuição (iid) normal com média zero e variância de 1000, dessa forma, é encontrada a distribuição conjunta a priori de $\boldsymbol{\beta}$ a $p(\boldsymbol{\beta})$ que será necessária mais a frente.

$$\boldsymbol{\beta}_j \sim \boldsymbol{N}\left(\mu_j = 0; \sigma_j^2 = 1000\right) \qquad \boldsymbol{\beta} \sim \boldsymbol{N}_p\left(\underline{\mu}; \boldsymbol{V}_\beta\right) \tag{3.20}$$

Onde $\underline{\mu}$ é um vetor nulo, $p \times 1$, e $\boldsymbol{V}_\beta$ é a matriz de covariâncias de $\boldsymbol{\beta}$, que é uma matriz diagonal $p \times p$ com valores da diagonal principal de acordo com a variância da cada β_j, 1000.

A seguir, é definida a função de verossimilhança, como as variáveis respostas são (iid), é o caso do produto da função de probabilidade de uma Bernoulli e a variável aleatória Y_i, definida na equação (3.5), aplicada a transformação linear da função **logito** no lugar do parâmetro de probabilidade π_i, como mostrado nas equações (3.6) e (3.8), chega-se na

equação (3.21) que é a nossa função dos dados observados, $p(Y = y \mid \boldsymbol{\beta})$.

$$l(\mathbf{y}; \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\left(\frac{\exp(\boldsymbol{x_i}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}\boldsymbol{\beta})} \right)^{y_i} \left(\frac{\exp(\boldsymbol{x_i}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}\boldsymbol{\beta})} \right)^{(1-y_i)} \right] \tag{3.21}$$

Agora, empregada a forma de inferência escolhida, definida na Seção 3.5, no modelo proposto está construída uma regressão logística que sob a inferência bayesiana é composta pela distribuição a priori, a função de verossimilhança e a distribuição a posteriori relacionadas de acordo com a equação (3.17). Com isso, é encontrada a equação (3.22), a seguir

$$\begin{aligned} P(\boldsymbol{\beta} \mid Y = y) \propto & \prod_{i=1}^{n} \left[\left(\frac{\exp(\boldsymbol{x_i}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}\boldsymbol{\beta})} \right)^{y_i} \left(\frac{\exp(\boldsymbol{x_i}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}\boldsymbol{\beta})} \right)^{(1-y_i)} \right] \\ & \times \prod_{j=0}^{p-1} \left[\frac{1}{\mid \boldsymbol{V}_{\beta} \mid^{-\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{\beta}^T \boldsymbol{V}_{\beta}^{-1} \boldsymbol{\beta}) \right) \right] \end{aligned} \tag{3.22}$$

Ao construir um MLG e buscar inferir sobre os parâmetros desconhecidos através da inferência bayesiana exige-se que a distribuição a posteriori tenha uma forma fechada para encontrar uma integral desta distribuição multivariada, logo, em muitos casos, não é possível definir a distribuição conjunta da posteriori de forma analítica, para isto, recorre-se aos métodos iterativos de Monte Carlo via Cadeias de Markov discutidos na Subseção 3.5.3, para amostrar a distribuição dos coeficientes do vetor $\boldsymbol{\beta}$ e encontrar as distribuições a posteriori de cada $\boldsymbol{\beta_j}$.

Agora, é descrita a parte computacional do processo, o programa que irá gerar as amostras para definir a distribuição dos parâmetros da posteriori é o JAGS (Plummer (2017)) que será gerenciado remotamente por um código programado no pacote estatístico R (R Core Team (2014)), através do pacote **rjags** elaborado por Plummer, Stukalov e Denwood (2019).

Para este processo de estimação de parâmetros é necessário definir além das prioris, os valores iniciais para cada parâmetro em cada cadeia, uma lista com valores observados e um arquivo de texto codificado com as informações do modelo.

Algumas características precisam ser citadas antes da observação dos arquivos com as etapas da estimação definidas, material disponível no Apêndice 2. No arquivo de texto a função *logit* indica a transformação **logito**, aplicada ao parâmetro π_i, da variável resposta y_i com distribuição Bernoulli. E para as prioris o parâmetro de escala da distribuição

normal é a precisão τ, definida como o inverso da variância.

A avaliação dada as cadeias partiu de uma amostragem piloto, para cada ano, com um aquecimento inicial de 1000 amostras e mais 5000 amostras sem saltos. O processo de testagem inicial para identificar os valores sugeridos de parâmetros da simulação das novas cadeias que seriam simuladas para obter a convergência e permitir a validação do processo foi realizada através dos pacotes **coda** e **boa** para a linguagem R. (Plummer et al. (2006)) (Smith (2007)).

Concluída a apresentação dos testes aplicados, o texto segue com a avaliação dos modelos estimados.

4 Análise dos Resultados

O estudo compreende três modelos diferentes associados aos períodos estudados e, na seção a seguir, é feita a apresentação das simulações realizadas em conjunto com os testes aplicados e a na Seção 4.2 são descritos os resultados estimados e a interpretação destes acompanhada de análises preliminares.

4.1 Análise das cadeias do modelo proposto

Pelo processo de inferência, para a estimação do modelo proposto aplicado nos três períodos, o primeiro passo é analisar as cadeias simuladas para estimação do efeito sobre cada fator do modelo proposto. Para isto, são considerados o traço e a distribuição associada as simulações, nas próximas figuras são expostas as estimações dos modelos para os períodos de 2012, 2016 e 2020.

As análises dos traços das cadeias simuladas de 2012, Figura 5, indicam convergência dos coeficientes, apesar de ter uma menor concentração em algumas cadeias, identificadas na Figura 6, pode-se ver um ponto com maior número de valores, os pontos de mais acúmulo de valores para os coeficientes da distribuição a posteriori no gráfico das distribuições. A primeira vista, os coeficientes de *area* e *sexo* apresentam grandes chances de não serem significativos, ou seja, o fator pode ter efeito zero no modelo.

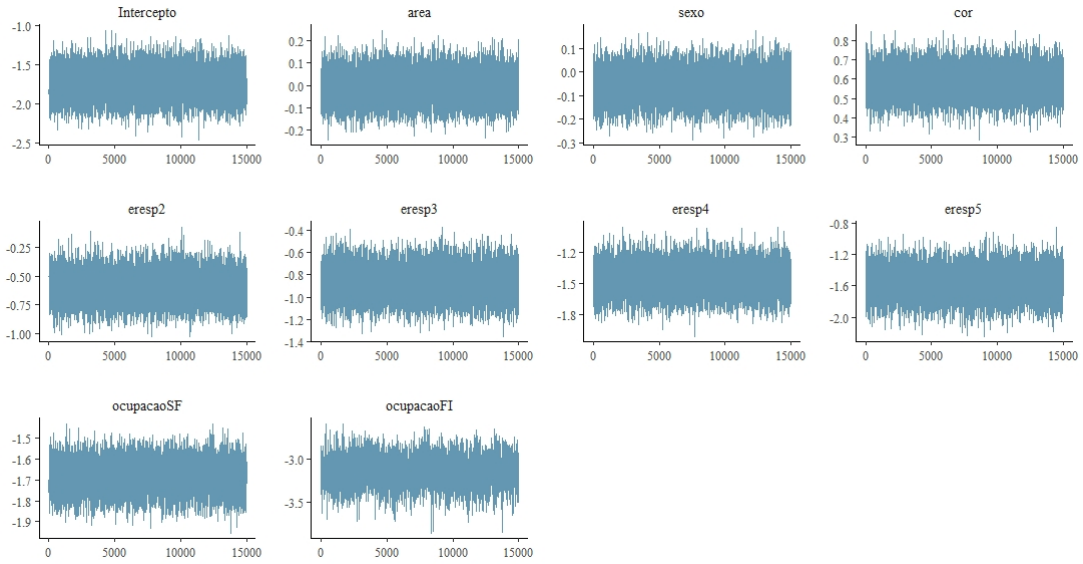

Figura 5: Gráfico dos traços das cadeias amostradas dos coeficientes da regressão de 2012

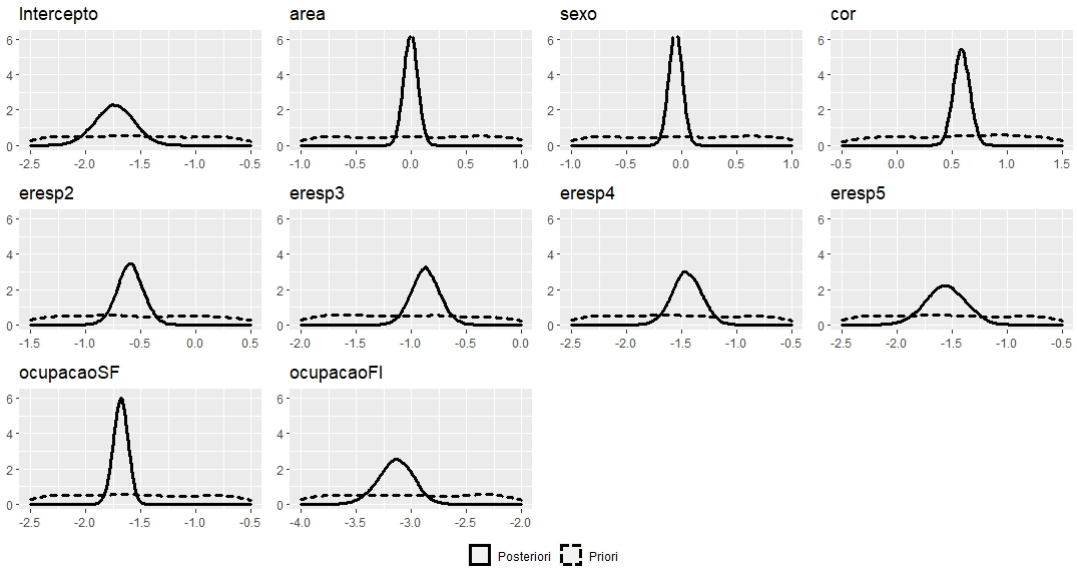

Figura 6: Gráfico das distribuições resultantes e da proposta inicial dos coeficientes da regressão de 2012

Considerando os traços das cadeias simuladas de 2016, Figura 7, elas apresentam convergência dos coeficientes. Para a maioria dos fatores, exceto para *ocupacaoFI*, é visto um grande acúmulo dos valores da posteriori permitindo identificar valores prováveis na Figura 8. Pelos gráficos, o coeficiente de *area* apresenta grandes chances de não ser significativo, ou seja, o fator pode ter efeito zero no modelo.

Retornando uma discussão, como visto na análise descritiva, a categoria *ocupacaoFI* que mede o efeito da categoria de ocupação no domicílio como ocupação formal e infor-

mal, mostra uma dispersão muito grande no traço o que fica claro também na Figura 8. Uma estatística pontual para representar o valor do efeito não tem muita eficiência para interpretação, mas será exibido nos resultados. Apesar deste fato, a distribuição do coeficiente adiciona valor as análises e a comparação principal do trabalho, a relação de formais e informais, medida pelo coeficiente *ocupacaoSF* mostra que não foi afetada, indicando bem os valores prováveis para o efeito, na Figura 8 e com um traço estável como visto na Figura 7.

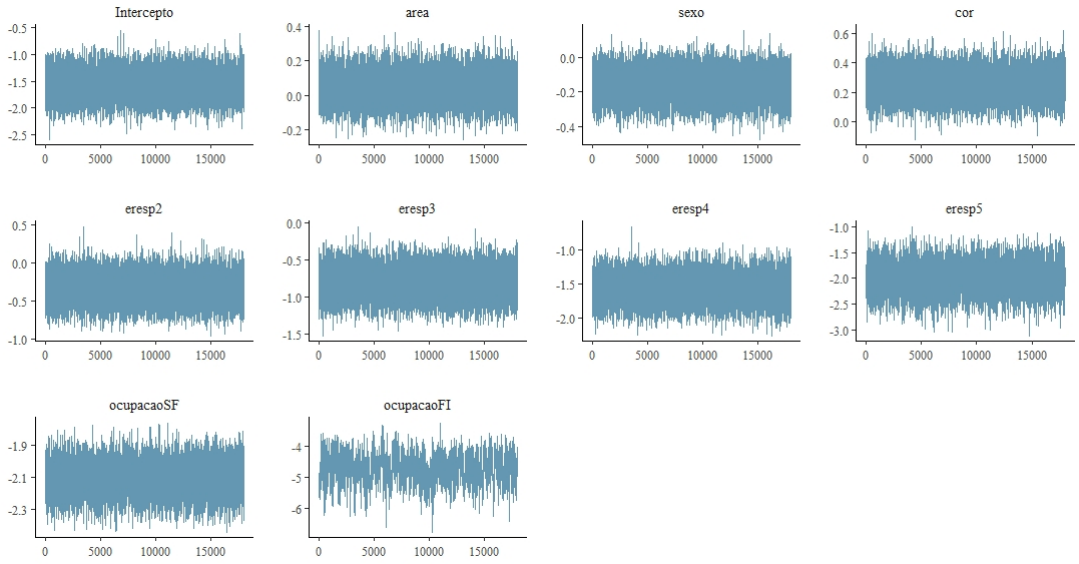

Figura 7: Gráfico dos traços das cadeias amostradas dos coeficientes da regressão de 2016

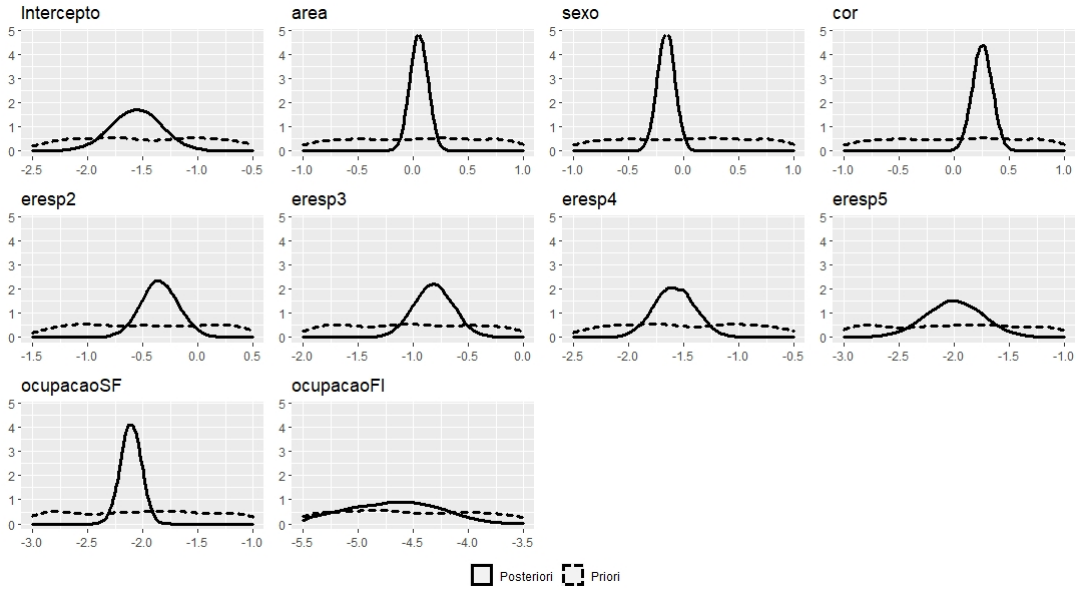

Figura 8: Gráfico das distribuições resultantes e da proposta inicial dos coeficientes da regressão de 2016

As análises dos traços das cadeias simuladas de 2020, Figura 9, indicam convergência dos coeficientes, apesar de uma maior dispersão em várias cadeias, identificadas na Figura 10, pode-se ver um ponto com maior número de valores, indicando valores prováveis para uma estimativa pontual dos coeficientes da distribuição a posteriori no gráfico das distribuições. A primeira vista, o coeficiente de *sexo* apresenta grandes chances de não ser significativo, ou seja, o fator pode ter efeito zero no modelo.

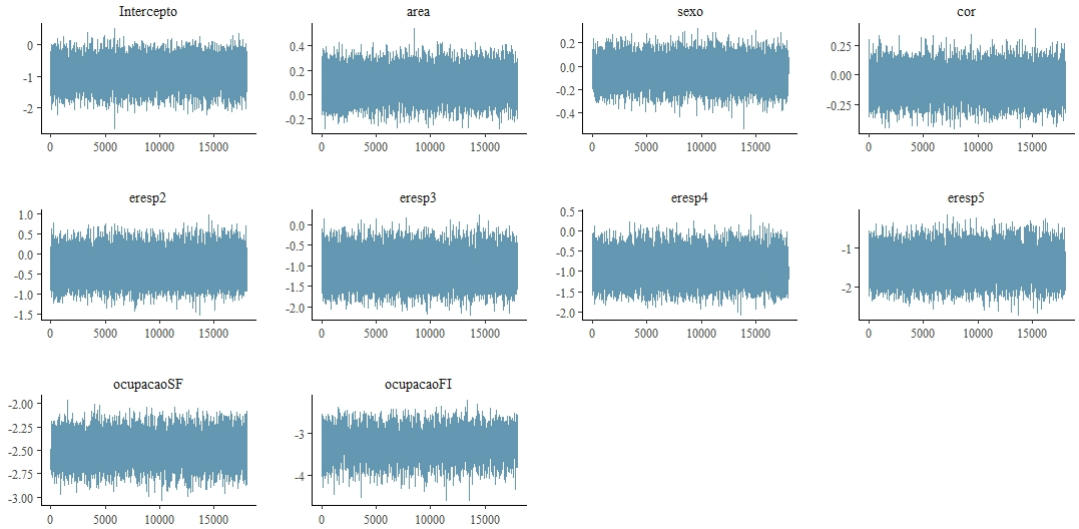

Figura 9: Gráfico dos traços das cadeias amostradas dos coeficientes da regressão de 2020

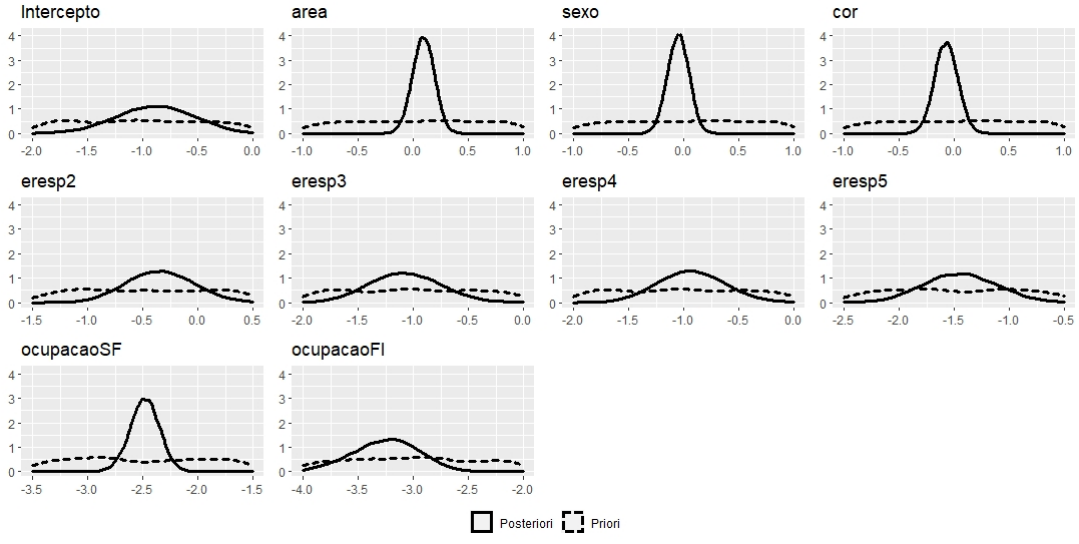

Figura 10: Gráfico das distribuições resultantes e da proposta inicial dos coeficientes da regressão de 2020

Com as cadeias resultantes estimadas, como definido na Seção 3.7. As Tabelas 10, 11 e 12 identificam os resultado dos testes aplicados nas simulações dos períodos estudados que concluíram o processo de estimação.

Baseado nos testes apresentados na Seção 3.6, nas simulações do período de 2012, descritos nas Tabela 10, verifica-se que a distribuição alvo foi encontrada, indica que, ao nível de significância de 5%, as sequências convergiram, sugere que não existem valores elevados do fator de dependência e que não rejeita-se o teste de que a cadeia provém de uma distribuição estacionária.

Tabela 10: Resultados dos testes para avaliação das simulações - 2012

Coeficiente	$\hat{R}$	Geweke (p-valor)		Fator de Dependência		Estacionariedade (H. e W.)	
		Cadeia 1	Cadeia 2	Cadeia 1	Cadeia 2	Cadeia 1	Cadeia 2
Intercepto	1	0.659	0.664	1.072	1.090	0.069	0.935
area	1	0.991	0.176	1.152	1.099	0.068	0.806
sexo	1	0.376	0.307	1.158	1.133	0.713	0.742
cor	1	0.169	0.342	1.213	1.158	0.438	0.650
eresp2	1	0.538	0.624	1.021	1.032	0.517	0.429
eresp3	1	0.509	0.405	1.072	1.043	0.931	0.456
eresp4	1	0.491	0.724	1.055	1.060	0.606	0.534
eresp5	1	0.421	0.571	2.372	1.315	0.067	0.498
ocupacaoSF	1	0.279	0.269	1.210	1.114	0.863	0.430
ocupacaoFI	1	0.322	0.577	4.340	3.687	0.686	0.817

Baseado nos testes apresentados na Seção 3.6, nas simulações do período de 2016, descritos nas Tabela 11, verifica-se que a distribuição alvo foi encontrada, indica que, ao nível de significância de 5%, as sequências convergiram e que não rejeita-se o teste de que a cadeia provém de uma distribuição estacionária. No entanto, houve um resultado, esperado pelo visto na análise descritiva, de valores elevados do fator de dependência para o coeficiente *ocupacaoFI*, para as duas cadeias amostradas. Este resultado não invalida as análises, porém os resultados somente para este fator no período de 2016 não são confiáveis.

Tabela 11: Resultados dos testes para avaliação das simulações - 2016

Coeficiente	$\hat{R}$	Geweke (p-valor)		Fator de Dependência		Estacionariedade (H. e W.)	
		Cadeia 1	Cadeia 2	Cadeia 1	Cadeia 2	Cadeia 1	Cadeia 2
Intercepto	1	0.518	0.397	1.125	1.136	0.868	0.999
area	1	0.175	0.672	1.228	1.257	0.834	0.535
sexo	1	0.147	0.808	1.347	1.222	0.215	0.675
cor	1	0.909	0.564	1.240	1.335	0.415	0.903
eresp2	1	0.104	0.448	1.009	1.013	0.448	0.905
eresp3	1	0.162	0.288	1.041	1.046	0.469	0.714
eresp4	1	0.168	0.900	1.125	1.060	0.536	0.751
eresp5	1	0.370	0.463	3.936	2.899	0.270	0.346
ocupacaoSF	1	0.621	0.390	2.266	2.320	0.215	0.836
ocupacaoFI	1	0.759	0.858	17.173	16.536	0.284	0.649

Baseado nos testes apresentados na Seção 3.6, nas simulações do período de 2020, descritos nas Tabela 12, verifica-se que a distribuição alvo foi encontrada, sugere que não existem valores elevados do fator de dependência e que não rejeita-se o teste de que a cadeia provém de uma distribuição estacionária. Ainda assim, na verificação de convergência das sequencias, os coeficientes de *area* e *ocupacaoFI* da cadeia 1 rejeitaram a hipótese de que as sequências convergiram.

Tabela 12: Resultados dos testes para avaliação das simulações - 2020

Coeficiente	$\hat{R}$	Geweke (p-valor)		Fator de Dependência		Estacionariedade (H. e W.)	
		Cadeia 1	Cadeia 2	Cadeia 1	Cadeia 2	Cadeia 1	Cadeia 2
Intercepto	1	0.891	0.350	1.061	1.060	0.558	0.860
area	1	0.029	0.391	1.152	1.146	0.336	0.441
sexo	1	0.918	0.877	1.100	1.146	0.207	0.658
cor	1	0.051	0.696	1.105	1.125	0.471	0.804
eresp2	1	0.356	0.565	1.022	0.990	0.974	0.664
eresp3	1	0.650	0.536	1.032	1.041	0.934	0.584
eresp4	1	0.360	0.724	0.999	1.004	0.945	0.616
eresp5	1	0.248	0.141	1.100	1.027	0.809	0.312
ocupacaoSF	1	0.950	0.711	2.463	2.608	0.114	0.811
ocupacaoFI	1	0.022	0.323	4.098	4.105	0.228	0.654

4.2 Estimativas

Os resultados apresentados são compostos, entre outros indicadores, pelas razões de chances (OR) dos parâmetros estimados e validados para todos os períodos do estudo. Com as Tabelas 13, 14 e 15 são possíveis as análises de saída dos coeficientes dos modelos.

Para chegada aos resultados dos três períodos, é necessário a escolha de uma cadeia para apresentação dos indicadores e, após a avaliação das cadeias, dos gráficos de distribuição e, principalmente, pelos resultados dos testes vistos na Seção 4.1, a cadeia de número 2 foi escolhida para a análise dos resultados de todos os períodos. A decisão foi individual para cada período. Dito isso, o trabalho segue para os resultados.

Previamente, indicadores negativos da média a posteriori para os coeficientes indicam redução das chances em relação a categoria base para fatores com 3 ou mais categorias. Com isso, é esperado que haja redução nas chances de ser pobre nas categorias de ocupação comparadas com indivíduos em domicílios somente com ocupados informais e nas categorias de escolaridade do responsável comparadas com indivíduos em domicílios com o responsável sem instrução. Foi realizada uma análise independente para cada coeficiente de cada período que são descritas após a respectiva tabela.

Tabela 13: Resultado da estimação do modelo de regressão logística - 2012

Coeficiente	Média a posteriori	Desvio padrão a posteriori	OR	Intervalo de credibilidade de OR		Probabilidade de significância
				2.5%	97.5%	
Intercepto	-1.74	0.17	-	-	-	1.00
area	-0.01	0.06	0.990	0.878	1.127	0.54
sexo	-0.06	0.06	0.942	0.835	1.062	0.83
cor	0.58	0.07	1.786	1.553	2.075	1.00
eresp2	-0.60	0.12	0.549	0.436	0.698	1.00
eresp3	-0.87	0.13	0.419	0.326	0.538	1.00
eresp4	-1.45	0.13	0.235	0.179	0.304	1.00
eresp5	-1.56	0.18	0.210	0.147	0.298	1.00
ocupacaoFI	-3.14	0.16	0.043	0.031	0.058	1.00
ocupacaoSF	-1.68	0.07	0.186	0.162	0.212	1.00

Para a análise das tabelas de resultados verifica-se o indicador da razão de chances, definido na Seção 3.4.1, ao realizar comparações entre as categorias das variáveis do modelo proposto, identificadas na Tabela 9.

Em prol de uma análise direta do indicador, em 2012 é medido para ocupação formal que morar em um domicílio somente com ocupado(s) no mercado formal (*ocupacaoSF*) reduz a chance de ser pobre, dado o recorte de pobreza do estudo, em $81,4\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal.

Ao analisar a Tabela 13, note que apesar dos intervalos de credibilidade das razões de chance das variáveis *area* e *sexo* conterem o valor unitário, a variável *sexo* tem alta probabilidade de significância. Isso mostra que a massa de probabilidade da distribuição a posteriori indica efeito significativo dessa variável. Portanto, é possível interpretar a variável *sexo* da seguinte forma: ser do sexo feminino reduz a chance de um indivíduo estar abaixo da linha de pobreza em $5,8\%$ em comparação com indivíduos do sexo masculino, com probabilidade de $0,83$ desse efeito ser, de fato, negativo.

Quanto ao coeficiente de *area*, residir na capital reduz a chance de um indivíduo ser pobre em $1,0\%$ em comparação com indivíduos que residem na RM, com probabilidade de $0,54$ desse efeito ser negativo. Como a variação da chance é muito pequena e tem um

resultado quase aleatório, probabilidade próxima de 0.50, o coeficiente de *area* pode ser considerado não significativo para o modelo de 2012.

Seguindo, para inferir sobre cor ou raça, o coeficiente *cor*, indica que ser preto ou pardo aumenta a chance de um indivíduo ser pobre em $78,6\%$ em comparação com indivíduos brancos.

Para os coeficientes de escolaridade do responsável, é encontrado que residir em um domicílio com a escolaridade do responsável como fundamental incompleto (*eresp2*) reduz a chance de estar abaixo da linha de pobreza em $45,1\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

Estar morando num domicilio cuja escolaridade do responsável é Fundamental completo (*eresp3*) reduz a chance do indivíduo ser pobre em $58,1\%$, quando comparado com indivíduos que residem em domicílios com os responsáveis sem instrução.

Viver em um domicílio com a escolaridade do responsável como médio completo (*eresp4*) reduz a chance de estar abaixo da linha de pobreza em $76,5\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

Morar em um domicílio com a escolaridade do responsável como superior completo (*eresp5*) reduz a chance de estar abaixo da linha de pobreza em $79,0\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

No caso da ocupação no domicílio, morar com ocupados no mercado formal e informal (*ocupacaoFI*) reduz a chance de ser pobre em $95,7\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal.

E, para revisar o resultado, viver com somente ocupado(s) no mercado formal (*ocupacaoSF*) reduz a chance de ser pobre em $81,4\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal.

Avançando para o resultado de 2016, é analisada a Tabela 14.

Tabela 14: Resultado da estimação do modelo de regressão logística - 2016

Coeficiente	Média a posteriori	Desvio padrão a posteriori	OR	Intervalo de credibilidade de OR		Probabilidade de significância
				2.5%	97.5%	
Intercepto	-1.56	0.23	-	-	-	1.00
area	0.05	0.08	1.051	0.896	1.234	0.74
sexo	-0.16	0.08	0.852	0.726	1.000	0.98
cor	0.25	0.09	1.284	1.083	1.537	1.00
eresp2	-0.35	0.17	0.705	0.507	0.990	0.98
eresp3	-0.82	0.18	0.440	0.310	0.631	1.00
eresp4	-1.58	0.19	0.206	0.142	0.298	1.00
eresp5	-2.00	0.27	0.135	0.079	0.228	1.00
ocupacaoSF	-2.11	0.09	0.121	0.100	0.145	1.00
ocupacaoFI	-4.74	0.46	0.009	0.003	0.020	1.00

Com a análise da Tabela 14, note que os intervalos de credibilidade das razões de chance das variáveis *area*, *sexo* e *eresp2* mostram uma probabilidade de significância alta, mas abaixo de 1. Então é possível interpretar a variável *area* da seguinte forma: residir na capital aumenta a chance de um indivíduo estar abaixo da linha de pobreza em $5,1\%$ em comparação com indivíduos que residem na RM, com probabilidade de $0,74$ desse efeito ser, de fato, positivo.

Então, ser do sexo feminino reduz a chance de um indivíduo estar abaixo da linha de pobreza em $14,8\%$ em comparação com indivíduos do sexo masculino, com probabilidade de $0,98$ desse efeito ser, de fato, negativo.

Para inferir sobre cor ou raça, o coeficiente *cor*, indica que ser preto ou pardo aumenta a chance de um indivíduo ser pobre em $28,4\%$ em comparação com indivíduos brancos.

Para os coeficientes de escolaridade do responsável, é encontrado que residir em um domicílio com a escolaridade do responsável como fundamental incompleto (*eresp2*) reduz a chance de estar abaixo da linha de pobreza em $29,5\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução, com probabilidade de $0,98$ desse efeito ser, de fato, negativo.

Estar morando num domicilio cuja escolaridade do responsável é Fundamental completo (*eresp3*) reduz a chance do indivíduo ser pobre em $56,0\%$, quando comparado com

indivíduos que residem em domicílios com os responsáveis sem instrução.

Viver em um domicílio com a escolaridade do responsável como médio completo (*eresp4*) reduz a chance de estar abaixo da linha de pobreza em $79,4\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

Morar em um domicílio com a escolaridade do responsável como superior completo (*eresp5*) reduz a chance de estar abaixo da linha de pobreza em $86,5\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

No caso da ocupação no domicílio, viver com somente ocupado(s) no mercado formal (*ocupacaoSF*) reduz a chance de ser pobre em $87,9\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal.

E, morar com ocupados no mercado formal e informal (*ocupacaoFI*) reduz a chance de ser pobre em $99,1\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal. Alerta-se para a qualidade do resultado, dada a complicação de baixa representatividade da categoria entre os pobres no período de 2016 identificada na análise descritiva e os resultados dos testes na Tabela 11.

Avançando para o resultado de 2020, é analisada a Tabela 15.

Tabela 15: Resultado da estimação do modelo de regressão logística - 2020

Coeficiente	Média a posteriori	Desvio padrão a posteriori	OR	Intervalo de credibilidade de OR		Probabilidade de significância
				2.5%	97.5%	
Intercepto	-0.90	0.36	-	-	-	1.00
area	0.09	0.10	1.094	0.896	1.323	0.81
sexo	-0.05	0.10	0.951	0.787	1.150	0.70
cor	-0.07	0.11	0.932	0.756	1.150	0.75
eresp2	-0.33	0.31	0.719	0.399	1.363	0.86
eresp3	-1.06	0.33	0.346	0.186	0.670	1.00
eresp4	-0.92	0.31	0.399	0.219	0.748	1.00
eresp5	-1.44	0.34	0.237	0.125	0.472	1.00
ocupacaoSF	-2.48	0.13	0.084	0.063	0.108	1.00
ocupacaoFI	-3.28	0.30	0.038	0.020	0.066	1.00

Ao analisar a Tabela 15, note que apesar dos intervalos de credibilidade das razões de chance das variáveis *area*, *sexo*, *cor* e *eresp2* conterem o valor unitário, há alta probabilidade de significância. Pode-se interpretar a variável *area* da seguinte forma: residir na capital aumenta a chance de um indivíduo estar abaixo da linha de pobreza em $9,4\%$ em comparação com indivíduos que residem na RM, com probabilidade de $0,81$ desse efeito ser, de fato, positivo.

Quanto ao coeficiente de *sexo*, ser mulher reduz a chance de um indivíduo ser pobre em $4,9\%$ em comparação com pessoas do sexo masculino, com probabilidade de 0.70 desse efeito ser negativo.

Para inferir sobre cor ou raça, o coeficiente *cor*, indica que ser preto ou pardo reduz a chance de um indivíduo ser pobre em $6,8\%$ em comparação com indivíduos brancos, com probabilidade de 0.75 desse efeito ser negativo.

Residir em um domicílio com a escolaridade do responsável como fundamental incompleto (*eresp2*) reduz a chance de estar abaixo da linha de pobreza em $28,1\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução, com probabilidade de $0,86$ desse efeito ser, de fato, negativo.

Para os demais coeficientes de escolaridade do responsável, é encontrado que residir num domicilio cuja escolaridade do responsável como Fundamental completo (*eresp3*) reduz a chance do indivíduo ser pobre em $65,4\%$, quando comparado com indivíduos que residem em domicílios com os responsáveis sem instrução.

Viver em um domicílio com a escolaridade do responsável como médio completo (*eresp4*) reduz a chance de estar abaixo da linha de pobreza em $60,1\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

Morar em um domicílio com a escolaridade do responsável como superior completo (*eresp5*) reduz a chance de estar abaixo da linha de pobreza em $76,3\%$ em comparação aos indivíduos que residem em domicílios com os responsáveis sem instrução.

No caso da ocupação no domicílio, viver com somente ocupado(s) no mercado formal (*ocupacaoSF*) reduz a chance de ser pobre em $91,6\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal.

E, morar com ocupados no mercado formal e informal (*ocupacaoFI*) reduz a chance de ser pobre em $96,2\%$ quando comparado com indivíduos que residem em domicílios somente com ocupado(s) fora do mercado formal.

Consolidam-se as informações dos resultados focadas no efeito da formalidade, que indica redução das chances de estar abaixo da linha da pobreza, onde em todos os períodos é mostrado um grande efeito para a redução das chances de pobreza dos residentes de domicílios com somente ocupados formais comparado com moradores de domicílios com ocupados informais. E, ao longo dos anos estudados, é aumentado este efeito sobre esta categoria. Em 2012 estava em $81,4\%$, para 2016, $87,9\%$ e em 2020 é medida uma taxa de $91,6\%$.

Ao comparar os resultados dos anos das pesquisas, os efeitos do fator de cor ou raça sinalizam um equilíbrio nas chances de ser pobre nas categorias do fator. De forma que em 8 anos, a comparação das chances de ser considerado pobre entre pretos ou pardos e brancos entrega uma distribuição que em 2012 têm valores que aumentam as chances de pretos ou pardos serem pobres comparado aos brancos $(78,6\%)$ e em 2020 é vista uma distribuição com chances equiparadas, com alta probabilidade de uma redução das chances em $6,8\%$.

O fator de tipo de área indica um leve movimento para que a chance de ser pobre na Capital aumente comparado aos moradores da RM. O fator de sexo manteve-se com uma redução das chances de ser pobre para as mulheres comparado aos homens.

Para que seja observada a efetividade dos modelos de regressão estimados foram construídas tabelas com a comparação entre os valores estimados pelas regressões e os observados nas bases para os 3 períodos estudados.

Na construção das 3 tabelas, para o valor estimado, é utilizado o resultado da variável resposta do modelo de regressão, que é uma função dos coeficientes. Os valores usados para cada coeficiente é média a posteriori de cada $\boldsymbol{\beta_j}$. O resultado dessa função de coeficientes é π_i a probabilidade de o indivíduo (i) ser considerado pobre. Com a probabilidade de ser pobre calculada para cada indivíduo, é possível ter uma amostra de uma distribuição de Bernoulli com o parâmetro π_i. E como os resultados serão os valores de 0 ou 1, para cada indivíduo, eles serão classificados com não pobres ou pobres, respectivamente. Por fim, serão comparadas as classificações encontradas pelo resultado do modelo de regressão e os valores observados da variável de condição de pobreza na base de dados do período de referência.

As Tabelas 16, 17 e 18 representam o resultado em percentual da comparação entre a classificação encontrada pelo valor esperado e a classificação propriamente observada nos dados.

Tabela 16: Resultado da comparação entre a estimação do modelo de regressão logística e o valor observado - 2012

	Estimado pobre	Estimado não pobre
Observado pobre	30,4%	12,3%
Observado não pobre	69,6%	87,7%

Tabela 17: Resultado da comparação entre a estimação do modelo de regressão logística e o valor observado - 2016

	Estimado pobre	Estimado não pobre
Observado pobre	27,3%	9,2%
Observado não pobre	72,7%	90,8%

Tabela 18: Resultado da comparação entre a estimação do modelo de regressão logística e o valor observado - 2020

	Estimado pobre	Estimado não pobre
Observado pobre	26,5%	9,7%
Observado não pobre	73,5%	90,3%

O resultado observado nas tabelas acima demonstram grande capacidade dos modelos de prever os não pobres, com valores acima de 85% para todos os períodos estudados, porém para os não pobres os modelos tiveram uma qualidade de ajuste muito fraca, prevendo em torno de 30%.

5 Conclusões

Este trabalho teve como principal objetivo comparar as chances de ser pobre entre fatores socioeconômicos dos moradores de domicílios com ao menos uma pessoa ocupada na Região Metropolitana do Rio de Janeiro utilizando um modelo de regressão logística sob o enfoque bayesiano.

O conjunto de informações utilizado foi a PNAD Contínua, com as pesquisas do quarto trimestre de 2012, 2016 e 2020. Após a aplicação dos filtros, as bases formavam em torno de 80% da amostra original, para cada período.

Posteriormente às análises descritivas e estimação dos modelos, os ajustes foram considerados válidos para análise, com a ressalva para a categoria de ocupação formal e informal no período de 2016, onde constatou-se que a categoria tem poucas observações entre os pobre neste período, complicando a convergência deste coeficiente em específico.

Acerca dos resultados, a redução das chances de ser pobre pela formalidade é realmente identificada no estudo e com a distância desta redução comparada aos informais ficando ainda maior ao longo dos anos.

O efeito da formalidade na pobreza mostra-se também mais relevante do que o efeito do fator de nível de escolaridade do responsável, considerando todos os períodos estudados. Comparando com as respectivas categorias base em 2016, por ter os valores mais próximos, a escolaridade do responsável como superior completo reduz as chances de ser pobre em $86,5\%$ e viver com somente ocupados no mercado formal reduz as chances em $87,9\%$.

O fator do nível de escolaridade do responsável pelo domicílio aponta o valor do esforço para a escolarização do indivíduo, a cada nível completado os resultados mostram uma redução da chance de ser pobre maior. Por exemplo, o responsável com ensino fundamental completo agrega uma redução na chance de ser pobre dele e de seus dependentes acima de 50% comparado com os moradores de domicílios com o responsável sem instrução.

As diferenças de condição de pobreza nas categorias do fator de cor ou raça indicadas

nas pesquisas e pelos resultados dos modelos mostram um equilíbrio de oportunidades na sociedade. Os resultados revelam uma tendência de neutralidade na variação das chances comparativas de ser pobre. Em 2012 é indicado um aumento das chances de ser pobre de preto ou pardo em relação aos brancos $(78,6\%)$, porém em 2020, é vista uma pequena redução $(6,8\%)$.

Os resultados obtidos pela estimação destes modelos podem não representar a real situação das populações estudadas, mesmo sendo a maior amostra recente disponível de informações domiciliares no Brasil. A ideia de comparação entre os domicílios de formais e informais quanto a condição de pobreza é apoiar políticas públicas que buscam aumentar os vínculos de emprego e incluir na discussão acadêmica métodos estatísticos de avaliação de condições do mercado de trabalho e de políticas sociais.

Discorrendo sobre o trabalho, são considerados alguns avanços para a pesquisa, analisar mais indicadores para caracterizar a pobreza, que podem ser dados pela publicação da pesquisa anual de 2020, assim como avaliar outros anos dentro da série, entre 2012 e 2020. Outras abordagens como medir a renda pelo consumo ou utilizar dados administrativos também podem ajudar. Em relação aos modelos, seria interessante aplicar um processo de seleção de variáveis para melhorar o processamento dos dados. O aprendizado aplicado sobre modelagem mostrou o grande "mundo"de técnicas que são os MLGs, sempre com novas ferramentas mais eficientes computacionalmente que permitem o avanço da inferência bayesiana, que por sua vez, entregam inovações na área da estatística.

Referências

ACQUAH, H. D.-G. Bayesian logistic regression modelling via markov chain monte carlo algorithm. *Journal of Social and Development Sciences*, v. 4, n. 4, p. 193–197, 2013.

ALKIRE, S.; FOSTER, J. Counting and multidimensional poverty measurement. Oxford: Oxford Poverty & Human Development Initiative, 2008. Disponível em: ⟨https://ophi.org.uk/⟩.

ATHIAS, L.; OLIVEIRA, L. Indicadores de padrão de vida e distribuição de renda - panorama nacional e internacional da produção de indicadores sociais. IBGE, Coordenação de População e Indicadores Sociais, 2016. p. 110-157, 2016. Disponível em: ⟨https://biblioteca.ibge.gov.br/visualizacao/livros/liv98624.pdf⟩.

BARROS, R. P. d.; MENDONçA, R. S. P. Pobreza, estrutura familiar e trabalho. IPEA, 1995.

BAYES, T. An essay towards solving in the doctrine of chances. Philosophical Transactions of the Royal Society London, 1763.

BIANCHINI, Z.; ALBIERI, S. E. Principais aspectos da amostragem da pesquisas domiciliares do ibge. IBGE, Departamento de Metodologia - Texto para discussão, 2002.

BOOTH, C. Life and labour of the people in london, 9 bde. *London/New York*, v. 1897, 1892.

BORGES, E. C. de C. Perfil dos pobres na região metropolitana de salvador: Uma análise para o ano de 2003. UFBA - Mestrado em Economia, 2005.

COBO, B.; ATHIAS, L.; MATTOS, G. G. d. A multidimensionalidade da pobreza a partir da efetivação de direitos sociais fundamentais: Uma proposta de análise. Revista Brasileira de Monitoramento e Avaliação, 2014.

CORDEIRO, G. M.; DEMéTRIO, C. G. B. *Modelos Lineares Generalizados e Extensões*. 1. ed. [s.n.], 2013. Disponível em: ⟨https://docs.ufpr.br/~taconeli/CE22517/LivClarice.pdf⟩.

DARLING, D. A. The kolmogorov-smirnov, cramer-von mises tests. Institute of Mathematical Statistics, v. 28, n. 4, 1957.

DOBSON, A. J.; BARNETT, A. G. *An Introduction to Generalized Linear Models*. 4. ed. [S.l.]: CRC Press, 2018.

GAMERMAN, D.; LOPES, H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2. ed. [S.l.]: Chapman and Hall, 2006.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, Taylor & Francis, v. 85, n. 410, p. 398–409, 1990.

GELMAN, A. et al. *Bayesian Data Analysis*. 3. ed. [S.l.]: Chapman and Hall, 2013.

GELMAN, A.; RUBIN, D. B. Inference from interative simulation using multiple sequences. *Statistical Science*, v. 7, p. 457–511, 1992.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distribution and bayesian restoration of images. IEE Transactions on Pattern Analysis and Machine Intelligence, 1984.

GEWEKE, J. Inference and prediction in the presence of uncertainty and determinism. *Statistical Science*, v. 7, n. 1, p. 94–101, 1992.

GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. *Markov Chain Monte Carlo in Practice*. 1. ed. [S.l.]: New York: Chapman and Hall, 1995.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. Oxford University Press, 1970.

HEIDELBERGER, P.; WELCH, P. D. Simulation run length control in the presence of an initial transient. *Operations Research*, v. 31, p. 1109–1144, 1983.

HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. 2. ed. [S.l.]: John Wiley & Sons, 2000.

IBGE. *Síntese de Indicadores Sociais: uma análise das condições de vida da população brasileira*. [S.l.]: Coordenação de População e Indicadores Sociais. IBGE, 2020.

KANSO, S. Utilização da regressão logística para a classificação de famílias quanto à condição de pobreza nas rms do rio de janeiro e recife nos anos de 1970, 1980 e 1991. IPEA, XIV Encontro Nacional de Estudos Populacionais, 2004.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. Journal of Chemical Physics, 1953.

MIGON, H.; GAMERMAN, D.; LOUZADA, F. *Statistical Inference: An Integrated Approach*. 2. ed. [S.l.]: CRC Press, 2014.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. Journal of the Royal Statistical Society, 1972.

ORSHANSKY, M. The shape of poverty in 1966. *Soc. Sec. Bull.*, HeinOnline, v. 31, p. 3, 1968.

PAULINO, C. D. et al. *Estatística Bayesiana*. 2. ed. [S.l.]: Fundação Calouste Gulbenkian, 2018.

PLUMMER, M. *JAGS Version 4.3.0 user manual*. [S.l.], 2017.

PLUMMER, M. et al. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, v. 6, n. 1, p. 7–11, 2006. Disponível em: ⟨https://journal.r-project.org/archive/⟩.

PLUMMER, M.; STUKALOV, A.; DENWOOD, M. *Pacote rjags, Bayesian Graphical Models using MCMC*. [s.n.], 2019. Disponível em: ⟨https://cran.r-project.org/web/packages/rjags/index.html⟩.

PNUD. *Relatório de Desenvolvimento Humano de 2019*. 1. ed. Programa das Nações Unidas para o Desenvolvimento, 2019. Disponível em: ⟨https://www.br.undp.org/content/brazil/pt/home/library/relatorio-do-desenvolvimento-humano-2019.html⟩.

PRAAG, T. G. B. V.; KAPTEYN, A. The poverty line–a pilot survey in europe. The MIT Press, 1980.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: ⟨http://www.R-project.org/⟩.

RAFTERY, A. E.; LEWIS, S. M. How many iterations in the gibbs samper? In: BERNARDO, J. et al. (Ed.). *Bayesian Statistics 4*. [S.l.]: University Press, Oxford, 1992. p. 763–773.

ROCHA, S. Pobreza no brasil: fatos básicos e implicações para política social. *Economia e Sociedade*, v. 5, n. 1, p. 141–151, 1996.

ROCHA, S. *Tranferências de renda no Brasil. O fim da Pobreza?* [S.l.]: Alta Books, 2019.

ROWNTREE, S. Poverty: A study of a town life. London: Macmillan & Co., 1901.

SMITH, B. J. boa: An r package for mcmc output convergence assessment and posterior inference. *Journal of Statistical Software*, v. 21, n. 11, p. 1–37, 2007.

SOARES, S. Metodologias para estabelecer a linha de pobreza: Objetivas, subjetivas, relativas e multidimensionais. IPEA, 2009.

SULIANO, D.; CARVALHO, M. Caracterização das pesquisas domiciliares com Ênfase na pnad contínua. Nota Técnica - Instituto de Pesquisa e Estratégia Econômica do Ceará, 2017.

APÊNDICE 1 – Revisão PNAD e PNAD Contínua

É buscada uma equivalência de comportamento entre a renda domiciliar per capita e a renda familiar per capita com a visualização gráfica. Com a Pesquisa Nacional por Amostra de Domicílios Contínua, base do trabalho, tem-se a limitação em não poder identificar as famílias, logo uma avaliação semelhante a realizada pelo Cadastro Único não será possível. Porém na PNAD, o grupo familiar era identificável. Então foram realizados gráficos da distribuição dessas variáveis, com valores nominais, dentro das pesquisas anuais da PNAD de 2001 e 2009 e as pesquisas anuais da PNAD Contínua de 2016 e 2019.

O resultado gráfico indica uma distribuição semelhante para os grupos ao longo dos anos.

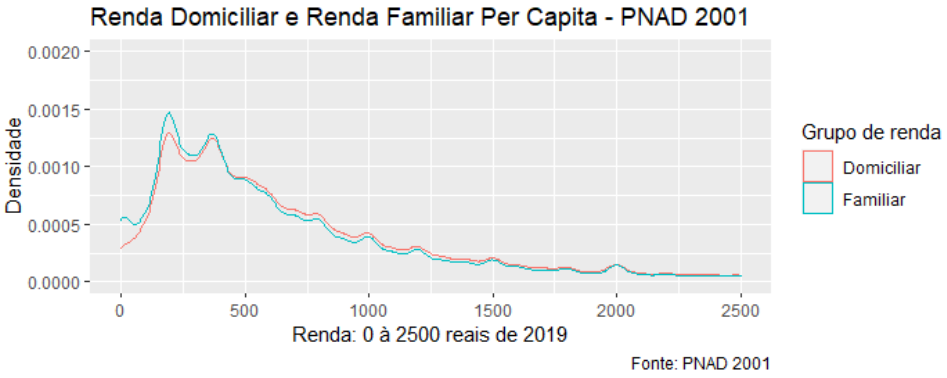

Figura 11: Distribuição de renda na PNAD 2001

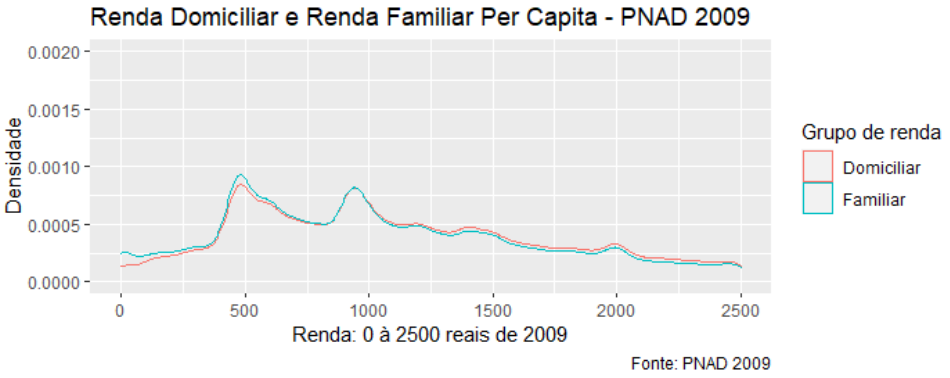

Figura 12: Distribuição de renda na PNAD 2009

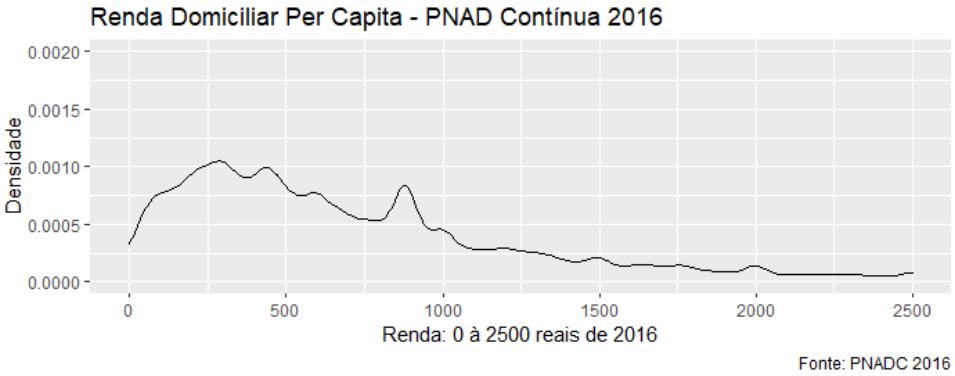

Figura 13: Distribuição de renda na PNAD Contínua 2016

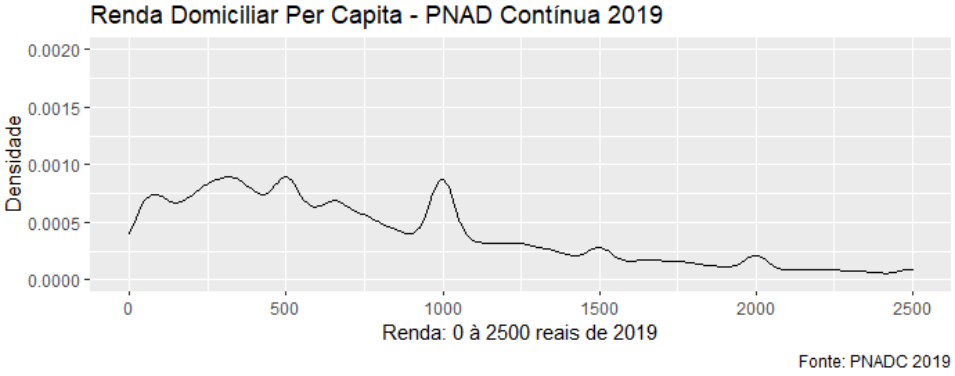

Figura 14: Distribuição de renda na PNAD Contínua 2019

APÊNDICE 2 – Códigos utilizados

Para a aplicação da parte computacional do modelo. A seguir, são expostos os códigos referentes ao processo de amostragem das cadeias dos parâmetros estimados, que consolidam a distribuição de cada β_j e validam a estimação.

Arquivo de texto aplicado na estimação dos modelos citados no Capítulo 4, temos as variáveis com o nome de seu respectivo efeito no modelo.

# arquivo_texto.txt
model {
for (i in 1:Num) { # Likelihood
$y[i] \sim dbern(p[i])$
$\text{logit}(p[i]) \leftarrow \text{Intercepto} + \text{area} * \text{x1}[i] + \text{sexo} * \text{x2}[i] + \text{cor} * \text{x3}[i]$
$+ \text{eresp2} * \text{x4}[i] + \text{eresp3} * \text{x5}[i] + \text{eresp4} * \text{x6}[i] + \text{eresp5} * \text{x7}[i]$
$+ \text{ocupacaoSF} * \text{x8}[i] + \text{ocupacaoFI} * \text{x9}[i]$
}

Prior
$\text{Intercepto} \sim dnorm(0, 0.001)$
$\text{area} \sim dnorm(0, 0.001)$
$\text{sexo} \sim dnorm(0, 0.001)$
$\text{cor} \sim dnorm(0, 0.001)$
$\text{eresp2} \sim dnorm(0, 0.001)$
$\text{eresp3} \sim dnorm(0, 0.001)$
$\text{eresp4} \sim dnorm(0, 0.001)$
$\text{eresp5} \sim dnorm(0, 0.001)$
$\text{ocupacaoSF} \sim dnorm(0, 0.001)$
$\text{ocupacaoFI} \sim dnorm(0, 0.001)$
}

Texto de estimação no R aplicado ao modelo 2020. As mudança para os outros

modelos são, além da variável *ANO*, o número de iterações e passo entre as observações.

Os números de iterações são de 60.000 para 2012 e de 90.000 para 2016 e 2020. O passo entre as observações para montar a cadeia final são de 4 em 2012 e 5 em 2016 e 2020.

```
    # estimacão dos parâmetros da posteriori do modelo 2020
ANO ← "2020"
library(rjags)
library(tidyverse)
# rlist, parallel e lubridate instalado
options(mc.cores = parallel::detectCores())
load.module("glm")
color_scheme_set("brightblue")


    #funcao apoio
save.chain ← function(chain,load.time) {
t=lubridate::ymd_hms(Sys.time())
time.string = paste0( lubridate::hour(t), lubridate::minute(t), "_", lubridate::day(t),
lubridate::month(t), lubridate::year(t))
out.string = paste0("cadeia_completa_",ANO,"_",time.string, ".rds")
vetor.covars = c(ANO, 'area', 'sexo', 'cor', 'vdocup', 'escresp', time.string, load.time)
n = length(chain)
chain[[n+1]] = vetor.covars
rlist::list.save(chain,out.string)      # salvar cadeias
}


    # base de dados observados
b ← read_rds("./../baseunificada.rds")


    # dados observados
data.list ← list(
y = as.double(b$pobre),
x1 = as.double(b$area),
x2 = as.double(b$sexo),
x3 = as.double(b$cor),
```

```
x4 = as.double(b$esc_resp2),
x5 = as.double(b$esc_resp3),
x6 = as.double(b$esc_resp4),
x7 = as.double(b$esc_resp5),
x8 = as.double(b$vd_ocupSF),
x9 = as.double(b$vd_ocupFI),
Num = length(b$pobre)
)

    # dados iniciais das cadeias: 2 cadeias
init.list ← list(
list(Intercepto=1, area=1, sexo=1, cor=1, eresp2=1, eresp3=1, eresp4=1,
eresp5=1, ocupacaoSF=1, ocupacaoFI=1),
list(Intercepto=-1, area=-1, sexo=-1, cor=-1, eresp2=-1, eresp3=-1, eresp4=-1,
eresp5=-1, ocupacaoSF=-1, ocupacaoFI=-1))

    # descrição dos efeitos: Beta_j
params ← c("Intercepto", "area", "sexo", "cor", "eresp2", "eresp3", "eresp4",
"eresp5", "ocupacaoSF", "ocupacaoFI")

    # preparando variável com descrição do modelo: m
m ← jags.model("arquivo_texto.txt", data.list, init.list, n.chains = 2)

    # primeiras amostras → burning 10k
update(m, n.iter=10000)

    # cadeias efetivas para avaliação → 90K
{
t.start ← Sys.time()
x ← coda.samples(m, params, n.iter=90000,thin = 5)
t.end ← Sys.time()
print(t.end - t.start)
save.chain(x,t.end - t.start)
}
```