UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA

MODELAGEM DOS REGISTROS DE NASCIMENTOS COM AS CONDIÇÕES DE VIDA NO SEMIÁRIDO BRASILEIRO

Lígia Maria de Oliveira Maia

Maio/2016

LÍGIA MARIA DE OLIVEIRA MAIA

MODELAGEM DOS REGISTROS DE NASCIMENTOS COM AS CONDIÇÕES DE VIDA NO SEMIÁRIDO BRASILEIRO

Orientador: Professor Dr. Neir Antunes Paes

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba como requisito parcial para obtenção do Grau de Bacharel.

João Pessoa
Maio de 2016

MODELAGEM DOS REGISTROS DE NASCIMENTO COM AS CONDIÇÕES DE VIDA NO SEMIÁRIDO BRASILEIRO

LÍGIA MARIA DE OLIVEIRA MAIA

Aprovada em 24/05/2016.

BANCA EXAMINADORA

Profº Dr. NEIR ANTUNES PAES
Orientador - UFPB

Profº Dr. MARCELO RODRIGO PORTELA FERREIRA
Examinador - UFPB

JOZEMAR PEREIRA DOS SANTOS
Examinador - UFPB

CONCEITO FINAL: _____

Dedico este trabalho a Deus, aos meus pais, Ana e Antonio, os quais tornaram possível a conclusão dessa etapa da minha vida e ao Professor Joab Lima (In Memorian), que sei que está olhando por mim.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me permitir viver intensamente cada momento dessa graduação. Agradeço por poder aprender, errar e consertar, por ter a oportunidade de estar realizando um dos maiores sonhos da minha vida. Sonho que não é só meu, sonho que antes de qualquer coisa, é o sonho dos meus pais (Tontonho e Ana Maia).

Peço muito obrigada a eles que tanto fizeram para que eu chegasse até aqui. Como foi difícil o início, mas tudo valeu a pena e luto todos os dias para que eu possa só trazer orgulho e alegria à vocês, Painho e Mainha. Obrigada, também, aos meus tios, tias e primos que apostaram em mim, e que me apoiaram em todos os momentos.

Agradeço aos meus amigos, Suellen, Clara, Salete, Igor e todos os que me acompanham desde muito tempo. Obrigada Williby por ser muito presente, por cada "Bom dia", por cada encontro, por todos os carões e por estar comigo sempre. Obrigada a você Vinícios Mendes por ter tanta paciência comigo e por me ensinar a ser "gente", como você mesmo diz. Obrigada pelas inúmeras vezes que me deu um pouco de você para que eu me tornasse uma pessoa melhor. Amo muito vocês.

Obrigada aos meus amigos da universidade. Obrigada àqueles do começo, Anderson, Mafferson, Jack. Obrigada aos do meio e fim, Elaine, Marina, Maizza, Michelle, Saul, Adenice, Diogo, Zé, Any, André, Diego, Geise, Everlane e todos os outros que estiveram presentes e me ajudaram muito, sem vocês tudo seria mais difícil. Muito Obrigada Alisson, meu companheiro de laboratório e amigo, sou muito grata por tudo que você fez por mim.

Tenho inúmeros motivos para agradecer aos professores do Departamento de Estatística, muito obrigada por estarem sempre presentes e dispostos a ajudar da melhor forma possível.

Obrigada Professor Hemílio por ter sido o primeiro a me acolher, obrigada por cada conselho, incentivo, descontrações, por cada puxão de orelha, por me cobrar, elogiar, defender, orientar. Obrigada por acreditar em mim, por ser um amigo, e por ter sempre aberto portas, tirado dúvidas nas madrugadas da vida, por ter me apresentado Joab (*In Memorian*),... Não tenho palavras suficientes para expressar minha gratidão a você.

Aprendi muito, não só como aluna, mas como pessoa.

Um grande Mestre que Deus me presenteou e levou para perto dele foi Professor Joab (*In Memorian*), este foi um paizão. Obrigada Professor. Obrigada por ter me dado as primeiras oportunidade do real contato com a profissão que escolhi para seguir. Obrigada por cuidar tanto de seus alunos, por ser simples e saber cativar-nos um a um. Obrigada por acreditar que eu era capaz e vou fazer valer aquele "Pingo de sucesso"que me foi confiado. Foi um prazer ser sua pupila.

Sou muito agradecida a Professora Tatiene pela primeira oportunidade de ser bolsista, o qual era um dos meus objetivos dentro da universidade. Obrigada Professora Tarciana pelos conselhos, que me ajudaram a ser uma pessoa melhor.

Não poderia deixar de agradecer imensamente aos professores que me ajudaram no decorrer deste trabalho. Obrigada Professora Maria Lídia por ter sido tão prestativa. Obrigada Professor João por todas as horas de que te procurei e o Senhor me presenteou com tanta sabedoria e experiência, sou muito feliz por ter pessoas como o Senhor em minha vida. Um exemplo de ser humano que Deus colocou na terra para ensinar as pessoas como ser melhor. Obrigada Professor Marcelo por tudo, todas as dúvidas, ideias, risadas, por solucionar meus problemas (que não eram poucos), por me permitir aprender com "o cara".

Não poderia deixar de agradece ao meu orientador, obrigada Professor Neir. Agradeço a Deus pela oportunidade de poder estar perto de uma pessoa com tamanha sabedoria, inteligência e muita experiência. Agradeço ao Senhor, por ter corrigido cada linha de cada trabalho com todo cuidado e toda dedicação. Obrigada por todas as lições de vida, por poder ser aprendiz e seguidora de Paes. Tenho muito orgulho em dizer que sou sua orientanda e espero corresponder a tanta responsabilidade.

"... a vida vai ser sempre essa roda gigante, e se você não aguentar o frio na barriga na hora da descida, não vai sentir o vento no rosto e a sensação única na hora da subida. E vai por mim, a vista lá de cima é INCRÍVEL!"

RESUMO

A região semiárida brasileira é a mais populosa do planeta e apresenta indicadores de desenvolvimento deficitários, bem como na atenção materna e infantil. No campo das estatísticas vitais existem problemas com o sub-registro dos nascimentos, cujos estudos são úteis para o monitoramento e melhoria da qualidade destes dados. Teve-se como objetivo modelar a variável registro de nascimento com as condições de vida do Semiárido Brasileiro por meio da regressão logística. Para tanto, fez-se uso dos microdados do Censo Demográfico de 2010 com relação ao item que investigou se as pessoas com idade até 10 anos possuíam documentos referentes aos seus nascimentos. Para a modelagem a variável registro de nascimento foi considerada como dependente e dezenove variáveis socioeconômicas e demográficas como independentes. Foram construídos dois bancos de dados: o Banco 1 que considerou a variável "ser beneficiária ou ter Rendimento Mensal do Programa Social Bolsa Família ou Programa de Erradicação do Trabalho Infantil", e o Banco 2, sem considerar esta variável. Para ambos os bancos, o nível de significância adotado foi de $p<0{,}05$. Para o Banco 1, as variáveis significativas foram: "moradia (esgoto e água)", "nível de instrução do respondente (ler)", e "poder aquisitivo (máquina, geladeira, celular)". Para o Banco 2, as variáveis significativas foram: local onde reside o entrevistado (água), nível de instrução (ler), aspecto monetário (computador, moto, TV, geladeira, celular) e demográfico (idade). Concluiu-se que o registro de nascimento pode ser modelado a partir das variáveis socioeconômicas e demográficas especificadas. Ou seja, em termos mais amplos, as pessoas com menos condições favoráveis ao registro de nascimento são aquelas sem instrução, com uma baixa renda e precárias condições do domicílio. Com estes resultados espera-se poder contribuir para a definição de estratégias e no fomento de políticas públicas que reduzam ou eliminem o sub-registro nos Estados do semiárido.

Palavras-chave: Registro de Nascimento, Regressão Logística, Semiárido.

ABSTRACT

The Brazilian semiarid region is the most populous on the planet and presents deficit of development indicators as well as maternal and child care. In the field of vital statistics there are problems with births underreporting, whose the studies, are useful for monitoring and improving the quality of data. The current work aimed modeling the variable birth record with the living conditions of the Brazilian Semiarid, by logistic regression. Therefore, it made use of micro data from the Census 2010 taking into consideration to the item investigated , whether people aged up to 10 years old have had documents relating to their births. As a model, the variable record birth was considered a dependent, and nineteen socio-economic and demographic models as independent. Two data bases banks were built: Bank 1- that considered being beneficiary or having income monthly from the Social Program "Bolsa Família" or Child Labor Eradication Program, and Bank 2, without considering this variable. For both banks, the level of significance was $p<0.05$. For the Bank 1, the significant variables were housing (sanitary sewage and water), respondent education level (read), and purchasing power (machine, refrigerator, cell phone). For the Bank 2, the significant variables were, where the respondent lives (water), education level (read), monetary aspect (computer, bike, TV, fridge, phone) and demographic (age). It has been concluded that the birth registration can be modeled from the socio-economic variables and demographic specified. That is, in broader terms, people with less favorable conditions for the birth registration are those respondents with no education, with a low income and poor household conditions. With these results, it is expected to contribute for the strategies definition and the promotion of public policies that reduce or eliminate the underreporting in semiarid States.

Keywords: Birth Registration, Logistic Regression, Semi-Arid.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

Siglas	Definição
ACF	Autocorrelação
AIC	Critério de Seleção de Akaike
BF	Programa Social Bolsa Família
DATASUS	Departamento de Informática do Sistema Único de Saúde
DN	Declaração de Nascido Vivo
MLG	Modelos Lineares Generalizados
IBGE	Instituto Brasileiro de Geografia e Estatística
IC	Intervalo de Confiança
LED	Laboratório de Estudos Demográficos
MS	Ministério da Saúde
PACF	Autocorrelação Parcial
PETI	Programa de Erradicação do Trabalho Infantil
RANI	Registro Administrativo de Nascimento Indígena
RCN	Registro Civil de Nascimento
SIM	Sistema de Informações sobre Mortalidade
SINASC	Sistema de Informações sobre Nascidos Vivos
SUS	Sistema Único de Saúde
UNICEF	Fundo das Nações Unidas para a Infância

CAPÍTULO 1

INTRODUÇÃO

O Brasil é um País de contrastes que se expressam em variadas formas, seja em sua economia, geografia, desigualdades sociais, acesso à educação, aos serviços de saúde, ao poder político, entre outros. As regiões refletem estes contrastes em diferentes graus de desigualdade que conduziu a sociedade brasileira a diferentes condições de vida. Entre estas regiões destaca-se a do Semiárido, com características singulares que apresenta diversos problemas sociais e econômicos (CASTRO; RODRIGUES-JUNIOR, 2012).

O Ministério da Integração Nacional define o "Semiárido brasileiro" como uma região que apresenta os seguintes padrões geográficos: precipitação média inferior a 800 mm; índice de aridez de até 0,5 calculado pelo balanço hídrico que relaciona as precipitações e a evapotranspiração potencial, no período compreendido entre os anos de 1961 e 1990; risco de seca maior que 60%, tomando-se por base o período entre os anos de 1970 e 1990 e território de 969.589,4 km^2. Segundo o IBGE (2016), em 2010, a população da região Nordeste era de 53 milhões de habitantes, sendo que, aproximadamente 25 milhões residiam na Região Semiárida, considerada a mais populosa do mundo.

Ao se estudar as estatísticas do Semiárido, depara-se com a deficiência de seus indicadores, entre elas estão as declarações das estatísticas vitais (óbitos e nascimentos). São poucos os trabalhos que se propuseram a enfocá-las . Geralmente, quando feitos, eles se detiveram a abordar regiões isoladas ou agregações de unidades geográficas com metodologias diferentes.

Os registros dos fatos vitais são informações que possibilitam, quando sua cobertura é adequada, a produção de indicadores tais como: taxas brutas de natalidade, taxas de fecundidade, taxas de mortalidade infantil e expectativa de vida. Eles se constituem em importantes elementos de referência para o planejamento de políticas públicas específicas, em áreas como demografia, saúde, entre outras.

Estes registros são motivo de preocupação, principalmente por parte dos demógrafos e epidemiologistas, devido à qualidade duvidosa dos dados, os quais se constituem em en-

traves na obtenção de indicadores confiáveis, dificultando, assim, as ações governamentais do País (PAES, 2009).

O registro de nascimento se constitui em um documento que oficializa a existência do indivíduo, logo funciona como a identidade formal do cidadão. A falta de informação ou má declaração sobre o nascido ocasiona problemas na identificação correta de suas características. Entre eles se destaca o sub-registro dos nascimentos, cuja estimação tem provocado discussões entre estudiosos. Estudos voltados para essa temática se não ausentes são escassos, particularmente para a Região Semiárida (PAES, 2010; CRESPO, 2012). Faz-se assim necessárias investigações que procurem explicar os fatores que determinaram seus registros.

Com base na literatura um conjunto de fatores socioeconômicos e demográficos foram considerados nesta investigação. Para tal propósito, os modelos de regressão são ferramentas uteis, podendo trazer aportes para o entendimento dos processos que regem o registro de nascimento no Semiárido brasileiro.

Neste sentido o Laboratório de estudos Demográficos (LED) da UFPB vem desenvolvendo pesquisas sobre estatísticas vitais do Semiárido brasileiro com uma perspectiva estatística e demográfica voltadas para estas preocupações.

Neste estudo, na busca de explicar o registro de nascimento, formulou-se os objetivos a seguir:

Geral

Modelar a variável Registro de Nascimento com as condições de vida do Semiárido Brasileiro por meio da regressão logística.

Específicos

- Gerar um modelo estatístico capaz de explicar a variável Registro de Nascimento com o banco de dados completo e o banco sem a presença da variável referente ao recebimento de auxílio de programas governamentais (Bolsa Família e PETI).
- Responder quais fatores influenciam no sub-registro de nascimento no Semiárido brasileiro.

CAPÍTULO 2

REFERENCIAL TEÓRICO

2.1 Caracterização do Semiárido brasileiro

O Semiárido brasileiro é representado pelos Estados do Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia e norte de Minas Gerais. Caracterizado por municípios com forte presença de população rural e indicadores socioeconômicos de baixo desempenho, a região semiárida é composta por 137 microrregiões e 1.133 municípios, que abrange uma área de 969.589,4 km^2, correspondendo a quase 90% da Região Nordeste. A semiaridez é identificada pela escassez e irregularidade das precipitações, com chuvas no verão e forte evaporação em consequência das altas temperaturas (CORTEZ et al., 2011) ; IBGE, 2016).

Sendo a região semiárida mais populosa do planeta, em 2010 residiam cerca de 22 milhões de pessoas, que representavam 11,8% da população brasileira. 58% da população pobre do País vive nesta região (BRASIL, 2013). Pesquisas realizadas pelo Fundo das Nações Unidas para a Infância (UNICEF) mostram que 67,4% das crianças e adolescentes no Semiárido são afetados pela pobreza, sendo assim, desprovidos dos direitos humanos e sociais básicos, e dos elementos indispensáveis ao seu desenvolvimento (ASA, 2016). Seus indicadores de desenvolvimento são os mais comprometidos do País. Do ponto de vista dos indicadores demográficos relacionados à natalidade apresentam níveis acima dos encontrados para as demais regiões, exceto a região Norte do Brasil. A obtenção destes últimos indicadores, por sua vez, depende, da disponibilidade de informações sobre as estatísticas vitais vigentes (MELO, 2014).

2.2 Estatísticas vitais

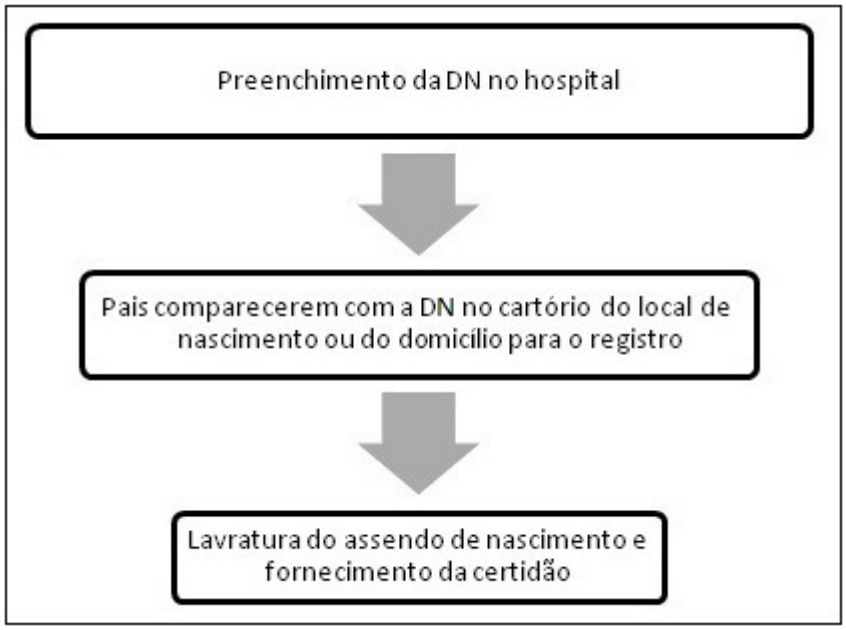
Uma das principais funções do Poder Público é a atribuição da cidadania aos integrantes da população. A Lei Federal 6.015 de 1973, que especifica quanto ao registro de

nascimento, dispõe em seu art. 50, que todo nascimento ocorrido em território nacional deve ser levado a registro, seja no lugar do parto ou no local de residência dos País. Assim, o registro civil de nascimento é um direito humano fundamental ao exercício da cidadania e dignidade da pessoa humana, a fim de conferir identidade à pessoa natural. As informações constantes do assento de nascimento são, assim indispensáveis para a perfeita identificação e individualização da pessoa cuja emissão da certidão é gratuita (CALTRAM, 2010).

Em regra, dentro do prazo legal (prazo de 15 dias, que é ampliado em até três meses para os lugares distantes mais de trinta quilômetros da sede do cartório), o registro civil de nascimento (RCN) deve ser feito na localidade onde a pessoa nasceu, na de residência dos genitores (pai, mãe) ou responsável legal (art. 50 da Lei 6.015/73). Fora do prazo legal, o RCN é feito unicamente no cartório da circunscrição da residência do interessado (art. 46 da Lei 6.015/73) (TJPB, 2015).

Para realizar o registro de nascimento é necessário que o responsável siga um passo a passo, ilustrado na Figura 2.1.

Figura 2.1: Fluxo do registro de nascimento no Brasil



A maior parte do conhecimento sobre a natalidade no Brasil vem de informações geradas pelos censos demográficos e registros vitais. Este último tem como principal fonte o IBGE e o Ministério da Saúde (MS). Além da produção de uma grande quantidade de informações informatizadas, o IBGE, (principal entidade da administração pública federal, e provedora de dados e informações do País) realizou o primeiro levantamento estatístico oficial brasileiro, o Censo do Império, em 1872. Desde sua criação, o IBGE cumpre a missão de contar a população, identificar, produzir e analisar o quadro do desenvolvimento da população do território brasileiro, mostrar dados econômicos e sociais, entre muitas

outras funções (NASCIMENTO, 2006).

O IBGE passou a coletar e a processar o número das declarações de nascido vivo no conjunto das informações declaradas no momento do registro civil, em todos os cartórios do País. Acrescenta-se o fato de que, em sendo o registro civil um ato fundamentalmente jurídico, as informações anotadas eram as enumeradas em lei e referiam-se a situações necessárias à comprovação legal do evento (JORGE; LAURENTI; GOTLIEB, 2007; IBGE, 2016).

Além do IBGE, o Ministério da Saúde responde pela produção dos registros sobre os nascimentos através do Sistema de Informações sobre Nascidos Vivos (SINASC). A declaração de nascido vivo foi implantada e padronizada, em todo o País, na década de 1990. O primeiro modelo de DN colocado no sistema era formado por oito blocos de informação, contendo dados sobre: o cartório onde o Registro Civil foi realizado, local de ocorrência; relativos ao recém-nascido: data do nascimento, sexo, peso ao nascer, índice de Apgar; duração da gestação, tipo de gravidez e tipo de parto; dados sobre a mãe: nome, idade, grau de instrução, residência e filhos tidos; nome do pai responsável pelo preenchimento da DN (WALDVOGEL, 2010; SÃO PAULO, 2011).

Posteriormente, foram realizadas alterações sucessivas modificando a forma de algumas perguntas ou foram introduzidas novas variáveis (raça/cor; presença de anomalia congênita) ou ainda suprimiram questões (nome do pai). Uma importante providência foi o contato com os Cartórios do Registro Civil, a fim de que eles pudessem participar do novo Sistema. Relativamente à implantação, foi estabelecido que ela devesse ser gradual, iniciando-se pelas capitais dos estados, estendendo-se, posteriormente, aos demais municípios. A descentralização administrativa e operacional surgida no País fez com que a maioria das Secretarias Municipais de Saúde assumisse as tarefas de coleta, processamento e análise dos dados, em nível municipal (WALDVOGEL, 2010).

Ademais do Ministério da Saúde, a partir de 2011, o Departamento de Informática do Sistema Único de Saúde (DATASUS) passou a integrar a Secretaria de Gestão Estratégica e Participativa, que trata da Estrutura Regimental do Ministério da Saúde, e desde então foi informatiza online os registros de nascimentos (DATASUS, 2015).

As Supervisões Técnicas de Saúde repassam e controlam os impressos de DN destinados aos estabelecimentos de saúde que realizam partos no âmbito de sua área de abrangência.

O processo da DN segue o fluxo da Figura 2.2 especificada em três situações: Nascimento em estabelecimento de saúde; Nascimento fora de estabelecimento de saúde com ausência de profissional de saúde e Nascimento domiciliar sem assistência de profissional de saúde. Esta DN é impressa em papel especial autocopiativo, em três vias, compondo um jogo com numeração sequencial em função das características do local de ocorrência do nascimento (BRASÍLIA, 2011). Cada via tem o seguinte destino:

1ª Via (branca) – De acordo com o responsável pela emissão da DN, seguirá o fluxo

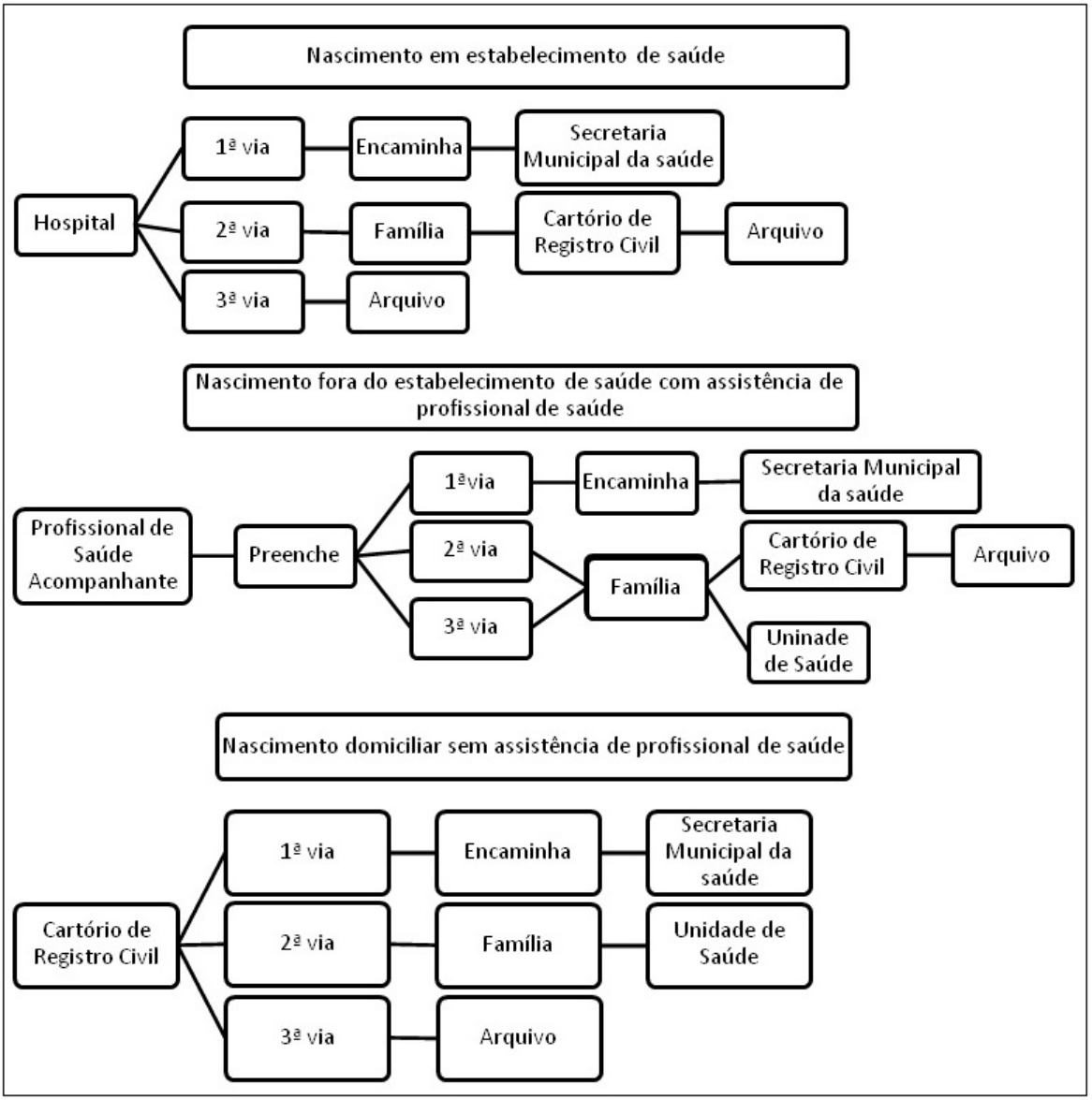
descrito:

- Estabelecimentos de Saúde que realizam partos, devem enviar para a Supervisão Técnica de Saúde de sua região.
- Cartórios de Registro Civil da capital e profissionais cadastrados que prestam assistência nos partos domiciliares, devem encaminhar para a Gerência do SINASC.

2ª Via (amarela) – Entregar ao pai ou responsável legal para assentamento do nascimento em cartório e obtenção da certidão de nascimento.

3ª Via (rosa) – Arquivar no prontuário da gestante ou do recém-nascido.

Figura 2.2: Fluxo da Declaração de Nascimento no Brasil



Fonte: Manual de Instruções para o preenchimento da Declaração de Nascido Vivo, Ministério da Saúde – Brasília, 2011.

No estudo da natalidade, importantes indicadores são produzidos para monitorar e verificar o estado de reprodução e de crescimento de uma população.

2.2.1 Estudo da natalidade e da reprodução

Para Paes (2009), enquanto a morte é um evento inevitável, fatal, independente da vontade do homem, o nascimento não somente passa pela questão da "inevitabilidade", mas também do desejo, da decisão e do controle. Enquanto o primeiro evento pode ser descrito por leis mais ou menos bem estabelecidas, o nascimento depende não somente de leis biológicas, mas também das comportamentais (papéis, valores, costumes, condições de vida, acesso às informações, etc). Desta maneira, a força da natalidade varia entre populações, entre diferentes grupos dentro de uma população, e historicamente, do passado ao presente. Pode-se afirmar que os modelos explicativos para entender a natalidade devem ser pensados como distintivos, cada um deles fornecendo uma peça diferenciada e/ou complementar para uma questão complexa.

O conhecimento da natalidade de um determinado lugar se constitui em um fator preponderante na dinâmica populacional das sociedades modernas, já que os nascimentos fazem parte da composição de inúmeros indicadores demográficos e epidemiológicos, como, por exemplo, taxas de mortalidade infantil, taxas de natalidade, taxa de fecundidade e de mortalidade materna, os quais se constituem em informações preciosas no planejamento e na delimitação das políticas públicas nas áreas da saúde materna e infantil (PAES, 2010; SOUZA, 2004).

As declarações censitárias, bem como do Ministério da Saúde relacionados à natalidade, estão sujeitas a erros que, dependendo de sua magnitude podem comprometer a veracidade dos indicadores que delas dependem.

2.2.2 Qualidade dos dados de nascimento

O sub-registro de nascimento é definido pelo IBGE como o conjunto de nascimentos ocorridos no ano de referência da pesquisa do registro civil e não registrados no próprio ano ou até o fim do primeiro trimestre do ano subsequente. A aplicação deste conceito se restringe a população nascida no ano para a qual se tem como referência os nascimentos estimados por métodos demográficos. Oportunamente, por ocasião do Censo Demográfico de 2010, fez-se, pela primeira vez, a opção por investigar para as pessoas com idade até 10 anos, a existência de documentos referentes ao nascimento desses indivíduos, sendo eles: o registro público feito em cartório, a Declaração de Nascido Vivo (DN) ou o Registro Administrativo de Nascimento Indígena (RANI) (CRESPO, 2012).

Apesar das ações realizadas para o incentivo do registro de nascimento, como a matrícula escolar, o Programa Bolsa Família, que tem como uma de suas condicionalidades o registro de nascimento da criança, a marcante deficiência dos registros de nascimentos no

Semiárido brasileiro se constitui em um dos entraves para um uso adequado e confiável. Sendo uma das regiões mais deficientes do País, nesta questão, ela apresenta problemas na cobertura, regularidade ou qualidade das informações. Consequentemente é necessário buscar meios para estimar a magnitude desses sub-registros (PAES, 2014); (PAIVA, 2016).

Informações coletadas pelos Cartórios de Registro Civil correspondentes aos nascimentos sofrem com a falta de cadastro dos nascidos vivos completos. Assim, uma parcela dos eventos fica fora de seus registros e, portanto, das estatísticas do IBGE e do MS. Isso se deve, em parte, ao atraso na realização do registro, que pode ocorrer quando os familiares residem distante dos cartórios ou desconhecem a necessidade e importância de tal registro. A evasão dos registros impossibilita o cálculo direto de alguns importantes indicadores demográficos, para os quais os estudiosos lançam mão de correções dos dados ou de metodologias especiais para sua obtenção (CRESPO BASTOS; CAVALCANTI, 2006).

No caso do censo demográfico, há dois tipos principais de erros que ocorrem na coleta dos dados: os que se referem à contagem, seja por omissão ou por contagem de um indivíduo inúmeras vezes, consequentes da má cobertura do censo, e os obtidos por falhas nas declarações, representados pela omissão ou declaração errônea.

Sendo estas variáveis de uso frequente na geração de indicadores demográficos e socioeconômicos, torna-se imprescindível uma avaliação da qualidade dessas declarações, da magnitude e do dimensionamento dos eventuais erros existentes (PAES, 1999).

Alguns fatores se destacam quando se estuda a omissão de registros de nascimentos, como o aspecto monetário, a filiação ilegítima, a falta de tempo, a ignorância sobre a importância do registro civil, o desconhecimento das leis, a negligência, a distância do domicílio ao cartório, principalmente em zonas rurais, o grau de instrução dos pais, a falta de fiscalização e controle , a falta do reconhecimento paterno, entre outros fatores (MELO, 2014).

Nas bases de dados produzidas pelos sistemas municipais de saúde, que compõem o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informações sobre Nascidos Vivos (SINASC), também se evidenciam algumas dificuldades. Uma delas refere-se, por exemplo, ao número inadequado ou parcial de nascidos vivos e de óbitos segundo o lugar de residência. Este problema resulta, principalmente, de dois tipos de erro: coleta insuficiente das ocorrências municipais e lugar de residência incorreto (WALDVOGEL, 2010).

2.3 Relacionamento entre a natalidade e as condições de vida

O Brasil é um país que apresenta em seu território grandes disparidades socioeconômicas. Algumas áreas são mais privilegiadas por aspectos naturais e por políticas de investimento em infraestrutura, fatos que promoveram um processo industrial mais avançado em determinadas regiões (BRASIL, 2016).

Nas últimas décadas diversas mudanças foram observadas nas condições de reprodução da população como a diminuição da fecundidade trazendo como consequência o aumento da esperança de vida ao nascer. Estas mudanças propiciaram melhores condições de vida e saúde da população, alterações nos padrões de relacionamento entre os membros da família e no papel da mulher dentro e fora do espaço doméstico, bem como no aumento de uniões consensuais (NASCIMENTO, 2006).

Analisar os fatores que contribuem direta ou indiretamente para as condições de nascimento demográficos e socioeconômicos significa estudar a associação da saúde e da qualidade de vida das famílias como a assistência pré-natal e ao parto, os desfechos gestacionais como prematuridade, baixo peso ao nascer e morte neonatal. Vários estudos trazem a influência das condições demográficas e socioeconômicas do País sobre estes aspectos (MELO, 2014; BARROS; NICOLAU, 2013; RAMOS; CUMAN, 2009).

O acesso a informações e a medidas de prevenção e promoção à saúde é fundamental, de forma que os baixos índices de escolaridade impedem as pessoas de adequarem sua vida pessoal e reprodutiva e promovam a busca por transformações sociais. A falta de políticas e de mecanismos que assegurem a todos os segmentos o acesso à escola, associada à evasão escolar e à repetência, está sempre associada às precárias condições socioeconômicas (RAMOS; CUMAN, 2009).

O estado civil também é um importante aspecto a ser levado em consideração, pois a ausência da figura paterna em geral pode trazer menor estabilidade financeira para a família, podendo se constituir em fator de risco para o nascimento.

A renda como aspecto econômico tem sido frequentemente associada com a saúde e o nascimento, tanto ao nível individual quanto ao coletivo. Nas famílias de menor renda, especialmente nos países em desenvolvimento, encontra-se uma alta frequência de desnutrição, de doenças transmissíveis e de condições ambientais deficientes, nível baixo de instrução e exercem ocupações que podem conter riscos apreciáveis para a saúde e consequentemente interferir na condição de nascimento e de reprodução (PEREIRA, 2003; PAULA, 2010).

2.4 Relacionamento entre a qualidade dos registros de nascimentos com as condições de vida

Há na literatura escassez de trabalhos que enfoquem o relacionamento entre a qualidade de registros de nascimentos e fatores associados as condições de vida. A maior dificuldade está da estimativa de indicadores de qualidade dos nascimentos. Se o interesse é voltado para as regiões com deficiência nos registros de nascimentos, então são exatamente nelas que estão as maiores dificuldades para dimensionar estas estimativas. Entre os poucos trabalhos elaborados nesta direção, se destacam os realizados por Paes (2010, 2014 e 2015).

Paes e Maia (2015) encontraram que pessoas que tinham registro de nascimentos declarados no censo de 2010 habitavam zonas urbanas com melhor infraestrutura de domicílios e melhores níveis sociais. Um segundo perfil é formado por pessoas com menos condições favoráveis ao registro de nascimento, que seriam aquelas sem instrução, habitando em áreas rurais, com uma baixa renda e sem esgotamento ou abastecimento de água nos domicílios. Portanto, são para as pessoas com o perfil do segundo fator, em que as prioridades para o registro de nascimentos deveriam ser reforçadas.

2.5 Modelos de Regressão

Para estudar o relacionamento das pessoas sem registro de nascimento com variáveis que permitam resumir os dados com as condições de vida, uma abordagem que pode abranger esta problemática é a análise de regressão. Esta, por sua vez, possibilita encontrar uma relação razoável entre as variáveis de entrada e saída, por meio de modelagem. O principal objetivo destes modelos é explorar a relação entre uma ou mais variáveis explicativas (ou independentes) e uma variável resposta (ou dependente). Nesta situação, a Regressão Linear Múltipla é um modelo desenhado para relacionar a variável resposta com mais de uma variável regressora.

Durante muitos anos os modelos normais lineares foram utilizados na tentativa de descrever a maioria dos fenômenos aleatórios. Sempre é útil conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer, sobre outras. Mesmo que não exista relação causal entre as variáveis pode-se relacioná-las por meio de uma expressão matemática, que pode servir para estimar o valor de uma das variáveis quando são conhecidos os valores das demais, sob determinadas condições.

Antes de iniciar um estudo de regressão logística é importante compreender que o objetivo de uma análise utilizando este método deverá ter pelo menos três qualidades fundamentais: parcimônia, generalidade e capacidade preditiva na relação entre a variável dependente e o conjunto de variáveis independentes (HOFFMANN, 2015).

Muitas técnicas estatísticas de construção destes modelos foram desenvolvidos nos

últimos dez anos e tem sido uma área de especial interesse em vários centros de pesquisas no mundo. A evolução do poder de processamento dos computadores pessoais contribuiu de maneira significativa para esta mudança nos métodos de análise de regressão, até então utilizados (PAIVA; FREIRE; CECATTI, 2010).

O que caracteriza um modelo de regressão logística é que a variável dependente é binária ou dicotômica. Ele segue, em geral, princípios usados na regressão linear. Em síntese, a Regressão Logística é uma técnica estatística que tem como objetivo modelar, a partir de um conjunto de observações, a relação "logística" entre uma variável resposta dicotômica e uma serie de variáveis explicativas numéricas (continuas, discretas) e/ou categóricas (CABRAL, 2013).

Os modelos lineares generalizados (MLG) tem sido amplamente estudados nos últimos anos devido às suas aplicações em diversas áreas do conhecimento incluindo: ciências atuariais, biologia, ciências biológicas, energia econômica, genômica, finanças, pescas, consumo de alimentos, crescimento da curvas de estimativa, investigação marinha, medicina, meteorologia, chuvas, vacinas, etc. (QUEIROZ, 2003).

CAPÍTULO 3

METODOLOGIA

3.1 Base de dados e Softwares

3.1.1 Fonte e organização da base de dados

Em 2010, o IBGE realizou o XII Censo Demográfico, que se constituiu na representação da população brasileira e das suas características socioeconômicas e, ao mesmo tempo, na base sobre a qual deverá se assentar todo o planejamento público e privado. O Censo 2010 é um retrato de corpo inteiro do país com o perfil da população e as características de seus domicílios. A coleta dos dados teve início em 1º de agosto de 2010, durando três meses. E os primeiros resultados foram divulgados em dezembro do mesmo ano (IBGE, 2016).

Os microdados consistes no menor nível de desagregação dos dados de uma pesquisa, retratando, sob a forma de códigos numéricos, o conteúdo dos questionários, preservando o sigilo das informações. Os microdados possibilitam aos usuários, com conhecimento de linguagens de programação ou *softwares* de cálculo, criar suas próprias tabelas. Os arquivos de microdados ora apresentados são acompanhados de uma documentação que fornece os nomes e os respectivos códigos das variáveis e suas categorias, a metodologia da pesquisa, e o instrumento de coleta.

A fonte oficial dos microdados de 2010 está disponibilizada na base censitária do IBGE Instituto Brasileiro de Geografia e Estatística (IBGE; http://www.ibge.gov.br). Para obter informações, os dados foram baixados em formato ".sav" o qual é utilizado no pacote estatístico *Statistical Package for the Social Sciences* (SPSS) versão 21.0.

O banco de dados original foi formado por 168 variáveis e 3.556.336 observações que são distribuídas por pessoa do Semiárido brasileiro. As 20 variáveis que foram utilizadas para este estudo foram pré-definidas com base na literatura com a expectativa de que pudessem ter alguma relação com a variável Registro de Nascimento.

A obtenção dos resultados, testes, e modelagens foram obtidos através da utilização do *software* R versão 3.2.4 (64 bit). Este *software* utiliza uma linguagem de programação orientada a objetos onde o usuário pode criar suas próprias funções e rotinas na análise de dados. Ele é uma importante ferramenta na manipulação e análise de dados, com testes paramétricos e não paramétricos, modelagem linear e não linear, análise de séries temporais, entre outros. O R[2] é um *software* livre para computação estatística, que é apresentado em versões de acordo com o sistema operacional *UNIX*, *Windows* ou *Macintosh*, apresenta código fonte aberto, podendo ser modificado ou implementado com novos procedimentos desenvolvidos por qualquer usuário a qualquer momento (R, 2016).

Para a geração da base de dados referente aos nascimentos do Semiárido brasileiro, fez-se a exportação dos dados para o *software Microsoft Office Excel* 2007 apenas das variáveis que foram utilizadas na modelagem. Para melhorar a interpretação das variáveis independentes, foram feitas modificações quanto as categorias, tal como está descrito na Seção 3.1.2.

3.1.2 Descrição das variáveis

Com a base de dados pronta, o banco[3] de microdados final apresentou 20 variáveis e 602.015 observações. A variável considerada como dependente é o Registro de nascimento e as demais apresentadas são as independentes.

- **Registro de nascimento** (no R: registro): Classificação da Informação:

  1 – Tem : Registro em cartório, Declaração de Nascido Vivo (DN) do hospital ou da maternidade e/ou Registro Administrativo de Nascimento Indígena (RANI)

  2 – Não tem: Não tem registro, não sabe, ignorado

  Branco: para as pessoas maiores de 10 anos de idade.

- **Idade calculada em anos** (no R: idade): Idade da pessoa em anos completos na data de referência da pesquisa.

- **Sexo**: Sexo da pessoa recenseada. Classificação da Informação:

  1 – Masculino

  2 – Feminino

- **Cor ou raça** (no R: cor): Cor ou raça conforme declaração da pessoa recenseada. Classificação da Informação:

  1 – Branca: para a pessoa que se declarou branca.

  2 – Não branca: para a pessoa que se declarou preta, amarela , parda ou indígena

[2] *Script* completo no Apêndice A

[3] Disponibilizado por email: `ligia__maia@hotmail.com`

- **Situação do domicílio** (no R: domicilio): Domicílio é o local estruturalmente separado e independente que se destina a servir de habitação a uma ou mais pessoas, ou que esteja sendo utilizado como tal na data de referência. Situação do domicílio em relação à sua localização quanto ao perímetro urbano do distrito, conforme estabelecido por lei municipal. Classificação da Informação:

  1 – Urbano: Área interna ao perímetro urbano de um distrito, composta por setores nas seguintes situações de setor

  2 – Rural: Área externa ao perímetro urbano de um distrito, composta por setores nas seguintes situações de setor.

- **Sabe ler e escrever** (no R: ler): Condição de alfabetização da pessoa. Classificação da Informação:

  1 – Sim: Para a pessoa que sabe ler e escrever um bilhete simples no idioma que conhece. Considerou-se também a pessoa alfabetizada que se tornou física ou mentalmente incapacitada de ler ou escrever.

  2 – Não: Para a pessoa que não sabe ler e escrever ou que apenas escreve o próprio nome. Considerou-se também como não sabendo ler e escrever a pessoa que aprendeu, mas esqueceu devido a ter passado por um processo de alfabetização que não se consolidou.

  Branco: para as pessoas menores de 5 anos de idade.

- **Rendimento domiciliar per capita em julho de 2010, em número de salários mínimos** (no R: renda): Rendimento bruto proveniente da divisão do rendimento mensal domiciliar pelo número de moradores do domicílio particular, exclusive aqueles cuja condição no domicílio fosse pensionista, empregado doméstico ou parente do empregado doméstico, em número de salários mínimos.

  1 - Menos de meio salário mínimo

  2 - Entre meio e um salário mínimo e meio

  3 - Mais de um salário mínimo e meio

- **Em julho de 2010, tinha rendimento mensal habitual de Programa Social Bolsa Família ou Programa de Erradicação do Trabalho Infantil – PETI (no R: bolsa)**: Este quesito destinava-se a captar se a pessoa tinha rendimento mensal habitual, no mês de julho de 2010, proveniente do Programa Social Bolsa Família (BF) ou do Programa de Erradicação do Trabalho Infantil (PETI). O programa Bolsa Família é um programa do governo federal, de transferência direta de rendimento com condicionalidades, que beneficia famílias em situação de pobreza. O programa de Erradicação do Trabalho Infantil-PETI é um programa do governo

federal que tem como objetivo contribuir para a erradicação de todas as formas de trabalho infantil no País, atendendo famílias cujas crianças e adolescentes com idade inferior a 16 anos se encontrem em situação de trabalho. Classificação da Informação:

1 - Sim

0 – Não

9 – Ignorado

Branco: para quem, na semana de 25 a 31 de julho de 2010 era menor de 10 anos de idade.

- **Banheiro de uso exclusivo, número** (no R: banheiro): Informação coletada somente para domicílios particulares permanentes. Banheiro é o cômodo que dispõe de chuveiro (ou banheira) e vaso sanitário (ou privada) e que seja de uso exclusivo dos moradores, inclusive os localizados no terreno ou na propriedade. Nota: Nos domicílios onde a instalação sanitária e o chuveiro ou banheira encontrem-se em compartimentos distintos, considera-se que o domicílio tem banheiro e os dois compartimentos onde o sanitário e o chuveiro se encontram são contados como um só cômodo. Classificação da informação:

  0 – zero banheiros

  1 – um banheiro

  2 – dois ou mais banheiros

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Esgotamento sanitário, tipo** (no R: esgoto): Informação coletada somente para domicílios particulares permanentes. Classificação da Informação:

  1 – Rede geral de esgoto ou pluvial: quando a canalização das águas servidas e dos dejetos, proveniente do banheiro ou sanitário, estava ligada a um sistema de coleta que os conduzia a um desaguadouro geral da área, região ou município, mesmo que o sistema não dispusesse de estação de tratamento da matéria esgotada.

  2 – Fossa séptica: quando a canalização do banheiro ou sanitário estava ligada a uma fossa séptica, ou seja, a matéria era esgotada para uma fossa próxima, onde passava por um processo de tratamento ou decantação, sendo, ou não, a parte líquida conduzida em seguida para um desaguadouro geral da área, região ou município.

  Branco: para domicílio particular improvisado, domicílio coletivo e domicílio particular permanente sem utilização de sanitário ou buraco para dejeções.

- **Abastecimento de água, canalização** (no R: agua): Informação coletada somente para domicílios particulares permanentes. Classificação quanto à existência

de canalização para a distribuição de água no domicílio. Classificação da Informação:

1 – Sim: Quando o domicílio for servido de água canalizada com distribuição interna para um ou mais cômodos ou quando a água chegar canalizada até a propriedade ou terreno sem haver distribuição interna no domicílio.

2 – Não: Quando não existir água canalizada no domicílio, na propriedade ou no terreno.

Branco: para domicílio particular improvisado e domicílio coletivo.

- **Energia elétrica, existência** (no R: energia): Informação coletada somente para domicílios particulares permanentes. Existência de energia elétrica no domicílio. Classificação da Informação:

  1 - Sim: Quando o domicílio for servido de energia elétrica de companhia distribuidora ou quando o domicílio for servido de energia elétrica proveniente de outras fontes, como: eólica, solar, gerador, etc.

  2 - Não existe energia elétrica: Quando o domicílio não possuir energia elétrica.

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Rádio, existência** (no R: radio): Informação coletada somente para domicílios particulares permanentes. Inclusive integrado a outro tipo de aparelho. Classificação da Informação:

  1 – Sim: quando houver no domicílio qualquer tipo de aparelho de rádio, inclusive à pilha ou integrado a outro tipo de aparelho.

  2 – Não: considere, também, neste item, o rádio integrado a aparelhos de uso pessoal, como telefone celular, mp3 player, etc.

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Televisão, existência** (no R: tv): Informação coletada somente para domicílios particulares permanentes. Existência de televisores tanto em cores como em preto e branco, plasma e LCD, desde que em condições de uso. Classificação da Informação:

  1 – Sim

  2 – Não

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Máquina de lavar roupa, existência** (no R: maquina): Informação coletada somente para domicílios particulares permanentes. Branco para domicílio particular improvisado e domicílio coletivo. Classificação da Informação:

1 – Sim

2 – Não: quando no domicílio não houver máquina de lavar roupa ou a máquina existente apenas lavar a roupa sem realizar as operações de enxágue e centrifugação (tanquinho e similares).

- **Geladeira, existência** (no R: geladeira): Informação coletada somente para domicílios particulares permanentes. Classificação da Informação:

  1 – Sim: quando no domicílio houver geladeira, mesmo que seja a gás ou querosene.

  2 – Não

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Telefone celular, existência** (no R: celular): Informação coletada somente para domicílios particulares permanentes. Classificação da Informação:

  1 – Sim: se pelo menos um morador possuir telefone celular.

  2 – Não Branco: para domicílio particular improvisado e domicílio coletivo.

- **Microcomputador, existência** (no R: computador): Informação coletada somente para domicílios particulares permanentes Classificação da Informação:

  1 – Sim: para o domicílio que possuir *desktop* (computador de mesa), laptop, (notebook) e netbook.

  2 – Não

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Motocicleta para uso particular, existência** (no R: moto): Informação coletada somente para domicílios particulares permanentes Classificação da Informação:

  1 – Sim: para o domicílio em que um de seus moradores possua: uma motocicleta para passeio ou locomoção de seus moradores para trabalho ou estudo, ou ainda, a motocicleta utilizada para desempenho profissional de ocupações como: moto-táxi, entregador de correspondências, pequenas encomendas, etc., desde que seja utilizada também para passeio ou locomoção dos moradores do domicílio.

  2 - Não

  Branco: para domicílio particular improvisado e domicílio coletivo.

- **Automóvel para uso particular, existência** (no R: carro): Informação coletada somente para domicílios particulares permanentes Classificação da Informação:

  1 – Sim: para o domicílio em que um de seus moradores possua um automóvel de passeio ou veículo utilitário para passeio ou locomoção dos seus moradores para

trabalho ou estudo, ou ainda o veículo utilizado para desempenho profissional de ocupações como motorista de táxi, vendedor que tem necessidade de transportar amostras de sua mercadoria para atender ou solicitar pedidos, etc., desde que este seja utilizado também para passeio ou locomoção dos moradores do domicílio.

2 – Não

Branco: para domicílio particular improvisado e domicílio coletivo.

3.2 Processo da Modelagem

Por ocasião do Censo Demográfico, fez-se a opção por investigar para as pessoas com idade até 10 anos, a existência de documentos referentes ao nascimento desses indivíduos, sendo eles: o registro público feito em cartório, a Declaração de Nascido Vivo (DN) ou o Registro Administrativo de Nascimento Indígena (RANI). As demais características se referem a "Não tem", "Não sabe" e "Ignorado". A seguinte pergunta foi feita no questionário "Tem Registro de Nascimento?".

Para Bob e Mikis (2008), modelagem estatística é a arte de construir modelos estatísticos parcimoniosos para uma melhor compreensão dos fenômenos de interesse. Para a modelagem referente à variável dependente Registro de Nascimento do Semiárido brasileiro foram realizadas as seguintes etapas (Figura 3.1):

A priori fez-se uso de todas as variáveis pré-selecionadas para realizar a modelagem da variável registro de nascimento, assim, teve-se uma variável dependente e 19 independentes.

Como a variável resposta é binária, utilizou-se a Regressão Logística Binomial. Foram selecionadas as variáveis para os Modelos Lineares Generalizados (MLG) com a finalidade de escolher os melhores parâmetros e as variáveis que melhor explicassem o fator ter ou não registro de nascimento.

O segundo passo da modelagem foi definir qual função de ligação utilizar: *Logit*, *Probit* ou *Cauchit*. Logo de inicio, a função de ligação *Cauchit* apresentou limitações para a modelagem da regressão logística, pois o algoritmo não convergiu. Desta forma, ela foi descartada.

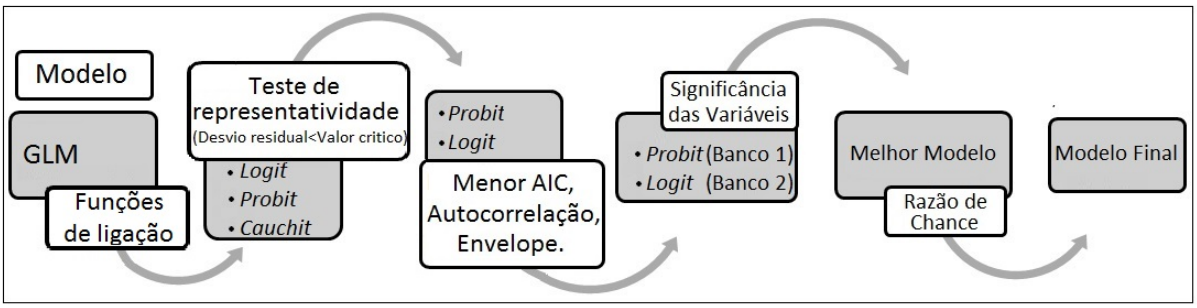
Para utilizar a melhor função de ligação, fez-se testes de representatividade do modelo, seleção do menor AIC (critério de Seleção de Akaike), autocorrelação e envelope. Após a definição, utilizou-se a função de ligação (Logit ou Probit) determinada para avaliar quão significativas foram as variáveis selecionadas em cada modelo. Assim, foi estimado o melhor modelo de regressão logística para a variável Registro de Nascimento, através da razão de chance.

A posteriori observou-se que a variável Bolsa (Em julho de 2010, tinha rendimento mensal habitual de Programa Social Bolsa Família ou Programa de Erradicação do Tra-

balho Infantil – PETI) considerou apenas pessoas com 10 anos de idade. Assim, notem que o Banco original que era composto por 602.015 observações reduziu-se a 54.774 observações, ou seja a aproximadamente 10% do número inicial de informações.

Sendo assim, tomou-se como objetivo, comparar os resultados obtidos para estes Bancos, nomeados: Banco 1 (considera a variável bolsa), e o Banco 2 (não considera a variável bolsa). O Banco 2 ficou formado por 293.399 (aproximadamente 49%) observações por excluir as linhas em que as demais variáveis tiveram informações em branco.

Figura 3.1: Fluxo das atividades para a modelagem da regressão logística



3.3 Modelos Lineares Generalizados (MLG)

3.3.1 Fundamentos básicos

A análise de regressão é usada para explicar ou modelar a relação entre uma única variável Y, chamada de resposta, de saída ou variável dependente; e um ou mais preditores, variáveis independentes ou explicativas, $X_1, X_2, ..., X_n$. Quando $n = 1$, tem-se a regressão simples, mas quando $n > 1$ a regressão múltipla ou regressão multivariada será a utilizada. A variável X é a variável independente da equação enquanto $Y = f(X)$ é a variável dependente das variações de X. O modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis. O modelo de regressão é múltiplo quando envolve uma relação causal com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por mais de uma variável independe $X_1, X_2, ..., X_n$ (FARAWAY, 2009).

3.3.2 Definição

Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLG), que permitem outras opções, além da distribuição normal, para a distribuição da variável resposta permitindo que a mesma pertença à família exponencial de distribuições, bem

como dar maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear.

A estrutura de um MLG é formada por três componentes:

- Aleatória: composta de uma variável aleatória Y com n observações independentes, um vetor de médias μ e uma distribuição de probabilidade pertencente à família exponencial;
- Sistemática: define o preditor linear $\eta = X\beta$;
- Função de ligação: relaciona as componentes anteriores.

3.3.3 Componente Aleatória

Seja $y = (y_1, ..., y_n)^T$ um vetor de observações referente às realizações da variável aleatória $Y = (Y_1, ..., Y_n)^T$, independentes e identicamente distribuídas, com vetor de médias $\mu = (\mu_1, ..., \mu_n)^T$, e com função de densidade da forma

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)\right\}, \tag{3.1}$$

onde $a(.)$, $b(.)$ e $c(.)$ são funções conhecidas, ϕ é o parâmetro de dispersão e θ_i é denominado parâmetro natural ou canônico, que caracteriza a distribuição em (3.1). Se ϕ é conhecido, a equação (3.1) representa a família exponencial uniparamétrica.

A log-verossimilhança é definida por

$$\ell(y_i; \theta_i, \phi) = \_i; \theta_i, \phi),$$

portanto,

$$\ell(y_i; \theta_i, \phi) = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi), \tag{3.2}$$

Derivando (3.2) sucessivamente com relação a θ_i temos

$$\frac{\partial \ell}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{a(\phi)} \tag{3.3}$$

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}. \tag{3.4}$$

Pode-se mostrar que $E\left[\frac{\partial \ell}{\partial \theta_i}\right] = 0$ e de (3.3) tem-se que

$$E\frac{[y_i - b'(\theta_i)]}{a(\phi)} = 0,$$

de onde tem-se que

$$E[Y_i] = \mu_i = b'(\theta_i).$$

Pode-se também mostrar que

$$E\left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right] + E\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = 0. \tag{3.5}$$

Então, a partir de (3.4) e (3.5) obtém-se

$$E\left\{-\frac{b''(\theta_i)}{a(\phi)}\right\} + E\left\{\frac{[y_i - b'(\theta_i)]}{a(\phi)}\right\}^2 = 0$$

$$-\frac{b''(\theta_i)}{a(\phi)} + \frac{1}{[a(\phi)]^2} E[y_i - E(y_i)]^2 = 0$$

$$\frac{1}{[a(\phi)]^2} Var(Y_i) = \frac{b''(\theta_i)}{a(\phi)},$$

logo, $Var(Y_i) = a(\phi)b''(\theta_i)$, que pode também ser escrita na forma $Var(Y_i) = a(\phi)V_i$, onde $V_i = \dfrac{d\mu}{d\theta_i}$ é chamado de variância.

Dentre as distribuições que pertencem à família exponencial pode-se elencar a Normal, Gamma, Poisson, Binomial.

Distribuição Binomial com parametrização na família exponencial

Seja Y uma variável aleatória binomial baseada em n repetições, denotada por $Y \sim Bin(n, \mu)$. Sua função de densidade de probabilidade é expressa por

$$f(y; \mu) = \binom{n}{y} \mu^y (1-\mu)^{n-y}, \tag{3.6}$$

onde $0 < \mu < 1$ e $y = 0, 1, 2, ...$

Podem-se expressar (3.6) na seguinte forma:

$$f(y) = \exp\left\{\log\binom{n}{y}\left(\frac{\mu}{1-\mu}\right) + n\log(1-\mu)\right\}. \tag{3.7}$$

Comparando (3.7) com a expressão que caracteriza a família exponencial tem-se que

$$\theta = \log\left(\frac{\mu}{1-\mu}\right), \quad b(\theta) = n\log(1-\mu), \quad a(\phi) = 1, \quad c(y, \phi) = \log\binom{n}{y}.$$

Portanto,

$$E(Y) = b'(\theta) = \theta = \mu, \quad V = b''(\theta) = 1, \quad Var(Y) = a(\phi)V = \sigma^2.$$

3.3.4 Componente Sistemática

A componente sistemática é formada pela estrutura linear de um modelo de regressão $\eta = X\beta$, onde $\eta = (\eta_1, ..., \eta_n)^T$, $\beta = (\beta_1, ..., \beta_p)^T$ e X é uma matriz modelo de dimensão $n \times p$ $(p < n)$ conhecida, de posto p. A função linear η dos parâmetros desconhecidos β é chamada de preditor linear e corresponde à parte sistemática de um MLG.

3.3.5 Função de Ligação

A média μ do vetor y é expressa por uma função g de η chamada de função de ligação. Esta, por sua vez, relaciona o componente aleatório a componente sistemática, ou seja, vincula a média ao preditor linear, isto é,

$$\mu_k = g^{-1}(\eta_k) \quad \text{ou} \quad \eta = g(\mu_k), \quad k = 1, ..., n,$$

sendo $g(.)$ uma função monótona e diferenciável.

Modelos que assumem a distribuição binomial para a variável resposta, onde $0 < \mu < 1$, o domínio da função de ligação deve, necessariamente, estar no intervalo $(0; 1)$, enquanto que seu contradomínio é o intervalo $(-\infty; +\infty)$. A escolha da função de ligação depende do problema em particular. Funções garantem esta condição para o modelo binomial:

- *Logit* (função de ligação canônica): $\eta = \log\left(\dfrac{\mu}{1-\mu}\right)$
- *Probit*: $\eta = \Phi^{-1}(\mu)$, onde $\Phi(.)$ é a função de distribuição acumulada da normal reduzida;
- *Cauchit*: $\eta = \dfrac{1}{\tilde{u}} \arctan(\mu) + \dfrac{1}{2}$

3.3.6 Estimação de β

O algoritmo de estimação dos parâmetros β's foi desenvolvido por Nelder e Wedderburn em 1972 e baseia-se no Método Escore de Fisher, semelhante ao de Newton-Raphson. A principal diferença em relação ao modelo clássico de regressão é que as equações de máxima verossimilhança são não-lineares. Seja $\ell(\beta)$ a log-verossimilhança como função de β e considere a função escore de Fisher

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta},$$

e a matriz de informação de Fisher

$$K = \left\{ -E\left(\frac{\partial^2 \ell(\beta)}{\partial \beta_j \partial \beta_s}\right)\right\} = -E\left(\frac{\partial U(\beta)}{\partial \beta}\right).$$

Expandindo a função escore em série de Taylor até primeira ordem obtém-se

$$U(\beta^{(m+1)}) = U(\beta^{(m)}) + \frac{\partial U(\beta)^{(m)}}{\partial \beta}[\beta^{(m+1)} - \beta^{(m)}] = 0$$

ou

$$\beta^{(m+1)} = \beta^{(m)} - \left[\frac{\partial U(\beta)^{(m)}}{\partial \beta}\right]^{-1} U(\beta^{(m)}),$$

onde o índice (m) significa o valor do termo na m-ésima iteração. O método escore de Fisher (1925) é obtido pela substituição de $-\dfrac{\partial U(\beta)}{\partial \beta}$ pelo seu valor esperado K, o que resulta no seguinte processo iterativo

$$\beta^{(m+1)} = \beta^{(m)} + K^{-1}(\beta^{(m)})U(\beta^{(m)}). \tag{3.8}$$

Trabalhando a expressão (3.8), chega-se a um processo iterativo de mínimos quadrados reponderados

$$\beta^{(m=1)} = (X^T W^{(m)} X)^{-1} X^T W^{(M)} z^{(m)}, \tag{3.9}$$

$m = 0, 1, ...,$ onde $z = \eta + W^{-1/2}V^{-1/2}(y - \mu)$. Note que z desempenha o papel de uma variável dependente modificada, enquanto W é uma matriz de pesos que muda a cada passo do processo iterativo. A convergência de (3.9) ocorre em um número finito de passos, independente dos valores iniciais utilizados. É usual iniciar (3.9) com $\eta^{(0)} = g(y)$. Para o modelo binomial logístico linear, tem-se

$$z = \eta + W^{-1/2}V^{-1/2}(y - \mu) = \eta + W^{-1}(y - \mu) = \eta + (y - n\mu)/n\mu(1 - \mu).$$

Sob condições de regularidade tem-se que $\hat{\beta}$ é um estimador consistente e eficiente de β e que

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, \phi^{-1}(\Sigma(\beta))^{-1}) \qquad \text{quando } n \to \infty,$$

onde

$$\Sigma(\beta) = \lim_{n\to\infty} \frac{K(\beta)}{n},$$

sendo $\Sigma(\beta)$ uma matriz definida positiva. E as condições para que exista $\Sigma(\beta)$ e seja definida positiva são

$$n_i/n \to a_i > 0, \qquad n \to \infty$$

e $\sum_{i=1}^{g} x_i x_i'$ seja de posto completo, onde $n = n_1 + \cdots + n_g$.

Sob certas condições de regularidade tem-se que

$$\sqrt{n}(\hat{\phi} - \phi) \to N(0, \sigma^2\phi),$$

quando $n \to \infty$, onde $\sigma_\phi^2 = \lim_{n\to\infty} -n\left[\sum_{i=1}^{n} c''(y_i, \phi)\right]^{-1}$, ou seja um estimador consis-

tente para $Var(\hat{\phi})$ é $[\sum_{i=1}^{n} -c''(y_i, \phi)]^{-1}$.

3.3.7 Função Desvio

Uma maneira de analisar discrepância ou bondade de ajuste é observar o desvio que equivale à diferença de log-verossimilhanças maximizadas. Seja o logaritmo da função de verossimilhança de $y = (y_1, ..., y_n)^T$, uma amostra aleatória com distribuição pertencente à família exponencial, expresso como função da média, isto é:

$$L(\mu; y) = \sum_{i=1}^{n} \ell(\mu_i, y_i),$$

onde $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = x_i'\beta$. Considerando o número de componentes do vetor de parâmetros $\beta(p)$ igual ao número de observações n, tem-se então o modelo saturado e a função $L(\mu; y)$ é estimada por

$$L(y; y) = \sum_{i=1}^{n} \ell(y_i, y_i).$$

Seja a estimativa de $L(\mu; y)$ dada por $L(\hat{\mu}; y)$, quando $p < n$. A estimativa de máxima verossimilhança de μ_i será dada por $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$, onde $\hat{\eta}_i = x_i'\hat{\beta}$, com $\hat{\beta}$, o estimador de máxima verossimilhança de β. A função desvio é definida da seguinte forma

$$D(y; \hat{\mu}) = \phi^{-1} D(y; \hat{\mu}) = 2L(y; y) - L(\hat{\mu}; y),$$

que é a distância entre o logaritmo da função de verossimilhança do modelo saturado e do modelo sob investigação (modelo com p parâmetros) avaliado na estimativa de máxima verossimilhança $\hat{\beta}$. Um valor pequeno para a função desvio indica que para um número menor de parâmetros , obtém-se um ajuste tão bom quanto o ajuste no modelo saturado. Sejam $\hat{\theta}_i = \theta_i(\hat{\mu})$ e $\tilde{\theta}_i = \theta_i(y)$ as estimativas dos parâmetros canônicos para o modelo em investigação e o modelo saturado, respectivamente. Tem-se que a função $D(y; \hat{\mu})$ fica dada por

$$D(y; \hat{\mu}) = 2\phi \sum_{i=1}^{n} \{y_i(\tilde{\theta}_i - \hat{\theta}_i) + [b(\hat{\theta}_i) - b(\tilde{\theta}_i)]\},$$

que é chamada função desvio para o modelo corrente.

$$D(y; \mu) \sim \chi^2_{n-p}$$

Se $D(y; \mu)/\phi \leq \chi^2_{n-p;1-\alpha}$, o modelo em investigação é aceito. Este teste equivale a um teste F num modelo de regressão.

Função desvio para a distribuição Binomial: Assumindo $Y_i \sim B(n_i, \mu_i)$, $i = 1, ..., k$, obtemos $\tilde{\theta}_i = \log\{y_i/(n_i - y_i)\}$ e $\tilde{\theta}_i = \log\{\hat{\mu}_i/(1 - \hat{\mu}_i)\}$ para $0 < y_i < n_i$. Logo, o desvio

assume a seguinte forma:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu_i}) + (n_i - y_i)\log[(n_i - y_i)/(n_i - \hat{\mu_i})]\}.$$

Todavia, quando $y_i = 0$ ou $y_i = n_i$, o i-ésimo termo de $D(y;\hat{\mu})$ vale $-2n_i \log(1 - \hat{\mu}_i)$ ou $-2n_i \log \hat{\mu}_i$, respectivamente. Portanto, os componentes do desvio no caso binomial assumem as seguintes formas:

$$\begin{array}{cc} y_i \log(y_i/n_i\hat{\mu_i}) + (n_i - y_i)\log[(1 - y_i/n_i)/(1 - \hat{\mu_i})] & \text{se } 0 < y_i < n_i; \\ -2n_i \log(1 - \hat{\mu}_i) & \text{se } y_i = 0; \\ -2n_i \log \hat{\mu}_i & \text{se } y_i = n_i. \end{array}$$

Um valor pequeno para a função de desvio, indica que, para um número menor de parâmetros, obtém-se um ajuste tão bom quanto o ajuste com modelo saturado.

3.3.8 Modelos para Dados Binários

Um dos casos particulares dos MLGs são os modelos para variáveis que apresentam apenas duas categorias ou que foram de alguma forma dicotomizadas. As variáveis que assumem valores 0 ou 1 chamadas *"dummy"* podem ser caracterizadas pela distribuição de Bernoulli. Comumente é chamado de sucesso (1) o resultado mais importante da resposta ou aquele que pretende-se relacionar com as demais variáveis de interesse (FERREIRA, 2004). Quando a função de ligação é a identidade (baseado na transformação *logit*) tem-se o modelo de probabilidade linear, porém isso as vezes não é conveniente e outras funções de ligação podem ser apropriadas, como é o caso da função *probit*.

A Análise de Regressão Logística tem sido mencionada como uma ferramenta poderosa de modelagem estatística principalmente para variáveis categóricas em diversos campos como epidemiologia, pesquisa médica, bancos, pesquisa de mercado, pesquisa social, etc. Mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns pesquisadores têm dicotomizado a resposta de modo que a probabilidade de sucesso possa ser ajustada através da regressão logística. Ela consiste em relacionar, através de um modelo, a variável resposta categórica (geralmente dicotômica, mas pode ser politômica, isto é, ter mais do que dois níveis de respostas) com as variáveis explanatórias (categóricas ou contínuas) que influenciam a ocorrência de determinado evento. Tudo isso se deve, principalmente, pela facilidade de interpretação dos parâmetros de um modelo logístico e também pela possibilidade do uso desse tipo de metodologia em análise discriminante (QUEIROZ, 2003; PAULA, 2004).

Modelo *Logit*

Seja o parâmetro natural $\theta_i(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, o logaritmo da razão das chances da resposta 1, é chamado o *logit* de π_i. Os MLGs que usam a ligação *logit* são chamados modelos *logits*, para os quais o preditor linear é dada por

$$\eta_i = x_i'\beta = \log\left(\frac{\pi_i}{1-\pi_i}\right),$$

onde $\pi_i(x)$ denota probabilidade de $y_i = 1$ para os valores das variáveis explicativas, ou seja, o grupo de preditores $x = (x_1, x_2, ..., x_p)$. No caso particular em que se tem $x = (1, x_1)$ e quando a relação entre x e $\pi_i(x)$ não é linear, ou seja, quando a relação se apresenta de forma curvilínea e monótona, então a função de ligação que relaciona o valor esperado de Y_i ao componente linear, isto é, $g(\mu_i) = x_i'\beta$, da forma

$$\pi_i(x) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}.$$

Daí,

$$\begin{aligned}
[1 + exp(\beta_0 + \beta_1 x_1)]\pi_i(x) &= \exp(\beta_0 + \beta_1 x_1) \\
\pi_i(x) + \exp(\beta_0 + \beta_1 x_1)\pi_i(x) &= \exp(\beta_0 + \beta_1 x_1) \\
\pi_i(x) &= \exp(\beta_0 + \beta_1 x_1) - exp(\beta_0 + \beta_1 x_1)\pi_i(x) \\
\pi_i(x) &= \exp(\beta_0 + \beta_1 x_1)[1 - \pi_i(x)] \\
\frac{\pi_i(x)}{[1 - \pi_i(x)]} &= \exp(\beta_0 + \beta_1 x_1),
\end{aligned}$$

ou, equivalentemente a

$$\log\frac{\pi_i(x)}{1 - \pi_i(x)} = \beta_0 + \beta_1 x_1 = x(\beta_0, \beta_1)^T,$$

que é chamada função de regressão logística. Assim, a função de ligação é o logaritmo das chances *(odds)*, o *logit*.

A fórmula fornece uma interpretação simples para β. As chances aumentam multiplicando por e^{β_1} para cada aumento na unidade em x_1. Para centrar o preditor em torno de 0 (isto é, substituindo x por $(x - \bar{x})$) β_0 torna-se o *logit* da média, e assim $e^{\beta_0}/(1 + e^{\beta_0}) = \pi(\bar{x})$. Generalizando para múltiplas variáveis explicativas, o modelo de regressão logística para valores $x = (x_1, x_2, ..., x_p)'$ de p variáveis explicativas é

$$\text{logit}[\pi_i(x)] = \log\frac{\pi_i(x)}{1 - \pi_i(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$

Assim quando a variável resposta é categorizada em duas categorias, modelos com ligação

logit são um caso particular dos modelos log-lineares.

Modelo *Probit*

Considerando a relação entre x e $\pi(x)$ da forma

$$\pi(x) = Pr(Z^* \leq \beta_0 + \beta_1 x_1) = \Phi(\beta_0 + \beta_1 x_1),$$

tem-se o denominado Modelo *Probit*, onde Z^* é uma variável aleatória com distribuição Normal Padrão e Φ é a densidade normal padrão acumulada, β_0 e β_1 são parâmetros desconhecidos a serem estimados. O modelo *probit* é uma função não linear para um conjunto linear de parâmetros e nesse caso a ligação do valor esperado com o preditor linear tem a seguinte forma

$$\eta = \Phi^{-1}(\pi(x)),$$

onde $\eta = \beta_0 + \beta_1 x_1$ e portanto,

$$\text{probit}[\pi(x)] = \Phi^{-1}[\pi(x)] = (\beta_0 + \beta_1 x_1),$$

é um MLG com função de ligação igual a inversa de Φ. Para este modelo, a curva para $\pi(x)$ ou para $[1 - \pi(x)]$, quando $\beta_1 < 0$ tem a aparência da densidade da normal com média $-\beta_0/\beta_1$ e desvio padrão $\sigma = 1/|\beta_1|$. O modelo *probit* é não-linear nos parâmetros e restringe $\pi(x)$ ao intervalo $(0, 1)$.

Regressão logística simples

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

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \alpha + \beta x,$$

em que α e β 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, 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 é dada por

$$\frac{\pi(1)}{1-\pi(1)} = e^{\alpha+\beta},$$

enquanto que a chance de desenvolvimento da doença para um indivíduo com ausência

do fator é simplesmente

$$\frac{\pi(0)}{1-\pi(0)} = e^{\alpha}.$$

Logo, a razão de chances é dada por

$$\psi = \frac{\pi(1)\{1-\pi(0)\}}{\pi(0)\{1-\pi(0)\}} = e^{\beta},$$

dependendo apenas do parâmetro β. Mesmo que a amostragem seja retrospectiva, isto é, são amostrados n_1 indivíduos registrados e n_2 indivíduos não registrados, o resultado acima continua valendo. Essa é uma das vantagens da regressão logística, a possibilidade de interpretação direta dos coeficientes como medidas de associação.

Regressão logística múltipla

Considerando o modelo geral de regressão logística, tem-se

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

em que $x = (1, x_2, ..., x_p)^T$ contém os valores observados de variáveis explicativas. O processo iterativo para obter $\hat{\beta}$ pode ser expresso como um processo iterativo de mínimos quadrados reponderados

$$\beta^{(m+1)} = (X^T V^{(m)} X)^{-1} X^T V^{(m)} z^{(m)},$$

em que $V = \text{diag}\{\pi_1(1-\pi_1), ..., \pi_n(1-\pi_n)\}$, $z = (z_1, ..., z_n)^T$ é a variável dependente modicada, $z_i = \eta_i + (y_i - \pi_i)/\pi_i(1-\pi_i)$, $m = 0, 1, ...$ e $i = 1, ..., n$. Para dados agrupados (k grupos), deve-se substituir n por k, $V = \text{diag}\{n_1\pi_1(1-\pi_1), ..., n_k\pi_k(1-\pi_k)\}$ e $z_i = \eta_i + (y_i - n_i\pi_i)/n_i\pi_i(1-\pi_i)$. Assintoticamente, $n \to \infty$ no primeiro caso e para $\frac{n_i}{n} \to a_i > 0$ no segundo caso, $\hat{\beta} - \beta \sim N_p(0, (X^T V X)^{-1})$.

Assim com o na Regressão Logística Linear, a razão de chances é dada por

$$\psi = e^{\beta}.$$

3.3.9 Seleção do melhor modelo

Uma vez definido o conjunto de covariáveis a ser incluído num modelo logístico, resta saber qual a melhor maneira de encontrar um modelo parcimonioso que inclua apenas as covaráveis e interações mais importantes para explicar a probabilidade de sucesso $\pi(x)$.

Método *stepwise*

Segundo Paula (2004), a seleção das variáveis do modelo logístico é realizada por um procedimento "passo a passo", conhecido como *stepwise*. O método baseia-se num algoritmo misto de inclusão e eliminação de variáveis explicativas segundo a importância das mesmas de acordo com algum critério estatístico. Esse grau de importância pode ser avaliado, por exemplo, pelo Critério Akaike (AIC) ou pelo nível de significância do teste da razão de verossimilhança entre os modelos que incluem ou excluem as variáveis em questão. Quanto menor forem esses níveis de significância ou AIC, mais importante será considerada a variável explicativa. Uma desvantagem do procedimento é exigir as estimativas de máxima verossimilhança em cada passo, o que sobrecarrega o trabalho computacional, principalmente quando há muitas variáveis explicativas (como é o caso deste trabalho).

Critério Akaike (AIC)

O método proposto por Akaike em 1974 tem como ideia básica selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança $L(\beta)$ cresce com o aumento do número de parâmetros do modelo, uma proposta razoável é encontrar o modelo com menor valor para a função (PAULA, 2004)

$$AIC = -2L(\hat{\beta}) + 2p,$$

onde p denota o número de parâmetros. Em função do desvio do modelo, tem-se:

$$AIC = D^*(y; \hat{\mu}) + 2p$$

em que $D^*(y; \hat{\mu})$ denota o desvio do modelo e p o número de parâmetros.

Autocorrelação

A autocorrelação é a correlação cruzada de uma observação com ela própria. É uma ferramenta estatística para encontrar padrões de repetição, tal como a presença de uma observação periódica obscurecida pelo ruído, ou para identificar a frequência fundamental em falta numa observação implícita pelas suas frequências harmônicas. As funções de autocorrelação (ACF) no *software* R calcula estimativas da função de autocovariância ou autocorrelação. Já para calcular a função para as autocorrelações parciais a função utilizada é PACF.

É comum que conjuntos de dados apresentem correlação, assim, estes conjuntos assumem um pressuposto necessário para a validade de mínimos quadrados com base em métodos de regressão, que os erros são independentes. Então, a seguir, como os mínimos

quadrados generalizados pode ser usado para ajustar modelos com erros autocorrelacionados. Os benefícios da transformação dos modelos de mínimos quadrados generalizado em modelos de mínimos quadrados, se trata de examinar o diagnósticos modelo (lSHEATHER, 2009).

Em vez de produzir gráficos de dispersão, uma prática comum estatística, consiste em olhar para os valores da correlação entre Y e os vários valores de Y defasado para diferentes períodos. Tais valores são chamados autocorrelações.

$$ACF(L) = \frac{\sum_{t=l+1}^{n}(y_t - \bar{y})(y_{t-l} - x\bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

As linhas tracejadas correspondem a $-2/\sqrt{n}$ e $+2/\sqrt{n}$, uma vez que autocorrelações são declarados como sendo estatisticamente diferentes de zero se forem menos de $-2/\sqrt{n}$ ou superior a $+2/\sqrt{n}$ (ou seja, se eles são mais do que dois desvios padrão de distância de zero).

3.3.10 Estudo dos Resíduos

Resíduos e, especialmente, parcelas de resíduos, desempenham um papel central na verificação de modelos estatísticos. Na regressão linear normal, os resíduos são normalmente distribuídos e pode ser padronizado para ter variâncias iguais. Em situações de regressão não-normais, tais como regressão logística ou análise de log-linear. Um problema particular ocorre quando a variável resposta é binomial. Em tal situação, os resíduos se encontram em curvas quase paralelas correspondentes aos valores da variável resposta, e essas curvas dificultam quaisquer que sejam as conclusões que possam ser observadas graficamente (DUNN; SMYTH, 1996).

A distribuição dos resíduos para os modelos de regressão não linear, é matematicamente trabalhosa e os critérios de diagnóstico são falhos. Para analisar os resíduos, pode-se padronizá-los afim de que estes se adequem melhor aos métodos de análise da bondade do modelo.

Com $y_1, ..., y_n$ respostas e para cada i, x_i um vetor de covariáveis, assumiu-se y_i independente e segue uma $P(\mu_i, \phi)$, onde $\mu_i = E(y_i)$ e ϕ é um vector parâmetro comum a todos os y_i . O μ_i dependem do x_i do vetor de parâmetros de regressão β. Tem-se particularmente um MLG em que a densidade de probabilidade é

$$f(y; \theta_i, \phi) = a(y, \phi)exp[\{y\theta_i - k(\theta_i)\}/\phi]$$

onde $a()$ e $k()$ são funções e $\mu_i = k'(\theta_i)$ conhecido. Neste modelo tem-se $Var(y_i) = \phi V(\mu_i)$ em que $V(\mu_i) = k''(\theta_i)$. É habitual assumir que $g(\mu_i) = x^T\beta$ onde $g()$ é uma função de ligação conhecida. O parâmetro ϕ é a constante de proporcionalidade na relação entre a média e a variância e é conhecido como o parâmetro de dispersão.

No contexto de MLG, duas definições de resíduos tem sido vulgarmente utilizados na prática. O Resíduo de Pearson é definido por

$$r_{p,i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)^{1/2}}$$

onde $\hat{\mu}_i$ é o valor enquadrados para μ_i . Este resíduo tem a vantagem de que a média e variância são exatamente zero e ϕ, respectivamente, se a variabilidade de amostragem em $\hat{\mu}_i$ é pequena.

Envelope

Atkinson (1981) propõe a construção por simulação de Monte Carlo de uma banda de confiança para os resíduos da regressão normal linear, a qual denominou envelope, e que permite uma melhor comparação entre os resíduos e os percentis da distribuição normal padrão. Os envelopes, no caso de MLGs com distribuições diferentes da normal, são construídos com os resíduos sendo gerados a partir do modelo ajustado.

Uma banda assintótica de confiança de coeficiente $1 - \alpha$ pode ser também construída para $\mu(z) = g^{-1}(z^T\beta)$, $\forall z \in \mathbb{R}^p$ generalizando os resultados da seção anterior. Assintoticamente tem-se que $\hat{\beta} - \beta \sim N_p(0, \phi - 1(X^T W X)^{-1})$. Logo, uma banda assintótica de confiança de coeficiente $1 - \alpha$ para o preditor linear $z^T\beta$, $\forall z \in \mathbb{R}^p$ fica dada por

$$g^{-1}[z^T\hat{\beta} \pm \sqrt{\phi^{-1}c_\alpha} z^T (X^T W X)^{-1} z^{1/2}] \quad \forall z \in \mathbb{R}^p.$$

Nota-se que z é um vetor $p \times 1$ que varia livremente no $\mathbb{R}^p$, enquanto X é uma matriz fixa com os valores das variáveis explicativas. As quantidades W e ϕ devem ser estimadas consistentemente.

3.4 Uso do R

3.4.1 Funções

A função *GLM* é usada para ajustar modelos lineares generalizados, especificadas, dando uma descrição simbólica do preditor linear e uma descrição da distribuição dos erros. O *software* estatístico R *stats* realiza a modelagem dos dados. Como a variável resposta y (Registro de Nascimento) é dicotômica, utilizou-se o modelo linear generalizado para dados binários e as variáveis independentes foram transformadas em fator e, assim, foram utilizadas no modelo.

```
reg = factor(REGISTRO, labels = c("Não Tem", "Tem"))
```

A fórmula é especificada para mlg como $y \sim x1 + \ldots + xn$ em que $x1, \ldots, xn$, são os nomes dos fatores. Tem-se como modelo inicial o que abrange todas as variáveis do banco

de dados.

```
maior=glm(y~cor+domicilio+sexo+ler+esgoto+agua+energia+renda+bolsa+
banheiro+computador+moto+carro+radio+tv+maquina+geladeira+celular,
family=binomial(link=logit),na.action=na.exclude)
```

O argumento *family* toma uma função que especifica a distribuição do erro e função de ligação a ser utilizada no modelo. Para a distribuição binomial, foi utilizada uma função da família *logit* e *probit*.

A seleção do melhor modelo foi feita através do *stepwise*, que seleciona um modelo baseado na fórmula pelo AIC. A seleção ou exclusão de variáveis de um modelo é baseado em um algoritmo que checa a importância das variáveis, incluindo ou excluindo-as do modelo se baseando em uma regra de decisão. Na regressão logística os erros seguem distribuição binomial e a significância é assegurada via Teste da Razão de Verossimilhança. Assim, em cada passo do procedimento a variável mais importante, em termos estatísticos, é aquela que produz a maior mudança no logaritmo da verossimilhança em relação ao modelo que não contém a variável.

```
fit=step(maior) # Melhor modelo com menor AIC

summary(fit) # Estimação ds parâmetros, Significância das Variáveis

summary(fit)$aic # AIC
```

Analisando a bondade do modelo, se $D(y;\mu)/\phi \leq \chi^2_{n-p;1-\alpha}$, o modelo em investigação é aceito.

```
phi<-summary(fit)$dispersion
D.phi=summary(fit)$deviance/phi
X2=qchisq(0.95,summary(fit)$df.residual)
```

Dessa forma, se `D.phi` $\leq$ `X2` o modelo em investigação é aceito.

Para analisar os resíduos normalizados, utilizou-se o pacote *statmod* para modelos lineares generalizados com a ideia de inverter a função de distribuição estimada para cada observação para obter resíduos normais exatamente padrão.

```
residuoNormal=qres.binom(fit)  # Padronização dos resíduos

qqPlot(residuoNormal) # Envelope
```

No caso de distribuições discretas, tais como a binomial e Poisson, são os resíduos de escolha para modelos lineares generalizados em grandes situações de quando os resíduos de Pearson podem ser grosseiramente não-normais. Os resíduos normalizados são os únicos úteis para dados binomiais ou de Poisson quando a resposta assume um pequeno número de valores distintos.

CAPÍTULO 4

RESULTADOS

4.1 Descrição do comportamento das variáveis

Nessa seção são apresentadas as estatísticas descritivas das variáveis utilizadas para a modelagem dos registros de nascimento no Semiárido Brasileiro. Os gráficos da Figura 4.1 apresentam as proporções do perfil sociodemográfico e econômico de pessoas até 10 anos de idade tendo ou não registro civil de nascimento, segundo o espaço geográfico do Semiárido dos Estados no ano 2010.

Os três primeiros gráficos são das variáveis registro de nascimento, rendimento do BF e PETI e Idade. Essas figuras estão destacadas por serem as variáveis que mais nortearam o estudo. 99% das pessoas tem registro de nascimento e assim, buscou-se uma possível justificativa para o sub-registro (1%).

O rendimento do BF e PETI foi responsável por reduzir o Banco 1 em 10% das observações do Banco original, pois abrange apenas pessoas de 10 anos de idade. Enquanto que o Banco 2 considera pessoas com idade de até 10 anos. As proporções das idades atingiram um patamar em média de 20%. O comportamento dos dados relacionados a BF é discrepante. Em média, 95% dos entrevistados não tem esses benefícios, não sendo diferenciado os com ou sem registro.

Os demais gráficos mostram a comparação das proporções entre o Banco 1 (em vermelho e cinza) e o Banco 2 (preto e branco). Em geral, pode-se observar que o padrão das barras para cada categoria das variáveis é o mesmo, diferindo nas magnitudes.

As seguintes categorias das variáveis apresentaram uma proporção aproximada:

- Acima de 90%: Existência de energia elétrica (Sim), existência de microcomputador (Não), existência de TV (Sim) e existência de máquina de lavar (Não).
- Entre 75% e 90%: Cor (Não Branca), rendimento domiciliar per capita (<0,5), número de banheiros (Um), existência de moto (Não), existência de carro (Não),

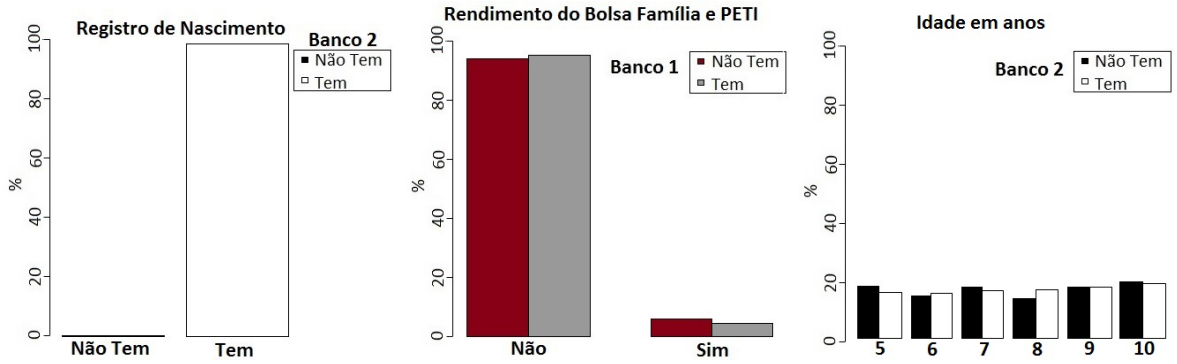
existência de rádio (Sim)

- Entre 50% e 75%: Situação do domicílio (Urbana), sexo (Masculino), existência de geladeira (Sim), existência de celular (Sim), tipo de esgotamento (Fossa séptica).
- Entre 25% e 50%: Cor (Branca), situação do domicílio (Rural), sexo (Feminino), existência de moto (Sim), existência de rádio (Não), existência de celular (Não).
- Menor que 25%: Saber ler/escrever (Não), existência de energia elétrica (Não), rendimento domiciliar per capita (0,5|-1,5; >1,5), número de banheiros (Nenhum, Dois ou mais),existência de microcomputador (Sim), existência de carro (Sim), existência de TV (Não), existência de máquina de lavar (Sim), existência de geladeira (Não), tipo de esgotamento (Rede geral de esgoto ou pluvial).

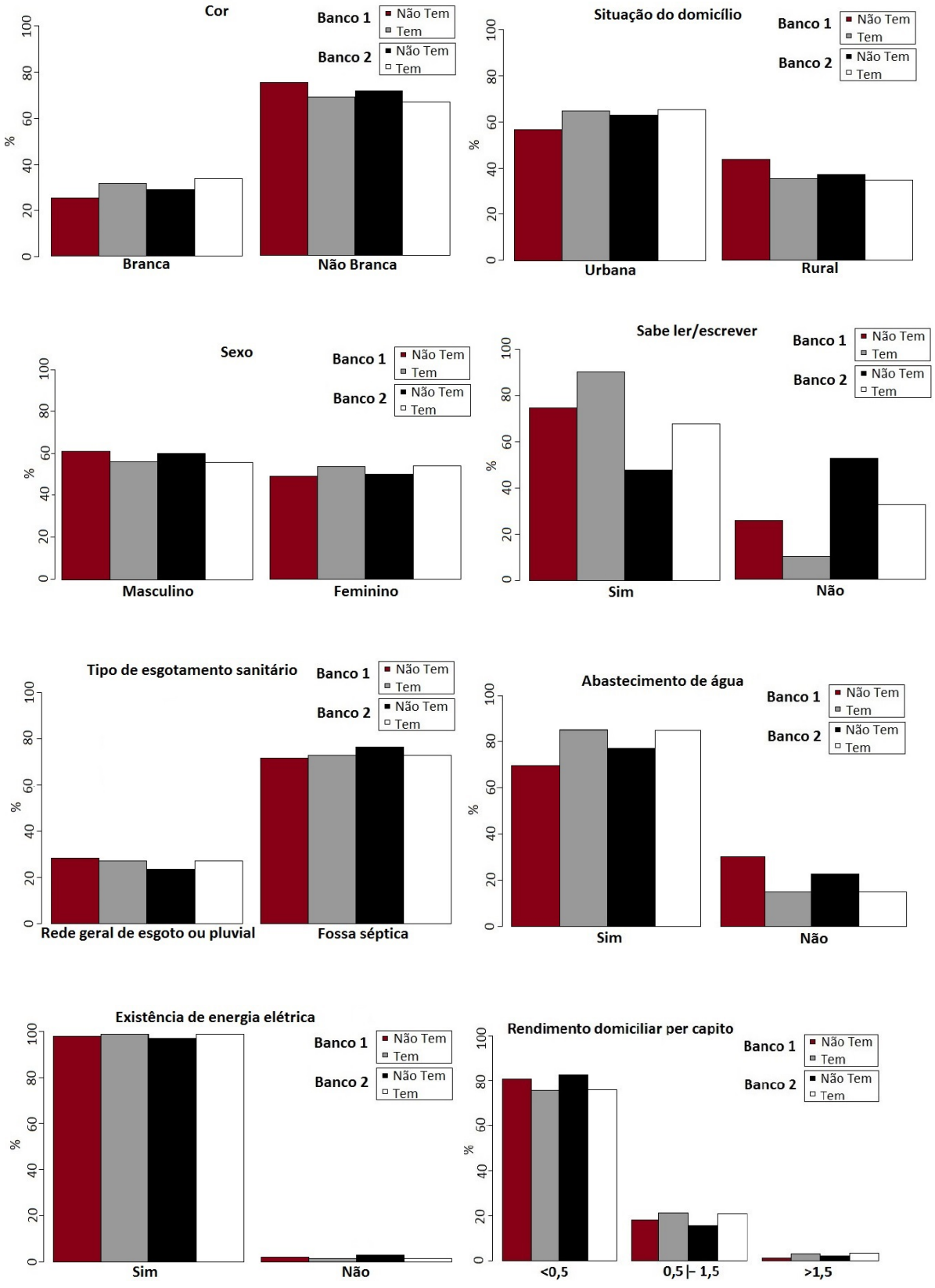
Analisando particularmente as pessoas registradas nota-se que se destaca um perfil com maiores porcentagens nos itens relacionados a riqueza, como: moto, carro, TV, computador, mais de um banheiro (aproximadamente 84%), mais de meio salário, máquina de lavar. Já os sem registros destacam-se nas categorias que indicam maior pobreza: não ter automotores (aproximadamente 18%), não tem banheiro na residência (quase 25%) e não tem abastecimento de água (em média 30%).

Figura 4.1: Proporção do perfil sociodemográfico e econômico de pessoas até 10 anos de idade tendo ou não registro civil de nascimento, segundo o espaço geográfico do Semiárido dos Estados, 2010.

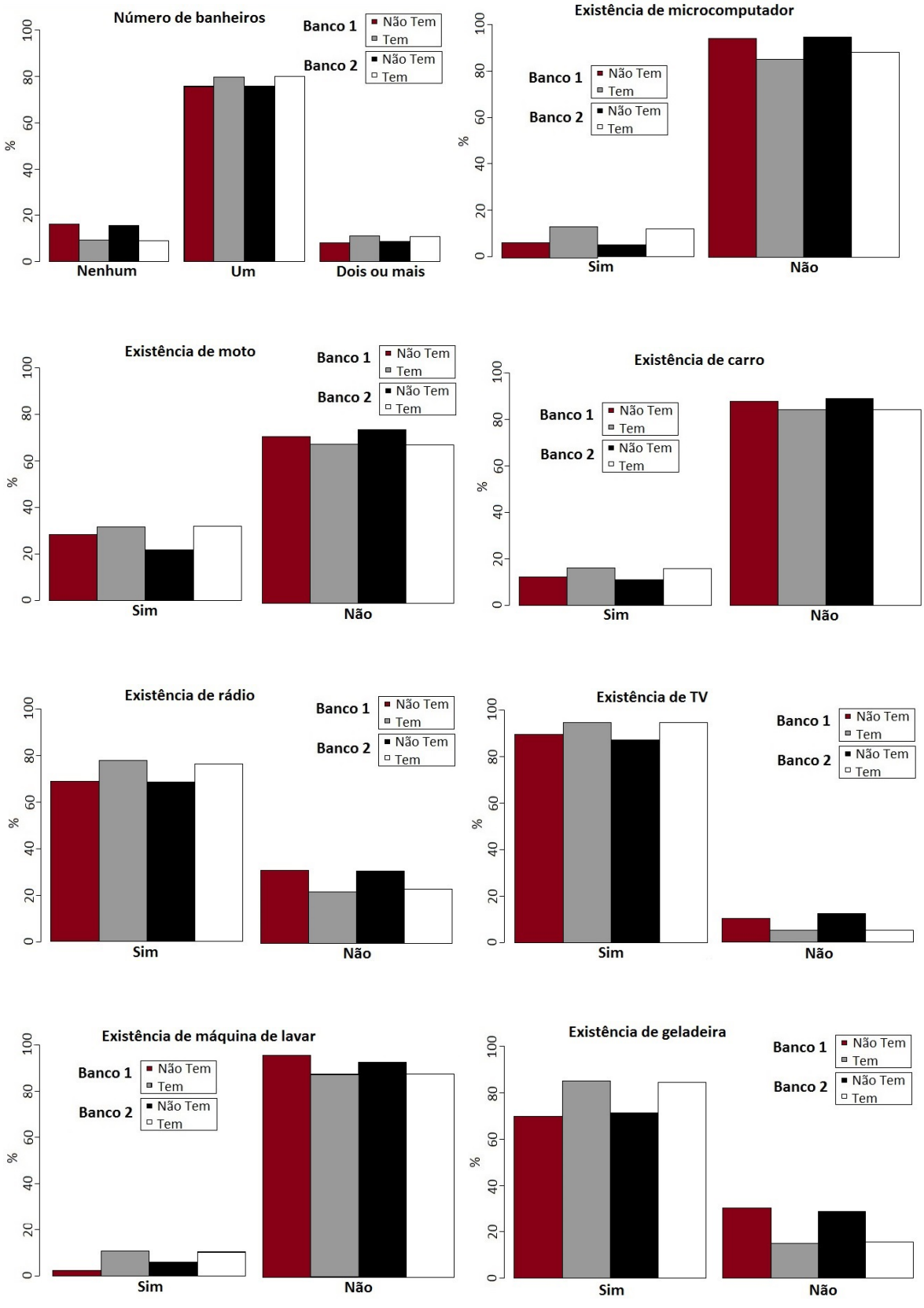
(Continua)



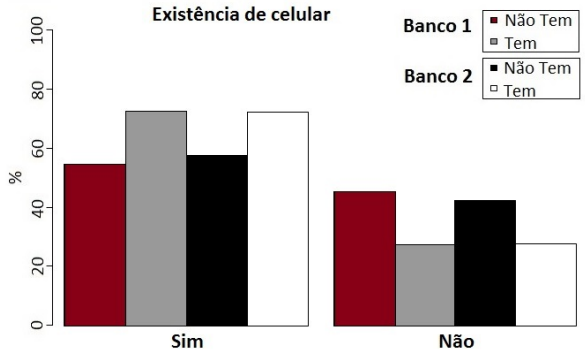
(Continua)


Cor
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Branca
Não Branca
Situação do domicílio
Urbana
Rural
Sexo
Masculino
Feminino
Sabe ler/escrever
Sim
Não
Tipo de esgotamento sanitário
Rede geral de esgoto ou pluvial
Fossa séptica
Abastecimento de água
Existência de energia elétrica
Rendimento domiciliar per capito
<0,5
0,5 |– 1,5
>1,5

(Continua)


Número de banheiros
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Nenhum
Um
Dois ou mais
Existência de microcomputador
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não
Existência de moto
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não
Existência de carro
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não
Existência de rádio
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não
Existência de TV
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não
Existência de máquina de lavar
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não
Existência de geladeira
Banco 1
Não Tem
Tem
Banco 2
Não Tem
Tem
%
0
20
40
60
80
100
Sim
Não

(Conclusão)



4.2 Banco 1

O Banco 1 considera todas as variáveis do banco de dados referente a idade específica de 10 anos. As variáveis são: y (Registro de Nascimento), cor, domicílio, sexo, ler, esgoto, água, energia, renda, bolsa, banheiro, computador, moto, carro, rádio, tv, máquina, geladeira, celular.

4.2.1 Seleção da função de ligação

Contemplando todas as variáveis do Banco 1, o *stepwise* mostrou 13 possíveis modelos para que o registro de nascimento fosse bem ajustado. Na Tabela 4.1 estão presentes os valores dos AIC's, desvios e estatísticas do qui-quadrado para definir o melhor modelo.

As duas funções de ligação, *logit* e *probit*, foram consideradas na análise. Pela comparação do AIC, o melhor modelo para o estudo é o primeiro, ajustado pela função de ligação *probit*, o qual se mostrou dessa forma até o 12º modelo.

A comparação do desvio com a estatística, permite dizer que os 13 modelos ajustados com cada função de ligação são representativos. Para a confirmação desta especulação, analisou-se os resíduos de cada um dos modelos em análise.

Tabela 4.1: Comparação entre as funções de ligação *Logit* e *Probit*, segundo AIC e análise dos desvios.

Modelo	*Logit*			*Probit*		
	AIC	$D(y;\mu)/\phi$	$\chi^2_{n-p;1-\alpha}*$	AIC	$D(y;\mu)/\phi$	$\chi^2_{n-p;1-\alpha}*$
1	1409,01	1395,01	55311,51	1408,59	1394,59	55311,51
2	1409,34	1393,34	55310,50	1409,15	1393,15	55310,50
3	1410,82	1392,82	55309,50	1410,59	1392,59	55309,50
4	1412,21	1392,21	55308,49	1411,90	1391,90	55308,49
5	1413,64	1391,64	55307,49	1413,34	1391,34	55307,49
6	1415,07	1391,07	55306,48	1414,80	1390,80	55306,48
7	1416,62	1390,62	55305,48	1416,43	1390,43	55305,48
8	1418,35	1390,35	55304,47	1418,22	1390,22	55304,47
9	1420,13	1390,13	55303,47	1420,07	1390,07	55303,47
10	1421,95	1389,95	55302,46	1421,91	1389,91	55302,46
11	1423,93	1389,93	55301,46	1423,88	1389,88	55301,46
12	1427,16	1389,16	55299,45	1427,12	1389,12	55299,45
13	1430,45	1388,45	55297,44	1430,49	1388,49	55297,44

Fonte dos dados básicos: Microdados - IBGE, 2010.
$\alpha = 0,05$

4.2.2 Estudo dos Resíduos

Para realizar análise dos resíduos tomou-se como ferramenta a dispersão da relação entre as observações e os resíduos (Figuras 4.2 e 4.3), o envelopamento dos resíduos padronizados (Figuras 4.4 e 4.5) que se mostram "perfeitos"pelo número de observações (Teorema Central do Limite), a autocorrelação (Figuras 4.6 e 4.7) e a autocorrelação parcial (Figuras 4.8 e 4.9). De acordo com os resultados, tanto do modelo inicial como final, utilizando as funções de ligação *logit* e *probit*, foram observados comportamentos similares, com exceção da autocorrelação parcial.

Os resíduos estão dispersos e quando envelopados, apresentam-se dentro do intervalo de confiança plotado. A autocorrelação permite afirmar que os resíduos são independentes e a autocorrelação parcial indica que o melhor modelo é o final da função de ligação *probit*, já que engloba uma maior proporção de defasagem dentro do intervalo de confiança de 95% evidenciando, assim, a independência entre os resíduos.

Na tentativa de encontrar uma relação adequada entre o registro de nascimento e as covariáveis citadas, vários modelos foram testados. O modelo escolhido é apresentado na seção seguinte.

Figura 4.2: Comparação dos resíduos padronizados dos modelos iniciais das funções de ligação *Logit* e *Probit*.

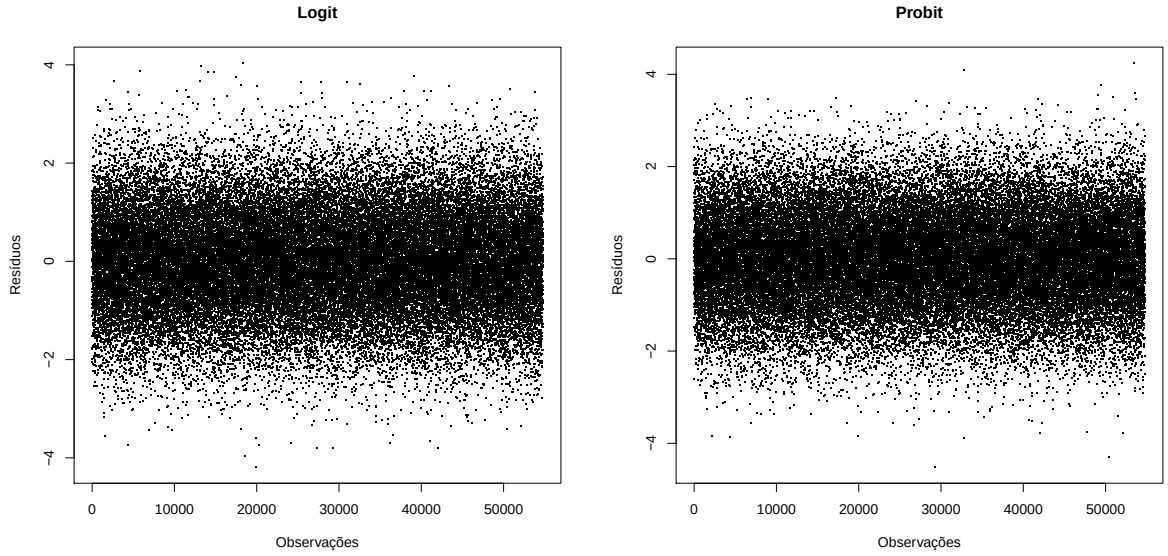


Figura 4.3: Comparação dos resíduos padronizados dos modelos finais das funções de ligação *Logit* e *Probit*.

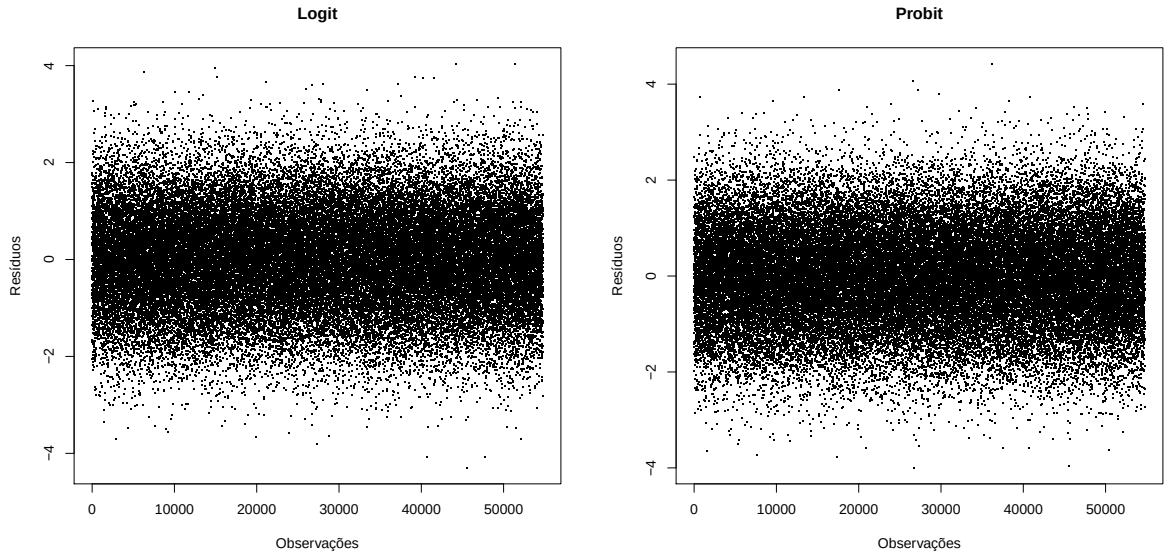


Figura 4.4: Envelope do modelo inicial das funções de ligação *Logit* e *Probit*.

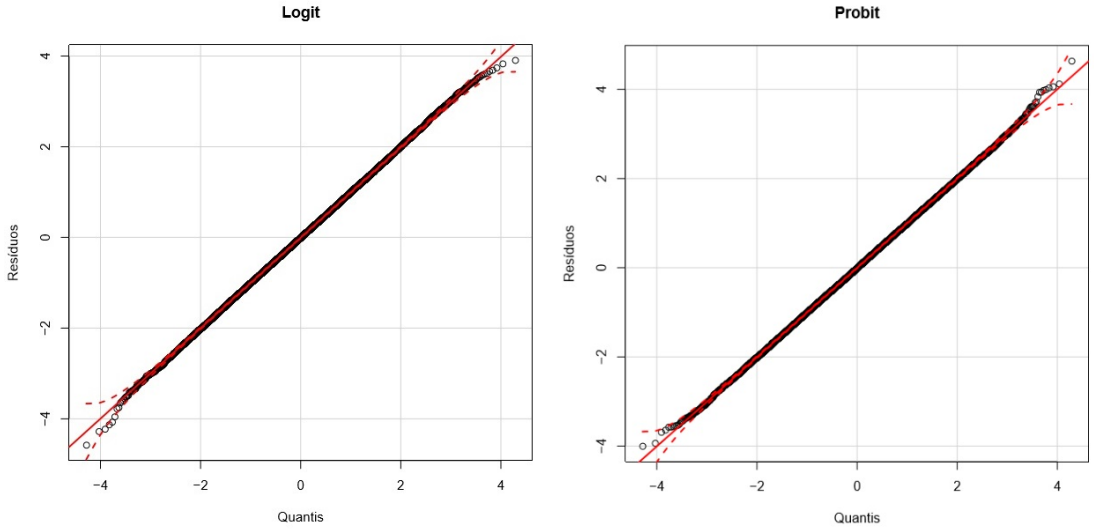


Figura 4.5: Envelope do modelo final das funções de ligação *Logit* e *Probit*.

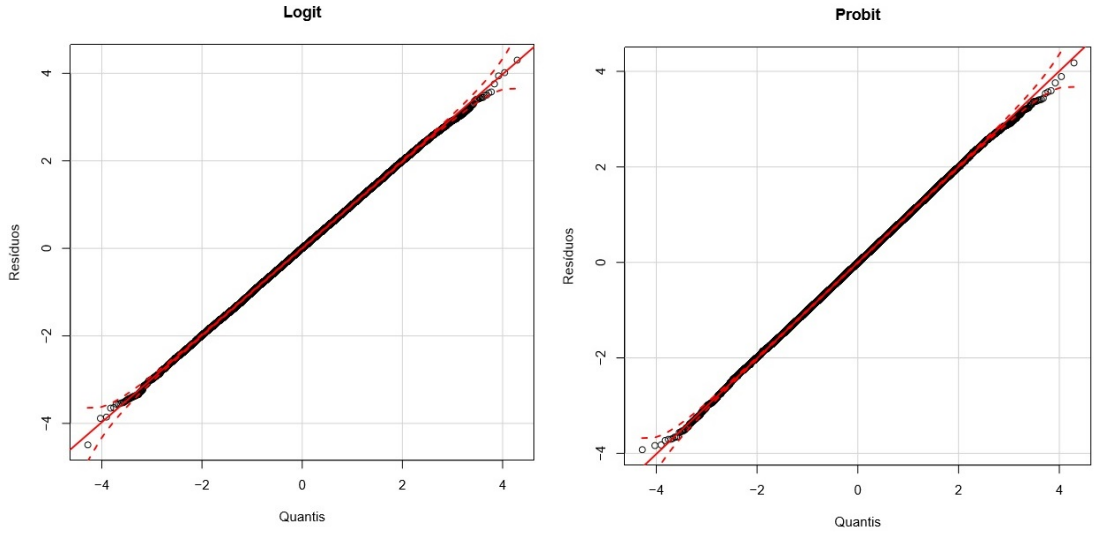


Figura 4.6: Autocorrelação do modelo inicial das funções de ligação *Logit* e *Probit*.

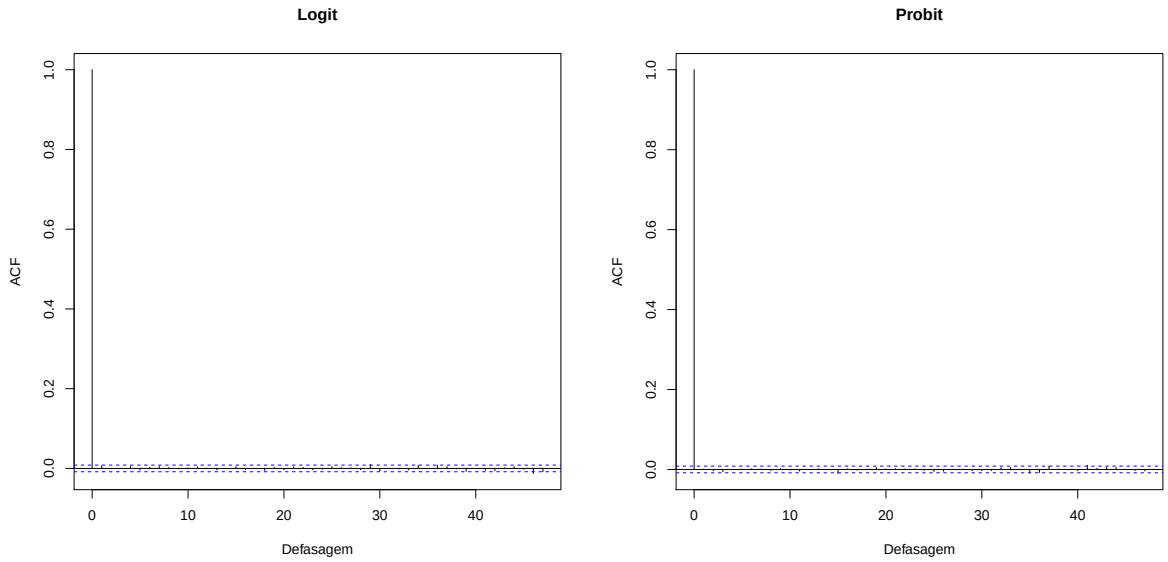


Figura 4.7: Autocorrelação dos modelos finais das funções de ligação *Logit* e *Probit*.

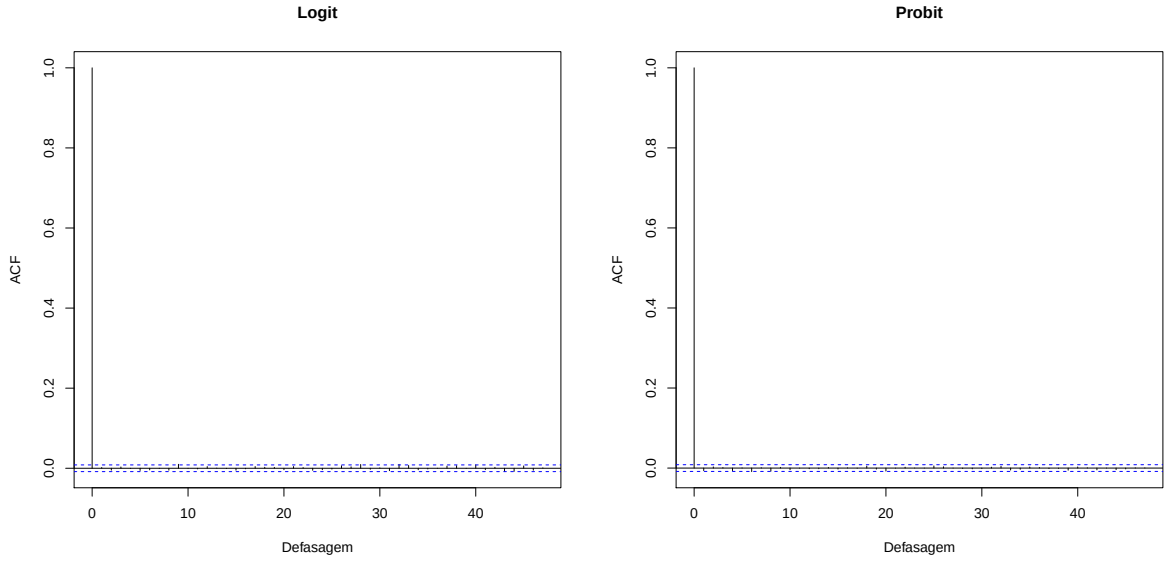


Figura 4.8: Autocorrelação Parcial dos modelos iniciais das funções de ligação *Logit* e *Probit*.

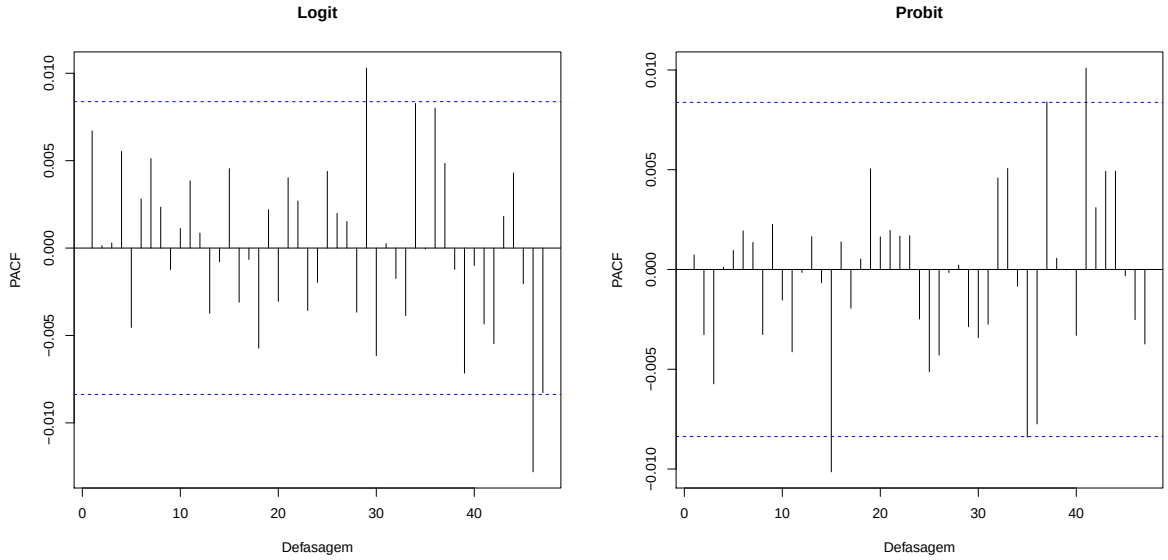
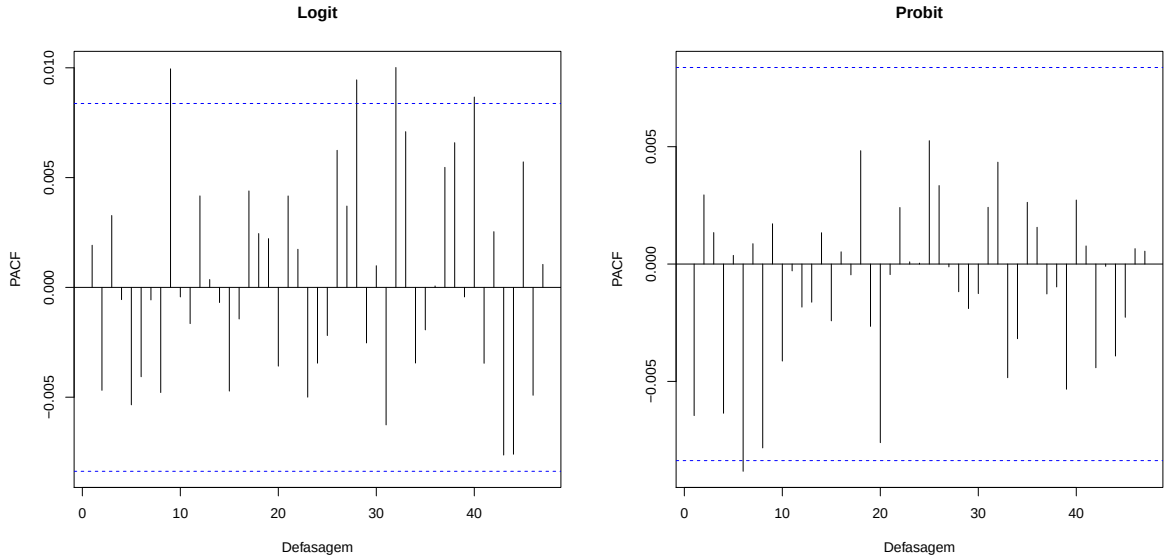


Figura 4.9: Autocorrelação Parcial dos modelos finais das funções de ligação *Logit* e *Probit*.



4.2.3 Melhor Modelo

Na Tabela 4.2 são mostrados os coeficientes, a significância de cada variável no modelo inicial e suas razões de chance. Já a Tabela 4.3 apresenta as variáveis significativas (P-valor<0,05) para a modelagem do registro de nascimento foram: Ler, Esgoto, Água, Máquina, Geladeira e Celular.

Analisando a influência que cada variável (Tabela 4.3) pode ter sobre o registro de nascimento, observou-se que ter fossa séptica pode aumentar aproximadamente 7 vezes a chance de ocorrer o registro. Enquanto que não ter celular ou geladeira diminui em média 6 vezes essa chance. Não saber ler (variável mais significativa) tem 3 vezes menos chance de ser registrado. Já para os que não possuem máquina de lavar a chance de não ter declarado o nascimento é de 1,94.

Tabela 4.2: Estimativa dos coeficientes, P-valor e Razão de Chance das variáveis do modelo inicial *Probit*

(Continua)

	Coeficientes	P-valor	Razão de Chance (IC)
Intercepto	3,4035	0,0000 ***	
Cor			
Branca			1
Não Branca	-0,0548	0,4577	19,00 (0,82; 98,00)
Domicílio			
Urbana			1
Rural	-0,0183	0,8236	49,00 (0,84; 86,00)
Sexo			
Masculino			1
Feminino	0,0299	0,6482	34,33 (0,91; 79,00)
Ler			
Sim			1
Não	-0,2809	0,0006 ***	3,17 (1,85; 8,09)
Esgoto			
Rede geral de esgoto ou pluvial			1
Fossa séptica	0,1644	0,0473 *	6,56 (1,01; 83,00)
Água			
Sim			1
Não	-0,2464	0,0040 **	3,63 (1,15; 69,00)
Energia			
Sim			1
Não	0,2839	0,2923	4,03(0,82; 54;00)
Renda			
<0,5			1
0,5/- 1,5	-0,06058	0,5019	15,67 (1,13; 78,00)
>1,5	0,0921	0,7681	11,00 (0,99; 57,00)
Bolsa			
Sim			1
Não	-0,1093	0,4316	9,00 (0,70; 55,00)

(Conclusão)

	Coeficientes	P-valor	Razão de Chance (IC)
Banheiro			
Nenhum			1
Um	0,0550	0,5705	6,69 (0,87;17,68)
Dois ou mais	-0,0333	0,8322	4,03 (0,71; 24,00)
Computador			
Sim			1
Não	-0,0823	0,5756	5,76 (0,68; 11,50)
Moto			
Sim			1
Não	0,0267	0,7205	6,24 (0,88; 34,33)
Carro			
Sim			1
Não	0,0567	0,5972	1,06 (0,85;49,00)
Rádio			
Sim			1
Não	-0,0875	0,2269	11,50 (0,80; 17,66)
TV			
Sim			1
Não	-0,1106	0,3990	2,33 (0,70; 9,00)
Máquina			
Sim			1
Não	-0,4119	0,0467 *	1,95 (1,02; 15,67)
Geladeira			
Sim			1
Não	-0,1482	0,0713 .	6,15 (2,84; 99,00)
Celular			
Sim			1
Não	-0,1496	0,0378 *	6,14 (3,00; 99,00)

Significância: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Tabela 4.3: Estimativa dos coeficientes, P-valor e Razão de Chance das variáveis do modelo final *Probit*

	Coeficientes	P-valor	Razão de Chance (IC)
Intercepto	3,3768	0,0000 ***	
Ler			
Sim			1
Não	-0,2889	0,0003 ***	3,00 (1,78; 4,22)
Esgoto			
Rede geral de esgoto ou pluvial			1
Fossa séptica	0,1555	0,0402 *	6,88 (3,85; 10,10)
Água			
Sim			1
Não	-0,2474	0,0013 **	3,55 (2,03; 10,11)
Máquina			
Sim			1
Não	-0,4148	0,0404 *	1,94 (1,01; 13,28)
Geladeira			
Sim			1
Não	-0,1517	0,0489 *	6,14 (2,84; 99,00)
Celular			
Sim			1
Não	-0,1583	0,0223 *	5,67 (3,00; 49,00)

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

4.3 Banco 2

O Banco 2 foi composto pelas variáveis: y (Registro de Nascimento), cor, domicilio, sexo, ler, esgoto , água, energia, renda, idade, banheiro, computador, moto, carro, rádio, TV, máquina, geladeira, celular.

A variável bolsa foi desconsiderada por restringir o banco à 10% dos dados originais com a finalidade de realizar a comparação entre as duas análises dos Banco 1 e 2.

4.3.1 Seleção da função de ligação

Considerando todas as variáveis do Banco 2, o *stepwise* mostrou 7 possíveis modelos para que o registro de nascimento fosse ajustado. Na Tabela 4.4 estão presentes os valores dos AIC's, desvios e estatísticas qui-quadrado para definir o melhor modelo.

As duas funções de ligação, *logit* e *probit*, foram consideradas na análise. Pela comparação do AIC, o melhor modelo para o estudo é o primeiro, ajustado pela função de ligação *logit*, ao qual se mostrou dessa forma até o 6º modelo.

A comparação do desvio com a estatística, permite dizer que os 7 modelos ajustados com cada função de ligação são representativos. Para a confirmação desta conclusão, analisou-se os resíduos de cada um dos modelos em análise.

Tabela 4.4: Comparação entre as funções de ligação *logit* e *probit*, segundo AIC e análise dos desvios.

Modelo	*Logit*			*Probit*		
	AIC	$D(y;\mu)/\phi$	$\chi^2_{n-p;1-\alpha}{}^*$	AIC	$D(y;\mu)/\phi$	$\chi^2_{n-p;1-\alpha}{}^*$
1	7366,96	7338,96	294645,11	7368,86	7342,86	294646,11
2	7367,62	7337,62	294644,10	7368,97	7340,97	294645,11
3	7368,79	7336,79	294643,10	7369,72	7339,72	294644,10
4	7370,09	7336,09	294642,10	7370,77	7338,77	294643,10
5	7372,09	7336,09	294641,10	7372,17	7338,17	294642,10
6	7374,09	7336,09	294640,10	7374,14	7338,14	294641,10
7	7377,46	7335,46	294638,09	7376,14	7338,14	294640,10

Fonte dos dados básicos: Microdados - IBGE, 2010.
$\alpha = 0.05$

4.3.2 Estudo dos Resíduos

Para realizar a análise dos resíduos tomou-se como ferramenta a dispersão da relação entre as observações e os resíduos (Figuras 4.10 e 4.11), o envelopamento dos resíduos padronizados (Figuras 4.12 e 4.13) que se mostram "perfeitos"pelo número de observações (Teorema Central do Limite), a autocorrelação (Figuras 4.14 e 4.15) e a autocorrelação parcial (Figuras 4.16 e 4.17). De acordo com os resultados, tanto do modelo inicial como final, utilizando as funções de ligação *logit* e *logit*, foram observados comportamentos similares, com exceção da auto correlação parcial.

Os resíduos estão dispersos e quando envelopados, apresentam-se dentro do intervalo de confiança plotado. A autocorrelação permite afirmar que os resíduos são independentes e a autocorrelação parcial indica que o melhor modelo é o final da função de ligação *logit*, pois este é o que proporcionou uma maior proporção de defasagem dentro do intervalo de confiança de 95%, o que evidencia a independência entre os resíduos. Na tentativa de encontrar uma relação adequada entre registro de nascimento e as covariáveis citadas, vários modelos foram testados. O modelo escolhido é apresentado na seção seguinte.

Figura 4.10: Comparação dos resíduos padronizados do modelo inicial das funções de ligação *Logit* e *Probit*.

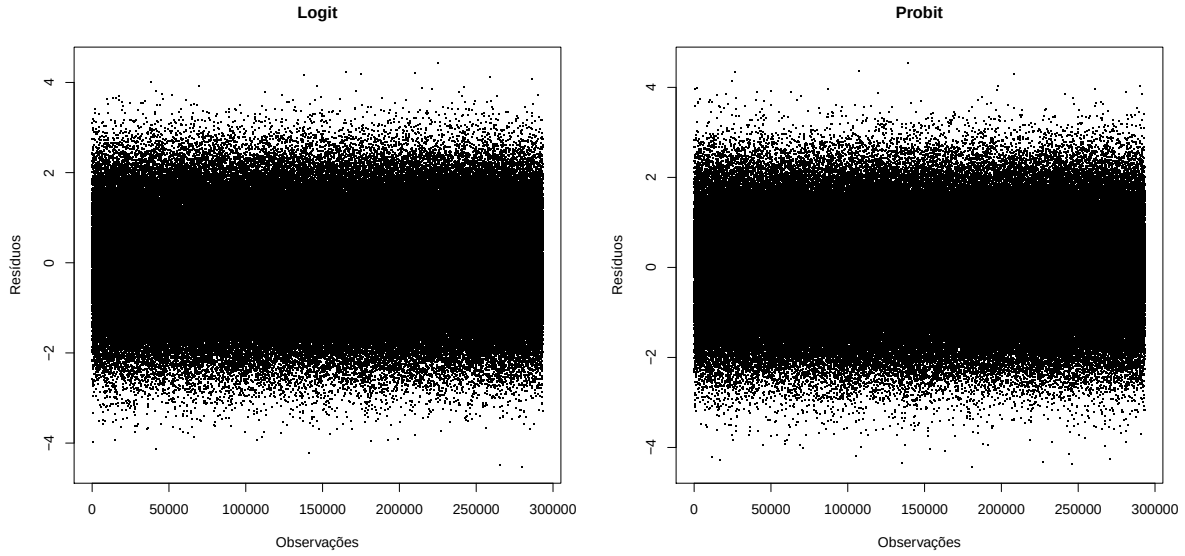


Figura 4.11: Comparação dos resíduos padronizados do modelo final das funções de ligação *Logit* e *Probit*.

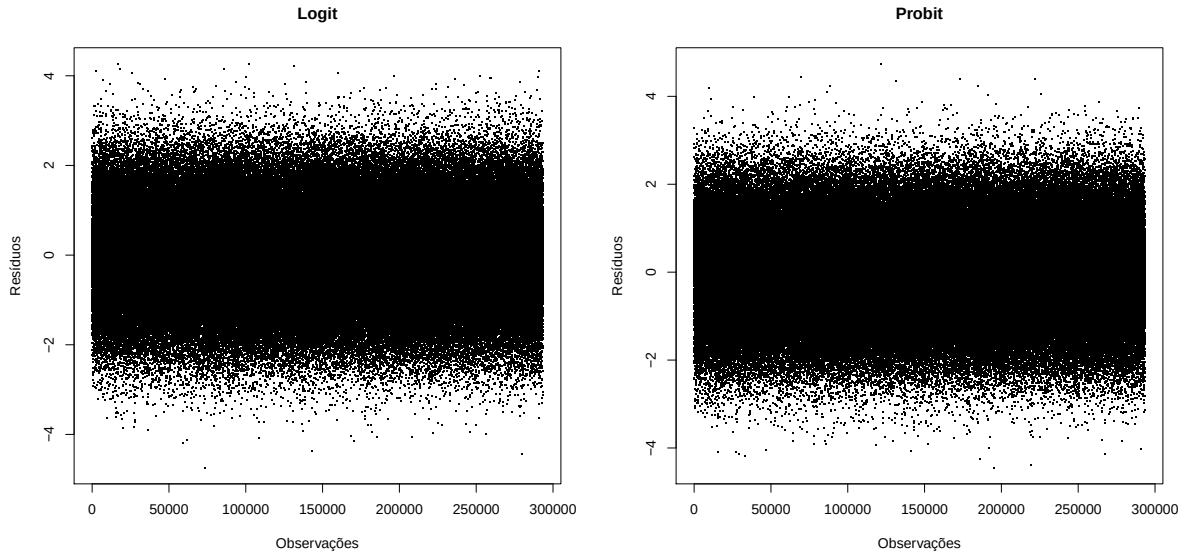


Figura 4.12: Envelope do modelo inicial das funções de ligação *Logit* e *Probit*.

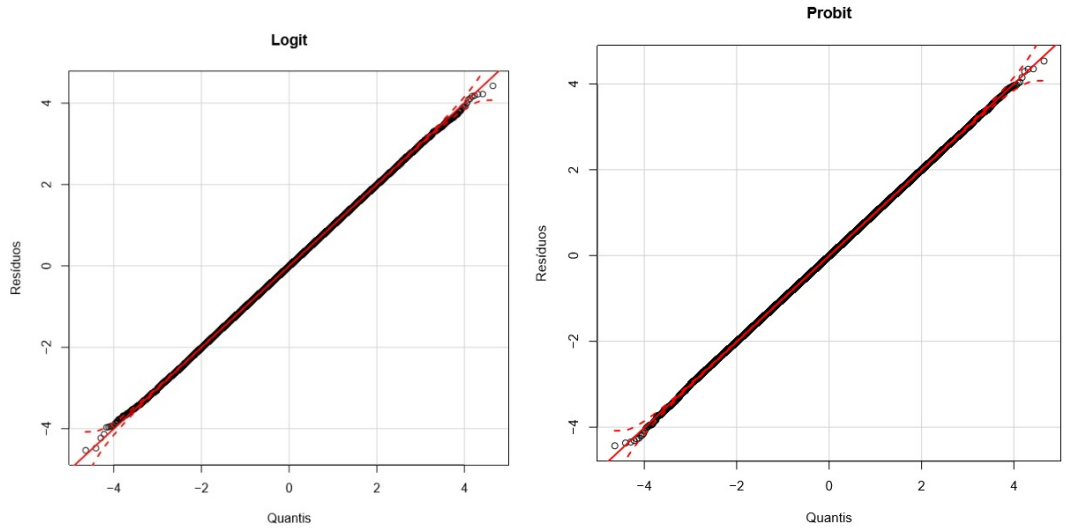


Figura 4.13: Envelope do modelo final das funções de ligação *Logit* e *Probit*.

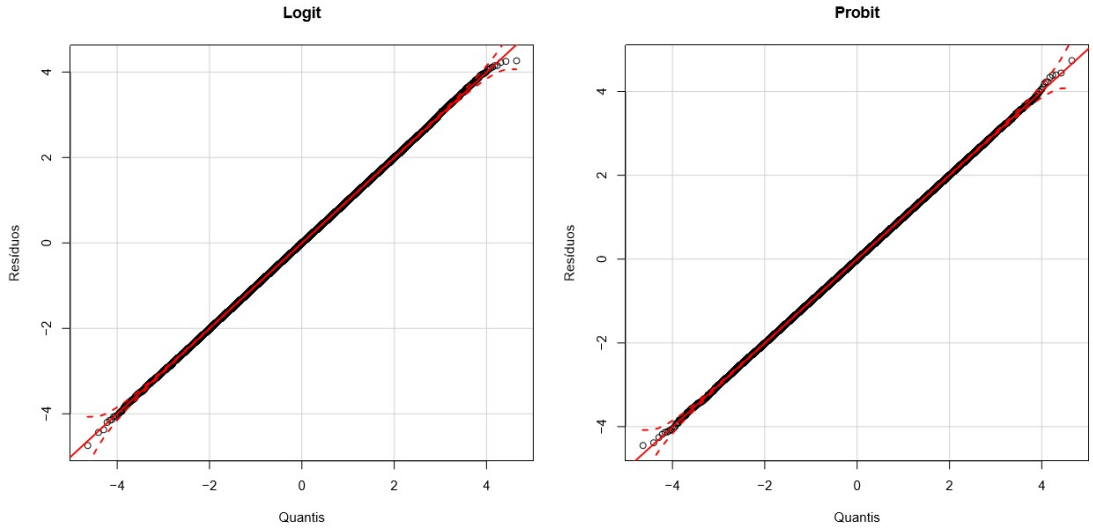


Figura 4.14: Autocorrelação do modelo inicial das funções de ligação *Logit* e *Probit*.

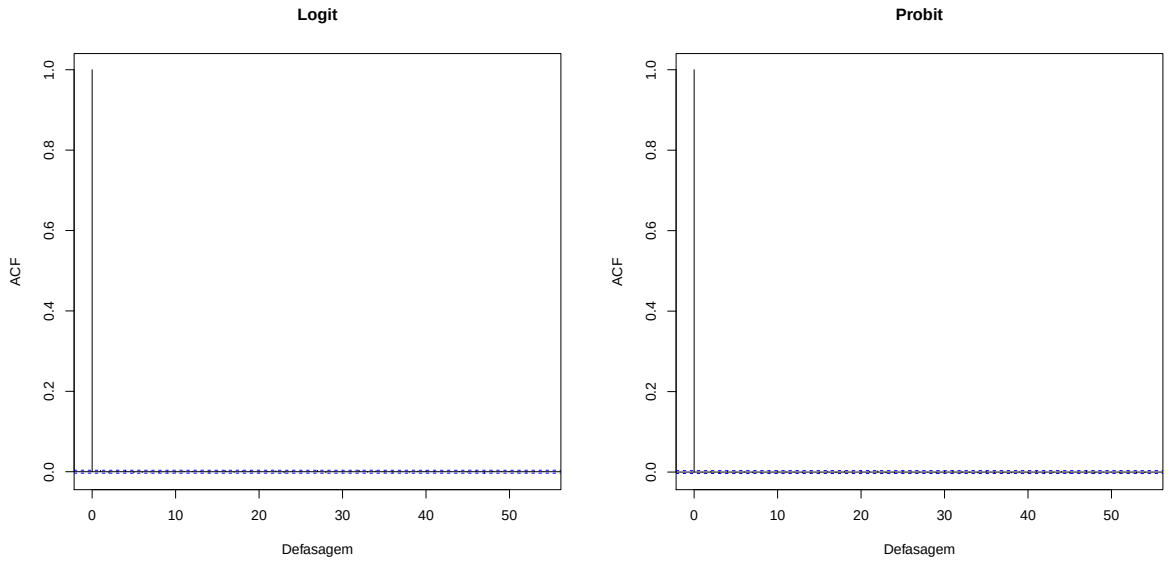


Figura 4.15: Autocorrelação do modelo final das funções de ligação *Logit* e *Probit*.

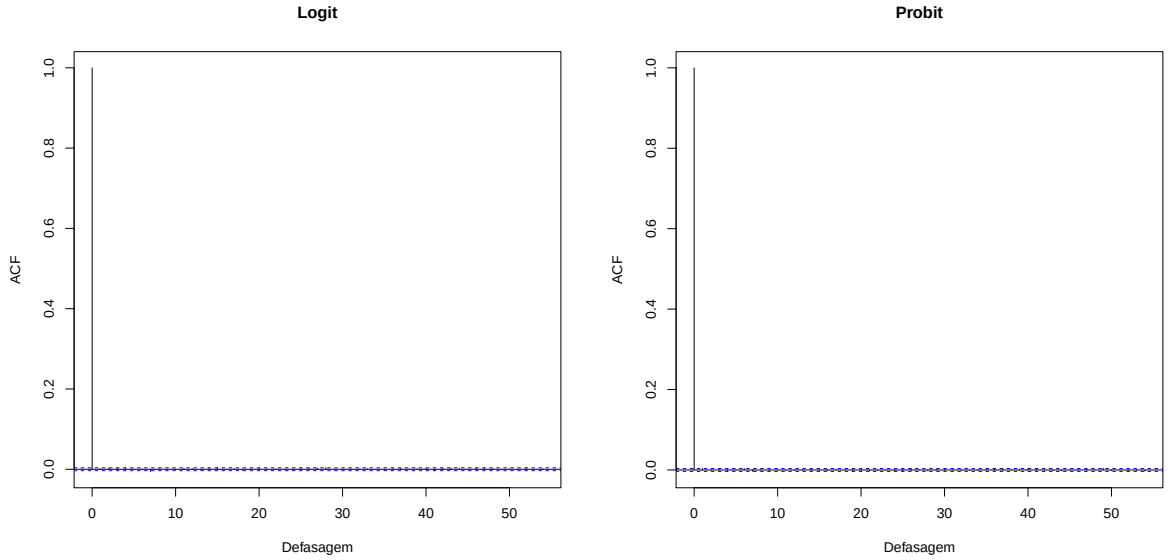


Figura 4.16: Autocorrelação Parcial do modelo inicial das funções de ligação *Logit* e *Probit*.

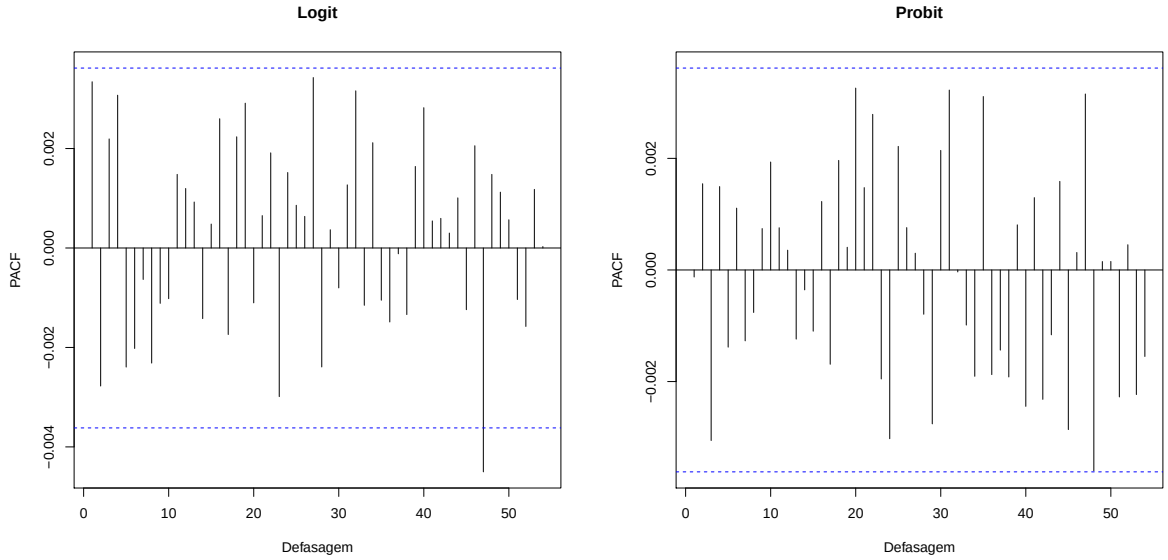
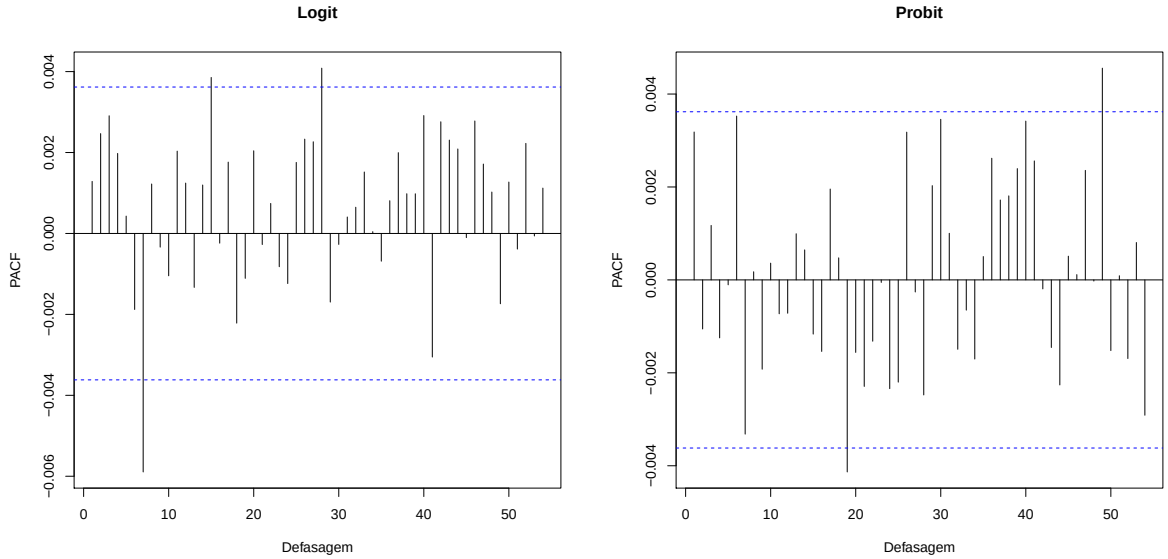


Figura 4.17: Autocorrelação Parcial do modelo final das funções de ligação *Logit* e *Probit*.



4.3.3 Melhor Modelo

Na Tabela 4.5 se encontram os coeficientes, a significância de cada variável no modelo inicial e suas razões de chance. As variáveis significativas (P-valor<0,05) para a modelagem do registro de nascimento foram apresentadas na Tabela 4.6. Estas são: Ler, Água, Idade, Computador, Moto, TV, Geladeira e Celular.

Estudando a influência que cada variável Tabela 4.6 pode ter sob o registro de nascimento, observou-se que não ter celular, geladeira, água encanada ou não saber ler, diminui em média 3 vezes a chance do registro do nascimento. As demais variáveis (computador, moto e TV) tiveram entre 1,50 e 2,90 menos razão de chance de registrar o nascimento, segundo a análise do P-valor.

Tabela 4.5: Estimativa dos coeficientes, P-valor e Razão de Chance das variáveis do modelo inicial *logit*

(Continua)

	Coeficientes	P-valor	Razão de Chance (IC)
Intercepto	8,5058	0,0000 ***	
Cor			
Branca			1
Não Branca	-0,0842	0,3932	11,50 (0,98; 47,00)
Domicílio			
Urbana			1
Rural	0,1888	0,0826 .	5,76 (1,06; 49,00)
Sexo			
Masculino			1
Feminino	0,0790	0,3756	13,50 (0,91; 38,65)
Ler			
Sim			1
Não	-0,8265	0,0000 ***	1,67 (1,56; 18,99)
Esgoto			
Rede geral de esgoto ou pluvial			1
Fossa séptica	0,0024	0,9830	1,00 (0,80; 5,00)
Água			
Sim			1
Não	-0,2963	0,0152 *	2,84 (1,44; 19,00)
Energia			
Sim			1
Não	0,3404	0,2552	1,41 (0,80; 57,89)
Renda			
<0,5			1
0,5|-1,5	-0,0555	0,6762	19,00 (0,73; 84,63)
>1,5	-0,2679	0,4426	3,17 (0,40; 67,28)
Idade	-0,0934	0,0010 ***	10,11 (6,15; 24,00)

(Conclusão)

	Coeficientes	P-valor	Razão de Chance (IC)
Banheiro			
Nenhum			1
Um	0,2375	0,0694 .	4,70 (3,25; 49,00)
Dois ou mais	-0,1405	0,5126	6,70 (0,57; 40,03)
Computador			
Sim			1
Não	-0,5226	0,0238 *	1,43 (0,50; 11,50)
Moto			
Sim			1
Não	-0,2996	0,0070 **	2,92 (1,43; 11,50)
Carro			
Sim			1
Não	-0,0113	0,9430	99,00 (0,72; 133;00)
Rádio			
Sim			1
Não	-0,1827	0,0618 .	5,18 (2,25; 58,98)
TV			
Sim			1
Não	-0,5998	0,0001 ***	1,22 (1,02; 3,17)
Máquina			
Sim			1
Não	-0,2953	0,1546	2,84 (0,96; 11,00)
Geladeira			
Sim			1
Não	-0,2967	0,0082 *	1,69 (1,50; 13,28)
Celular			
Sim			1
Não	-0,3292	0,0008 ***	2,57 (1,44; 6,70)

Significância: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Tabela 4.6: Estimativa dos coeficientes, P-valor e Razão de Chance das variáveis do modelo final *logit*

	Coeficientes	P-valor	Razão de Chance (IC)
Intercepto	8,4218	0,0000 ***	
Ler			
Sim			1
Não	-0,8361	0,0000 ***	3,00 (0,56; 6,13)
Água			
Sim			1
Não	-0,2829	0,0199 *	3,00 (1,50; 24,00)
Idade	-0,0954	0,0008 ***	10,00 (6,14; 24,00)
Computador			
Sim			1
Não	-0,4775	0,0278 *	1,63 (1,03; 13,29)
Moto			
Sim			1
Não	-0,3052	0,0058 **	2,84 (1,43; 10,11)
TV			
Sim			1
Não	-0,5184	0,0003 ***	1,50 (1,02; 4,00)
Geladeira			
Sim			1
Não	-0,2816	0,0115 *	3,00 (1,57; 15,70)
Celular			
Sim			1
Não	-0,3303	0,0007 ***	2,57 (1,44; 6,70)

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

4.4 Comparação de Resultados

Realizando a comparação dos resultados obtidos à partir do estudo dos dois bancos de dados, pode-se descrever os melhores modelos para cada um ajustados da seguinte forma:

Os modelos finais obtidos para os dois Bancos de dados foram:

- Modelo final do Banco 1:

$$y = 3.3768 - 0.2889x_1 + 0.1555x_2 - 0.2474x_3 - 0.4148x_4 - 0.1517x_5 - 0.1583x_6$$

onde,

y = Registro, x_1 = Ler (Não), x_2 = Esgoto (Fossa séptica) x_3 = Água (Não), x_4 = Máquina (Não), x_5 = Geladeira (Não), x_6 = Celular (Não).

- Modelo final do Banco 2:

$$y = 8.4218 - 0.8361x_1 - 0.2829x_2 - 0.0954x_3 - 0.4775x_4 - 0.3052x_5$$

$$-0.5184x_6 - 0.2816x_7 - 0.3303x_8$$

onde,

y = Registro, x_1 = Ler (Não), x_2 = Água (Não), x_3 = Idade, x_4 = Computador (Não), x_5 = Moto (Não), x_6 = TV (Não), x_7 = Geladeira (Não), x_8 = Celular (Não).

Assim, é possível afirmar que com exceção das variáveis esgoto e máquina, todas as demais contempladas no Banco 1 também o foram no Banco 2, só que neste último, ainda acrescentadas seis covariáveis. Os diagnósticos da modelagem foram muito parecidos, com a diferença da escolha da melhor função de ligação (*probit* para o Banco 1 e *logit* para o Banco 2).

Como este trabalho teve como objetivo modelar a variável registro de nascimento de acordo com as variáveis socioeconômicas e demográficas, pode-se observar que para o Banco 1 os pontos mais relevantes podem ser resumidos em variáveis relacionadas a moradia (esgoto e água), nível de instrução do respondente (ler), e poder aquisitivo (maquina, geladeira, celular).

O mesmo cenário pode ser visto no Banco 2, pois as variáveis selecionadas, se interpretadas da mesma forma, podem ser agrupadas em: moradia (água), nível de instrução (ler), aspecto monetário (computador, moto, TV, geladeira, celular) e demográfico (idade).

CAPÍTULO 5

CONCLUSÕES

Estudos relacionados aos registros de nascimento são importantes ferramentas de monitoramento da qualidade de vida, principalmente para a região do Semiárido brasileiro, onde ainda existe a presença de sub-registro. Estes estudos são escassos e há uma necessidade de fortalecer o debate sobre a atenção governamental para as pessoas que moram nessa parte do País. Com isso, foi realizado, pela primeira vez, a coleta de dados referentes ao nascimento desses indivíduos para as pessoas com idade até 10 anos no Censo 2010.

Buscando descrever, através de modelos matemáticos, o relacionamento da variável registro de nascimento com variáveis que representam as condições de vida, a análise de regressão logística possibilitou encontrar associação significativa entre a variável dependente e as independentes, por meio de modelagem.

A priori foram considerados dois bancos de dados, Banco 1 (idade 10 anos), Banco 2 (desconsiderando a variável bolsa). A posteriori analisou-se as possíveis funções de ligação para cada modelo. No Banco 1 a melhor foi a *probit*, enquanto que no Banco 2 o melhor ajuste foi dado pela função *logit*.

Para definir os melhores modelos, em ambos os casos, foi realizada a análise do critério AIC e o estudo residual por meio de gráficos de dispersão, envelope, autocorrelação e autocorrelação parcial. Desses critérios, os que foram essenciais para o ajuste final foi o menor valor AIC e melhor plotagem da autocorrelação parcial, pois nas demais comparações os resultados foram similares de modelo para modelo.

Além de concluir que o registro de nascimento pode ser modelado a partir de variáveis relacionadas ao rendimento domiciliar, nível de instrução e situação da moraria, as pessoas que teriam menos condições favoráveis ao registro de nascimento são pessoas sem instrução, com uma baixa renda e sem esgotamento ou abastecimento de água adequados.

Estudos voltados para estas perspectivas que foram exploradas aqui são ausentes na literatura, apesar de necessários. Espera-se que este trabalho contribua com aportes para o entendimento do sub-registro de nascimentos no semiárido brasileiro. Espera-se, ainda,

poder contribuir para a definição de estratégias que melhorem o sistema de registros de nascimentos, que possa subsidiar o planejamento e fomentar políticas públicas nas áreas de saúde materna e infantil nos Estados do semiárido e suas respectivas microrregiões.

Como sugestão para trabalhos futuros, a análise do perfil das pessoas sem registros pode ser ajustada por métodos multivariados. Explorações neste sentido podem ser feitas através da análise de componentes principais, análise fatorial ou análise de agrupamento para identificar extratos, nos quais poderiam ser dadas prioridades nas ações de planejamento. Ainda como sugestão, a realização dos estudos ecológicos usando como unidade o município, para que possa auxiliar os gestores na tomada de decisões globais.

REFERÊNCIAS BIBLIOGRÁFICAS

ASA; **Articulação no Semiárido brasileiro**; Disponível em: `http://www.asabrasil.org.br`; Acessado em: 15/02/2016.

ATKINSON, A. C.; **Two graphical display for outlying and influential observations in regression**; Biometrika 68, 13-20, 1981.

BARROS, M. A. R.; NICOLAU, A. I. O.; **Fatores socioeconômicos da gestante associados ao peso do recém-nascido**; Rev enferm UFPE on line., Recife- PE, 2013.

BRASIL ESCOLA; **As Desigualdades Socioeconômicas no Brasil**; Disponível em: `http://educador.brasilescola.uol.com.br/estrategias-ensino/as-desigualdades-socioeconomicas-no-brasil.htm`; Acesso em: 10/01/2016.

BRASÍLIA, **Manual de Instruções para o preenchimento da Declaração de Nascido Vivo**; Ministério da Saúde – Brasília, 2011.; Disponível em: `http://www.uff.br/epidemiologia2/blog/wp-content/uploads/2012/10/Manual-de-DNV-4ed-2011.pdf`; Acessado em: 17/03/2016.

CABRAL; C. I. S.; **Aplicação do Modelo de Regressão Logística num Estudo de Mercado; Projeto Mestrado em Matemática Aplicada à Economia e à Gestão**; Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, 2013.

CALTRAM, G. A. F.; **O registro de nascimento como direito fundamental ao pleno exercício da cidadania**; UNIMEP – Universidade Metodista DE Piracicaba, Curso Mestrado – Direito Piracicaba/SP, 2010.

CASTRO, J. M.; RODRIGUES-JUNIOR, A. L.; *A influência da mortalidade por causas externas no desenvolvimento humano na Faixa de Fronteira brasileira*; Cad. Saúde Pública, Rio de Janeiro, 29(1), 195-200, Jan, 2012.

CORTEZ, J. W. et al.; **Atributos físicos do argissolo amarelo do semiárido nordestino sob sistemas de preparo**; Revista Brasileira de Ciência do Solo, v. 35, n. 4, p. 1207-1216, 2011.

CRESPO, C. D.; BASTOS, A.A; CAVALCANTI, W.A.; **A Pesquisa do Registro Civil: condicionantes do subregistro de nascimento e perspectivas de melhorias da cobertura**; In: XV Encontro Nacional de Estudos Populacionais, 2006, Caxambu - MG. Desafios e Oportunidades do crescimento zero, 2006.

CRESPO, C. D.; **Diferenciais Socioespaciais da População sem Registro Civil de Nascimento: uma análise das informações do Censo Demográfico 2010**; Trabalho apresentado no XVIII Encontro Nacional de Estudos Populacionais, ABEP, realizado em Águas de Lindóia/SP – Brasil, 2012.

DATASUS; **DEPARTAMENTO DE INFORMÁTICA DO SUS**, 2015. Disponível em: `http://www.http://datasus.saude.gov.br`; Acesso em: 10/01/2016.

DUNN, P. K.; SMYTH, G. K;. **Randomized quantile residuals**; Journal of Computational and Graphical Statistics, v. 5, n. 3, p. 236-244, 1996.

FARAWAY, J. J.;**Linear Models with R**; Edição publicada no Taylor& Francis e-Library, 2009.

FERREIRA, M. R. P. ; **Modelos para resposta binária: uma aplicação a dados biológicos**; Relatório apresentado ao Departamento de Estatística da UFPE para obtenção de conceito na disciplina Estágio Supervisionado Obrigatório,2004.

HOFFMANN, R. et al.; **Análise de regressão: uma introdução à econometria**;Biblioteca Digital da Produção Intelectual - BDPI - USP, 2015.

IBGE; **Instituto Brasileiro de Geografia e Estatística**; Disponível em: `http://www.ibge.gov.br`; Acesso em: 10/01/2016.

JORGE, M. H. P. M.; LAURENTI, R.; GOTLIEB, S. L. D.; **Análise da qualidade das estatísticas vitais brasileiras: a experiência de implantação do SIM e do SINASC**; Ciênc Saúde Coletiva, v. 12, n. 3, p. 643-54, 2007.

MELO, I. R. S.; **As estatísticas de nascimento e sua relação com os fatores socioeconômicos das microregiões do semiárido brasileiro**; Trabalho apresentado ao Curso de Graduação em Estatística da Universidade Federal da Paraíba, como requisito parcial para obtenção do título de Bacharel, 2014.

NASCIMENTO, A. M.; **População e família brasileira: ontem e hoje**;Trabalho apresentado no XV Encontro Nacional de Estudos Populacionais, ABEP, realizado em Caxambú- MG – Brasil, 2006.

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

PAES, N. A.; SANTOS, C.S.A.; **As estatísticas de nascimento e os fatores maternos e da criança nas microrregiões do Nordeste brasileiro: uma investigação usando análise fatorial**; Cad Saude Publica, v. 26, n. 2, p. 311-22, 2010.

PAES, N. A.; ALBUQUEQUE, M. E. E. **Avaliação da qualidade dos dados populacionais e cobertura dos registros de óbitos para as regiões brasileiras**; Rev. Saúde Pública, 33 (1): 33-43. 1999.; Disponível em: `www.fsp.usp.br/~rsp`; Acessado em: 16/03/2016.

PAES, N. A.;**Demografia Estatística da Saúde** - Livro: Congresso da RBRAS e SEAGRO, 2009.

PAES, N. A.; **O potencial das estatísticas vitais para a construção de indicadores de mortalidade e de natalidade no semiárido brasileiro**; Projeto para Inscrição no Processo Seletivo 2014-2015 (PIBIC, PIBITI, PIBIC-AF, PIVIC e PIVITI), 2014.

PAES, N.A.; MAIA, L. M. O.; **O potencial das estatísticas vitais para a construção de indicadores de natalidade no semiárido brasileiro**; Relatório de Execução – PIBIC/CNPq/UFPB, 2015.

PAIVA, J. P. L.; **Plano Nacional Para O Registro Civil de Nascimento (Sugestões)**;Disponível em : `http://registrodeimoveis1zona.com.br/?p=233`; Acessado em: 22/03/2016.

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G.; **Modelos Aditivos Generalizados para Posição, Escala e Forma (GAMLSS) na Modelagem de Curvas de Referência**; Revista Brasileira de Ciências da Saúde, v. 12, n. 3, p. 289-310, 2010.

Paula, G. A.; **"Modelos de regressão com apoio computacional"**; São Paulo: IME/USP, 2004.

PAULA, C. G. et al.; **Baixo peso ao nascer: fatores socioeconômicos, assistência pré-natal e nutricional** – uma revisão.; Augustus, v. 14, n. 29, p. 55-65, 2010.

PEREIRA, M. G.; **Epidemiologia: teoria e prática**. Rio de Janeiro: Ed. Guanabara Koogan, 2003.

QUEIROZ, N. M. O. B.; **Regressão logística – uma estimativa Bayesiana aplicada na identificação de fatores de risco para HIV, em doadores de sangue**; Dissertação (Mestrado em Biometria) – Universidade Federal Rural de Pernambuco. Departamento de Física e mate- mática, 2004.

O Projeto R para computação estatística; Acessado em: 15/01/2016 https://www.r-project.org/

RAMOS, H. A. C.; CUMAN, R. K. N.; **Fatores de risco para prematuridade: pesquisa documental**; Esc Anna Nery Rev Enferm, v. 13, n. 2, p. 297-304, 2009.

RIGBY, R.A.; STASINOPOULOS, D.M.; **Instructions on how to use the gamlss package in R.** Second Edition, 2008; Disponível em: http://www.gamlss.com/; Acessado em: 11/04/2016.

SÃO PAULO; **Manual de preenchimento da Declaração de Nascido Vivo**; São Paulo: Secretaria Municipal da Saúde, 2011.; Disponível em: http://www.prefeitura.sp.gov.br/cidade/secretarias/upload/saude/arquivos/publicacoes/Manual_DN_02fev2011.pdf; Acessado em: 18/03/2016.

SHEATHER, S. J.;**A Modern Approach to Regression with R**; Book: Springer Texts in Statistics, 2009

SOUZA, L. M.; **Avaliação do Sistema de Informação sobre Nascidos Vivos**; Trabalho apresentado no XIV Encontro Nacional de Estudos Populacionais, ABEP, Caxambu, Minas Gerais, Setembro de 2004.

TJPB; **TRIBUNAL DE JUSTIÇA DA PARAÍBA**, 2015; Disponível em: http://www.tjpb.jus.br; Acessado em: 22/12/2015.

WALDVOGEL, B. C. et al.; **Integração das bases de estatísticas vitais: Uma realidade possível**; Trabalho apresentado no XVII Encontro Nacional de Estudos Populacionais, realizado em Caxambu - MG – Brasil, 2010.

APÊNDICE

Script no R

```
#=====================================================================#
#                              Dados
#=====================================================================#
rm(list=ls())
setwd("E:/UFPB/TCC/Banco de Dados")       # Diretório
dados=read.csv2("Banco_Lígia_Final.csv")
attach(dados)
names(dados)
length(names(dados))                      # Tamanho
dados1= na.exclude(dados)
dim(dados1)
dim(dados)                                # Linhas e Colunas
summary(dados)
detach(dados)
attach(dados1)
#=====================================================================#
#               Classificando as variáveis como fator
#=====================================================================#
reg = factor(REGISTRO, labels = c("Não Tem", "Tem"))
idade=IDADE
cor= factor(COR, labels = c("Branca", "Não Branca"))
domicilio= factor(DOMICILIO, labels = c("Urbana", "Rural"))
sexo= factor(SEXO, labels = c("Masculino", "Feminino"))
ler= factor(LER, labels = c("Sim", "Não"))
esgoto= factor(ESGOTO, labels = c("Rede geral de esgoto ou pluvial"
, "Fossas, vala e rio"))
agua= factor(AGUA, labels = c("Sim", "Não"))
```

```
energia= factor(ENERGIA, labels = c("Sim", "Não"))
renda= factor(RENDA, labels = c("<0.5", "0.5|-1.5",">1.5"))
bolsa= factor(BOLSA, labels = c( "Não","Sim"))
banheiro= factor(BANHEIRO, labels = c("Nenhum", "1","2 ou mais"))
computador = factor(COMPUTADOR, labels = c("Sim", "Não"))
moto= factor(MOTO, labels = c("Sim", "Não"))
carro= factor(CARRO, labels = c("Sim", "Não"))
radio= factor(RADIO, labels = c("Sim", "Não"))
tv= factor(TV, labels = c("Sim", "Não"))
maquina= factor(MAQ, labels = c("Sim", "Não"))
geladeira= factor(GELADEIRA, labels = c("Sim", "Não"))
celular= factor(CELULAR, labels = c("Sim", "Não"))
y=dados1$REGISTRO
y= factor(y)                                  # Variável Dependente
#=====================================================================#
#                     Descritiva das Variáveis
#=====================================================================#
summary(reg)
summary(cor)
summary(domicilio)
summary(sexo)
summary(ler)
summary(esgoto)
summary(agua)
summary(energia)
summary(renda)
summary(bolsa)
summary(banheiro)
summary(computador)
summary(moto)
summary(carro)
summary(radio)
summary(tv)
summary(maquina)
summary(geladeira)
summary(celular)
#=====================================================================#
#                                 Gráfico
#=====================================================================#
plot(reg)
plot(cor)
```

```
plot(domicilio)
plot(sexo)
plot(ler)
plot(esgoto)
plot(agua)
plot(energia)
plot(renda)
plot(bolsa)
plot(banheiro)
plot(computador)
plot(moto)
plot(carro)
plot(radio)
plot(tv)
plot(maquina)
plot(geladeira)
plot(celular)
#=====================================================================#
#                    Tabela de frequencia relativa
#=====================================================================#
library(xtable)
table(preg,cor)
py=summary(y)/length(y);py
pcor=summary(cor)/length(cor);pcor
pdomicilio=summary(domicilio)/length(domicilio);pdomicilio
psexo=summary(sexo)/length(sexo);psexo
pler=summary(ler)/length(ler);pler
pesgoto=summary(esgoto)/length(esgoto);pesgoto
pagua=summary(agua)/length(agua);pagua
penergia=summary(energia)/length(energia);penergia
prenda=summary(renda)/length(renda);prenda
pbolsa=summary(bolsa)/length(bolsa);pbolsa
pbanheiro=summary(banheiro)/length(banheiro);pbanheiro
pcomputador=summary(computador)/length(computador);pcomputador
pmoto=summary(moto)/length(moto);pmoto
pcarro=summary(carro)/length(carro);pcarro
pradio=summary(radio)/length(radio);pradio
ptv=summary(tv)/length(tv);ptv
pmaquina=summary(maquina)/length(maquina);pmaquina
pgeladeira=summary(geladeira)/length(geladeira);pgeladeira
pcelular=summary(celular)/length(celular);pcelular
```

```
#=====================================================================#
#                          Tabelas cruzadas
#=====================================================================#
ycor=table(y,cor);ycor
a1=ycor[1]/sum(ycor[1],ycor[3])
b1=ycor[2]/sum(ycor[2],ycor[4])
c1=ycor[3]/sum(ycor[1],ycor[3])
d1=ycor[4]/sum(ycor[2],ycor[4])
mat1=100*matrix(c(a1,b1,c1,d1),2,2);mat1

ydomicilio=table(y,domicilio);ydomicilio
a2=ydomicilio[1]/sum(ydomicilio[1],ydomicilio[3])
b2=ydomicilio[2]/sum(ydomicilio[2],ydomicilio[4])
c2=ydomicilio[3]/sum(ydomicilio[1],ydomicilio[3])
d2=ydomicilio[4]/sum(ydomicilio[2],ydomicilio[4])
mat2=100*matrix(c(a2,b2,c2,d2),2,2);mat2

ysexo=table(y,sexo);ysexo
a3=ysexo[1]/sum(ysexo[1],ysexo[3])
b3=ysexo[2]/sum(ysexo[2],ysexo[4])
c3=ysexo[3]/sum(ysexo[1],ysexo[3])
d3=ysexo[4]/sum(ysexo[2],ysexo[4])
mat3=100*matrix(c(a3,b3,c3,d3),2,2);mat3

yler=table(y,ler);yler
a4=yler[1]/sum(yler[1],yler[3])
b4=yler[2]/sum(yler[2],yler[4])
c4=yler[3]/sum(yler[1],yler[3])
d4=yler[4]/sum(yler[2],yler[4])
mat4=100*matrix(c(a4,b4,c4,d4),2,2);mat4

yesgoto=table(y,esgoto);yesgoto
a5=yesgoto[1]/sum(yesgoto[1],yesgoto[3])
b5=yesgoto[2]/sum(yesgoto[2],yesgoto[4])
c5=yesgoto[3]/sum(yesgoto[1],yesgoto[3])
d5=yesgoto[4]/sum(yesgoto[2],yesgoto[4])
mat5=100*matrix(c(a5,b5,c5,d5),2,2);mat5

yagua=table(y,agua);yagua
a6=yagua[1]/sum(yagua[1],yagua[3])
b6=yagua[2]/sum(yagua[2],yagua[4])
```

```
c6=yagua[3]/sum(yagua[1],yagua[3])
d6=yagua[4]/sum(yagua[2],yagua[4])
mat6=100*matrix(c(a6,b6,c6,d6),2,2);mat6

yenergia=table(y,energia);yenergia
a7=yenergia[1]/sum(yenergia[1],yenergia[3])
b7=yenergia[2]/sum(yenergia[2],yenergia[4])
c7=yenergia[3]/sum(yenergia[1],yenergia[3])
d7=yenergia[4]/sum(yenergia[2],yenergia[4])
mat7=100*matrix(c(a7,b7,c7,d7),2,2);mat7

yrenda=table(y,renda);yrenda
a8=yrenda[1]/sum(yrenda[1],yrenda[3],yrenda[5])
b8=yrenda[2]/sum(yrenda[2],yrenda[4],yrenda[6])
c8=yrenda[3]/sum(yrenda[1],yrenda[3],yrenda[5])
d8=yrenda[4]/sum(yrenda[2],yrenda[4],yrenda[6])
e8=yrenda[5]/sum(yrenda[1],yrenda[3],yrenda[5])
f8=yrenda[6]/sum(yrenda[2],yrenda[4],yrenda[6])
mat8=100*matrix(c(a8,b8,c8,d8,e8,f8),2,3);mat8

ybolsa=table(y,bolsa);ybolsa
a9=ybolsa[1]/sum(ybolsa[1],ybolsa[3])
b9=ybolsa[2]/sum(ybolsa[2],ybolsa[4])
c9=ybolsa[3]/sum(ybolsa[1],ybolsa[3])
d9=ybolsa[4]/sum(ybolsa[2],ybolsa[4])
mat9=100*matrix(c(a9,b9,c9,d9),2,2);mat9

ybanheiro=table(y,banheiro);ybanheiro
a10=ybanheiro[1]/sum(ybanheiro[1],ybanheiro[3],ybanheiro[5])
b10=ybanheiro[2]/sum(ybanheiro[2],ybanheiro[4],ybanheiro[6])
c10=ybanheiro[3]/sum(ybanheiro[1],ybanheiro[3],ybanheiro[5])
d10=ybanheiro[4]/sum(ybanheiro[2],ybanheiro[4],ybanheiro[6])
e10=ybanheiro[5]/sum(ybanheiro[1],ybanheiro[3],ybanheiro[5])
f10=ybanheiro[6]/sum(ybanheiro[2],ybanheiro[4],ybanheiro[6])
mat10=100*matrix(c(a10,b10,c10,d10,e10,f10),2,3);mat10

ycomputador=table(y,computador);ycomputador
a11=ycomputador[1]/sum(ycomputador[1],ycomputador[3])
b11=ycomputador[2]/sum(ycomputador[2],ycomputador[4])
c11=ycomputador[3]/sum(ycomputador[1],ycomputador[3])
d11=ycomputador[4]/sum(ycomputador[2],ycomputador[4])
```

```
mat11=100*matrix(c(a11,b11,c11,d11),2,2);mat11

ymoto=table(y,moto);ymoto
a12=ymoto[1]/sum(ymoto[1],ymoto[3])
b12=ymoto[2]/sum(ymoto[2],ymoto[4])
c12=ymoto[3]/sum(ymoto[1],ymoto[3])
d12=ymoto[4]/sum(ymoto[2],ymoto[4])
mat12=100*matrix(c(a12,b12,c12,d12),2,2);mat12

ycarro=table(y,carro);ycarro
a13=ycarro[1]/sum(ycarro[1],ycarro[3])
b13=ycarro[2]/sum(ycarro[2],ycarro[4])
c13=ycarro[3]/sum(ycarro[1],ycarro[3])
d13=ycarro[4]/sum(ycarro[2],ycarro[4])
mat13=100*matrix(c(a13,b13,c13,d13),2,2);mat13

yradio=table(y,radio);yradio
a14=yradio[1]/sum(yradio[1],yradio[3])
b14=yradio[2]/sum(yradio[2],yradio[4])
c14=yradio[3]/sum(yradio[1],yradio[3])
d14=yradio[4]/sum(yradio[2],yradio[4])
mat14=100*matrix(c(a14,b14,c14,d14),2,2);mat4

ytv=table(y,tv);ytv
a15=ytv[1]/sum(ytv[1],ytv[3])
b15=ytv[2]/sum(ytv[2],ytv[4])
c15=ytv[3]/sum(ytv[1],ytv[3])
d15=ytv[4]/sum(ytv[2],ytv[4])
mat15=100*matrix(c(a15,b15,c15,d15),2,2);mat15

ymaquina=table(y,maquina);ymaquina
a16=ymaquina[1]/sum(ymaquina[1],ymaquina[3])
b16=ymaquina[2]/sum(ymaquina[2],ymaquina[4])
c16=ymaquina[3]/sum(ymaquina[1],ymaquina[3])
d16=ymaquina[4]/sum(ymaquina[2],ymaquina[4])
mat16=100*matrix(c(a16,b16,c16,d16),2,2);mat16

ygeladeira=table(y,geladeira);ygeladeira
a17=ygeladeira[1]/sum(ygeladeira[1],ygeladeira[3])
b17=ygeladeira[2]/sum(ygeladeira[2],ygeladeira[4])
c17=ygeladeira[3]/sum(ygeladeira[1],ygeladeira[3])
```

```
d17=ygeladeira[4]/sum(ygeladeira[2],ygeladeira[4])
mat17=100*matrix(c(a17,b17,c17,d17),2,2);mat17

ycelular=table(y,celular);ycelular
a18=ycelular[1]/sum(ycelular[1],ycelular[3])
b18=ycelular[2]/sum(ycelular[2],ycelular[4])
c18=ycelular[3]/sum(ycelular[1],ycelular[3])
d18=ycelular[4]/sum(ycelular[2],ycelular[4])
mat18=100*matrix(c(a18,b18,c18,d18),2,2);mat18

rownames(mat1)=c("Não Tem","Tem")
colnames(mat1)=c("Branca","Não Branca")
rownames(mat2)=c("Não Tem","Tem")
colnames(mat2)=c("Urbana", "Rural")
rownames(mat3)=c("Não Tem","Tem")
colnames(mat3)=c("Masculino", "Feminino")
rownames(mat4)=c("Não Tem","Tem")
colnames(mat4)=c("Sim", "Não")
rownames(mat5)=c("Não Tem","Tem")
colnames(mat5)=c("Rede*", "Fossas, valas e rio")
rownames(mat6)=c("Não Tem","Tem")
colnames(mat6)=c("Sim", "Não")
rownames(mat7)=c("Não Tem","Tem")
colnames(mat7)=c("Sim", "Não")
rownames(mat8)=c("Não Tem","Tem")
colnames(mat8)=c("<0.5", "0.5|-1.5",">1.5")
rownames(mat9)=c("Não Tem","Tem")
colnames(mat9)=c("Não","Sim")
rownames(mat10)=c("Não Tem","Tem")
colnames(mat10)=c("Nenhum", "1","2 ou mais")
rownames(mat11)=c("Não Tem","Tem")
colnames(mat11)=c("Sim", "Não")
rownames(mat12)=c("Não Tem","Tem")
colnames(mat12)=c("Sim", "Não")
rownames(mat13)=c("Não Tem","Tem")
colnames(mat13)=c("Sim", "Não")
rownames(mat14)=c("Não Tem","Tem")
colnames(mat14)=c("Sim", "Não")
rownames(mat15)=c("Não Tem","Tem")
colnames(mat15)=c("Sim", "Não")
rownames(mat16)=c("Não Tem","Tem")
```

```
colnames(mat16)=c("Sim", "Não")
rownames(mat17)=c("Não Tem","Tem")
colnames(mat17)=c("Sim", "Não")
rownames(mat18)=c("Não Tem","Tem")
colnames(mat18)=c("Sim", "Não")
#=====================================================================#
#                Gráficos das Prorpoçoes (cruzadas)
#=====================================================================#
pdf("mat1.pdf")
barplot(mat1,main=c("Cor"),beside=T,legend=c("Não Tem" ,"Tem"),ylim=c
(0,100),ylab="%")
dev.off()
pdf("mat2.pdf")
barplot(mat2,main=c("Situação do domicilio"),beside=T,legend=c("Não
Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat3.pdf")
barplot(mat3,main=c("Sexo"),beside=T,legend=c("Não Tem" ,"Tem"),ylim=c
(0,100),ylab="%")
dev.off()
pdf("mat4.pdf")
barplot(mat4,main=c("Sabe ler/escrever"),beside=T,legend=c("Não Tem"
,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat5.pdf")
barplot(mat5,main=c("Tipo de esgotamento sanitário"),beside=T,legend
=c("Não Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat6.pdf")
barplot(mat6,main=c("Abastecimento de Água"),beside=T,legend=c("Não
Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat7.pdf")
barplot(mat7,main=c("Existência de energia elétrica"),beside=T,legend
=c("Não Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat8.pdf")
barplot(mat8,main=c("Rendimento domiciliar per capito"),beside=T,
legend=c("Não Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat9.pdf")
```

```
barplot(mat9,main=c("Rendimento de Bolsa Família e PETI"),beside=T,
legend=c("Não Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat10.pdf")
barplot(mat10,main=c("Número de Banheiros"),beside=T,legend=c("Não
Tem","Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat11.pdf")
barplot(mat11,main=c("Existência de microcomputador"),beside=T,legend
=c("Não Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat12.pdf")
barplot(mat12,main=c("Existência de moto"),beside=T,legend=c("Não Tem"
,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat13.pdf")
barplot(mat13,main=c("Existência de carro"),beside=T,legend=c("Não Tem"
,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat14.pdf")
barplot(mat14,main=c("Existência de radio"),beside=T,legend=c("Não Tem"
,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat15.pdf")
barplot(mat15,main=c("Existência de tv"),beside=T,legend=c("Não Tem"
,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat16.pdf")
barplot(mat16,main=c("Existência de máquina de lavar"),beside=T,legend
=c("Não Tem" ,"Tem"),ylim=c(0,119),ylab="%")
dev.off()
pdf("mat17.pdf")
barplot(mat17,main=c("Existência de geladeira"),beside=T,legend=c("Não
Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat18.pdf")
barplot(mat18,main=c("Existência de celular"),beside=T,legend=c("Não
Tem" ,"Tem"),ylim=c(0,100),ylab="%")
dev.off()
#=====================================================================#
#                     Seleção de Modelo GLM (logit)
```

```
#=====================================================================#
maior<-glm(y~cor+domicilio+sexo+ler+esgoto+agua+
energia+renda+bolsa+banheiro+computador+
moto+carro+radio+tv+maquina+geladeira+celular
, family = binomial(link=logit), na.action = na.exclude)
fit_logit=step(maior);fit_logit
#=====================================================================#
#      Teste de modelos GLM  fução de ligação "logit" fit(*)
#=====================================================================#
fit1 = glm(y ~ ler + esgoto + agua + maquina + geladeira + celular,
family = binomial(link = logit))
fit2 = glm(y ~ ler + esgoto + agua + radio + maquina + geladeira +
celular, family = binomial(link = logit))
fit3 = glm(y ~ ler + esgoto + agua + energia + radio + maquina +
geladeira + celular, family = binomial(link = logit))
fit4 = glm(y ~ ler + esgoto + agua + energia + radio + tv + maquina +
geladeira + celular, family = binomial(link = logit))
fit5 = glm(y ~ ler + esgoto + agua + energia + bolsa + radio + tv +
maquina + geladeira + celular, family = binomial(link = logit))
fit6 = glm(y ~  cor + ler + esgoto + agua + energia + bolsa + radio +
tv + maquina + geladeira + celular, family = binomial(link = logit))
fit7 = glm(y ~ cor + ler + esgoto + agua + energia + bolsa + carro +
radio + tv + maquina + geladeira + celular, family =
binomial(link = logit))
fit8 = glm(y ~  cor + ler + esgoto + agua + energia + bolsa + moto +
carro + radio + tv + maquina + geladeira + celular,family =
binomial(link = logit))
fit9 = glm(y ~ cor + ler + esgoto + agua + energia + bolsa +
computador + moto + carro + radio + tv + maquina + geladeira +
celular,family = binomial(link = logit))
fit10 = glm(y ~ cor + sexo + ler + esgoto + agua + energia + bolsa +
computador + moto + carro + radio + tv + maquina + geladeira +
celular,family = binomial(link = logit))
fit11 = glm(y ~  cor + domicilio + sexo + ler + esgoto + agua +
energia + bolsa + computador + moto + carro + radio + tv + maquina
+ geladeira + celular, family = binomial(link = logit))
fit12 = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua +
energia + bolsa + banheiro + computador + moto + carro + radio +
tv + maquina + geladeira + celular, family = binomial(link = logit))
fit13 = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ renda + bolsa + banheiro + computador + moto + carro + radio +
```

```
tv + maquina + geladeira + celular, family = binomial(link = logit))
#=========================================================================#
#                  Representatividade do Modelo GLM
#=========================================================================#
phi1<-summary(fit1)$dispersion
dp1=summary(fit1)$deviance/phi1
x2.1=qchisq(0.95,summary(fit1)$df.residual)

phi2<-summary(fit2)$dispersion
d2=summary(fit2)$deviance/phi2
x2.2=qchisq(0.95,summary(fit2)$df.residual)

phi3<-summary(fit3)$dispersion
d3=summary(fit3)$deviance/phi3
x2.3=qchisq(0.95,summary(fit3)$df.residual)

phi4<-summary(fit4)$dispersion
d4=summary(fit4)$deviance/phi4
x2.4=qchisq(0.95,summary(fit4)$df.residual)

phi5<-summary(fit5)$dispersion
d5=summary(fit5)$deviance/phi5
x2.5=qchisq(0.95,summary(fit5)$df.residual)

phi6<-summary(fit6)$dispersion
d6=summary(fit6)$deviance/phi6
x2.6=qchisq(0.95,summary(fit6)$df.residual)

phi7<-summary(fit7)$dispersion
d7=summary(fit7)$deviance/phi7
x2.7=qchisq(0.95,summary(fit7)$df.residual)

phi8<-summary(fit8)$dispersion
d8=summary(fit8)$deviance/phi8
x2.8=qchisq(0.95,summary(fit8)$df.residual)

phi9<-summary(fit9)$dispersion
d9=summary(fit9)$deviance/phi9
x2.9=qchisq(0.95,summary(fit9)$df.residual)

phi10<-summary(fit10)$dispersion
```

```
d10=summary(fit10)$deviance/phi10
x2.10=qchisq(0.95,summary(fit10)$df.residual)

phi11<-summary(fit11)$dispersion
d11=summary(fit11)$deviance/phi11
x2.11=qchisq(0.95,summary(fit11)$df.residual)

phi12<-summary(fit12)$dispersion
d12=summary(fit12)$deviance/phi12
x2.12=qchisq(0.95,summary(fit12)$df.residual)

phi13<-summary(fit13)$dispersion
d13=summary(fit13)$deviance/phi13
x2.13=qchisq(0.95,summary(fit13)$df.residual)

D=c(d1,d2,d3,d4,d5,d6,d7,d8,d9,d10,d11,d12,d13);D
x2=c(x2.1,x2.2,x2.3,x2.4,x2.5,x2.6,x2.7,x2.8,x2.9,x2.10,
x2.11,x2.12,x2.13);x2
desvio=cbind(D,x2);desvio
library(xtable)
xtable(desvio)
#=========================================================================#
#                  Seleção de Modelo GLM (probit)
#=========================================================================#
maior1<-glm(y~ cor+domicilio+sexo+ler+esgoto+agua+
energia+renda+bolsa+banheiro+computador+
moto+carro+radio+tv+maquina+geladeira+celular
, family = binomial(link=probit), na.action = na.exclude)
fit_probit=step(maior1);fit_probit
#=========================================================================#
#       Teste de modelos GLM fução de ligação "probit" fit(*1)
#=========================================================================#
fit1p = glm(y ~ ler + esgoto + agua + maquina + geladeira + celular,
family = binomial(link = probit))
fit2p = glm(y ~ ler + esgoto + agua + radio + maquina + geladeira +
celular,family = binomial(link = probit))
fit3p = glm(y ~ ler + esgoto + agua + energia + radio + maquina +
geladeira + celular, family = binomial(link = probit))
fit4p = glm(y ~ ler + esgoto + agua + energia + radio + tv + maquina
+ geladeira + celular, family = binomial(link = probit))
fit5p = glm(y ~ ler + esgoto + agua + energia + bolsa + radio + tv +
```

```
maquina +geladeira + celular, family = binomial(link = probit))
fit6p = glm(y ~ cor + ler + esgoto + agua + energia + bolsa + radio +
tv + maquina + geladeira + celular, family = binomial(link = probit))
fit7p = glm(y ~ cor + ler + esgoto + agua + energia + bolsa + carro +
radio + tv + maquina + geladeira + celular,family = binomial(link =
probit))
fit8p = glm(y ~ cor + sexo + ler + esgoto + agua + energia + bolsa +
carro + radio + tv + maquina + geladeira + celular,family = binomial
(link = probit))
fit9p = glm(y ~ cor + sexo + ler + esgoto + agua + energia + bolsa
+ computador + carro + radio + tv + maquina + geladeira + celular,
family = binomial(link = probit))
fit10p = glm(y ~ cor + sexo + ler + esgoto + agua + energia + bolsa +
computador + moto + carro + radio + tv + maquina + geladeira + celular,
family = binomial(link = probit))
fit11p = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ bolsa + computador + moto + carro + radio + tv + maquina +
geladeira + celular, family = binomial(link = probit))

fit12p = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ bolsa + banheiro + computador + moto + carro + radio + tv +
maquina + geladeira + celular, family = binomial(link = probit))

fit13p = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ renda + bolsa + banheiro + computador + moto + carro + radio +
tv + maquina + geladeira + celular, family = binomial(link = probit))
#=====================================================================#
#                       Representatividade do modelo
#=====================================================================#
phi1p<-summary(fit1p)$dispersion
d1p=summary(fit1p)$deviance/phi1p
x2.1p=qchisq(0.95,summary(fit1p)$df.residual)

phi2p<-summary(fit2p)$dispersion
d2p=summary(fit2p)$deviance/phi2p
x2.2p=qchisq(0.95,summary(fit2p)$df.residual)

phi3p<-summary(fit3p)$dispersion
d3p=summary(fit3p)$deviance/phi3p
x2.3p=qchisq(0.95,summary(fit3p)$df.residual)
```

```
phi4p<-summary(fit4p)$dispersion
d4p=summary(fit4p)$deviance/phi4p
x2.4p=qchisq(0.95,summary(fit4p)$df.residual)

phi5p<-summary(fit5p)$dispersion
d5p=summary(fit5p)$deviance/phi5p
x2.5p=qchisq(0.95,summary(fit5p)$df.residual)

phi6p<-summary(fit6p)$dispersion
d6p=summary(fit6p)$deviance/phi6p
x2.6p=qchisq(0.95,summary(fit6p)$df.residual)

phi7p<-summary(fit7p)$dispersion
d7p=summary(fit7p)$deviance/phi7p
x2.7p=qchisq(0.95,summary(fit7p)$df.residual)

phi8p<-summary(fit8p)$dispersion
d8p=summary(fit8p)$deviance/phi8p
x2.8p=qchisq(0.95,summary(fit8p)$df.residual)

phi9p<-summary(fit9p)$dispersion
d9p=summary(fit9p)$deviance/phi9p
x2.9p=qchisq(0.95,summary(fit9p)$df.residual)

phi10p<-summary(fit10p)$dispersion
d10p=summary(fit10p)$deviance/phi10p
x2.10p=qchisq(0.95,summary(fit10p)$df.residual)

phi11p<-summary(fit11p)$dispersion
d11p=summary(fit11p)$deviance/phi11p
x2.11p=qchisq(0.95,summary(fit11p)$df.residual)

phi12p<-summary(fit12p)$dispersion
d12p=summary(fit12p)$deviance/phi12p
x2.12p=qchisq(0.95,summary(fit12p)$df.residual)

phi13p<-summary(fit13p)$dispersion
d13p=summary(fit13p)$deviance/phi13p
x2.13p=qchisq(0.95,summary(fit13p)$df.residual)

Dp=c(d1p,d2p,d3p,d4p,d5p,d6p,d7p,d8p,d9p,d10p,d11p,d12p,d13p);Dp
```

```
x2p=c(x2.1p,x2.2p,x2.3p,x2.4p,x2.5p,x2.6p,x2.7p,x2.8p,x2.9p,x2.10p,
x2.11p,x2.12p,x2.13p);x2p
desviop=cbind(Dp,x2p);desviop
library(xtable)
xtable(desviop)
#=========================================================================#
#                         Seleção de Modelo (cauchit)
#=========================================================================#
maior2<-glm(y~cor+domicilio+sexo+ler+esgoto+agua+
energia+renda+bolsa+banheiro+computador+
moto+carro+radio+tv+maquina+geladeira+celular
, family = binomial(link=cauchit), na.action = na.exclude)
modelo2=step(menor2,scope=list(upper=maior2),direction="both");modelo2
fit_cauchit=step(maior2);fit_cauchit
#=========================================================================#
#         Teste de modelos:  Fução de ligação "cauchit" fit(*2)
#=========================================================================#
#Warning messages:
#1: glm.fit: algorithm did not converge
#2: glm.fit: fitted probabilities numerically 0 or 1 occurred
#=========================================================================#
#                   Significância das variáveis
#=========================================================================#
# GLM Logit
xtable(summary(maior))
xtable(summary(fit1))
xtable(summary(fit2))
xtable(summary(fit3))
xtable(summary(fit4))
xtable(summary(fit5))
xtable(summary(fit6))
xtable(summary(fit7))
xtable(summary(fit8))
xtable(summary(fit9))
xtable(summary(fit10))
xtable(summary(fit11))
xtable(summary(fit12))
xtable(summary(fit13))
# GLM Probit
xtable(summary(maior1))
xtable(summary(fit1p))
```

```
xtable(summary(fit2p))
xtable(summary(fit3p))
xtable(summary(fit4p))
xtable(summary(fit5p))
xtable(summary(fit6p))
xtable(summary(fit7p))
xtable(summary(fit8p))
xtable(summary(fit9p))
xtable(summary(fit10p))
xtable(summary(fit11p))
xtable(summary(fit12p))
xtable(summary(fit13p))
#=========================================================================#
#                          Comparação dos AIC's
#=========================================================================#
# Logit
aic=c(
summary(fit1)$aic,
summary(fit2)$aic,
summary(fit3)$aic,
summary(fit4)$aic,
summary(fit5)$aic,
summary(fit6)$aic,
summary(fit7)$aic,
summary(fit8)$aic,
summary(fit9)$aic,
summary(fit10)$aic,
summary(fit11)$aic,
summary(fit12)$aic,
summary(fit13)$aic
);aic
# Probit
aicp=c(
summary(fit1p)$aic,
summary(fit2p)$aic,
summary(fit3p)$aic,
summary(fit4p)$aic,
summary(fit5p)$aic,
summary(fit6p)$aic,
summary(fit7p)$aic,
summary(fit8p)$aic,
```

```
summary(fit9p)$aic,
summary(fit10p)$aic,
summary(fit11p)$aic,
summary(fit12p)$aic,
summary(fit13p)$aic
);aicp
require(xtable)
library(xtable)
aics=cbind(aic,aicp);aics
xtable(aics,digits=3) # para latex
comparacao=cbind(aic,D,x2,aicp,Dp,x2p)
xtable(comparacao)
#=========================================================================#
#                     Padronização dos resíduos
#=========================================================================#
library(car)
library(statmod)
# Logit
res1=qres.binom(fit1)
res2=qres.binom(fit2)
res3=qres.binom(fit3)
res4=qres.binom(fit4)
res5=qres.binom(fit5)
res6=qres.binom(fit6)
res7=qres.binom(fit7)
res8=qres.binom(fit8)
res9=qres.binom(fit9)
res10=qres.binom(fit10)
res11=qres.binom(fit11)
res12=qres.binom(fit12)
res13=qres.binom(fit13)
resm=qres.binom(maior)
# Probit
res1p=qres.binom(fit1p)
res2p=qres.binom(fit2p)
res3p=qres.binom(fit3p)
res4p=qres.binom(fit4p)
res5p=qres.binom(fit5p)
res6p=qres.binom(fit6p)
res7p=qres.binom(fit7p)
res8p=qres.binom(fit8p)
```

```
res9p=qres.binom(fit9p)
res10p=qres.binom(fit10p)
res11p=qres.binom(fit11p)
res12p=qres.binom(fit12p)
res13p=qres.binom(fit13p)
resm1=qres.binom(maior1)
#=========================================================================#
#                     Resíduos
#=========================================================================#
# Logit
pdf("res1.pdf")
plot(res1,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res2.pdf")
plot(res2,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res3.pdf")
plot(res3,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res4.pdf")
plot(res4,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res5.pdf")
plot(res5,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res6.pdf")
plot(res6,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res7.pdf")
plot(res7,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res8.pdf")
plot(res8,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
```

```
dev.off()
pdf("res9.pdf")
plot(res9,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res10.pdf")
plot(res10,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res11.pdf")
plot(res11,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res12.pdf")
plot(res12,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res13.pdf")
plot(res13,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("resm.pdf")
plot(resm,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
# Probit
pdf("res1p.pdf")
plot(res1p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res2p.pdf")
plot(res2p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res3p.pdf")
plot(res3p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res4p.pdf")
plot(res4p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
```

```
dev.off()
pdf("res5p.pdf")
plot(res5p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res6p.pdf")
plot(res6p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res7p.pdf")
plot(res7p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res8p.pdf")
plot(res8p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res9p.pdf")
plot(res9p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res10p.pdf")
plot(res10p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res11p.pdf")
plot(res11p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res12p.pdf")
plot(res12p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res13p.pdf")
plot(res13p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("resm1.pdf")
plot(resm1,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
```

```
#=====================================================================#
# Autocorrelação
#=====================================================================#
pdf("acf1.pdf")
acf(res1,main="Logit",xlab=c("Defasagem"))
dev.off()
pdf("acf1p.pdf")
acf(res1p,main="Probit",xlab=c("Defasagem"))
dev.off()
pdf("acfm.pdf")
acf(resm,main="Logit",xlab=c("Defasagem"))
dev.off()
pdf("acfm1.pdf")
acf(resm1,main="Probit",xlab=c("Defasagem"))
dev.off()
#=====================================================================#
#   Autocorrelação Parcial
#=====================================================================#
pdf("pacf1.pdf")
pacf(res1,main="Logit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
pdf("pacf1p.pdf")
pacf(res1p,main="Probit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
pdf("pacfm.pdf")
pacf(resm,main="Logit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
pdf("pacfm1.pdf")
pacf(resm1,main="Probit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
#=====================================================================#
#      Envelope
#=====================================================================#
# Logit
pdf("env1.pdf")
qqPlot(res1,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env2.pdf")
qqPlot(res2,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env3.pdf")
```

```
qqPlot(res3,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env4.pdf")
qqPlot(res4,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env5.pdf")
qqPlot(res5,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env6.pdf")
qqPlot(res6,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env7.pdf")
qqPlot(res7,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env8.pdf")
qqPlot(res8,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env9.pdf")
qqPlot(res9,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env10.pdf")
qqPlot(res10,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env11.pdf")
qqPlot(res11,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env12.pdf")
qqPlot(res12,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env13.pdf")
qqPlot(res13,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("envm.pdf")
qqPlot(resm,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
# Probit
pdf("env1p.pdf")
qqPlot(res1p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env2p.pdf")
qqPlot(res2p,main="Probit", ylab="Resíduos",xlab="Quantis")
```

```
dev.off()
pdf("env3p.pdf")
qqPlot(res3p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env4p.pdf")
qqPlot(res4p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env5p.pdf")
qqPlot(res5p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env6p.pdf")
qqPlot(res6p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env7p.pdf")
qqPlot(res7p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env8p.pdf")
qqPlot(res8p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env9p.pdf")
qqPlot(res9p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env10p.pdf")
qqPlot(res10p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env11p.pdf")
qqPlot(res11p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env12p.pdf")
qqPlot(res12p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env13p.pdf")
qqPlot(res13p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("envmp.pdf")
qqPlot(resm1,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
#=====================================================================#
#                          Razão de chance
#=====================================================================#
OR1=cbind(Estimativa=exp(summary(fit1)$coefficients)[,1],
```

```
exp(confint(fit1)))
ORm=cbind(Estimativa=exp(summary(maior)$coefficients)[,1],
exp(confint(maior)))
OR1p=cbind(Estimativa=exp(summary(fit1p)$coefficients)[,1],
exp(confint(fit1p)))
ORmp=cbind(Estimativa=exp(summary(maior1)$coefficients)[,1],
exp(confint(maior1)))
xtable(OR1)
xtable(ORm)
xtable(OR1p)
xtable(ORmp)
#*********************************************************************#
#=====================================================================#
#                          Dados sem bolsa
#=====================================================================#
rm(list=ls())
setwd("E:/UFPB/TCC/Banco de Dados")       # Diretório
dados=read.csv2("Banco_Lígia_Final.csv")
attach(dados)
names(dados)
length(names(dados))                      # Tamanho
dados1=dados[,-8] # retirando a variável bolsa
names(dados1)
dados1= na.exclude(dados1)
dim(dados1)
dim(dados)                                # Linhas e Colunas
summary(dados)
detach(dados)
attach(dados1)
#=====================================================================#
#              Classificando as variáveis como fator
#=====================================================================#
reg = factor(REGISTRO, labels = c("Não Tem", "Tem"))
idade=IDADE
cor= factor(COR, labels = c("Branca", "Não Branca"))
domicilio= factor(DOMICILIO, labels = c("Urbana", "Rural"))
sexo= factor(SEXO, labels = c("Masculino", "Feminino"))
ler= factor(LER, labels = c("Sim", "Não"))
esgoto= factor(ESGOTO, labels = c("Rede geral de esgoto ou pluvial"
, "Fossas, vala e rio"))
agua= factor(AGUA, labels = c("Sim", "Não"))
```

```
energia= factor(ENERGIA, labels = c("Sim", "Não"))
renda= factor(RENDA, labels = c("<0.5", "0.5|-1.5",">1.5"))
banheiro= factor(BANHEIRO, labels = c("Nenhum", "1","2 ou mais"))
computador = factor(COMPUTADOR, labels = c("Sim", "Não"))
moto= factor(MOTO, labels = c("Sim", "Não"))
carro= factor(CARRO, labels = c("Sim", "Não"))
radio= factor(RADIO, labels = c("Sim", "Não"))
tv= factor(TV, labels = c("Sim", "Não"))
maquina= factor(MAQ, labels = c("Sim", "Não"))
geladeira= factor(GELADEIRA, labels = c("Sim", "Não"))
celular= factor(CELULAR, labels = c("Sim", "Não"))
y=dados1$REGISTRO
y= factor(y)                                  # Variável Dependente
#=====================================================================#
#                     Descritiva das Variáveis
#=====================================================================#
summary(idade)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  5.000   6.000   8.000   7.609   9.000  10.000
#=====================================================================#
#                    Tabela de frequencia relativa
#=====================================================================#
library(xtable)
table(preg,cor)
barplot(p,y)
py=summary(y)/length(y);py
pcor=summary(cor)/length(cor);pcor
pdomicilio=summary(domicilio)/length(domicilio);pdomicilio
psexo=summary(sexo)/length(sexo);psexo
pler=summary(ler)/length(ler);pler
pesgoto=summary(esgoto)/length(esgoto);pesgoto
pagua=summary(agua)/length(agua);pagua
penergia=summary(energia)/length(energia);penergia
prenda=summary(renda)/length(renda);prenda
pidade=table(idade)/length(idade);pidade
pbanheiro=summary(banheiro)/length(banheiro);pbanheiro
pcomputador=summary(computador)/length(computador);pcomputador
pmoto=summary(moto)/length(moto);pmoto
pcarro=summary(carro)/length(carro);pcarro
pradio=summary(radio)/length(radio);pradio
ptv=summary(tv)/length(tv);ptv
```

```
pmaquina=summary(maquina)/length(maquina);pmaquina
pgeladeira=summary(geladeira)/length(geladeira);pgeladeira
pcelular=summary(celular)/length(celular);pcelular
#=====================================================================#
#                          Tabelas cruzadas
#=====================================================================#
yy=table(y);yy
a=yy[1]/sum(yy[1],yy[2])
b=yy[2]/sum(yy[1],yy[2])
mat=100*matrix(c(a,b),1,2);mat

ycor=table(y,cor);ycor
a1=ycor[1]/sum(ycor[1],ycor[3])
b1=ycor[2]/sum(ycor[2],ycor[4])
c1=ycor[3]/sum(ycor[1],ycor[3])
d1=ycor[4]/sum(ycor[2],ycor[4])
mat1=100*matrix(c(a1,b1,c1,d1),2,2);mat1

ydomicilio=table(y,domicilio);ydomicilio
a2=ydomicilio[1]/sum(ydomicilio[1],ydomicilio[3])
b2=ydomicilio[2]/sum(ydomicilio[2],ydomicilio[4])
c2=ydomicilio[3]/sum(ydomicilio[1],ydomicilio[3])
d2=ydomicilio[4]/sum(ydomicilio[2],ydomicilio[4])
mat2=100*matrix(c(a2,b2,c2,d2),2,2);mat2

ysexo=table(y,sexo);ysexo
a3=ysexo[1]/sum(ysexo[1],ysexo[3])
b3=ysexo[2]/sum(ysexo[2],ysexo[4])
c3=ysexo[3]/sum(ysexo[1],ysexo[3])
d3=ysexo[4]/sum(ysexo[2],ysexo[4])
mat3=100*matrix(c(a3,b3,c3,d3),2,2);mat3

yler=table(y,ler);yler
a4=yler[1]/sum(yler[1],yler[3])
b4=yler[2]/sum(yler[2],yler[4])
c4=yler[3]/sum(yler[1],yler[3])
d4=yler[4]/sum(yler[2],yler[4])
mat4=100*matrix(c(a4,b4,c4,d4),2,2);mat4

yesgoto=table(y,esgoto);yesgoto
a5=yesgoto[1]/sum(yesgoto[1],yesgoto[3])
```

```
b5=yesgoto[2]/sum(yesgoto[2],yesgoto[4])
c5=yesgoto[3]/sum(yesgoto[1],yesgoto[3])
d5=yesgoto[4]/sum(yesgoto[2],yesgoto[4])
mat5=100*matrix(c(a5,b5,c5,d5),2,2);mat5

yagua=table(y,agua);yagua
a6=yagua[1]/sum(yagua[1],yagua[3])
b6=yagua[2]/sum(yagua[2],yagua[4])
c6=yagua[3]/sum(yagua[1],yagua[3])
d6=yagua[4]/sum(yagua[2],yagua[4])
mat6=100*matrix(c(a6,b6,c6,d6),2,2);mat6

yenergia=table(y,energia);yenergia
a7=yenergia[1]/sum(yenergia[1],yenergia[3])
b7=yenergia[2]/sum(yenergia[2],yenergia[4])
c7=yenergia[3]/sum(yenergia[1],yenergia[3])
d7=yenergia[4]/sum(yenergia[2],yenergia[4])
mat7=100*matrix(c(a7,b7,c7,d7),2,2);mat7

yrenda=table(y,renda);yrenda
a8=yrenda[1]/sum(yrenda[1],yrenda[3],yrenda[5])
b8=yrenda[2]/sum(yrenda[2],yrenda[4],yrenda[6])
c8=yrenda[3]/sum(yrenda[1],yrenda[3],yrenda[5])
d8=yrenda[4]/sum(yrenda[2],yrenda[4],yrenda[6])
e8=yrenda[5]/sum(yrenda[1],yrenda[3],yrenda[5])
f8=yrenda[6]/sum(yrenda[2],yrenda[4],yrenda[6])
mat8=100*matrix(c(a8,b8,c8,d8,e8,f8),2,3);mat8

yidade=table(y,idade);yidade
a9=yidade[1]/sum(yidade[1],yidade[3],yidade[5],yidade[7],yidade[9],
yidade[11])
b9=yidade[2]/sum(yidade[2],yidade[4],yidade[6],yidade[8],yidade[10],
yidade[12])
c9=yidade[3]/sum(yidade[1],yidade[3],yidade[5],yidade[7],yidade[9],
yidade[11])
d9=yidade[4]/sum(yidade[2],yidade[4],yidade[6],yidade[8],yidade[10],
yidade[12])
e9=yidade[5]/sum(yidade[1],yidade[3],yidade[5],yidade[7],yidade[9],
yidade[11])
f9=yidade[6]/sum(yidade[2],yidade[4],yidade[6],yidade[8],yidade[10],
yidade[12])
```

```
g9=yidade[7]/sum(yidade[1],yidade[3],yidade[5],yidade[7],yidade[9],
yidade[11])
h9=yidade[8]/sum(yidade[2],yidade[4],yidade[6],yidade[8],yidade[10],
yidade[12])
i9=yidade[9]/sum(yidade[1],yidade[3],yidade[5],yidade[7],yidade[9],
yidade[11])
j9=yidade[10]/sum(yidade[2],yidade[4],yidade[6],yidade[8],yidade[10],
yidade[12])
k9=yidade[11]/sum(yidade[1],yidade[3],yidade[5],yidade[7],yidade[9],
yidade[11])
l9=yidade[12]/sum(yidade[2],yidade[4],yidade[6],yidade[8],yidade[10],
yidade[12])
mat9=100*matrix(c(a9,b9,c9,d9,e9,f9,g9,h9,i9,j9,k9,l9),2,6);mat9

ybanheiro=table(y,banheiro);ybanheiro
a10=ybanheiro[1]/sum(ybanheiro[1],ybanheiro[3],ybanheiro[5])
b10=ybanheiro[2]/sum(ybanheiro[2],ybanheiro[4],ybanheiro[6])
c10=ybanheiro[3]/sum(ybanheiro[1],ybanheiro[3],ybanheiro[5])
d10=ybanheiro[4]/sum(ybanheiro[2],ybanheiro[4],ybanheiro[6])
e10=ybanheiro[5]/sum(ybanheiro[1],ybanheiro[3],ybanheiro[5])
f10=ybanheiro[6]/sum(ybanheiro[2],ybanheiro[4],ybanheiro[6])
mat10=100*matrix(c(a10,b10,c10,d10,e10,f10),2,3);mat10

ycomputador=table(y,computador);ycomputador
a11=ycomputador[1]/sum(ycomputador[1],ycomputador[3])
b11=ycomputador[2]/sum(ycomputador[2],ycomputador[4])
c11=ycomputador[3]/sum(ycomputador[1],ycomputador[3])
d11=ycomputador[4]/sum(ycomputador[2],ycomputador[4])
mat11=100*matrix(c(a11,b11,c11,d11),2,2);mat11

ymoto=table(y,moto);ymoto
a12=ymoto[1]/sum(ymoto[1],ymoto[3])
b12=ymoto[2]/sum(ymoto[2],ymoto[4])
c12=ymoto[3]/sum(ymoto[1],ymoto[3])
d12=ymoto[4]/sum(ymoto[2],ymoto[4])
mat12=100*matrix(c(a12,b12,c12,d12),2,2);mat12

ycarro=table(y,carro);ycarro
a13=ycarro[1]/sum(ycarro[1],ycarro[3])
b13=ycarro[2]/sum(ycarro[2],ycarro[4])
c13=ycarro[3]/sum(ycarro[1],ycarro[3])
```

```
d13=ycarro[4]/sum(ycarro[2],ycarro[4])
mat13=100*matrix(c(a13,b13,c13,d13),2,2);mat13

yradio=table(y,radio);yradio
a14=yradio[1]/sum(yradio[1],yradio[3])
b14=yradio[2]/sum(yradio[2],yradio[4])
c14=yradio[3]/sum(yradio[1],yradio[3])
d14=yradio[4]/sum(yradio[2],yradio[4])
mat14=100*matrix(c(a14,b14,c14,d14),2,2);mat4

ytv=table(y,tv);ytv
a15=ytv[1]/sum(ytv[1],ytv[3])
b15=ytv[2]/sum(ytv[2],ytv[4])
c15=ytv[3]/sum(ytv[1],ytv[3])
d15=ytv[4]/sum(ytv[2],ytv[4])
mat15=100*matrix(c(a15,b15,c15,d15),2,2);mat15

ymaquina=table(y,maquina);ymaquina
a16=ymaquina[1]/sum(ymaquina[1],ymaquina[3])
b16=ymaquina[2]/sum(ymaquina[2],ymaquina[4])
c16=ymaquina[3]/sum(ymaquina[1],ymaquina[3])
d16=ymaquina[4]/sum(ymaquina[2],ymaquina[4])
mat16=100*matrix(c(a16,b16,c16,d16),2,2);mat16

ygeladeira=table(y,geladeira);ygeladeira
a17=ygeladeira[1]/sum(ygeladeira[1],ygeladeira[3])
b17=ygeladeira[2]/sum(ygeladeira[2],ygeladeira[4])
c17=ygeladeira[3]/sum(ygeladeira[1],ygeladeira[3])
d17=ygeladeira[4]/sum(ygeladeira[2],ygeladeira[4])
mat17=100*matrix(c(a17,b17,c17,d17),2,2);mat17

ycelular=table(y,celular);ycelular
a18=ycelular[1]/sum(ycelular[1],ycelular[3])
b18=ycelular[2]/sum(ycelular[2],ycelular[4])
c18=ycelular[3]/sum(ycelular[1],ycelular[3])
d18=ycelular[4]/sum(ycelular[2],ycelular[4])
mat18=100*matrix(c(a18,b18,c18,d18),2,2);mat18


colnames(mat)=c("Não Tem","Tem")
rownames(mat1)=c("Não Tem","Tem")
```

```
colnames(mat1)=c("Branca","Não Branca")
rownames(mat2)=c("Não Tem","Tem")
colnames(mat2)=c("Urbana", "Rural")
rownames(mat3)=c("Não Tem","Tem")
colnames(mat3)=c("Masculino", "Feminino")
rownames(mat4)=c("Não Tem","Tem")
colnames(mat4)=c("Sim", "Não")
rownames(mat5)=c("Não Tem","Tem")
colnames(mat5)=c("Rede*", "Fossas, valas e rio")
rownames(mat6)=c("Não Tem","Tem")
colnames(mat6)=c("Sim", "Não")
rownames(mat7)=c("Não Tem","Tem")
colnames(mat7)=c("Sim", "Não")
rownames(mat8)=c("Não Tem","Tem")
colnames(mat8)=c("<0.5", "0.5|-1.5",">1.5")
rownames(mat9)=c("Não Tem","Tem")
colnames(mat9)=c("5", "6","7","8","9","10")
rownames(mat10)=c("Não Tem","Tem")
colnames(mat10)=c("Nenhum", "1","2 ou mais")
rownames(mat11)=c("Não Tem","Tem")
colnames(mat11)=c("Sim", "Não")
rownames(mat12)=c("Não Tem","Tem")
colnames(mat12)=c("Sim", "Não")
rownames(mat13)=c("Não Tem","Tem")
colnames(mat13)=c("Sim", "Não")
rownames(mat14)=c("Não Tem","Tem")
colnames(mat14)=c("Sim", "Não")
rownames(mat15)=c("Não Tem","Tem")
colnames(mat15)=c("Sim", "Não")
rownames(mat16)=c("Não Tem","Tem")
colnames(mat16)=c("Sim", "Não")
rownames(mat17)=c("Não Tem","Tem")
colnames(mat17)=c("Sim", "Não")
rownames(mat18)=c("Não Tem","Tem")
colnames(mat18)=c("Sim", "Não")
#=====================================================================#
#                Gráficos das Prorpoçoes (cruzadas)
#=====================================================================#
pdf("mat.pdf")
barplot(mat,main=c("Registro de Nascimento"),beside=T,legend=c("Não
Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
```

```
dev.off()
pdf("mat1b.pdf")
barplot(mat1,main=c("Cor"),beside=T,legend=c("Não Tem" ,"Tem"),col=
c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat2b.pdf")
barplot(mat2,main=c("Situação do domicilio"),beside=T,legend=c("Não
Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat3b.pdf")
barplot(mat3,main=c("Sexo"),beside=T,legend=c("Não Tem" ,"Tem"),col=
c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat4b.pdf")
barplot(mat4,main=c("Sabe ler/escrever"),beside=T,legend=c("Não Tem"
,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat5b.pdf")
barplot(mat5,main=c("Tipo de esgotamento sanitário"),beside=T,legend
=c("Não Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat6b.pdf")
barplot(mat6,main=c("Abastecimento de Água"),beside=T,legend=c("Não
Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat7b.pdf")
barplot(mat7,main=c("Existência de energia elétrica"),beside=T,legend
=c("Não Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat8b.pdf")
barplot(mat8,main=c("Rendimento domiciliar per capito"),beside=T,
legend=c("Não Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),
ylab="%")
dev.off()
pdf("mat9b.pdf")
barplot(mat9,main=c("Idade (em anos)"),beside=T,legend=c("Não Tem" ,
"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat10b.pdf")
barplot(mat10,main=c("Número de Banheiros"),beside=T,legend=c("Não Tem"
,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
```

```
dev.off()
pdf("mat11b.pdf")
barplot(mat11,main=c("Existência de microcomputador"),beside=T,legend=
c("Não Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat12b.pdf")
barplot(mat12,main=c("Existência de moto"),beside=T,legend=c("Não Tem"
,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat13b.pdf")
barplot(mat13,main=c("Existência de carro"),beside=T,legend=c("Não Tem"
,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat14b.pdf")
barplot(mat14,main=c("Existência de radio"),beside=T,legend=c("Não Tem"
,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat15b.pdf")
barplot(mat15,main=c("Existência de tv"),beside=T,legend=c("Não Tem" ,
"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat16b.pdf")
barplot(mat16,main=c("Existência de máquina de lavar"),beside=T,legend
=c("Não Tem" ,"Tem"),col=c("black","white"),ylim=c(0,119),ylab="%")
dev.off()
pdf("mat17b.pdf")
barplot(mat17,main=c("Existência de geladeira"),beside=T,legend=c("Não
Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
pdf("mat18b.pdf")
barplot(mat18,main=c("Existência de celular"),beside=T,legend=c("Não
Tem" ,"Tem"),col=c("black","white"),ylim=c(0,100),ylab="%")
dev.off()
#=====================================================================#
#                    Seleção de Modelo GLM (logit)
#=====================================================================#
maior<-glm(y~cor+domicilio+sexo+ler+esgoto+agua+
energia+renda+idade+banheiro+computador+
moto+carro+radio+tv+maquina+geladeira+celular
, family = binomial(link=logit), na.action = na.exclude)
fit_logit=step(maior);fit_logit
```

```
#=====================================================================#
#      Teste de modelos GLM  fução de ligação "logit" fit(*)
#=====================================================================#
fit1 = glm(y ~ domicilio + ler + agua + idade + banheiro + computador
+moto + radio + tv + geladeira + celular, family = binomial(link =
logit))
fit2 = glm(y ~ domicilio + ler + agua + energia + idade + banheiro +
computador + moto + radio + tv + maquina + geladeira + celular, family
= binomial(link = logit))
fit3 = glm(y ~ domicilio + sexo + ler + agua + energia + idade +
banheiro + computador + moto + radio + tv + maquina + geladeira +
celular, family = binomial(link = logit))
fit4 = glm(y ~ cor + domicilio + sexo + ler + agua + energia + idade
+ banheiro + computador + moto + radio + tv + maquina + geladeira +
celular, family = binomial(link = logit))
fit5 = glm(y ~ cor + domicilio + sexo + ler + agua + energia + idade
+ banheiro + computador + moto + carro + radio + tv + maquina +
geladeira + celular, family = binomial(link = logit))
fit6 = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ idade + banheiro + computador + moto + carro + radio + tv +
maquina + geladeira + celular, family = binomial(link = logit))
fit7 = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ renda + idade + banheiro + computador + moto + carro + radio +
tv + maquina + geladeira + celular, family = binomial(link = logit))
#=====================================================================#
#                  Representatividade do Modelo GLM
#=====================================================================#
phi1<-summary(fit1)$dispersion
d1=summary(fit1)$deviance/phi1
x2.1=qchisq(0.95,summary(fit1)$df.residual)

phi2<-summary(fit2)$dispersion
d2=summary(fit2)$deviance/phi2
x2.2=qchisq(0.95,summary(fit2)$df.residual)

phi3<-summary(fit3)$dispersion
d3=summary(fit3)$deviance/phi3
x2.3=qchisq(0.95,summary(fit3)$df.residual)

phi4<-summary(fit4)$dispersion
d4=summary(fit4)$deviance/phi4
```

```
x2.4=qchisq(0.95,summary(fit4)$df.residual)

phi5<-summary(fit5)$dispersion
d5=summary(fit5)$deviance/phi5
x2.5=qchisq(0.95,summary(fit5)$df.residual)

phi6<-summary(fit6)$dispersion
d6=summary(fit6)$deviance/phi6
x2.6=qchisq(0.95,summary(fit6)$df.residual)

phi7<-summary(fit7)$dispersion
d7=summary(fit7)$deviance/phi7
x2.7=qchisq(0.95,summary(fit7)$df.residual)

D=c(d1,d2,d3,d4,d5,d6,d7);D
x2=c(x2.1,x2.2,x2.3,x2.4,x2.5,x2.6,x2.7);x2
desvio=cbind(D,x2);desvio
library(xtable)
xtable(desvio)
#=========================================================================#
#                            Seleção de Modelo GLM (probit)
#=========================================================================#
maior1<-glm(y~ cor+domicilio+sexo+ler+esgoto+agua+
energia+renda+idade+banheiro+computador+
moto+carro+radio+tv+maquina+geladeira+celular
, family = binomial(link=probit), na.action = na.exclude)
fit_probit=step(maior1);fit_probit
#=========================================================================#
#      Teste de modelos GLM fução de ligação "probit" fit(*1)
#=========================================================================#
fit1p = glm(y ~ domicilio + ler + agua + idade + banheiro + computador
+moto + radio + tv + geladeira + celular, family = binomial(link =
probit))
fit2p = glm(y ~ domicilio + ler + agua + idade + banheiro + computador
+ moto + radio + tv + maquina + geladeira + celular, family = binomial
(link = probit))
fit3p = glm(y ~ domicilio + ler + agua + energia + idade + banheiro +
computador + moto + radio + tv + maquina + geladeira + celular, family
= binomial(link = probit))
fit4p = glm(y ~ domicilio + sexo + ler + agua + energia + idade +
banheiro + computador + moto + radio + tv + maquina + geladeira +
```

```
celular, family = binomial(link = probit))
fit5p = glm(y ~ cor + domicilio + sexo + ler + agua + energia + idade
+ banheiro + computador + moto + radio + tv + maquina + geladeira +
celular, family = binomial(link = probit))
fit6p = glm(y ~ cor + domicilio + sexo + ler + agua + energia + idade
+ banheiro + computador + moto + carro + radio + tv + maquina +
geladeira + celular, family = binomial(link = probit))
fit7p = glm(y ~ cor + domicilio + sexo + ler + esgoto + agua + energia
+ idade + banheiro + computador + moto + carro + radio + tv +
maquina + geladeira + celular, family = binomial(link = probit))
#=========================================================================#
#                       Representatividade do modelo
#=========================================================================#
phi1p<-summary(fit1p)$dispersion
d1p=summary(fit1p)$deviance/phi1p
x2.1p=qchisq(0.95,summary(fit1p)$df.residual)

phi2p<-summary(fit2p)$dispersion
d2p=summary(fit2p)$deviance/phi2p
x2.2p=qchisq(0.95,summary(fit2p)$df.residual)

phi3p<-summary(fit3p)$dispersion
d3p=summary(fit3p)$deviance/phi3p
x2.3p=qchisq(0.95,summary(fit3p)$df.residual)

phi4p<-summary(fit4p)$dispersion
d4p=summary(fit4p)$deviance/phi4p
x2.4p=qchisq(0.95,summary(fit4p)$df.residual)

phi5p<-summary(fit5p)$dispersion
d5p=summary(fit5p)$deviance/phi5p
x2.5p=qchisq(0.95,summary(fit5p)$df.residual)

phi6p<-summary(fit6p)$dispersion
d6p=summary(fit6p)$deviance/phi6p
x2.6p=qchisq(0.95,summary(fit6p)$df.residual)

phi7p<-summary(fit7p)$dispersion
d7p=summary(fit7p)$deviance/phi7p
x2.7p=qchisq(0.95,summary(fit7p)$df.residual)
```

```
Dp=c(d1p,d2p,d3p,d4p,d5p,d6p,d7p);Dp
x2p=c(x2.1p,x2.2p,x2.3p,x2.4p,x2.5p,x2.6p,x2.7p);x2p
desviop=cbind(Dp,x2p);desviop
library(xtable)
xtable(desviop)

#=====================================================================#

#                        Seleção de Modelo (cauchit)

#=====================================================================#
maior2<-glm(y~cor+domicilio+sexo+ler+esgoto+agua+
energia+renda+idade+banheiro+computador+
moto+carro+radio+tv+maquina+geladeira+celular
, family = binomial(link=cauchit), na.action = na.exclude)
modelo2=step(menor2,scope=list(upper=maior2),direction="both");modelo2
fit_cauchit=step(maior2);fit_cauchit
#=====================================================================#
#        Teste de modelos:  Fução de ligação "cauchit" fit(*2)
#=====================================================================#
#Warning messages:
#1: glm.fit: algorithm did not converge
#2: glm.fit: fitted probabilities numerically 0 or 1 occurred

#=====================================================================#
#                    Significância das variáveis
#=====================================================================#
# GLM Logit
xtable(summary(maior))
xtable(summary(fit1))
xtable(summary(fit2b))
xtable(summary(fit3b))
xtable(summary(fit4b))
xtable(summary(fit5b))
xtable(summary(fit6b))
xtable(summary(fit7b))
# GLM Probit
xtable(summary(maior1))
xtable(summary(fit1p))
xtable(summary(fit2p))
xtable(summary(fit3p))
```

```
xtable(summary(fit4p))
xtable(summary(fit5p))
xtable(summary(fit6p))
xtable(summary(fit7p))
#=========================================================================#
#                          Comparação dos AIC's
#=========================================================================#
# Logit
aic=c(
summary(fit1)$aic,
summary(fit2)$aic,
summary(fit3)$aic,
summary(fit4)$aic,
summary(fit5)$aic,
summary(fit6)$aic,
summary(fit7)$aic
);aic
# Probit
aicp=c(
summary(fit1p)$aic,
summary(fit2p)$aic,
summary(fit3p)$aic,
summary(fit4p)$aic,
summary(fit5p)$aic,
summary(fit6p)$aic,
summary(fit7p)$aic
);aicp
require(xtable)
library(xtable)
aics=cbind(aic,aicp);aics
xtable(aics,digits=3) # para latex
comparacao=cbind(aic,D,x2,aicp,Dp,x2p)
xtable(comparacao)
#=========================================================================#
#                      Padronização dos resíduos
#=========================================================================#
library(car)
library(statmod)
# Logit
res1=qres.binom(fit1)
res2=qres.binom(fit2)
```

```
res3=qres.binom(fit3)
res4=qres.binom(fit4)
res5=qres.binom(fit5)
res6=qres.binom(fit6)
res7=qres.binom(fit7)
resm=qres.binom(maior)
# Probit
res1p=qres.binom(fit1p)
res2p=qres.binom(fit2p)
res3p=qres.binom(fit3p)
res4p=qres.binom(fit4p)
res5p=qres.binom(fit5p)
res6p=qres.binom(fit6p)
res7p=qres.binom(fit7p)
resm1=qres.binom(maior1)
#=====================================================================#
#                    Resíduos
#=====================================================================#
# Logit
pdf("res1b.pdf")
plot(res1,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res2b.pdf")
plot(res2,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res3b.pdf")
plot(res3,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res4b.pdf")
plot(res4,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res5b.pdf")
plot(res5,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("res6b.pdf")
plot(res6,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
```

```
"Logit")
dev.off()
pdf("res7b.pdf")
plot(res7,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
pdf("resmb.pdf")
plot(resm,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Logit")
dev.off()
# Probit
pdf("res1pb.pdf")
plot(res1p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res2pb.pdf")
plot(res2p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res3pb.pdf")
plot(res3p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res4pb.pdf")
plot(res4p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res5pb.pdf")
plot(res5p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res6pb.pdf")
plot(res6p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("res7pb.pdf")
plot(res7p,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
"Probit")
dev.off()
pdf("resm1b.pdf")
plot(resm1,xlab=c("Observações"), ylab=c("Resíduos"),pch=".",main=
```

```
"Probit")
dev.off()
#=========================================================================#
# Autocorrelação
#=========================================================================#
pdf("acf1b.pdf")
acf(res1,main="Logit",xlab=c("Defasagem"))
dev.off()
pdf("acf1pb.pdf")
acf(res1p,main="Probit",xlab=c("Defasagem"))
dev.off()
pdf("acfmb.pdf")
acf(resm,main="Logit",xlab=c("Defasagem"))
dev.off()
pdf("acfm1b.pdf")
acf(resm1,main="Probit",xlab=c("Defasagem"))
dev.off()
#=========================================================================#
#   Autocorrelação Parcial
#=========================================================================#
pdf("pacf1b.pdf")
pacf(res1,main="Logit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
pdf("pacf1pb.pdf")
pacf(res1p,main="Probit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
pdf("pacfmb.pdf")
pacf(resm,main="Logit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
pdf("pacfm1b.pdf")
pacf(resm1,main="Probit",xlab=c("Defasagem"),ylab="PACF")
dev.off()
#=========================================================================#
# Envelope
#=========================================================================#
# Logit
pdf("env1b.pdf")
qqPlot(res1,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env2b.pdf")
qqPlot(res2,main="Logit", ylab="Resíduos",xlab="Quantis")
```

```
dev.off()
pdf("env3b.pdf")
qqPlot(res3,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env4b.pdf")
qqPlot(res4,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env5b.pdf")
qqPlot(res5,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env6b.pdf")
qqPlot(res6,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env7b.pdf")
qqPlot(res7,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("envmb.pdf")
qqPlot(resm,main="Logit", ylab="Resíduos",xlab="Quantis")
dev.off()
# Probit
pdf("env1pb.pdf")
qqPlot(res1p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env2pb.pdf")
qqPlot(res2p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env3pb.pdf")
qqPlot(res3p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env4pb.pdf")
qqPlot(res4p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env5pb.pdf")
qqPlot(res5p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env6pb.pdf")
qqPlot(res6p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
pdf("env7pb.pdf")
qqPlot(res7p,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
```

```
pdf("envmpb.pdf")
qqPlot(resm1,main="Probit", ylab="Resíduos",xlab="Quantis")
dev.off()
#=========================================================================#
#                           Razão de chance
#=========================================================================#
OR1=cbind(Estimativa=exp(summary(fit1)$coefficients)[,1],
exp(confint(fit1)))
ORm=cbind(Estimativa=exp(summary(maior)$coefficients)[,1],
exp(confint(maior)))
OR1p=cbind(Estimativa=exp(summary(fit1p)$coefficients)[,1],
exp(confint(fit1p)))
ORmp=cbind(Estimativa=exp(summary(maior1)$coefficients)[,1],
exp(confint(maior1)))
xtable(OR1)
xtable(ORm)
xtable(OR1p)
xtable(ORmp)
#*************************************************************************#
```