

Clarissa Cardoso Oesselmann

**Absenteísmo escolar na Amazônia: uma
aplicação com dados de contagem**

Niterói - RJ, Brasil

16 de Dezembro de 2013

Clarissa Cardoso Oesselmann

**Absenteísmo escolar na Amazônia:
uma aplicação com dados de
contagem**

Trabalho de Conclusão de Curso

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

Orientadora: Profa. Ludmilla da Silva Viana Jacobson

Niterói - RJ, Brasil

16 de Dezembro de 2013

Clarissa Cardoso Oesselmann

**Absenteísmo escolar na Amazônia: uma
aplicação com dados de contagem**

Monografia de Projeto Final de Graduação sob o título “*Absenteísmo escolar na Amazônia: uma aplicação com dados de contagem*”, defendida por Clarissa Cardoso Oesselmann e aprovada em 16 de Dezembro de 2013, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Ludmilla da Silva Viana Jacobson
Orientadora
Departamento de Estatística – UFF

Profa. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Prof. Jony Arrais Pinto Junior
Departamento de Estatística – UFF

Niterói, 16 de Dezembro de 2013

Oesselmann, Clarissa Cardoso

Absenteísmo escolar na Amazônia: uma aplicação com dados de contagem / Clarissa Cardoso Oesselmann; Ludmilla da Silva Viana Jacobson, orientadora. Niterói, 2013.

64 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2013.

1. Absenteísmo escolar. 2. Tangará da Serra. 3. Modelo de regressão de Poisson. 4. Superdispersão. 5. Modelo Quasi-Poisson. I. Jacobson, Ludmilla da Silva Viana, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Resumo

O absenteísmo escolar é definido como o ato de um estudante se ausentar da escola. A literatura cita vários tipos de absenteísmo, por exemplo, por doença, voluntário, legal e compulsório. O absenteísmo por doença é o que ocorre com mais frequência e o que é mais estudado, pois pode ser visto como um indicador de condição de saúde. O objetivo desse trabalho é avaliar os fatores de risco associados ao absenteísmo escolar em uma escola da Amazônia Brasileira, especialmente os efeitos da poluição atmosférica. Para analisar esses fatores foram utilizados dados de um estudo de painel realizado na cidade de Tangará da Serra, Mato Grosso. O desfecho, ou variável dependente, foi o número de faltas no período escolar durante os meses de seca da região. As variáveis explicativas investigadas se referiam às características individuais, socioeconômicas, de saúde e ambientais. Foi proposto o uso do Modelo de Regressão de Poisson com função de ligação *log*. A superdispersão foi um problema observado nos dados, corrigida utilizando um Modelo “Quasi-Poisson”. Na modelagem dos dados foram propostos seis modelos. Após o ajuste e o diagnóstico dos modelos, decidiu-se pelo modelo com as seguintes variáveis explicativas: Número de comôdos da casa (1 comôdo e 1 banheiro: Referência; Mais comôdos: $\hat{RR} = 0,842$ e $IC95\% = (0,478; 1,485)$); Idade de quem cuida ($\hat{RR} = 0,991$ e $IC95\% = (0,984; 0,997)$); Mora perto de uma plantação (Sim: Referência; Não: $\hat{RR} = 0,749$ e $IC95\% = (0,623; 0,899)$); Nível médio do Pico de Fluxo Expiratório diário ($\hat{RR} = 1,002$ e $IC95\% = (1,002; 1,003)$); Média da Tosse com lag 1 ($\hat{RR} = 1,230$ e $IC95\% = (1,173; 1,291)$); e Nível do poluente PM_{10} com lag 3 ($\hat{RR} = 1,005$ e $IC95\% = (1,0004; 1,010)$). Destes resultados, pode-se concluir que os sintomas respiratórios, principalmente a ocorrência de tosse, mostraram forte associação com o absenteísmo escolar, assim como o poluente PM_{10} .

Palavras-chaves: Absenteísmo Escolar, Tangará da Serra, Modelo de Regressão de Poisson, Superdispersão, Modelo “Quasi-Poisson”.

*Se as coisas são inatingíveis... ora!
Não é motivo para não querê-las...
Que tristes os caminhos, se não fora
A presença distante das estrelas!*

Mário Quintana

Agradecimentos

Agradecer é um sentimento de reconhecimento. Por isso quero reconhecer algumas pessoas, alguns momentos e alguns lugares que me fizeram chegar onde estou.

Quero começar agradecendo à minha vó Osmary, que além de todo o apoio e acolhimento que me deu em todos esses anos, não só de faculdade, mas de vida, me deu palavras de sabedoria e carinho que foram muito importantes pra me guiar na vida!

Sem menos importância, agradeço à minha mãe Nancy e ao meu irmão Daniel pelas conversas engraçadas, sérias e lindas. Sempre estive com vocês, em toda a minha vida, então foi difícil começar num lugar sem vocês por perto, mas sei que lidamos com isso da melhor maneira, e agradeço muito por isso!

À minha dinda linda, ou se preferir Tia Lê, agradeço por tudo, mas principalmente pela amizade! La e Lu eu agradeço por serem as primas chatas de sempre comigo!

Quero agradecer ao meu pai Gabriele, que nessa fase que passou, me desculpou, cuidou, guiou e acrescentou mais o carinho de duas pessoas: Monique e Dante. Obrigada pelas ideias, pelos passeios e pela presença!

Quero agradecer em geral, às minhas famílias, de longe e de perto, presentes e não presentes, por todo o carinho e o apoio.

Agora, vem a parte difícil, que é agradecer a todos que foram importantes para a conclusão desse trabalho e até desse curso. Espero não esquecer de ninguém, por isso, desde já agradeço a todos que sentiram que fizeram parte dessa minha jornada!

O que não é possível esquecer de agradecer é aos professores e professoras que foram tão importantes para a minha formação. À Ludmilla, minha maravilhosa orientadora, eu agradeço pelos conselhos e pela ajuda, não só durante esse trabalho, mas sempre quando precisei.

Também quero agradecer a alguns professores, Jony, Jéssica, Valentin, Leo, Núbia e muitos outros que durante o meu curso me ensinaram, ajudaram, aconselharam, fizeram sorrir e se importaram comigo.

É difícil colocar no papel todos os agradecimentos que tenho que dar aos meus amigos e colegas desse curso de Estatística. Por exemplo, gostaria de agradecer pelos momentos que passamos juntos, nas salas de aula no IME no Valonguinho, nos famosos degraus escada, na sala do sétimo, no laboratório do 5º andar, no plaza, no cinema, na cantareira, no bandejão e até no DCE! Em todos esse lugares tenho lembranças que estão marcadas para sempre!

Sei que não é possível lembrar de todos, mas aqui estão algumas pessoas que estiveram comigo nesse período de curso e que eu agradeço demais, começando pelas meninas: Paola, Nadine, Dani, Thalita, Leo, Carol, Marcela, Kiese, Luciana, Keilane e Andressa, e agora os meninos: Gui, Everson, Bruno, Pablo, Natan, Caio, Sérgio, Welington, Lucas (Nariz), Lucas (Orelha), Renato, China, Mayro, Victor, Guilherme, Fábio, Evandro, Riba, Cidade, Barrientos e Rapha. Todos acima são importantes pra mim e fazem parte das lembranças que tenho da graduação.

Quero agradecer em especial ao meu grupo de amigos que fiz, que são muito importantes e especiais, por todos os rodízios, por todas as saídas, por todas as idas à minha casa, por todos os jogos que já jogamos, por todas as idas ao cinema, por todas discussões sobre onde ir ou quando ir, por todas as conversas, por todas as montagens e por todas as vezes que vocês me fizeram rir e me apoiaram! Obrigada por tudo! Vocês tornaram os meus anos nessa faculdade muito especiais e espero poder levar vocês pra sempre comigo!

Também agradeço às pessoas que conheci na seção 34 do CHM da Marinha, onde fiz o meu primeiro estágio. Obrigada pelos ensinamentos, pelas risadas e pelo o carinho que recebi de todos! E quero agradecer também por terem me dado dois novos amigos que fiz Jéssica e Rafael!

Outras pessoas que não poderia esquecer de agradecer, são os amigos que tenho, que me ajudaram a chegar aqui e que ainda me alegram e me acompanham!

E por último, mas não menos importante, quero agradecer a Deus, por ter me dado todas as pessoas e momentos que já descrevi acima, por ter me dado paciência, sabedoria e coragem para ter feito cada escolha que fiz e que ainda vou fazer.

E obrigada a você que teve paciência para ler o agradecimento até aqui!

Sumário

Lista de Figuras

Lista de Tabelas

1	Introdução	p. 12
1.1	Justificativa	p. 15
2	Objetivos	p. 17
2.1	Geral	p. 17
2.2	Específicos	p. 17
3	Material e Métodos	p. 18
3.1	Absenteísmo na Amazônia	p. 18
3.2	Modelos Lineares Generalizados	p. 23
3.2.1	Definição	p. 23
3.2.2	Modelo de Regressão de Poisson	p. 25
3.2.3	Medidas de Qualidade de Ajuste	p. 27
3.2.4	Diagnóstico do Modelo	p. 30
3.2.5	Superdispersão	p. 32
4	Resultados	p. 35
4.1	Análise Exploratória	p. 35
4.2	Modelos de Regressão	p. 36
4.2.1	Comparação entre os modelos ajustados - Envelope Simulado . .	p. 44

5 Conclusão	p. 48
Referências	p. 50
Anexo A - Análise Exploratória	p. 52
Anexo B - Modelo de Regressão Quasi-Poisson	p. 60

Lista de Figuras

1	Modelo teórico para avaliar o absentismo escolar.	p. 23
2	Gráficos de diagnóstico do modelo base	p. 38
3	Gráficos do diagnóstico do modelo base com a variável de sintoma. . . .	p. 40
4	Efeitos do poluente PM_{10} no absentismo, para o incremento de 10 unidades no nível do poluente, considerando as defasagens. Tangará da Serra, 2008.	p. 42
5	Gráficos do diagnóstico do modelo com a variável de poluição	p. 43
6	Gráficos dos Envelopes Simulados para os modelos estimados.	p. 45
7	Efeitos do Modelo 6.	p. 47

Lista de Tabelas

1	Resumo da construção das variáveis Absenteísmo, Exposições, Sintomas Respiratórios e Pico de Fluxo Expiratório	p. 22
2	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para o modelo base com e sem <i>outlier</i> . Tangará da Serra, 2008.	p. 38
3	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para a adição das variáveis de Sintomas Respiratórios no modelo base com e sem <i>outlier</i> . Tangará da Serra, 2008.	p. 39
4	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para o modelo base com a variável de Sintoma, com e sem <i>outlier</i> . Tangará da Serra, 2008.	p. 40
5	Efeitos da Poluição no absenteísmo escolar para o incremento de uma unidade nos níveis dos poluentes, considerando modelos com e sem <i>outlier</i> .	p. 41
6	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para o novo modelo com a variável de poluição com e sem <i>outlier</i> . Tangará da Serra, 2008.	p. 43
7	Estatísticas Descritivas das Variáveis Quantitativas de Condições de Saúde e da Variável Resposta Falta. Tangará da Serra, 2008.	p. 52
8	Estatísticas Descritivas das Variáveis Contínuas Ambientais e de Poluição. Tangará da Serra, 2008.	p. 53
9	Estatísticas Descritivas das Variáveis Contínuas de Características Individuais. Tangará da Serra, 2008.	p. 53
10	Correlações entre as variáveis quantitativas. Tangará da Serra, 2008. . .	p. 54
11	Continuação das correlações entre as variáveis quantitativas. Tangará da Serra, 2008.	p. 54

12	Frequências das Variáveis Categóricas. Tangará da Serra, 2008.	p. 55
13	Frequências das Variáveis Categóricas 2. Tangará da Serra, 2008.	p. 56
14	Médias das Variáveis Ambientais e de Poluição com lag 1 por quantidade de faltas. Tangará da Serra, 2008.	p. 57
15	Médias dos valores da soma das Variáveis de Condições de Saúde. Tangará da Serra, 2008.	p. 57
16	Média das médias ponderadas das Variáveis de Condições de Saúde. Tangará da Serra, 2008.	p. 58
17	Média das médias ponderadas com lag 1 das Variáveis de Condições de Saúde. Tangará da Serra, 2008.	p. 58
18	Média dos valores das Variáveis de Espirometria de acordo com a quantidade de Faltas. Tangará da Serra, 2008.	p. 59
19	Média dos valores das Variáveis de Características Individuais de acordo com a quantidade de Faltas. Tangará da Serra, 2008.	p. 59
20	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor do Teste de Wald. Tangará da Serra, 2008.	p. 60
21	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor do Teste de Wald 2. Tangará da Serra, 2008.	p. 61
22	Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor do Teste de Wald 3. Tangará da Serra, 2008.	p. 62
23	Tabela de Correlação das Variáveis Quantitativas e Significantes no Teste de Wald. Tangará da Serra, 2008.	p. 63

1 Introdução

Na palavra absenteísmo o sufixo “ísmo”, que designa exagero ou excesso, indica a repetição de faltas em locais cuja frequência é necessária ou obrigatória (Neto, 2006 [1]). Então, o absenteísmo pode ser definido como o ato de um trabalhador ou um estudante, se ausentar do trabalho ou da escola.

Segundo Neto (2006) [1], as causas do absenteísmo podem ser divididas de tais formas: absenteísmo por doença (qualquer doença diagnosticada por um procedimento médico; é o tipo de absenteísmo mais grave devido ao volume de faltas que proporciona), absenteísmo voluntário (por razões particulares de exclusiva responsabilidade da pessoa), absenteísmo por patologia profissional (acidentes ou doenças desenvolvidas no ambiente de trabalho), absenteísmo legal (amparados pela lei, como licença maternidade, doação de sangue e serviço militar) e absenteísmo compulsório (imposto por lei, mesmo que a pessoa não deseje, como suspensão e prisão).

Como absenteísmo por doença é o que ocorre com mais frequência, o número de estudos sobre esse tipo de abenteísmo é também maior. As doenças que mais aparecem em estudos são aquelas relacionadas aos problemas respiratórios, tanto para adultos quanto para crianças. Não só é estudado a associação entre o absenteísmo e as doenças, como também o absenteísmo com as causas dessas doenças.

Entre os adultos, um exemplo de estudo que verificou o absenteísmo por motivo de doença é o de Ervasti et al. (2012) [2]. Os autores utilizaram três grupos de covariáveis para estudar o absenteísmo de professores de diversas escolas, a saber: a qualidade do ar, como uma forma de prever problemas de saúde; a satisfação com o emprego; e a satisfação com o ambiente em que trabalha. Foi constatado que houve uma diminuição no risco de absenteísmo entre os professores que trabalhavam em escolas com uma qualidade do ar considerada boa (Razão de Chance, RC = 0,6), quando comparados com professores que trabalhavam em escolas com qualidade do ar ruim. Neste estudo, algumas escolas sofreram intervenções e, em seguida, as mudanças foram avaliadas pelos professores e alunos.

Os resultados mostraram que as escolas que passaram por intervenção e apresentaram mudanças negativas segundo a percepção dos alunos, aumentaram o risco ($RC = 1,8$) a curto prazo do absenteísmo entre os professores, quando comparados com aqueles professores das escolas que não passaram por mudanças. Da mesma forma, o bullying ($RC = 1,5$) entre os alunos também foi fator de risco para o absenteísmo dos professores. O ambiente de trabalho não teve associação com as faltas a longo prazo.

Entre as crianças, uma grande parte dos estudos é focado no absenteísmo por doenças respiratórias. Existem os estudos que utilizaram como variáveis explicativas do absenteísmo somente as variáveis ambientais (Makino K., 2000 [3]; Gilliland et al., 2001 [4] e Berhane et al., 2002 [5]), como por exemplo as concentrações de poluentes, temperatura e umidade relativa. Outras pesquisas, além de levarem em conta as variáveis ambientais, também consideraram os problemas respiratórios da criança, como asma, sinusite, pneumonia, tosse, problemas nasais e outros. Existem ainda outras pesquisas que além das variáveis ambientais e sintomáticas incluíram as variáveis socioeconômicas, como sexo, idade, características do ambiente em que vive, etc. (Hwang et al., 2000 [6]). Há estudos que associam as doenças respiratórias com o tabagismo (Gilliland et al., 2003 [7]; Levy et al., 2011 [8]) e outros que além do tabagismo também associam o absenteísmo com as características socioeconômicas (Meng et al., 2012 [9]).

Makino (2000) [3] afirmou que a prevalência de absenteísmo teve uma significativa relação com os poluentes MPS (Material Particulado em Suspensão) e NO_2 , umidade relativa e temperatura. Os poluentes e a umidade relativa mostraram uma relação positiva com o absenteísmo. Por outro lado, a temperatura apresentou uma relação negativa.

Os resultados encontrados por Gilliland et al. (2001) [4] mostraram que uma mudança a curto prazo do nível do poluente ozônio (O_3) tem uma associação com o aumento do absenteísmo escolar. A cada aumento de 20 ppb de O_3 , o absenteísmo aumentou em 62,9%. Enquanto as mudanças dos níveis dos poluentes NO_2 e material particulado com diâmetro aerodinâmico menor que $10 \mu m$ (PM_{10}) não apresentaram associação significativa com o absenteísmo.

Berhane et al. (2002) [5] fizeram a avaliação do absenteísmo escolar tanto com a exposição aguda aos poluentes atmosféricos quanto com a exposição crônica. Quanto ao efeito agudo, os resultados mostraram que a incidência do absenteísmo por doenças está associada positivamente com aumentos diários de O_3 , enquanto que não há associação com o aumento diário do nível do PM_{10} . No que se refere ao efeito crônico, comunidades com maiores valores do PM_{10} a longo prazo tendem a ter maiores faltas relacionadas a

doenças.

No estudo de Hwang et al. (2000) [6] foram utilizadas três abordagens para a modelagem dos dados. Na primeira, os autores estimaram um modelo de Poisson, cujos dados foram agregados segundo o dia. Na segunda, também foi estimado um modelo de Poisson, porém os dados foram agregados por indivíduo. E na última, o modelo estimado pela segunda abordagem foi ajustado por características individuais. Somente a última abordagem apresentou resultados estatisticamente significativos. O resultado mostrou uma associação entre absenteísmo e os poluentes óxido de nitrogênio (NO) e dióxido de nitrogênio (NO_2). O modelo estimou um risco relativo de absenteísmo escolar por doença de 1.11 e 1.23 para um aumento de 10-ppb de exposição de cada poluente, respectivamente. Também verificou-se associação com as variáveis temperatura, notas escolares, sintomas pessoais como tosse de dia ou de noite e histórico familiar de doenças respiratórias.

O estudo de Gilliland et al. (2003) [7] demonstrou que a exposição a ambientes com fumaça de cigarro está associada com o risco de absenteísmo escolar devido a doenças respiratórias (Risco Relativo, $RR = 1,27$). O estudo também mostrou que crianças que moravam em uma casa com 2 ou mais fumantes aumentaram o risco de absenteísmo ($RR = 1,75$). Quando comparadas com crianças não asmáticas que viviam em um ambiente sem exposição à fumaça de cigarro, crianças com asma aumentaram o risco de absenteísmo escolar por doenças respiratórias quando expostas a 1 ($RR = 2,35$) ou ≥ 2 ($RR = 4,45$) fumantes em casa. Crianças não asmáticas também apresentaram um risco alto se expostas a dois ou mais fumantes em casa ($RR = 1,44$). Por outro lado, o estudo de Levy et al. (2011) [8] não encontrou associação significativa entre o fumo no domicílio e o absenteísmo escolar.

O modelo encontrado por Meng et al. (2012) [9] indicou que crianças com famílias de baixa renda estão mais propensas a faltar aula por causa de asma, do que as crianças de alta renda. Além disso, as crianças mais novas, que tiveram sintomas de asma com mais frequência ou aquelas usando medicamentos para asma estão mais propensas ao absenteísmo escolar.

Os métodos utilizados para encontrar esses resultados foram Modelos de Regressão de Poisson (Hwang et al., 2000 [6]; Makino K., 2000 [3]; Gilliland et al., 2001 [4]; Gilliland et al., 2003 [7]), Modelos de Regressão Logística (Makino K., 2000 [3]; Levy et al., 2011 [8]; Ervasti et al., 2012 [2]), Modelos de Regressão Linear (Makino K., 2000 [3]; Gilliland et al., 2003 [7]; Levy et al., 2011 [8]) e Séries Temporais (Berhane et al., 2002 [5]).

Os estudos apresentados foram realizados nos EUA, Ásia e Europa. No Brasil, um estudo em São Paulo avaliou o absenteísmo no trabalho (Santos et al., 1992 [10]). Não foram encontrados estudos sobre o absenteísmo escolar, principalmente na região Amazônica.

1.1 Justificativa

A maior parte da região Amazônica possui um clima sazonal, caracterizado pelas altas temperaturas o ano inteiro, pela pequena amplitude térmica anual e pelos períodos de seca (Junho, Julho, Agosto e Setembro) e chuva (Dezembro, Janeiro, Fevereiro e Março) bem marcados. É nesse período de seca que se verifica um dos maiores problemas ambientais, as queimadas, que é considerada a maior fonte de produção de gases poluentes (Gonçalves et al. 2012 [11]).

Ribeiro et al. (2002) [12] destacou o papel dos poluentes das queimadas nos agravos à saúde das pessoas que convivem nesse ambiente, sendo os problemas respiratórios um dos principais. Apesar de toda a literatura que indica a associação entre os poluentes e a saúde, quando se fala sobre os efeitos da fumaça das queimadas na saúde os estudos ainda são escassos, principalmente na região Amazônica. E isso pode ser explicado, pois existem inúmeras variáveis que podem influenciar a ocorrência de problemas na saúde e separá-las para determinar o efeito isolado das queimadas é bastante difícil (Gonçalves et al., 2012 [11]). Relacionar então o efeito dos poluentes com o absenteísmo escolar é mais complicado ainda.

De acordo com Jacobson (2013) [13], na Amazônia já foram estudados os efeitos da poluição atmosférica com os seguintes desfechos: pico de fluxo expiratório, atendimentos de emergência por doenças respiratórias, internações hospitalares por doenças respiratórias, prevalência de doenças respiratórias e frequência de micronúcleos em células esfoliadas da mucosa bucal.

Por não existir estudos na Amazônia Brasileira que relacionam o absenteísmo escolar com as variáveis ambientais, sintomáticas e socioeconômicas, que esse trabalho se faz necessário.

Os estudos de Jacobson (2012) [13] e Oliveira et al. (2012) [14], avaliaram o efeito da poluição em uma amostra de crianças residentes em Tangará da Serra (MT). Essas crianças foram acompanhadas diariamente durante 4 meses, e registrados além do pico de fluxo expiratório, informações sobre o absenteísmo escolar, sintomas respiratórios e dados

socioeconômicos. Por ser um estudo de painel ¹, Jacobson (2013) [13] utilizou modelos de efeitos mistos.

Na análise dos dados sobre o absentismo escolar propõem-se o uso de Modelos de Regressão de Poisson, através da agregação dos dados da amostra acompanhada por Jacobson (2012) [13] e Oliveira et al. (2013) [14], como proposto por Hwang et al. (2000) [6].

¹Os dados de estudo de painel possuem estrutura hierárquica, onde as medidas repetidas são aninhadas em grupos.

2 Objetivos

2.1 Geral

Avaliar os fatores de risco associados ao absenteísmo escolar em uma escola da Amazônia Brasileira, acompanhada durante os meses de seca da região, especialmente os efeitos da poluição atmosférica.

2.2 Específicos

- Estimar Modelos de Regressão de Poisson, considerando como variável resposta o número de faltas no período letivo;
- Explorar a melhor construção das variáveis explicativas;
- Explorar a defasagem das variáveis explicativas, no que se refere as questões ambientais e de sintomas respiratórios;
- Avaliar a qualidade do ajuste do Modelo de Regressão de Poisson Múltiplo;
- Avaliar a superdispersão dos dados.

3 Material e Métodos

3.1 Absenteísmo na Amazônia

O estudo de painel foi realizado em Tangará da Serra durante os meses de Agosto a Dezembro na Escola Pedro Alberto Tayano, situada no bairro Vila Esmeralda, de 11 de Agosto de 2008 a 12 de Dezembro de 2008. Foi considerado um universo amostral com crianças e adolescentes de 6 a 15 anos de idade. Foram selecionados 25 alunos de cada idade, no entanto somente 234 estudantes iniciaram o estudo. Os responsáveis dos 234 alunos assinaram o Termo de Consentimento Livre e Esclarecido.

Ao trabalhar com o banco de dados, algumas crianças foram excluídas do estudo, aquelas que não responderam ao questionário e aquelas que saíram da escola durante o período de estudo. Ao final, o número de crianças do estudo passou a ser 211.

O banco de dados contém as seguintes variáveis que foram usadas neste projeto:

- Absenteísmo escolar, isto é, se a criança faltou no dia em questão a aula, e é classificada como: 0 (quando foi a aula), 1 (quando faltou um dia de aula) e 2 (quando era um dia sem aula, por exemplo finais de semana, feriado e dia letivo sem atividade escolar);
- Variáveis de poluição e meteorológicas (como níveis dos poluentes PM_{10} , BC e O_3 , temperatura e umidade);
- Variáveis de sintomas respiratórios (tosse, dor de cabeça, coceira no ouvido, coceira na garganta, lacrimejamento, coriza, uso de medicamento, visita a um posto médico);
- Exames subclínicos (Pico de fluxo expiratório diário e Espirometria);
- Características individuais (obtidas através de um questionário).

Inicialmente, o banco de dados foi preparado para este projeto. A primeira mudança foi retirar os dados relacionados aos dias do mês de dezembro, por ser um mês de final de período letivo e o número de faltas ser grande. Uma vez que foi possível perceber que a incidência de absenteísmo escolar nesta amostra é baixa, decidiu-se trabalhar com os dados agregados por aluno.

Para agregar os dados definiu-se dois períodos de tempos diferentes. Um em que são considerados todos os dias de acompanhamento do estudo da criança i (de tamanho t_{1i}) e outro em que são considerados somente os dias de aulas, excluindo o fim de semana e dias sem aula (de tamanho t_{2i}). Neste estudo, o tempo t_{2i} é importante para avaliar a defasagem no tempo das variáveis relacionadas aos poluentes atmosféricos, às meteorológicas e aos sintomas respiratórios.

O absenteísmo escolar é definido como y_{it} e será 1 se o i -ésimo indivíduo faltou no dia t . A variável resposta para o banco agregado é o n° total de faltas do indivíduo i , ou seja, a soma das faltas no período t_2 :

$$y_{i+} = \sum_t^{t_{2i}} y_{it}.$$

Temos p variáveis de poluição, temperatura e umidade, que serão denotadas da seguinte forma:

$$\mathbf{X}_i = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{t1} & x_{t2} & \dots & x_{tj} & \dots & x_{tp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{t_{2i}1} & x_{t_{2i}2} & \dots & x_{t_{2i}j} & \dots & x_{t_{2i}p} \end{pmatrix}_{t_{2i} \times p},$$

sendo que x_{tj} representa uma covariável j no instante t , com $t = 1, 2, \dots, t_{2i}$ e $j = 1, 2, \dots, p$.

Para agregar as variáveis de poluição, temperatura e umidade $\mathbf{x}_i^T = (x_{t1}, x_{t2}, \dots, x_{tp})$ foi utilizada a mesma metodologia do estudo de Hwang et al. (2000) [6]. Foi calculado o nível de exposição da variável \mathbf{x}_j , $j = 1, 2, \dots, p$, para cada indivíduo i , definido como z_{ij} , ponderado pelos dias em que o indivíduo faltou, ou seja, uma média dos valores da variável \mathbf{x}_j ponderado pelos dias em que o indivíduo i faltou:

$$z_{ij} = \frac{1}{\sum_t^{t_{2i}} y_{it}} \sum_t^{t_{2i}} (y_{it} x_{tj}).$$

Para medir o nível de exposição da covariável j da criança i , é feita uma média dos valores da covariável x_{tj} dos dias em que a criança i faltou. Por exemplo, suponhamos que a criança i apresentou os seguintes dados:

t	Faltou (y_{it})	Covariável j (x_{tj})
1	0	29,80
2	1	28,30
3	1	28,60
4	0	24,30
5	1	23,50

em que t é o dia do estudo, y_{it} é a variável que indica se a criança faltou ($y_{it} = 1$) ou não ($y_{it} = 0$) no dia t e x_{tj} é o valor da covariável j no dia t . Nesse caso o período letivo da criança seria $t_{2i} = 5$ e o nível de exposição ficaria da seguinte forma:

$$z_{ij} = \frac{1}{\sum_t^5 y_{it}} \sum_t^5 (y_{it} x_{tj})$$

$$z_{ij} = \frac{1}{(0 + 1 + 1 + 0 + 1)} [(0 \times 29,80) + (1 \times 28,30) + (1 \times 28,60) + (0 \times 24,30) + (1 \times 23,50)]$$

$$z_{ij} = \frac{1}{3} (28,30 + 28,60 + 23,50) = \frac{80,40}{3} = 26,80.$$

Então esse seria o nível de exposição da covariável j para a criança i .

Não necessariamente os valores dos níveis da exposição que poderiam influenciar o absenteísmo são do dia em que ocorreu a falta. Neste sentido, foi também calculado o nível de exposição considerando uma defasagem (lag) de 0 a 5 dias para as variáveis de poluição e de 0 a 2 dias para as metereológicas.

Por exemplo, se formos calcular o nível de exposição da covariável j com lag 1 para a criança i seria da seguinte forma:

$$z_{ij} = \frac{1}{\sum_t^5 y_{it}} \sum_t^5 (y_{it} x_{(t-1)j})$$

$$z_{ij} = \frac{1}{(0 + 1 + 1 + 0 + 1)} [0 + (1 \times 29,80) + (1 \times 28,30) + (0 \times 28,60) + (1 \times 24,30)]$$

$$z_{ij} = \frac{1}{3} (29,80 + 28,30 + 24,30) = \frac{82,40}{3} = 27,47.$$

Para as crianças que não faltaram nenhum dia, ou seja, que o número de faltas conta-

biliza zero, foi considerado como nível de exposição a média da variável em todos os dias de aula.

Para as variáveis de sintomas respiratórios, o procedimento para agregação foi parecido com o que foi feito para o absenteísmo escolar. Foi criado um indicador (s_{it}) que terá valor 1, se o indivíduo i teve sintoma e faltou no mesmo dia e depois realizar o somatório desse indicador:

$$s_i = \sum_t^{t_{2i}} s_{it}.$$

Então se temos uma criança i com os seguintes dados:

t	Faltou (y_{it})	Variável Sintomática	Indicador s_{it}
1	0	1	0
2	1	1	1
3	1	0	0
4	0	0	0
5	1	1	1

em que t é o dia do estudo, y_{it} é a variável que indica se a criança faltou ($y_{it} = 1$) ou não ($y_{it} = 0$) no dia t , a coluna de variável sintomática indica se a criança teve ou não o sintoma em questão no dia t e o indicador $s_{it} = 1$ se a criança i faltou e teve o sintoma em questão no dia t . Logo, para essa criança teríamos o seguinte valor para o indicador:

$$s_i = \sum_t^5 s_{it} = (0 + 1 + 0 + 0 + 1) = 2.$$

Para essas variáveis também foi adotada uma abordagem defasada, ou seja, considerando se o indivíduo teve sintomas em outros dias e faltou hoje. Outro tipo de abordagem sobre as variáveis de sintomas respiratórios foi a de usar o total de vezes que a criança teve o sintoma durante o período do estudo.

O exame subclínico utilizado neste estudo foi a Espirometria, que tem como parâmetros a Capacidade Vital Forçada (CVF), o Volume Expiratório Forçado no 1º segundo (FEV1), a Razão entre o FEV1 e a CVF (FEV1/CVF), o Pico de Fluxo Expiratório (PEFR) e o Fluxo Expiratório Forçado entre 25 e 75% da CVF (FEF25-75%). Essas medidas possuem um valor previsto que cada criança deveria alcançar, criado a partir de características pessoais.

Foi feito um exame com cada criança no começo do estudo e o valor de cada parâmetro

que a criança alcançou foi anotado. Para cada parâmetro da Espirometria, as medidas usadas para cada criança correspondeu a porcentagem do valor que a criança obteve no exame sobre o valor que deveria ser alcançado. Assim, para cada criança tem-se a porcentagem que foi alcançada em cada parâmetro.

Além disso, o estudo coletou diariamente, através de aparelho portátil, três medidas sequenciais do Pico de Fluxo Expiratório de cada criança. Porém, a maior entre as três medidas foi a escolhida. Para esta variável foram calculadas a média simples, a mediana do período do estudo e a média ponderada pelos dias em que ela faltou (através do mesmo método utilizado para o nível de exposição).

A Tabela 1 apresenta o resultado final da construção das variáveis.

Tabela 1: Resumo da construção das variáveis Absenteísmo, Exposições, Sintomas Respiratórios e Pico de Fluxo Expiratório

Variável Original	Variável Transformada	Explicação da Variável Transformada
Se faltou aula ou não	Y_i	Número de faltas que uma criança teve no período de aula.
Valores diários de Temperatura Média e Umidade Relativa (com lag 0, 1 e 2)	Z_i	Média das variáveis ambientais em que a i -ésima criança foi exposta nos dias em que faltou.
Valores diários de PM10 Médio, BC Médio e O3 Máximo (com lag 0, 1, 2, 3, 4 e 5)	Z_i	Nível de exposição da i -ésima criança aos poluentes nos dias em que faltou. Um valor para cada poluente.
Se teve Tosse, Coceira no Ouvido, Coceira na Garganta, Coriza, Dor de Cabeça e Lacrimejamento, se foi ao Posto Médico ou se fez uso de Remédio	S_i	Número total ou médio de ocorrências do sintoma nos dias em que a i -ésima criança faltou. Um valor para cada sintoma.
Medidas diárias de Pico de Fluxo	PF_i	Média Simples ou Mediana dos valores máximos diários do Pico de Fluxo da i -ésima criança. Média do Pico de Fluxo máximo nos dias em que a i -ésima criança faltou.

As variáveis relacionadas às características individuais de cada criança, foram coletadas através de um questionário com diversas questões, que envolviam informações tanto sobre o ambiente em que reside, quanto sobre o histórico de doenças. Foram selecionadas para este estudo aquelas variáveis que, de acordo com a literatura, se relacionavam com o absenteísmo escolar, como, por exemplo, as condições: de moradia, de saneamento e infraestrutura, econômicas, de saúde e ambientais.

A Figura 1 apresenta o modelo teórico para avaliar os fatores de risco para o absenteísmo escolar a partir das variáveis agregadas.

Como o estudo tem como variável resposta uma contagem, ou seja o número de faltas escolares que cada criança obteve, é possível considerar que a variável possui uma distribuição de Poisson. Portanto, um Modelo de Regressão de Poisson é indicado para a

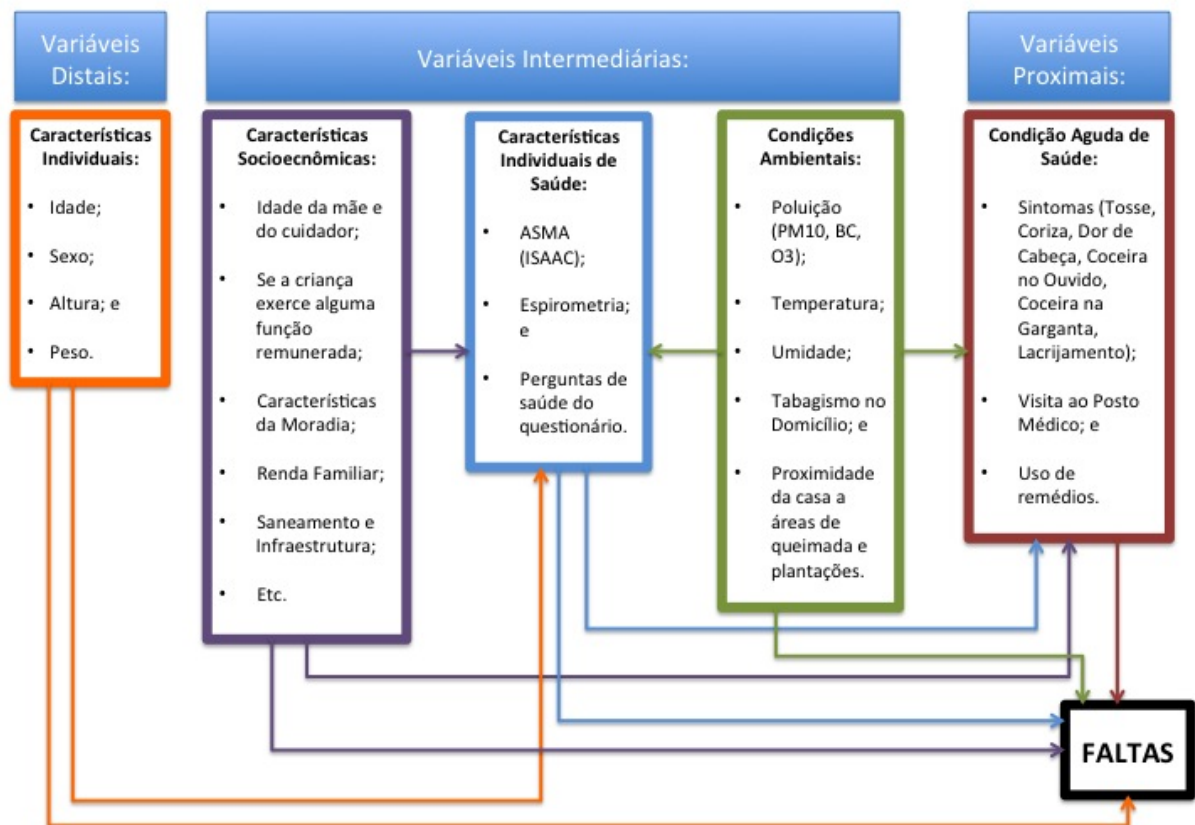


Figura 1: Modelo teórico para avaliar o absenteísmo escolar.

modelagem dos dados.

3.2 Modelos Lineares Generalizados

3.2.1 Definição

Os Modelos Lineares Generalizados (MLG), desenvolvidos por Nelder & Wedderburn (1972), são definidos por uma distribuição de probabilidade para a variável resposta \mathbf{Y} que está associada a um conjunto de variáveis explicativas x_1, \dots, x_p e uma função de ligação. Será útil representar as variáveis resposta e explicativas na forma matricial:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1} \quad e \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n \times p} .$$

Um MLG é formado por três componentes:

1. **Componente Aleatória** é composta por uma variável aleatória (v.a.) \mathbf{Y} com n valores independentes, um vetor de média $\boldsymbol{\mu} = E(\mathbf{Y})$ e uma distribuição de probabilidade pertencente a família exponencial. Notação: Seja um vetor de valores $y = (y_1, y_2, \dots, y_n)^T$ referentes às realizações das v.a.'s \mathbf{Y} independentes e identicamente distribuídas (i.i.d.), com vetor de médias $\boldsymbol{\mu} = E(\mathbf{Y}) = (\mu_1 = E(y_1), \mu_2 = E(y_2), \dots, \mu_n = E(y_n))$.
2. **Componente Sistemática** composta por variáveis explicativas (ou independentes) $x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ (i -ésima coluna da matriz \mathbf{X}), que produzem um preditor linear η_i , dado por:

$$\eta_i = x_i^T \boldsymbol{\beta} = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \sum_{j=0}^p \beta_j x_{ij},$$

sendo $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$ um vetor de parâmetros desconhecidos.

3. **Função de ligação** é uma função monotônica diferenciável, denotada por $g(\mu_i)$, que relaciona a média da variável resposta $\mu_i = E(Y_i)$ ao preditor linear $\eta_i = x_i^T \boldsymbol{\beta}$. Isto é:

$$g(\mu_i) = \eta_i = x_i^T \boldsymbol{\beta}, \forall i = 1, 2, \dots, n.$$

Os Modelos Lineares Generalizados têm como importantes características a hipótese de independência das observações e a distribuição da variável resposta que tem de ser um membro da família exponencial.

Dizemos que uma v.a. contínua \mathbf{Y} com função de densidade de probabilidade $f(y, \theta)$ pertence à família exponencial se a sua função de densidade de probabilidade puder ser expressa da forma:

$$f(y, \theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}.$$

Para modelar dados através de um MLG existem três etapas que devem ser seguidas: formulação de dados, ajustamento do modelo e seleção e validação do modelo.

Para **formulação dos dados** de um modelo MLG basta especificar:

- A distribuição de probabilidade para a variável resposta. É fundamental analisar os dados para observar algumas características importantes;
- Um conjunto de variáveis explicativas quantitativas e/ou qualitativas apropriadas com o problema específico do estudo; e
- Uma função de ligação que relacione a variável resposta com as variáveis explicativas.

A fase do **ajustamento do modelo** passa pela estimação dos parâmetros do modelo, ou seja, dos coeficientes β associados às variáveis explicativas.

Existem diferentes métodos para estimar os parâmetros de modelos estatísticos, sendo que os métodos mais conhecidos são:

- Método de Mínimos Quadrados (MQ); e
- Método de Máxima Verossimilhança (MV).

O método mais usado no ajuste de MLG's é o de Máxima Verossimilhança baseado num método de estimação conhecido como Método de Escore de Fisher. Para encontrar os parâmetros estimados neste método, basta igualar a zero a função de Escore:

$$U = \frac{\partial \log L(\hat{\beta})}{\partial \beta_j} = 0.$$

A fase de **seleção e validação do modelo** tem como objetivo encontrar um modelo que tenha um número de parâmetros que sejam adequados aos dados, detectar as diferenças entre os dados e os valores preditos, verificar a existência de *outliers*, etc. O modelo a ser escolhido como o melhor deve ser um que consiga atingir um equilíbrio entre os fatores: adequabilidade, parcimônia e interpretação.

3.2.2 Modelo de Regressão de Poisson

Neste trabalho propõem-se um modelo em que a componente aleatória \mathbf{Y} tem distribuição de Poisson, ou seja, é uma contagem de acontecimentos aleatórios. Então se \mathbf{Y} tem distribuição de Poisson, com valores em \mathbb{N}_0 , sua função de probabilidade poderá ser escrita da seguinte forma:

$$P(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!}, y = 0, 1, 2, \dots \text{ e } \lambda > 0.$$

Como a média de uma variável aleatória com distribuição de Poisson assume o valor do próprio parâmetro pode-se afirmar que nesse caso $\boldsymbol{\mu} = \boldsymbol{\lambda}$. Então a função de ligação para o Modelo de Regressão de Poisson será uma função g tal que:

$$g(\lambda_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}, i = 1, \dots, n,$$

onde $\mathbf{x}_i^T \boldsymbol{\beta}$ é a componente sistemática do modelo. Ao escrever a função de probabilidade da Poisson na forma da família exponencial é possível encontrar o parâmetro natural da distribuição, que é $b(\lambda) = \ln(\lambda)$:

$$P(Y = y) = \exp(y \ln(\lambda) - \ln(y!) - \lambda),$$

$$a(y) = y; \quad b(\lambda) = \ln(\lambda); \quad c(\lambda) = -\lambda; \quad d(y) = \ln(y!).$$

Assim, a função de ligação canônica para uma componente aleatória com distribuição de Poisson ($\mathbf{Y} \sim Poisson(\lambda)$) é a função de ligação logarítmica:

$$g(\lambda_i) = \ln(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta} \Leftrightarrow \lambda = e^{\mathbf{x}_i^T \boldsymbol{\beta}}.$$

Um modelo assim definido é chamado de Modelo Log-Linear. A função de ligação logarítmica apenas permite valores positivos do parâmetro λ , o que está estruturalmente de acordo com a distribuição de Poisson. Então, podemos escrever o Modelo de Regressão de Poisson da seguinte forma:

$$\ln \lambda_i = \ln n_i + \mathbf{x}_i^T \boldsymbol{\beta}.$$

O termo incluído $\ln n_i$ é chamado de *offset*, ou seja, é uma constante conhecida incorporada ao modelo, que é utilizada como uma forma de padronizar a variável resposta. Por exemplo, para o estudo em questão o *offset* será o tempo letivo de cada criança, ou seja, o t_{2i} que é o período que a criança i participou do estudo sem considerar os finais de semana, feriados e dias sem aula.

No Modelo de Regressão de Poisson, a interpretação para uma variável explicativa x_k é a de que o aumento em uma unidade resulta em um efeito multiplicativo de e^{β_k} na medida μ .

Para estimação dos parâmetros $\boldsymbol{\beta}$ através do método de Máxima Verossimilhança, é preciso conhecer a sua função de verossimilhança:

$$L(\boldsymbol{\lambda}|\mathbf{Y}) = \prod_{i=1}^n e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} = e^{-\sum_{i=1}^n \lambda_i} \prod_{i=1}^n (\lambda_i^{y_i}) \prod_{i=1}^n \left(\frac{1}{y_i!}\right),$$

e a sua log-verossimilhança:

$$\ln L(\boldsymbol{\lambda}|\mathbf{Y}) = \ln(L(\boldsymbol{\lambda}|\mathbf{Y})) = -\sum_{i=1}^n \lambda_i + \sum_{i=1}^n (y_i \ln(\lambda_i)) + \sum_{i=1}^n \ln\left(\frac{1}{y_i!}\right).$$

Substituindo $\lambda_i = e^{x_i^T \boldsymbol{\beta}}$, temos:

$$\ln L(\boldsymbol{\beta}|\mathbf{Y}) = -\sum_{i=1}^n e^{x_i^T \boldsymbol{\beta}} + \sum_{i=1}^n (y_i x_i^T \boldsymbol{\beta}) + \sum_{i=1}^n \frac{1}{y_i!},$$

encontrando a estatística de Escore:

$$U = \frac{\partial \ln L(\boldsymbol{\beta}|\mathbf{Y})}{\partial \beta_j} = -\sum_{i=1}^n x_i^T e^{x_i^T \boldsymbol{\beta}} + \sum_{i=1}^n y_i x_i^T = \sum_{i=1}^n [x_i^T (y_i - e^{x_i^T \boldsymbol{\beta}})].$$

Ao igualar a estatística de Escore a zero, obtém-se um sistema com p equações não-lineares, que são resolvidas por métodos numéricos iterativos, como a Aproximação de Newton-Raphson. Essa aproximação consiste que a partir de um valor inicial suposto $\boldsymbol{\beta}^{(1)}$ sucessivas aproximações são obtidas usando a fórmula abaixo, até a convergência do processo iterativo:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \frac{U^{(m-1)}}{U'^{(m-1)}},$$

ou pode ser reescrito da seguinte forma, como é mostrado em Paula (2013)[15] :

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

em que $z = \boldsymbol{\eta} + W^{-\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} (\mathbf{Y} - \boldsymbol{\lambda})$, $\mathbf{V} = \text{diag}(V_1, \dots, V_n)$, com $V_i = V(\lambda_i) = \frac{d\lambda_i}{d\beta_i}$ e $W = \text{diag}(w_1, \dots, w_n)$, com $w_i = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \lambda_i}{\partial \eta_i}\right)^2$.

Esse método de encontrar as estimativas de $\boldsymbol{\beta}$ é chamado de Método de Escore de Fisher.

3.2.3 Medidas de Qualidade de Ajuste

Existem diversas formas de verificar se o modelo ajustado é bom e adequado suficientemente para os dados. Através desse diagnóstico é possível melhorar o modelo, até encontrar o melhor ajuste.

Uma das formas é verificar se existe significância individual de cada parâmetro do modelo, onde testa-se as hipóteses de que um parâmetro β_j é igual ou diferente de zero. Essas hipóteses podem ser testadas usando a distribuição da estatística de Wald:

$$\xi_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathfrak{S} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_p^2,$$

sendo que $\hat{\boldsymbol{\beta}}$ é um vetor com as estimativas dos parâmetros do modelo, e \mathfrak{S} é a matriz de Informação de Fisher que é definida da seguinte forma:

$$\mathfrak{S} = E \left[-\frac{\partial^2 \ln L(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} \right].$$

A estatística de Wald sob a hipótese $H_0 : \beta_j = 0$ será:

$$\xi_W = \frac{\hat{\beta}_j^2}{VAR(\hat{\beta}_j)} \sim \chi_1^2.$$

Uma outra abordagem que pode ser adotada consiste em especificar as hipóteses sob os modelos a serem avaliados e então comparar as estatísticas de qualidade do ajuste (*Deviances*) dos modelos concorrentes.

A *Deviance* pode ser definida como uma comparação entre a estatística do log da razão de verossimilhanças (ln) para o modelo maximal e o modelo de interesse, na qual segue uma distribuição χ^2 com $(n - p)$ graus de liberdade sob a hipótese de que o modelo de interesse é tão adequado quanto o maximal:

$$D = 2 \ln \lambda = 2 \ln \left[\frac{L(\hat{\boldsymbol{\beta}}_{MAX}|\mathbf{Y})}{L(\hat{\boldsymbol{\beta}}|\mathbf{Y})} \right] = 2[\ln L(\hat{\boldsymbol{\beta}}_{MAX}|\mathbf{Y}) - L(\hat{\boldsymbol{\beta}}|\mathbf{Y})] \sim \chi_{n-p}^2.$$

Se o modelo é adequado, espera-se que o valor da *Deviance* seja próximo da média da distribuição. Então, se um modelo com p parâmetros produz uma boa descrição das n observações dos dados, então $D \sim \chi_{n-p}^2$ e se espera que $D \cong n - p$. É possível concluir que se $D \leq \chi_{n-p;\alpha}^2$ então o modelo se ajusta bem aos dados, ao nível de significância de $\alpha\%$.

O Teste de comparabilidade de modelos encaixados, isto é, que diferem no número de parâmetros tendo a mesma distribuição e função de ligação, usa a *Deviance* como base da sua estatística de teste. As hipóteses do teste são as seguintes (com $q < p < n$) (Dobson,

2002 [16]) :

$$\left\{ \begin{array}{l} H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix} \\ H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \end{array} \right.$$

Pode-se testar H_0 contra H_1 de duas formas:

1. Usando a diferença entre as *Deviances* de ambos os modelos:

$$D_0 = 2[\ln L(\hat{\boldsymbol{\beta}}_{MAX}|\mathbf{Y}) - L(\hat{\boldsymbol{\beta}}_0|\mathbf{Y})] \sim \chi_{n-q}^2$$

$$D_1 = 2[\ln L(\hat{\boldsymbol{\beta}}_{MAX}|\mathbf{Y}) - L(\hat{\boldsymbol{\beta}}_1|\mathbf{Y})] \sim \chi_{n-p}^2$$

$$\Delta D = D_0 - D_1 = 2[\ln L(\hat{\boldsymbol{\beta}}_1|\mathbf{Y}) - L(\hat{\boldsymbol{\beta}}_0|\mathbf{Y})] \sim \chi_{p-q}^2$$

2. Usando a estatística F, que é definida da seguinte forma:

$$F = \frac{\frac{\Delta D}{D_1}}{\frac{p-q}{n-p}} = \frac{\frac{D_0 - D_1}{D_1}}{\frac{p-q}{n-p}} \sim F_{p-q; n-p}$$

Para ambos os casos, se a hipótese nula H_0 é rejeitada ao nível de significância de $\alpha\%$ concluí-se que o modelo de interesse ($\boldsymbol{\beta}_0$) não é tão adequado quanto o modelo completo ($\boldsymbol{\beta}_1$), caso contrário, dizemos que o modelo reduzido é tão adequado quanto o modelo completo.

Além dessas abordagens, outros critérios de seleção de modelos podem ser empregados, como o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC).

O AIC é uma medida de qualidade de ajuste, que mede o grau de informação que se perde ao adotar determinado modelo, levando em consideração o número de parâmetros (p), definido da seguinte forma:

$$AIC = -2[\ln L(\boldsymbol{\beta}) - p].$$

O BIC também mede o grau de informação que se perde ao adotar determinado modelo, levando em conta além do número de parâmetros (p), também o tamanho da amostra (n):

$$BIC = -2 \ln L(\boldsymbol{\beta}) + p \ln(n).$$

Cada modelo de ajuste possui o seu valor do AIC e BIC, quanto menor esse valor melhor é considerado o ajuste.

3.2.4 Diagnóstico do Modelo

Um passo importante na análise do ajuste de um Modelo é verificar afastamentos sérios das suposições feitas para o modelo e detectar observações atípicas que destoam do conjunto com alguma interferência nos resultados do ajuste. Diversas técnicas de diagnóstico para Modelos Lineares Generalizadas são apresentadas em Paula (2013) [15].

Uma das técnicas de diagnóstico do ajuste do modelo é verificar as observações *outliers*, ou seja, que se encontram significativamente mais afastadas que as demais. Essas observações são classificadas em três grupos:

- **Pontos Aberrantes** (observações mal ajustadas, com resíduos altos);
- **Pontos de Alavanca** (observações posicionadas em regiões remotas com alta influência no próprio valor ajustado); e
- **Pontos influentes** (observações que exercem uma influência desproporcional nas estimativas dos coeficientes do modelo).

A ideia do Ponto de Alavanca é verificar a influência de y_i sobre o valor ajustado \hat{y}_i . Para os MLG's uma sugestão para encontrar os Pontos de Alavanca é usar os elementos da diagonal (h_{ii}) da matriz \mathbf{H} , definida abaixo, e através do gráfico desses elementos contra os valores do modelo ajustados (\hat{y}_i) verificar os pontos que se destacam.

$$\mathbf{H} = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}},$$

em que $W = \text{diag}(w_1, \dots, w_n)$, com $w_i = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$.

Supondo que todos os pontos exerçam a mesma influência sobre os valores ajustados, o esperado é que h_{ii} se aproxime do valor $\frac{\text{tr}(\mathbf{H})}{n} = \frac{p}{n}$. Portanto os Pontos de Alavanca são os pontos tais que $h_{ii} \geq \frac{2p}{n}$.

Outra técnica de diagnóstico de modelo é a de comparar os resíduos. O resíduo ordinário para os Modelos Lineares Generalizados pode ser definido como:

$$r^* = \hat{W}^{\frac{1}{2}}(z - \hat{\eta}), \text{ com } z = \hat{\eta} + \hat{W}^{-\frac{1}{2}} \hat{V}^{-\frac{1}{2}}(y - \hat{\mu}).$$

No entanto, esses resíduos não possuem uma variância constante como é indicado por Paula (2013) [15], logo com a finalidade de comparar os resíduos deve-se utilizar os resíduos padronizados de Pearson, que são definidos da seguinte forma:

$$t_{S_i} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{V}_i(1 - \hat{h}_{ii})}}.$$

Com esse resíduo padronizado de Pearson, é possível encontrar as observações que possuem um resíduo superior as outras, assim é possível caracterizar esse ponto como Aberrante.

Influência é uma técnica de diagnóstico de modelos, que visa verificar a influência que uma determinada observação tem nos parâmetros estimados $\hat{\beta}$. Os pontos influentes são aqueles que, quando retirados do modelo, mudam significativamente as estimativas ou até a significância dos coeficientes do modelo. O método para detectar esses pontos é o da deleção dos pontos. Nesse método um ponto é retirado e é verificado se houve alguma mudança nas estimativas.

Uma medida de influência é o afastamento da verossimilhança. Ou seja, quando retiramos uma observação i o afastamento é denotado da seguinte forma:

$$LD_i = 2 \left\{ L(\hat{\beta}) - L(\hat{\beta}_{(i)}) \right\},$$

em que $\hat{\beta}_{(i)}$ é a estimativa de β com exclusão da i -ésima observação. Através de uma aproximação, que é mostrada em Paula (2013) [15], se chega na seguinte expressão, conhecida como Distância de Cook:

$$LD_i \approx \left(\frac{\hat{h}_{ii}}{1 - \hat{h}_{ii}} \right) t_{S_i}^2.$$

Após terem sido detectados os pontos Aberrantes, de Alavanca e Influentes, é necessária a realização de uma análise confirmatória. Essa análise consiste em retirar cada ponto, reajustar o modelo e verificar o efeito que é exercido sobre as estimativas.

Outras técnicas que são utilizadas no processo de diagnóstico do modelo que facilitam a localização dos problemas e na resolução são as técnicas gráficas. Os gráficos que são recomendados para os MLGs são:

1. Gráfico dos resíduos *Deviance* padronizados (t_{D_i}) versus a ordem das observações (para encontrar pontos Aberrantes);
2. Gráfico dos resíduos *Deviance* padronizados (t_{D_i}) versus os valores ajustados (para verificar aleatoriedade dos resíduos);
3. Gráfico das probabilidades dos resíduos *Deviance* padronizados (t_{D_i}) versus os percentis da distribuição com Envelope;
4. Gráfico de \hat{z}_i versus $\hat{\eta}_i$ (para verificar adequação da função de ligação);
5. Gráfico de h_{ii} versus a ordem das observações (para encontrar pontos de Alavanca-gem); e
6. Gráfico de LD_i versus ordem das observações (para encontrar pontos de influência).

Para o gráfico do item (3) é difícil avaliar se os percentis da distribuição estão afastados da reta ajustada, então é aconselhado a construção de um tipo de banda de confiança para os resíduos e assim é possível detectar afastamentos sérios. É sugerido (Paula (2013) [15]) a construção de um tipo de banda de confiança através de simulações chamado de Envelope. Basicamente o procedimento consiste em gerar resíduos que tenham média zero e matriz de variância-covariância ($I_n - \mathbf{H}$).

3.2.5 Superdispersão

Na distribuição de Poisson se uma variável aleatória (Y) tem média λ , então sua variância também deve ser λ , ou seja:

$$E[Y] = Var[Y] = \lambda$$

No entanto, alguns dados de contagem violam esta suposição. Quando a variância da variável aleatória é maior do que sua média, diz-se que estes dados tem superdispersão (Stroup, 2013 [17]).

O ajuste incorreto da superdispersão pode causar a subestimação do erro padrão dos estimadores e a inflação do nível de significância.

A superdispersão é um fenômeno muito comum e por isso já foram desenvolvidos diversos modelos para esses dados, incluindo os modelos “Quasi-Poisson”, Binomial Negativa, Poisson Generalizada e com Zeros Inflacionados (Hoef et al., 2007)[18]. Apesar de existirem todos esses modelos, o “Quasi-Poisson” e a Binomial Negativa são usados com mais frequência, principalmente por causa do seu fácil acesso em softwares e por sua generalização para regressões. Mas, como é sugerido no artigo de Hoef et al. (2007), a “Quasi-Poisson” pareceu ser um melhor ajuste para os dados. Então para este trabalho a superdispersão será corrigida através desse modelo.

O modelo “Quasi-Poisson” pode ser definido da mesma forma que qualquer MLG. Seja Y uma Variável Aleatória tal que:

$$E(Y) = \lambda \quad \text{Var}(Y) = \phi\lambda,$$

onde $E(Y)$ é a esperança de Y , $Var(Y)$ é a variância de Y , λ é o parâmetro da distribuição e tem que ser maior do que 0 e ϕ o parâmetro de dispersão, que para o caso de superdispersão assume valores maiores que 1.

A relação próxima dos valores da esperança e variância com os da distribuição de Poisson e o uso de uma função de ligação logarítmica justificam o porque do modelo se chamar “Quasi-Poisson”, denotado por $Y \sim Poi(\lambda, \phi)$

As vantagens do uso de um modelo “Quasi” é que os parâmetros são deixados em um estado natural e interpretável e que é permitido o uso das técnicas de diagnósticos padrão sem perda de eficiência.

Assim como para o Modelo de Regressão de Poisson, para o modelo “Quasi-Poisson” a média é um parâmetro que varia como uma função das covariáveis. Ou seja, assumindo que $y_i \sim Poi(\lambda_i, \phi)$, a média λ_i varia como uma função das covariáveis do modelo, e para esse trabalho essa função de ligação será a \ln :

$$\ln(\lambda_i) = \ln(n_i) + \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Uma das mudanças que ocorre utilizando o modelo “Quasi-Poisson” na estimação dos

parâmetros do modelo β , é na matriz dos pesos W , que fica em função das médias e do parâmetro de dispersão da seguinte forma:

$$W = \text{diag}(w_1, w_2, \dots, w_n) = \text{diag}\left(\frac{\lambda_1}{\phi}, \frac{\lambda_2}{\phi}, \dots, \frac{\lambda_n}{\phi}\right).$$

O parâmetro de superdispersão ϕ deve ser estimado através da estatística do χ^2 de Pearson, que é definida da seguinte forma:

$$\chi_p^2 = \sum_{i=1}^n \frac{(y_i - \lambda_i)^2}{\text{var}(y_i)} = \sum_{i=1}^n \frac{(y_i - \lambda_i)^2}{\phi \lambda_i}.$$

Então, a estimação de ϕ segue como:

$$\hat{\phi} = \frac{\chi_p^2}{n - p},$$

em que p é o número de parâmetros do modelo e n o número de observações.

Outras mudanças que, de acordo com Burnham et al. (2002) [19], ocorrem ao se utilizar um modelo “Quasi” são:

- Com a superdispersão a log-verossimilhança do parâmetro β deve ser a seguinte:

$$\frac{\ln L(\beta|\mathbf{Y})}{\hat{\phi}};$$

- O número de parâmetros K passa a ser agora o número de parâmetros do modelo $p + 1$ adicionado 1 que indica o parâmetro de dispersão:

$$K = (p + 1) + 1;$$

- As estimativas do P-valor passam a ser baseadas na estatística t ao invés da estatística z , que é usada no Modelo de Regressão de Poisson;
- Modificações no cálculo do AIC, que passa a se chamar de QAIC:

$$QAIC = -\frac{2\ln L(\beta|\mathbf{Y})}{\hat{\phi}} + 2K.$$

4 Resultados

4.1 Análise Exploratória

Uma análise exploratória é um primeiro passo nas análises de um banco de dados, pois é importante primeiro saber o comportamento das variáveis do estudo e suas principais características.

Então, em primeiro lugar observamos as variáveis quantitativas, e para analisá-las foram calculadas as seguintes medidas: tamanho da amostra, médias, desvio padrão, mínimo, máximo e o P-valor para o teste de Normalidade de Kolmogorov-Smirnov. A ideia de realizar o teste de normalidade é que com as variáveis que possuem uma distribuição normal, é possível calcular a correlação de Pearson entre elas.

Os resultados iniciais das variáveis contínuas, vistas no Anexo A indicam que as variáveis que apresentaram distribuição normal são as variáveis de ambiente e poluição (Tabelas 7, 8 e 9). E a correlação entre essas variáveis indicam baixa relação com a variável resposta Y (nº de faltas) e que a relação entre o poluente O_3 e as demais variáveis aparenta ser inversa, ou seja, a correlação entre esses é negativa (Tabelas 10 e 11).

Para avaliar a superdispersão dos dados, verificou-se na Tabela 7 no Anexo A que a média da variável aleatória número total de faltas é menor do que a variância. Sendo assim, estes resultados indicam a superdispersão.

Com as variáveis qualitativas foram calculadas as frequências das suas categorias, com a ideia de mostrar como se comportam em relação ao total da amostra.

Também foram estimadas as médias dos valores das variáveis quantitativas de acordo com o número de faltas categorizadas. Ou seja, a variável resposta Y foi separada em categorias que indicam o número de faltas, e para cada categoria foi estimada a média dos valores das covariáveis para as crianças que se encaixam naquela categoria. E o objetivo dessa tabela é perceber se existe alguma tendência nos valores das variáveis conforme aumenta o número de faltas.

As variáveis que apresentaram resultados interessantes, no Anexo A, foi a média ponderada, com e sem lag, das variáveis de Condições de Saúde (Tabelas 16 e 17). Conforme o número de faltas aumentava, os valores dessas médias das variáveis de sintomas também aumentaram, o que pode indicar alguma relação entre a variável resposta e estas co-variáveis.

Outra tabela no Anexo A, que apesar de não apresentar nenhuma tendência significativa, apresentou um resultado interessante, foi a tabela com as variáveis de Espirometria (Tabela 18). Mostra que as variáveis que indicam uma porcentagem do quanto a criança deveria alcançar daquela valor, para as crianças que não faltaram nenhum dia, estão todas acima de 100. Ou seja, todas as médias daqueles valores para as crianças que não faltaram nenhum dia está acima do que se esperava das crianças, o que pode ser um indicativo de que essas crianças são saudáveis.

4.2 Modelos de Regressão

O objetivo dessa seção é encontrar o melhor modelo que ajuste aos dados. Foi utilizada a função *glm* do software R versão 2.15.1.

O banco de dados era composto de 89 variáveis e, como esse número é grande, foram criados critérios para o ajuste do modelo. No primeiro passo, decidiu-se estimar modelos simples da variável resposta com cada variável do banco de dados e, assim, através do teste de Wald, selecionar aquelas variáveis significativas a um nível de 20%.

Desde o começo das análises, os modelos foram ajustados pela família “Quasi-Poisson”, com função de ligação \ln e com o *offset* t_{2i} .

As Tabelas 20, 21 e 22 do Anexo B apresentam os resultados dos modelos simples. Como é possível verificar as variáveis de sintomas respiratórios foram quase todas significantes; algumas variáveis do questionário foram significantes; somente a média ponderada do PF foi significativa das variáveis dos Exames Subclínicos; e nenhuma variável ambiental e pessoal foi significativa.

O próximo passo foi ajustar o modelo de acordo com as variáveis mais distais até as mais proximais, segundo o modelo teórico. Como nenhuma variável pessoal foi aceita, que representam as variáveis distais, primeiramente foram adicionadas as variáveis do questionário, que representam um grupo das variáveis intermediárias de acordo com a Figura 1.

Então, o primeiro modelo era composto da variável do questionário que apresentou o menor P-valor no teste de Wald. A partir desse modelo foram criados outros modelos adicionando as outras variáveis do questionário. Neste momento, era feita a comparação entre esses novos modelos com o modelo simples.

Para comparação de Modelos de Regressão de Poisson, a literatura sugeria o método da razão de verossimilhança e/ou avaliar o AIC. No entanto, como a família utilizada nos modelos deste estudo era a “Quasi-Poisson” e essa família não possui uma verossimilhança e sim uma “quasi-verossimilhança”, a função *glm* do R não calculava nem a verossimilhança e nem a medida AIC. Mas, o programa calculava as *Deviances* dos modelos. Então, para comparar os modelos foi utilizada a diferença entre as *Deviances*.

No final, foi selecionado um modelo base, com as variáveis intermediárias significantes de acordo com o teste de Wald e que apresentou uma maior ΔD .

O Modelo Base selecionado foi o seguinte:

$$\ln(\hat{\lambda}_i) = \ln(t_{2i}) + \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i},$$

onde

- X_{1i} é a variável que indica quantos comôdos existem na casa da criança i ;
- X_{2i} indica a idade de quem cuida da criança i ;
- X_{3i} indica se a criança i mora perto de uma plantação;
- X_{4i} indica a média do PF diário da criança i ponderada pelas faltas.

Esse modelo foi calculado para uma amostra de $n = 203$ crianças, pois foram excluídas as observações que eram *missing*. Os resultados do modelo estão apresentados na Tabela 2. Para melhor interpretação na tabela estão apresentados os valores dos Riscos Relativos estimados que são definidos da seguinte forma:

$$\hat{RR}_{X_i} = \exp(\hat{\beta}_i).$$

Na tabela também são apresentados os valores do Intervalo de Confiança dos RR com 95% de confiança, além do P-valor do teste de Wald.

O passo seguinte foi verificar o diagnóstico do modelo e a qualidade de ajuste. Como diagnóstico serão apresentados os gráficos da Distância de Cook e o gráfico do preditor linear pelos Resíduos de Pearson. Os gráficos da Figura 2 indicam esse diagnóstico.

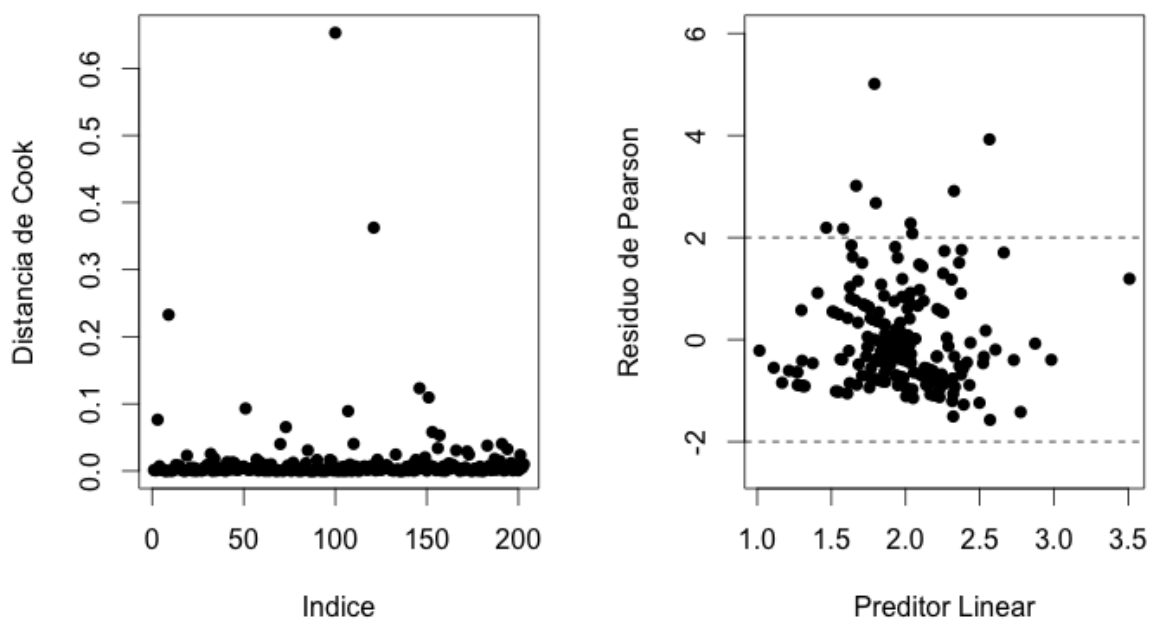


Figura 2: Gráficos de diagnóstico do modelo base

A Figura 2 indica a presença de alguns *outliers*. Como análise de Sensibilidade, foram retirados os *outliers* e reestimado o Modelo 1 (Modelo 2), como é mostrado na Tabela 2. Considerou-se *outliers* aqueles resíduos padronizados cujo o módulo era maior que 2. Observou-se que as estimativas pontuais dos RR foram semelhantes, no entanto a estimativa do parâmetro de dispersão diminuiu.

Tabela 2: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para o modelo base com e sem *outlier*. Tangará da Serra, 2008.

Variáveis	Modelo 1 Com <i>outlier</i> ($n = 203$)			Modelo 2 Sem <i>outlier</i> ($n = 194$)		
	\hat{RR}	IC95%	P-valor	\hat{RR}	IC95%	P-valor
Intercepto	0,286	(0,159 ; 0,517)	<0,001	0,240	(0,147 ; 0,390)	<0,001
Número de comodos						
1 comodo e 1 banheiro	1	-	-	1	-	-
Mais comodos	0,362	(0,224 ; 0,585)	<0,001	0,307	(0,208 ; 0,454)	<0,001
Idade de quem cuida	0,988	(0,980 ; 0,997)	0,007	0,993	(0,985 ; 1,000)	0,046
Mora perto de uma plantação						
Sim	1	-	-	1	-	-
Não	0,759	(0,603 ; 0,955)	0,020	0,750	(0,615 ; 0,915)	0,005
Nível de PF	1,002	(1,001 ; 1,003)	<0,001	1,002	(1,002 ; 1,003)	<0,001
Parâmetro de Dispersão		4,495			2,908	

Como pode ser visto na Tabela 23 no Anexo B, existe uma grande correlação entre as variáveis de sintomas, por isso, para evitar colinearidade, foram criados diversos modelos

com a adição de cada uma das variáveis de sintomas ao modelo base, com e sem *outlier*, e os resultados estão apresentados na Tabela 3.

Tabela 3: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para a adição das variáveis de Sintomas Respiratórios no modelo base com e sem *outlier*. Tangará da Serra, 2008.

Variáveis	Com <i>outlier</i>			Sem <i>outlier</i>		
	\hat{RR}	IC95%	P-valor	\hat{RR}	IC95%	P-valor
Sintomas Geral 1	0,997	(0,991 ; 1,002)	0,239	0,998	(0,993 ; 1,003)	0,473
Soma Tosse	0,998	(0,991 ; 1,006)	0,693	1,000	(0,993 ; 1,006)	0,959
Média Tosse	1,168	(1,098 ; 1,242)	< 0,001	1,113	(1,046 ; 1,185)	0,001
Média Tosse L1	1,189	(1,124 ; 1,258)	< 0,001	1,158	(1,092 ; 1,228)	< 0,001
Média Posto Médico	1,257	(1,020 ; 1,548)	0,033	1,242	(1,030 ; 1,498)	0,025
Média Posto Médico L1	1,453	(1,082 ; 1,951)	0,014	1,229	(0,857 ; 1,762)	0,263
Média Coceira no Ouvido	1,263	(1,131 ; 1,411)	< 0,001	1,233	(1,062 ; 1,431)	0,007
Média Coceira no Ouvido L1	1,324	(1,201 ; 1,459)	< 0,001	1,234	(1,065 ; 1,430)	0,006
Soma Coceira na Garganta	0,993	(0,977 ; 1,009)	0,383	0,990	(0,975 ; 1,004)	0,162
Média Coceira na Garganta	1,153	(1,057 ; 1,259)	0,002	1,036	(0,931 ; 1,152)	0,515
Média Coceira na Garganta L1	1,149	(1,059 ; 1,246)	0,001	1,039	(0,939 ; 1,151)	0,457
Soma Coriza	0,999	(0,991 ; 1,007)	0,771	1,001	(0,994 ; 1,008)	0,740
Média Coriza	1,239	(1,157 ; 1,327)	< 0,001	1,163	(1,081 ; 1,251)	< 0,001
Média Coriza L1	1,224	(1,136 ; 1,318)	< 0,001	1,167	(1,085 ; 1,255)	< 0,001
Média Dor de Cabeça	1,131	(1,021 ; 1,252)	0,019	1,121	(1,022 ; 1,230)	0,016
Média Dor de Cabeça L1	1,151	(1,047 ; 1,264)	0,004	1,112	(1,014 ; 1,220)	0,025
Média Lacrimejamento L1	1,168	(1,044 ; 1,307)	0,007	1,102	(0,973 ; 1,249)	0,129
Soma Remédio	0,996	(0,985 ; 1,007)	0,501	1,001	(0,992 ; 1,009)	0,871
Média Remédio	1,158	(1,030 ; 1,300)	0,015	1,189	(1,070 ; 1,321)	0,002

O que é interessante perceber nos modelos com adição das variáveis de sintomas, é que as variáveis de soma dos sintomas não foram significantes, enquanto que as médias ponderadas foram. E também é possível perceber a diferença quando a média é defasada no modelo com *outlier*, ou seja, a média com lag 1 aparenta apresentar um risco maior do que a média sem lag. Enquanto que no modelo sem *outlier* os riscos se aproximam, e o valor do risco da média do sintoma com e sem defasagem são próximos.

A partir desse resultado, foi escolhida para entrar no modelo somente a variável de sintoma que foi apresentou o menor P-valor no Teste de Wald e no teste de diferenças de *Deviances*. E, dessa forma, foi estimado um novo modelo, que corresponde às variáveis do modelo base mais a variável de sintoma:

$$\ln(\hat{\lambda}_i) = \ln(t_{2i}) + \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i},$$

onde a variável X_{5i} representa a média ponderada do sintoma Tosse com lag 1 e as outras covariáveis são as mesmas do modelo base.

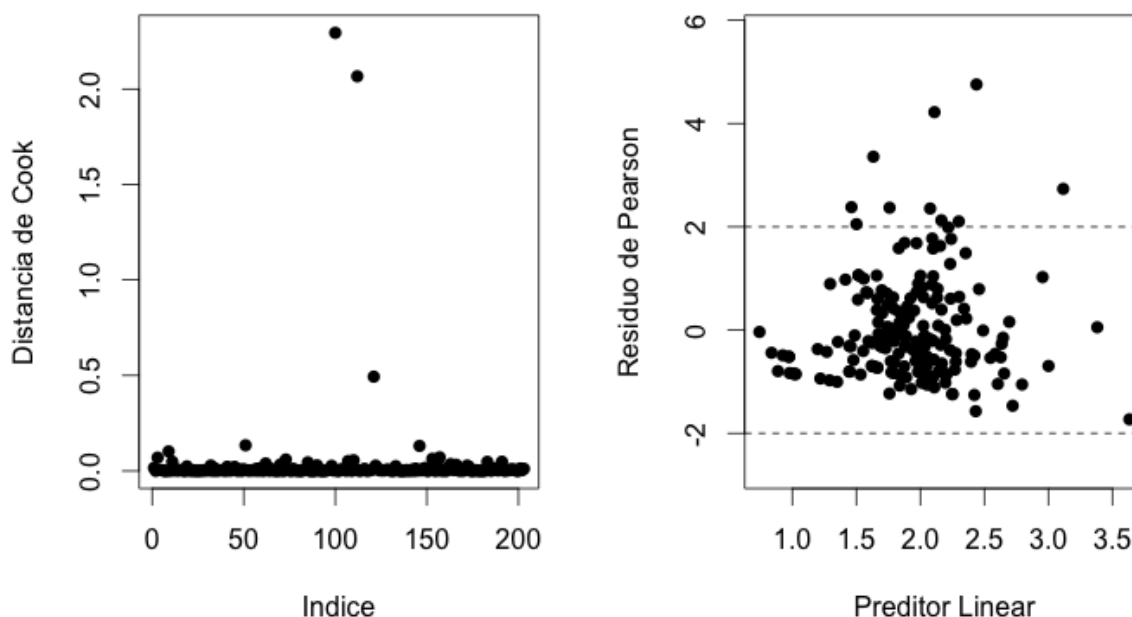


Figura 3: Gráficos do diagnóstico do modelo base com a variável de sintoma.

A partir desse modelo foi feito um novo diagnóstico, para verificar o novo ajuste como é mostrado na Figura 3, e a partir desse diagnóstico foram retirados os *outliers* da mesma forma como foi feito anteriormente. Os RR estimados dos dois modelos, com e sem *outlier*, estão apresentados na Tabela 4.

Tabela 4: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para o modelo base com a variável de Sintoma, com e sem *outlier*. Tangará da Serra, 2008.

Variáveis	Modelo 3 Com <i>outlier</i> (n = 203)			Modelo 4 Sem <i>outlier</i> (n = 193)		
	\hat{RR}	IC95%	P-valor	\hat{RR}	IC95%	P-valor
Intercepto	0,165	(0,089 ; 0,303)	< 0,001	0,075	(0,039 ; 0,144)	< 0,001
Número de comodors						
1 comodo e 1 banheiro	1	-	-	1	-	-
Mais comodors	0,471	(0,298 ; 0,745)	0,002	0,903	(0,517 ; 1,577)	0,720
Idade de quem cuida	0,989	(0,982 ; 0,997)	0,007	0,990	(0,984 ; 0,997)	0,005
Mora perto de uma plantação						
Sim	1	-	-	1	-	-
Não	0,732	(0,591 ; 0,906)	0,005	0,786	(0,654 ; 0,945)	0,011
Nível de PF	1,002	(1,002 ; 1,003)	< 0,001	1,002	(1,001 ; 1,003)	< 0,001
Média Tosse L1	1,189	(1,124 ; 1,258)	< 0,001	1,237	(1,180 ; 1,297)	< 0,001
Parâmetro de Dispersão		3,838			2,309	

Da mesma forma como aconteceu com o modelo sem a variável de sintoma, a retirada dos *outliers* não mudou substancialmente a estimativa dos RR. No entanto, o modelo sem

outliers indicou mudança no parâmetro de dispersão e a variável X_{1i} , que é o número de cômodos, deixou de ser significativa ao nível de 5%.

Como o objetivo do trabalho é avaliar o efeito dos poluentes no absenteísmo escolar, a partir do modelo novo ajustado (Modelo 3) foram adicionadas as variáveis dos níveis de poluição com lags de 1 a 5 e os resultados são apresentados na Tabela 5.

Tabela 5: Efeitos da Poluição no absenteísmo escolar para o incremento de uma unidade nos níveis dos poluentes, considerando modelos com e sem *outlier*.

Variáveis	Com <i>outlier</i>			Sem <i>outlier</i>		
	\hat{RR}	IC95%	P-valor	\hat{RR}	IC95%	P-valor
Média PM10 L1	1,006	(0,999 ; 1,013)	0,077	1,005	(0,999 ; 1,010)	0,110
Média PM10 L2	1,004	(0,999 ; 1,010)	0,128	1,003	(0,998 ; 1,008)	0,185
Média PM10 L3	1,005	(1,000 ; 1,011)	0,059	1,005	(1,000 ; 1,009)	0,060
Média PM10 L4	1,004	(0,998 ; 1,011)	0,165	1,004	(0,998 ; 1,009)	0,176
Média PM10 L5	1,006	(0,999 ; 1,013)	0,084	1,003	(0,998 ; 1,009)	0,262
Média BC L1	1,059	(0,628 ; 1,786)	0,831	1,006	(0,649 ; 1,559)	0,979
Média BC L2	1,080	(0,657 ; 1,776)	0,761	0,921	(0,605 ; 1,401)	0,699
Média BC L3	1,189	(0,696 ; 2,030)	0,527	1,023	(0,653 ; 1,605)	0,920
Média BC L4	1,151	(0,712 ; 1,860)	0,567	1,006	(0,672 ; 1,504)	0,978
Média BC L5	1,244	(0,796 ; 1,944)	0,339	1,036	(0,710 ; 1,512)	0,854
Média O3 L1	0,984	(0,961 ; 1,008)	0,194	0,993	(0,974 ; 1,013)	0,520
Média O3 L2	0,991	(0,964 ; 1,017)	0,489	0,991	(0,969 ; 1,014)	0,431
Média O3 L3	0,999	(0,975 ; 1,024)	0,930	0,995	(0,975 ; 1,016)	0,639
Média O3 L4	1,004	(0,977 ; 1,031)	0,786	1,002	(0,980 ; 1,024)	0,888
Média O3 L5	1,000	(0,976 ; 1,025)	0,994	0,997	(0,977 ; 1,018)	0,773

Como é apresentado na tabela, os poluentes BC e O_3 não foram significantes para explicar o absenteísmo escolar ao nível de 5% de significância. Já o poluente PM_{10} , no modelo com *outlier* apresentou significância a 10% nos lags 1, 3 e 5. No modelo sem *outlier* os resultados da estimação do RR para o poluente PM_{10} não mudam, no entanto, o único nível do poluente significativa a 10% foi o lag 3.

A Tabela 5 apresenta os RR para o incremento de uma unidade nos níveis dos poluentes, no entanto, para o poluente PM_{10} a interpretação é melhor para incrementos de 10 unidades. Por isso é apresentado na Figura 4 os valores dos RR, para incrementos de 10 unidades, para cada lag desse poluente.

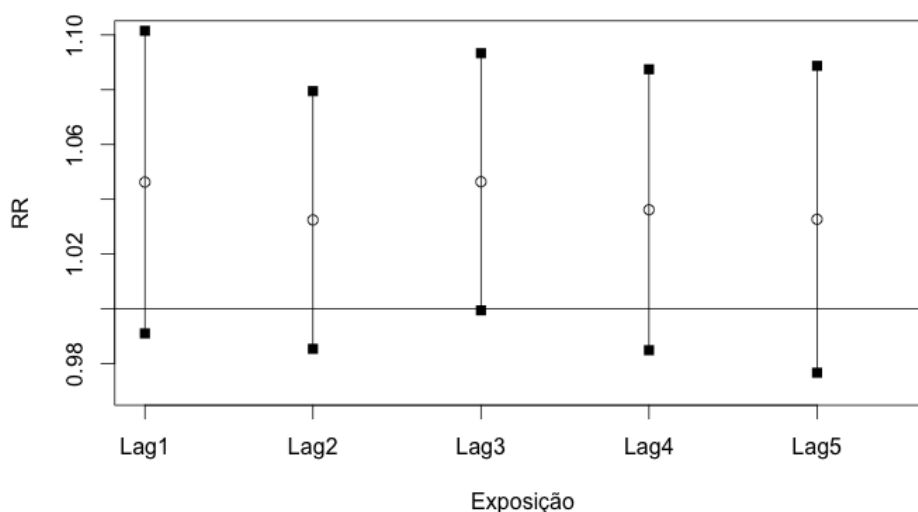


Figura 4: Efeitos do poluente PM_{10} no absenteísmo, para o incremento de 10 unidades no nível do poluente, considerando as defasagens. Tangará da Serra, 2008.

O que é possível perceber a partir da Figura 4 é que o incremento de 10 unidades do nível do poluente PM_{10} com lag 3 aumenta o risco de absenteísmo escolar em aproximadamente 5%.

Ao modelo base mais a variável de sintoma foi adicionado uma variável de poluição. Nesses modelos a variável de poluição que foi adicionada X_{6i} foi a que apresentou menor P-valor no teste de Wald, que é o nível de exposição do poluente PM_{10} com lag 3 para a criança i :

$$\ln(\hat{\lambda}_i) = \ln(t_{2i}) + \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_6 X_{6i}.$$

A partir da adição dessa variável também foi feito um diagnóstico (Figura 5) do modelo, e os *outliers* encontrados foram retirados usando o mesmo critério dos outros modelos.

E a Tabela 6 apresenta os valores dos RR para o Modelo com a variável de poluição com e sem *outlier*.

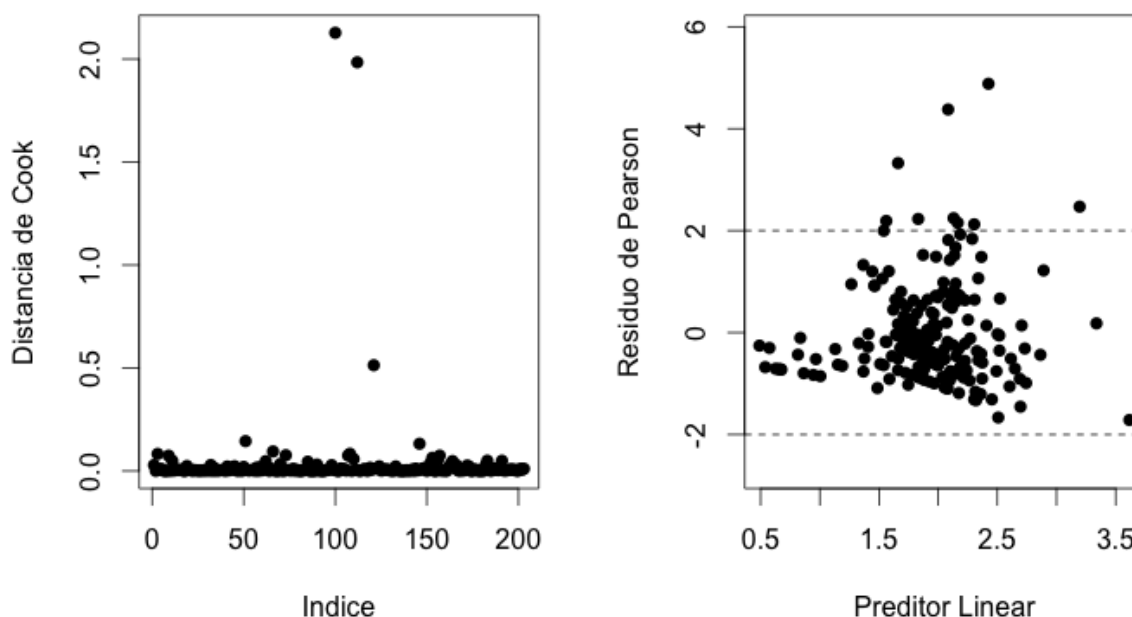


Figura 5: Gráficos do diagnóstico do modelo com a variável de poluição

Tabela 6: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor para o novo modelo com a variável de poluição com e sem *outlier*. Tangará da Serra, 2008.

Variáveis	Modelo 5 Com <i>outlier</i> ($n = 203$)			Modelo 6 Sem <i>outlier</i> ($n = 194$)		
	\hat{RR}	IC95%	P-valor	\hat{RR}	IC95%	P-valor
Intercepto	0,125	(0,064 ; 0,246)	< 0,001	0,060	(0,030 ; 0,121)	< 0,001
Número de comodos						
1 comodo e 1 banheiro	1	-	-	1	-	-
Mais comodos	0,458	(0,290 ; 0,723)	0,001	0,842	(0,478 ; 1,485)	0,554
Idade de quem cuida	0,989	(0,982 ; 0,997)	0,007	0,991	(0,984 ; 0,997)	0,006
Mora perto de uma plantação						
Sim	1	-	-	1	-	-
Não	0,722	(0,584 ; 0,893)	0,003	0,749	(0,623 ; 0,899)	0,002
Nível de PF	1,003	(1,002 ; 1,004)	< 0,001	1,002	(1,002 ; 1,003)	< 0,001
Média Tosse L1	1,187	(1,122 ; 1,256)	< 0,001	1,230	(1,173 ; 1,291)	< 0,001
Média PM10 L3	1,005	(1,000 ; 1,011)	0,059	1,005	(1,000 ; 1,010)	0,035
Parâmetro de Dispersão	3,800			2,361		

A partir da Tabela 6 é interessante verificar que o poluente PM_{10} com lag 3 passa a ser significativo ao nível de 5% quando retiramos os *outliers* do modelo.

4.2.1 Comparação entre os modelos ajustados - Envelope Simulado

Depois de estimados os efeitos dos poluentes, foi verificado se esse modelo se ajusta bem aos dados. Uma forma de verificar essa afirmação é através dos gráficos de Envelope Simulado.

Nessa seção serão comparados 6 modelos. Os modelos base com e sem *outlier* (Modelos 1 e 2) que são apresentados na Tabela 2, os modelos base mais a variável de sintoma, com e sem *outlier* (Modelos 3 e 4), que são apresentados na Tabela 4 e os modelos base mais a variável de sintoma e a variável de poluição, com e sem *outlier* (Modelos 5 e 6) apresentados na Tabela 6.

Os gráficos de envelope para cada um dos modelos está apresentado na Figura 6. Esses gráficos foram feitos com os Resíduos de Pearson e os intervalos do Envelope foram construídos com um nível de 95% de confiança.

Pelos gráficos apresentados é possível perceber que os modelos não estão em perfeito ajuste.

Apesar de os modelos não estarem em perfeito ajuste, é possível perceber que os modelos da coluna da direita, que são os modelos sem *outliers*, apresentam um melhor ajuste.

Além disso, com a adição das variáveis no modelo base é possível perceber como a cauda superior se ajusta melhor ao intervalo.

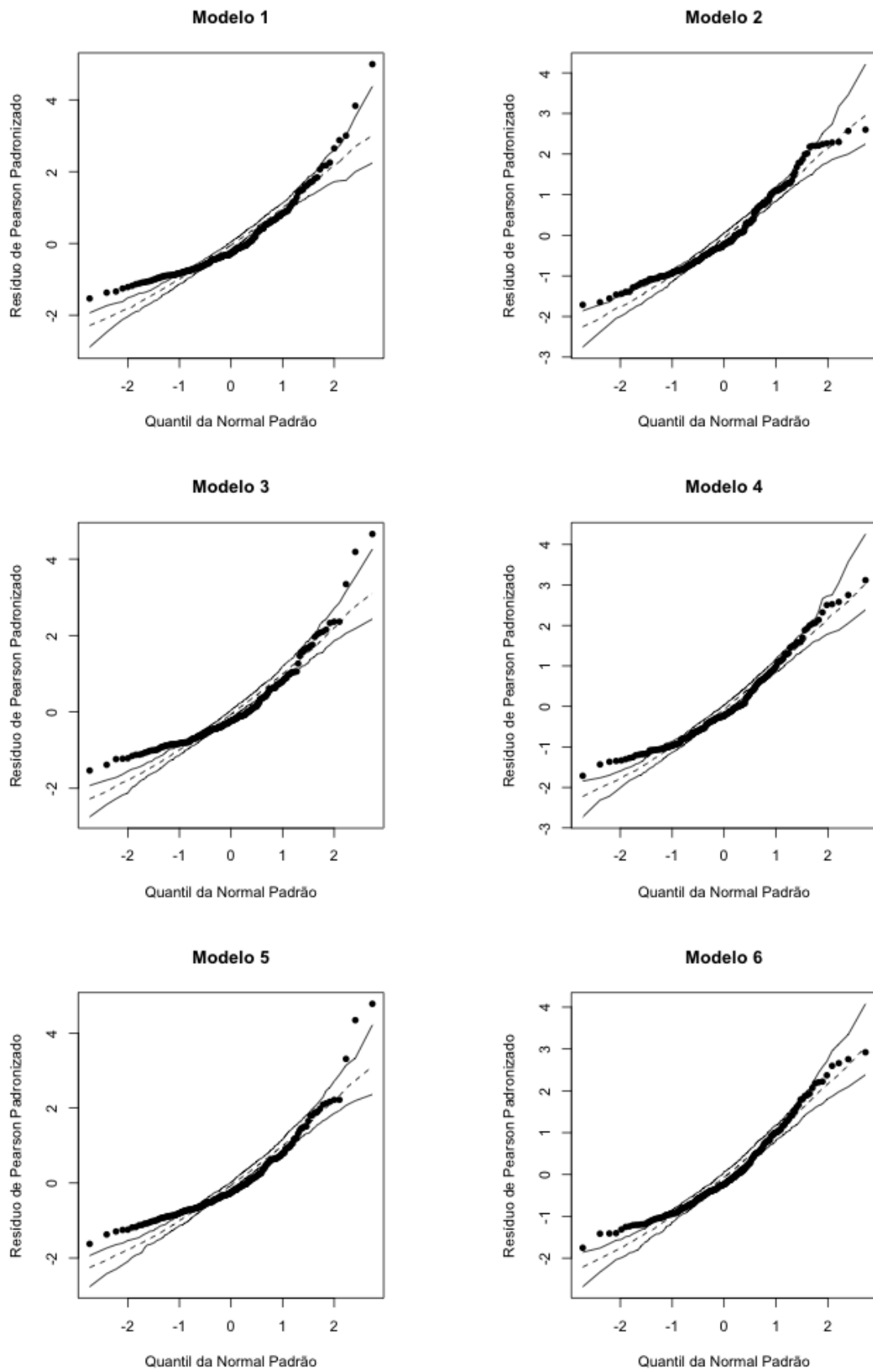


Figura 6: Gráficos dos Envelopes Simulados para os modelos estimados.

No geral é possível perceber que a retirada dos *outliers* melhora a qualidade do ajuste do modelo. Sendo assim, o Modelo 6 foi escolhido como o modelo final deste trabalho, e pode ser representado da seguinte forma, em função dos valores dos parâmetros $\hat{\beta}$:

$$\ln(\hat{\lambda}_i) = \ln(t_{2i}) - 2,077 - 0,782X_{1i} - 0,011X_{2i} - 0,326X_{3i} + 0,003X_{4i} + 0,171X_{5i} + 0,005X_{6i},$$

onde

- X_{1i} é a variável que indica quantos comôdos existem na casa da criança i ;
- X_{2i} indica a idade de quem cuida da criança i ;
- X_{3i} indica se a criança i mora perto de uma plantação;
- X_{4i} indica a média do PF diário da criança i ponderada pelas faltas;
- X_{5i} indica a média de Tosse com lag 1 para a criança i ;
- X_{6i} indica o nível de poluente PM_{10} com lag 3 para a criança i .

A interpretação da estimação dos RR das variáveis desse modelo segue abaixo:

- $\widehat{RR}_{X_{1i}} = 0,842$

Risco relativo da variável que indica a quantidade de comôdos na casa, o que pode ser um indicativo da condição socioeconômica da criança. Podemos perceber que em relação ao grupo de crianças que moram em uma casa com 1 comôdo e 1 banheiro, as crianças que moram em casas com mais comôdos têm um risco 15,8% menor de absenteísmo escolar.

- $\widehat{RR}_{X_{2i}} = 0,991$

Risco relativo da variável que indica a idade de quem cuida. Essa variável também apresentou um risco relativo inferior a 1, o que indica que conforme aumentamos a idade de quem cuida da criança, o absenteísmo diminui. O que mostra que crianças que são cuidadas por pessoas mais velhas correm um risco menor de faltar a escola.

- $\widehat{RR}_{X_{3i}} = 0,749$

Risco relativo da variável que indica se a criança mora perto de uma plantação. Com um risco inferior a 1 é possível verificar que em relação a crianças que moram perto de uma plantação (cana-de-açúcar e soja), as crianças que não vivem perto de uma plantação tem um risco 25,1% menor de absenteísmo escolar.

- $\widehat{RR}_{X_{4i}} = 1,002$

Risco relativo da variável que indica a média ponderada de PF. O efeito dessa variável mostra que o aumento do nível de pico de fluxo também aumenta o risco da criança faltar a escola. Como o Pico de Fluxo é uma medida que varia de acordo com as características de cada criança, acredita-se que esse efeito possa ser um indicativo de outro resultado. Por exemplo, o PF depende da idade da criança, ou seja, crianças mais velhas tendem a ter um PF mais elevado do que crianças mais novas. Então o que o efeito dessa variável pode estar demonstrando é que crianças mais velhas, que tem um PF mais elevado, tem maior risco de faltar à aula.

- $\widehat{RR}_{X_{5i}} = 1,230$

Risco relativo da variável Média de Tosse com lag 1. Essa variável representa o conjunto das variáveis de sintomas. Verifica-se que o aumento de 1 unidade da média de Tosse defasada aumenta em 23,0% o risco de absenteísmo. O que mostra uma grande relação entre os sintomas e o absenteísmo escolar.

- $\widehat{RR}_{X_{6i}} = 1,005$

Risco relativo da variável Média de PM_{10} com lag 3. O efeito dessa variável indica um aumento no risco de absenteísmo escolar, quando a média dos valores do poluente de 3 dias anteriores à falta da criança aumenta em 1 unidade.

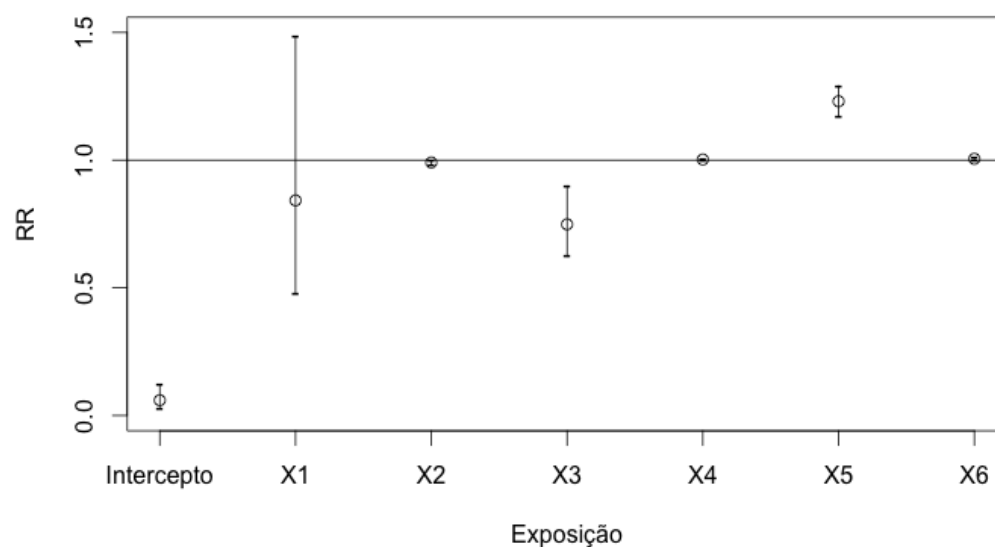


Figura 7: Efeitos do Modelo 6.

5 Conclusão

Com o modelo final selecionado, podemos verificar que existe uma relação entre a poluição e o absentismo escolar. Apesar de a maior parte dos poluentes com lags de 1 a 5 não terem sido estatisticamente significante, ainda tivemos um resultado significante para o poluente PM_{10} fazendo com que possa ser afirmada a relação entre o aumento desse poluente com a falta na escola.

Esse resultado se assemelha com o que foi encontrado no estudo de Berhane et al. (2002) [5], que encontrou uma relação com o poluente PM_{10} a longo prazo.

No entanto, os estudos de Gilliland et al. (2001) [4] e Berhane et al. (2002) [5] encontraram uma relação entre o absentismo escolar e o poluente O_3 , enquanto no nosso trabalho não encontramos nenhuma relação.

Mesmo que a maior parte das variáveis de poluição não apresentassem significância, foi possível perceber uma grande quantidade de variáveis relacionadas a sintomas significantes, cujo risco de absentismo escolar era grande, principalmente entre as variáveis defasadas. A entrada da variável Tosse com lag 1 condiz com o resultado do artigo de Hwang et al. (2000) [6] que mostra que a tosse diurna e noturna esteve associada com o absentismo.

E outro resultado similar foi com o artigo do Meng et al. (2012) [9], no qual mostrou que crianças com família de baixa renda faltam mais à escola. Esta é a mesma interpretação que se pode ter da variável número de comôdos na casa.

Outro resultado importante verificado é o do parâmetro de dispersão, que pelo livro de Burnham et al. (2002) [19] é esperado ser maior que 1, quando há a superdispersão. No entanto, se o modelo está bem adequado, o valor não deve ultrapassar 4. Se o parâmetro supera esse valor, é um indicativo de que não é somente a superdispersão que está afetando o valor do parâmetro, ou seja, o modelo está com uma estrutura inadequada.

Pelos resultados obtidos, é possível ver nas Tabelas 2, 4 e 6 que os valores de $\hat{\phi}$ ficam

adequados, exceto para o Modelo 1, principalmente quando os *outliers* são retirados. Ou seja, é possível afirmar que os *outliers* estavam afetando a dispersão do modelo.

Portanto de acordo com o modelo final estimado, que pode ser visto na Tabela 6, concluímos que as variáveis *idade de quem cuida* e *mora perto de uma plantação* podem ser consideradas fatores de proteção, enquanto que as variáveis *nível de Pico de Fluxo*, *média de Tosse com lag 1* e *média de PM_{10} com lag 3* podem ser consideradas fatores de risco.

Referências

- [1] MARQUES, A. *Absenteísmo nas Empresas*. Tese (Doutorado) — Centro Universitário de Brasília, 2006.
- [2] ERVASTI, J. et al. School environment as predictor of teacher sick leave: data-linked prospective cohort study. *BMC Public Health*, v. 12, p. 770, 2012.
- [3] MAKINO, K. Association of school absence with air pollution in areas around arterial roads. *J Epidemiol*, v. 10, n. 5, p. 292–299, Sep 2000.
- [4] GILLILAND, F. D. et al. The effects of ambient air pollution on school absenteeism due to respiratory illnesses. *Epidemiology*, v. 12, n. 1, p. 43–54, Jan 2001.
- [5] BERHANE, K.; THOMAS, D. C. A two-stage model for multiple time series data of counts. *Biostatistics*, v. 3, n. 1, p. 21–32, Mar 2002.
- [6] HWANG, J.-S. et al. Subject-domain approach to the study of air pollution effects on schoolchildren’s illness absence. *American journal of epidemiology*, Oxford Univ Press, 152, n. 1, p. 67–74, 2000.
- [7] GILLILAND, F. D. et al. Environmental tobacco smoke and absenteeism related to respiratory illness in schoolchildren. *Am. J. Epidemiol.*, v. 157, n. 10, p. 861–869, May 2003.
- [8] LEVY, D. E.; WINICKOFF, J. P.; RIGOTTI, N. A. School absenteeism among children living with smokers. *Pediatrics*, v. 128, n. 4, p. 650–656, Oct 2011.
- [9] MENG, Y. Y.; BABEY, S. H.; WOLSTEIN, J. Asthma-related school absenteeism and school concentration of low-income students in California. *Prev Chronic Dis*, v. 9, p. E98, 2012.
- [10] SANTOS, U. P. et al. Síndrome dos edifícios doentes em bancários. In: SciELO Public Health. *Publicação financiada pela FAPESP. Processo Saúde Coletiva*. [S.l.], 1992. 91, p. 4994–0.
- [11] HACON, S. de S.; GONÇALVES, K. dos S.; CASTRO, H. A. de. As queimadas na região amazônica eo adoecimento respiratório. *Ciência & Saúde Coletiva*, Associação Brasileira de Pós-Graduação em Saúde Coletiva, v. 17, n. 6, p. 1523–1532, 2012.
- [12] RIBEIRO, H.; ASSUNÇÃO, J. V. d. Efeitos das queimadas na saúde humana. *Estudos Avançados*, SciELO Brasil, 16, n. 44, p. 125–148, 2002.
- [13] JACOBSON, L. d. a. S. et al. Association between fine particulate matter and the peak expiratory flow of schoolchildren in the Brazilian subequatorial Amazon: a panel study. *Environ. Res.*, v. 117, p. 27–35, Aug 2012.

- [14] OLIVEIRA, B. F. de et al. Risk assessment of PM(2.5) to child residents in Brazilian Amazon region with biofuel production. *Environ Health*, v. 11, p. 64, 2012.
- [15] PAULA, G. A. *Modelos de Regressão: com apoio computacional*. [S.l.]: IME-USP Sao Paulo:, 2013.
- [16] DOBSON, A. J. *An introduction to generalized linear models. 2*. [S.l.]: CRC press, 2002.
- [17] STROUP, W. W. *Generalized linear mixed models: Modern concepts, methods and applications*. [S.l.]: CRC Press, 2012.
- [18] HOEF, J. M. V.; BOVENG, P. L. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, Eco Soc America, 88, n. 11, p. 2766–2772, 2007.
- [19] BURNHAM, K. P.; ANDERSON, D. R. *Model selection and multi-model inference: a practical information-theoretic approach*. [S.l.]: Springer, 2002.

ANEXO A - Análise Exploratória

Tabela 7: Estatísticas Descritivas das Variáveis Quantitativas de Condições de Saúde e da Variável Resposta Falta. Tangará da Serra, 2008.

Variáveis	n	Média	Desvio Padrão	Mínimo	Máximo	P-valor*
Faltas	211	7,498	6,770	0	48	<0,001
Soma Tosse	211	14,801	15,580	0	88	<0,001
Média Tosse	211	0,744	1,299	0	7	<0,001
Média Tosse L1	211	0,782	1,352	0	7	<0,001
Soma Posto Médico	211	0,938	1,735	0	13	<0,001
Média Posto Médico	211	0,166	0,454	0	3	<0,001
Média Posto Médico L1	211	0,052	0,262	0	2	<0,001
Soma Coceira no Ouvido	211	3,924	6,347	0	40	<0,001
Média Coceira no Ouvido	211	0,237	0,655	0	6	<0,001
Média Coceira no Ouvido L1	211	0,228	0,666	0	6	<0,001
Soma Coceira na Garganta	211	4,374	7,474	0	50	<0,001
Média Coceira na Garganta	211	0,337	0,854	0	6	<0,001
Média Coceira na Garganta L1	211	0,256	0,857	0	7	<0,001
Soma Coriza	211	12,062	14,997	0	83	<0,001
Média Coriza	211	0,597	1,135	0	6	<0,001
Média Coriza L1	211	0,574	1,099	0	6	<0,001
Soma Dor de Cabeça	211	5,683	6,961	0	52	<0,001
Média Dor de Cabeça	211	0,417	0,882	0	6	<0,001
Média Dor de Cabeça L1	211	0,337	0,826	0	7	<0,001
Soma Lacrimejamento	211	2,948	6,054	0	43	<0,001
Média Lacrimejamento	211	0,161	0,588	0	5	<0,001
Média Lacrimejamento L1	211	0,204	0,705	0	5	<0,001
Soma Uso de Remédio	211	5,972	12,918	0	112	<0,001
Média Uso de Remédio	211	0,303	0,770	0	5	<0,001
Média Uso de Remédio L1	211	0,256	0,670	0	5	<0,001

* Os P-valores em vermelho são aqueles significantes a um nível de 5% para o teste de normalidade de Kolmogorov-Smirnov.

Tabela 8: Estatísticas Descritivas das Variáveis Contínuas Ambientais e de Poluição. Tangará da Serra, 2008.

Variáveis	n	Média	Desvio Padrão	Mínimo	Máximo	P-valor*
Média Temperatura	211	25,944	1,213	21,4	29,05	0,148
Média Temperatura L1	211	25,696	1,163	21,8	28,8	0,168
Média Temperatura L2	211	25,593	1,211	21,8	29,1	0,088
Média Umidade Relativa	211	61,497	8,485	34,1	87,1	0,494
Média Umidade Relativa L1	211	63,174	8,627	34,55	87,3	0,142
Média Umidade Relativa L2	211	63,763	9,050	34,1	84,9	0,324
Média PM10	211	53,341	16,905	14,78	125,22	0,060
Média PM10 L1	211	47,767	16,344	14,89	140,4	0,015
Média PM10 L2	211	50,334	19,000	16,29	142,34	0,037
Média PM10 L3	211	51,635	18,068	18,110	150,480	0,013
Média PM10 L4	211	51,819	16,929	14,740	135,010	0,113
Média PM10 L5	211	52,814	15,653	15,070	135,710	0,082
Média BC	211	0,975	0,232	0,36	1,59	0,271
Média BC L1	211	0,860	0,193	0,29	1,48	0,466
Média BC L2	211	0,781	0,210	0,27	1,62	0,082
Média BC L3	211	0,766	0,199	0,27	1,54	0,082
Média BC L4	211	0,793	0,222	0,27	1,63	0,223
Média BC L5	211	0,861	0,242	0,36	1,84	0,057
Média O3	211	36,353	5,116	19,7	49,8	0,117
Média O3 L1	211	34,996	4,206	24,5	49	0,007
Média O3 L2	211	33,027	3,798	21,9	50,8	0,228
Média O3 L3	211	32,851	4,437	18	53,2	0,404
Média O3 L4	211	33,019	3,934	18	48,23	0,148
Média O3 L5	211	33,284	4,488	18	49,15	0,057

* Os P-valores em vermelho são aqueles significantes a um nível de 5% para o teste de normalidade de Kolmogorov-Smirnov.

Tabela 9: Estatísticas Descritivas das Variáveis Contínuas de Características Individuais. Tangará da Serra, 2008.

Variáveis	n	Média	Desvio Padrão	Mínimo	Máximo	P-valor*
FVC %	211	95,943	17,089	0	137	0,001
FEV1 %	211	97,891	17,331	0	135	0,014
FEV1/FVC %	211	99,431	6,022	78	113	0,236
PEFR %	211	95,664	20,690	47	157	0,577
FEF2575 %	211	101,692	27,019	0	169	0,483
Média do PF	211	291,472	81,944	145	525,36	0,042
Mediana do PF	211	292,287	82,878	150	520	0,002
Altura	211	140,085	14,942	111	173	0,239
Peso	211	36,208	13,395	17	82,2	0,023
Idade	211	10,260	2,526	6	15	0,002

* Os P-valores em vermelho são aqueles significantes a um nível de 5% para o teste de normalidade de Kolmogorov-Smirnov.

Tabela 10: Correlações entre as variáveis quantitativas. Tangará da Serra, 2008.

	Faltas	Média Temp	Média Temp L1	Média Temp L2	Média UR	Média UR L1	Média UR L2	Média PM10	Média PM10 L1	Média PM10 L2	Média PM10 L3	Média PM10 L4	Média PM10 L5
Faltas	1												
Média Temp	-0,060	1											
Média Temp L1	0,000	0,669	1										
Média Temp L2	0,049	-0,004	0,581	1									
Média UR	0,079	-0,640	-0,495	-0,116	1								
Média UR L1	0,045	-0,524	-0,703	-0,387	0,891	1							
Média UR L2	0,020	-0,239	-0,632	-0,697	0,659	0,855	1						
Média PM10	-0,010	0,274	0,125	0,082	-0,157	-0,041	-0,051	1					
Média PM10 L1	0,057	0,116	0,074	0,134	-0,072	-0,024	-0,021	0,777	1				
Média PM10 L2	0,029	0,166	0,045	0,071	-0,060	0,025	0,035	0,714	0,890	1			
Média PM10 L3	0,216	-0,007	0,059	0,162	-0,140	-0,150	-0,156	0,323	0,396	0,427	1		
Média PM10 L4	0,188	0,093	0,127	0,138	-0,188	-0,200	-0,199	0,251	0,328	0,352	0,889	1	
Média PM10 L5	0,187	0,152	0,250	0,175	-0,209	-0,280	-0,287	0,243	0,270	0,269	0,762	0,883	1
Média BC	-0,037	0,297	0,247	0,151	-0,349	-0,300	-0,307	0,712	0,415	0,389	0,323	0,320	0,359
Média BC L1	0,011	0,131	0,308	0,274	-0,286	-0,343	-0,312	0,638	0,517	0,405	0,349	0,258	0,277
Média BC L2	0,005	0,082	0,296	0,457	-0,230	-0,304	-0,357	0,524	0,584	0,629	0,510	0,384	0,290
Média BC L3	0,006	0,046	0,194	0,334	-0,272	-0,313	-0,334	0,319	0,508	0,553	0,652	0,565	0,408
Média BC L4	0,006	0,146	0,239	0,275	-0,336	-0,375	-0,394	0,255	0,429	0,463	0,573	0,727	0,598
Média BC L5	0,036	0,201	0,366	0,283	-0,300	-0,427	-0,452	0,248	0,348	0,353	0,490	0,653	0,787
Média O3	-0,096	0,504	0,229	-0,012	-0,614	-0,463	-0,295	0,462	0,217	0,214	0,056	0,102	0,090
Média O3 L1	-0,092	0,471	0,397	0,179	-0,570	-0,516	-0,411	0,592	0,309	0,260	0,050	0,045	0,113
Média O3 L2	-0,047	0,198	0,367	0,410	-0,516	-0,576	-0,597	0,479	0,468	0,398	0,265	0,149	0,147
Média O3 L3	-0,026	0,247	0,321	0,357	-0,429	-0,424	-0,413	0,384	0,497	0,446	0,314	0,317	0,152
Média O3 L4	0,020	0,031	0,122	0,248	-0,401	-0,419	-0,427	0,327	0,416	0,363	0,374	0,345	0,227
Média O3 L5	0,000	0,193	0,308	0,350	-0,245	-0,333	-0,388	0,226	0,456	0,504	0,447	0,559	0,498

Tabela 11: Continuação das correlações entre as variáveis quantitativas. Tangará da Serra, 2008.

	Média BC	Média BC L1	Média BC L2	Média BC L3	Média BC L4	Média BC L5	Média O3	Média O3 L1	Média O3 L2	Média O3 L3	Média O3 L4	Média O3 L5
Faltas												
Média Temp												
Média Temp L1												
Média Temp L2												
Média UR												
Média UR L1												
Média UR L2												
Média PM10												
Média PM10 L1												
Média PM10 L2												
Média PM10 L3												
Média PM10 L4												
Média PM10 L5												
Média BC	1											
Média BC L1	0,757	1										
Média BC L2	0,486	0,734	1									
Média BC L3	0,230	0,439	0,782	1								
Média BC L4	0,277	0,307	0,576	0,831	1							
Média BC L5	0,339	0,329	0,441	0,612	0,818	1						
Média O3	0,674	0,352	0,096	-0,068	0,059	0,056	1					
Média O3 L1	0,671	0,606	0,326	0,034	0,051	0,139	0,765	1				
Média O3 L2	0,540	0,607	0,560	0,366	0,255	0,242	0,517	0,660	1			
Média O3 L3	0,270	0,475	0,640	0,585	0,473	0,252	0,219	0,375	0,564	1		
Média O3 L4	0,154	0,351	0,480	0,590	0,515	0,352	0,141	0,310	0,553	0,612	1	
Média O3 L5	0,176	0,205	0,453	0,612	0,708	0,677	0,055	0,103	0,356	0,542	0,489	1

Tabela 12: Frequências das Variáveis Categóricas. Tangará da Serra, 2008.

Variável		n	%
Sexo	Homem	100	47,4
	Mulher	111	52,6
	Total	211	100,0
Trabalha fora da escola	Sim	41	19,4
	Não	170	80,6
	Total	211	100,0
Tipo de moradia	Madeira	65	31,3
	Alvenaria	140	67,3
	Outros	3	1,4
	Total	208	100,0
Tamanho da casa	Menor de 30m ²	58	27,8
	De 30m² a 80m²	135	64,6
	Maior que 80m ²	16	7,7
	Total	209	100,0
Casa possui forro	Sim	58	27,6
	Não	152	72,4
	Total	210	100,0
Número de comodors	1 comodo e 1 banheiro	4	1,9
	mais comodors	207	98,1
	Total	211	100,0
Quantas pessoas moram na residencia	Menos que 3	8	3,8
	De 3 a 5	133	63,6
	Mais que 5	68	32,5
	Total	209	100,0
Destino do lixo doméstico	Queimado	3	1,4
	Enterrado	2	0,9
	Coleta Pública	202	95,7
	Jogado em terreno baldio	4	1,9
	Total	211	100,0
Destino do lixo do quintal	Queimado	20	9,6
	Enterrado	38	18,2
	Coleta Pública	151	72,2
	Total	209	100,0
Destino dos esgotos da casa	Céu aberto	4	1,9
	Fossa negra	15	7,1
	Fossa séptica	191	91,0
	Total	210	100,0
Casa possui energia elétrica	Sim	207	98,1
	Não	4	1,9
	Total	211	100,0
De onde vem a água para beber	Rede encanada	183	87,6
	Cisterna	1	0,5
	Mineral	3	1,4
	Poço individual	17	8,1
	Outra	5	2,4
	Total	209	100,0

Tabela 13: Frequências das Variáveis Categóricas 2. Tangará da Serra, 2008.

Variável		n	%
Rua é asfaltada	Sim	99	47,4
	Não	110	52,6
	Total	209	100,0
Renda total da família	Menos de 1 salario min	7	3,4
	1 a 2 salarios min	107	51,7
	3 a 4 salarios min	78	37,7
	5 ou mais salarios min	15	7,2
	Total	207	100,0
Possui alguma doença atualmente	Sim	55	26,1
	Não	156	73,9
	Total	211	100,0
Faz algum tratamento para a saúde	Sim	23	11,3
	Nao	172	84,7
	Não se aplica	8	3,9
	Total	203	100,0
Membro da família tem problema de saúde	Sim	106	50,5
	Não	104	49,5
	Total	210	100,0
Nasceu com algum problema de saúde	Sim	25	12,1
	Não	182	87,9
	Total	207	100,0
Nasceu ou desenvolveu um problema fisico	Sim	45	21,3
	Não	166	78,7
	Total	211	100,0
Algum fumante em casa	Sim	72	34,1
	Não	124	58,8
	Sim, mas parou de fumar	15	7,1
	Total	211	100,0
Alérgico a algum alimento	Sim	8	3,8
	Não	202	96,2
	Total	210	100,0
Alérgico a algum produto de limpeza	Sim	15	7,1
	Não	196	92,9
	Total	211	100,0
Mora perto de uma plantação	Sim	51	25,0
	Não	153	75,0
	Total	204	100,0
Ocorrem queimadas proximo a sua casa	Sim	144	68,9
	Não	65	31,1
	Total	209	100,0
Tem celular	Sim	63	30,0
	Não	147	70,0
	Total	210	100,0
Possui asma	Sim	39	18,5
	Não	172	81,5
	Total	211	100,0

Tabela 14: Médias das Variáveis Ambientais e de Poluição com lag 1 por quantidade de faltas. Tangará da Serra, 2008.

Faltas	n	%	Temp	UR	PM10	BC	O3
0	10	4,7	25,925	61,762	51,911	0,871	34,107
1	10	4,7	25,680	63,700	31,048	0,737	33,381
2	20	9,5	26,125	60,775	48,424	0,867	35,678
3	17	8,1	25,384	63,678	47,721	0,840	34,177
4	20	9,5	25,800	63,214	46,236	0,828	36,983
5	24	11,4	25,199	65,205	50,286	0,919	36,166
6	15	7,1	25,899	63,383	48,168	0,862	34,287
7	17	8,1	25,824	62,116	48,480	0,900	35,683
8	17	8,1	25,565	62,210	45,415	0,863	34,684
9	11	5,2	25,469	62,595	50,301	0,879	35,520
10 a 13	22	10,4	25,937	64,153	48,031	0,823	34,081
14 a 17	14	6,6	25,538	62,988	53,721	0,900	34,934
18 ou mais	14	6,6	25,762	64,415	46,938	0,839	33,351

Tabela 15: Médias dos valores da soma das Variáveis de Condições de Saúde. Tangará da Serra, 2008.

Faltas	n	%	Tosse	Posto Méd.	Coc. Ouvido	Coc. Garganta	Coriza	Dor Cabeça	Lacrimenj.	Remédio
0	10	4,7	18,300	0,900	3,500	7,800	16,100	8,300	5,600	7,700
1	10	4,7	18,300	0,400	5,400	4,900	8,300	6,300	3,700	14,200
2	20	9,5	17,500	0,850	4,800	4,350	14,250	5,700	0,950	4,800
3	17	8,1	13,529	0,647	2,647	5,882	11,059	5,177	1,000	2,588
4	20	9,5	12,700	0,300	3,350	4,100	9,950	8,450	3,550	4,100
5	24	11,4	13,500	0,667	3,708	4,875	11,083	3,542	2,708	5,458
6	15	7,1	15,200	1,867	7,333	5,800	18,867	7,267	4,067	11,267
7	17	8,1	10,941	0,824	2,471	1,059	9,235	5,235	2,059	7,882
8	17	8,1	21,882	1,588	5,235	5,941	17,706	6,000	6,000	4,765
9	11	5,2	14,273	1,636	3,455	3,364	13,000	6,636	2,727	9,546
10 a 13	22	10,4	14,046	1,091	4,318	3,500	9,136	4,818	3,227	6,773
14 a 17	14	6,6	17,429	1,429	2,357	3,286	14,714	5,357	2,000	2,071
18 ou mais	14	6,6	7,357	0,286	2,500	3,143	5,143	3,071	2,143	1,500

Tabela 16: Média das médias ponderadas das Variáveis de Condições de Saúde. Tangará da Serra, 2008.

Faltas	n	%	Tosse	Posto Méd.	Coc. Ouvido	Coc. Garganta	Coriza	Dor Cabeça	Lacrimaj.	Remédio
0	10	4,7	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1	10	4,7	0,200	0,100	0,000	0,100	0,100	0,100	0,000	0,100
2	20	9,5	0,200	0,050	0,100	0,250	0,250	0,000	0,000	0,000
3	17	8,1	0,412	0,059	0,000	0,177	0,177	0,235	0,000	0,118
4	20	9,5	0,350	0,050	0,050	0,250	0,350	0,300	0,150	0,100
5	24	11,4	0,625	0,292	0,167	0,208	0,250	0,167	0,083	0,375
6	15	7,1	1,067	0,267	0,333	0,267	0,667	0,667	0,067	0,400
7	17	8,1	0,529	0,177	0,177	0,059	0,647	0,412	0,235	0,294
8	17	8,1	1,235	0,177	0,353	0,471	1,000	0,882	0,529	0,412
9	11	5,2	0,636	0,364	0,455	0,818	1,091	1,000	0,455	0,909
10 a 13	22	10,4	1,046	0,136	0,455	0,364	0,818	0,727	0,182	0,636
14 a 17	14	6,6	1,714	0,429	0,357	0,643	1,357	0,357	0,286	0,214
18 ou mais	14	6,6	1,571	0,071	0,643	0,929	1,214	0,643	0,143	0,357

Tabela 17: Média das médias ponderadas com lag 1 das Variáveis de Condições de Saúde. Tangará da Serra, 2008.

Faltas	n	%	Tosse	Posto Méd.	Coc. Ouvido	Coc. Garganta	Coriza	Dor Cabeça	Lacrimaj.	Remédio
0	10	4,7	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1	10	4,7	0,000	0,000	0,000	0,000	0,100	0,000	0,000	0,100
2	20	9,5	0,150	0,000	0,000	0,050	0,100	0,100	0,000	0,050
3	17	8,1	0,235	0,000	0,059	0,118	0,118	0,177	0,059	0,000
4	20	9,5	0,300	0,050	0,000	0,150	0,300	0,250	0,100	0,100
5	24	11,4	0,333	0,083	0,250	0,125	0,250	0,042	0,083	0,250
6	15	7,1	1,200	0,000	0,267	0,200	0,933	0,667	0,267	0,467
7	17	8,1	0,824	0,059	0,118	0,000	0,588	0,118	0,059	0,471
8	17	8,1	1,294	0,000	0,471	0,529	1,118	0,706	0,471	0,235
9	11	5,2	1,091	0,182	0,455	0,818	1,091	1,091	0,455	1,000
10 a 13	22	10,4	1,409	0,046	0,409	0,136	0,727	0,364	0,273	0,409
14 a 17	14	6,6	1,857	0,143	0,214	0,643	1,429	0,500	0,286	0,214
18 ou mais	14	6,6	1,500	0,143	0,714	0,857	0,929	0,643	0,714	0,143

Tabela 18: Média dos valores das Variáveis de Espirometria de acordo com a quantidade de Faltas. Tangará da Serra, 2008.

Faltas	n	%	Média PF	Mediana PF	FVC %	FEV1 %	FEV1/FVC %	PEFR %	FEF 25-75 %
0	10	4,7	290,418	293,000	100,500	106,600	103,400	113,400	124,100
1	10	4,7	275,733	277,000	99,900	105,100	101,500	97,900	108,000
2	20	9,5	281,484	281,000	99,650	100,650	98,550	100,300	102,350
3	17	8,1	278,709	279,412	94,706	96,059	98,882	91,000	97,706
4	20	9,5	319,730	322,250	89,950	90,900	99,400	86,450	92,850
5	24	11,4	298,414	297,917	99,542	100,958	98,833	97,500	103,750
6	15	7,1	278,845	278,333	93,600	96,800	100,600	99,800	109,533
7	17	8,1	299,138	301,765	99,941	101,824	99,824	93,529	108,000
8	17	8,1	274,502	274,706	93,588	95,412	99,177	90,294	97,941
9	11	5,2	254,620	257,500	89,636	92,727	98,182	90,727	92,455
10 a 13	22	10,4	310,789	310,909	93,591	94,046	99,000	95,909	93,455
14 a 17	14	6,6	269,471	270,714	96,071	98,357	99,571	107,643	106,071
18 ou mais	14	6,6	326,395	326,429	97,500	98,571	98,214	86,643	98,500

Tabela 19: Média dos valores das Variáveis de Características Individuais de acordo com a quantidade de Faltas. Tangará da Serra, 2008.

Faltas	n	%	Altura	Peso	Idade
0	10	4,7	141,200	35,500	9,900
1	10	4,7	133,300	31,780	9,400
2	20	9,5	137,550	31,655	9,650
3	17	8,1	139,471	35,847	10,240
4	20	9,5	148,550	41,790	11,300
5	24	11,4	142,208	38,883	10,960
6	15	7,1	137,200	34,327	9,600
7	17	8,1	139,882	37,706	10,470
8	17	8,1	140,588	34,129	10,120
9	11	5,2	130,182	30,027	8,640
10 a 13	22	10,4	143,273	38,596	10,550
14 a 17	14	6,6	132,571	31,371	9,290
18 ou mais	14	6,6	145,786	42,914	11,710

ANEXO B – Modelo de Regressão Quasi-Poisson

Tabela 20: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor do Teste de Wald. Tangará da Serra, 2008.

Variáveis	<i>RR</i>	IC95%	P-valor
Sintomas Gerais	0,993313	(0,987191 ; 0,999473)	0,034571
Soma Tosse	0,994353	(0,986087 ; 1,002688)	0,185052
Média Tosse	1,173943	(1,089982 ; 1,264373)	0,000034
Média Tosse L1	1,190629	(1,113503 ; 1,273098)	0,000001
Soma Posto Médico	1,017554	(0,951034 ; 1,088728)	0,614443
Média Posto Médico	1,197930	(0,945944 ; 1,517041)	0,135434
Média Posto Médico L1	1,490240	(1,059422 ; 2,096251)	0,022927
Soma Coceira no Ouvido	0,990687	(0,970308 ; 1,011494)	0,378637
Média Coceira no Ouvido	1,298470	(1,156390 ; 1,458007)	0,000016
Média Coceira no Ouvido L1	1,307435	(1,172103 ; 1,458392)	0,000003
Soma Coceira na Garganta	0,987464	(0,969374 ; 1,005892)	0,182583
Média Coceira na Garganta	1,211516	(1,096548 ; 1,338537)	0,000211
Média Coceira na Garganta L1	1,206054	(1,099856 ; 1,322507)	0,000094
Soma Coriza	0,994259	(0,985669 ; 1,002924)	0,194885
Média Coriza	1,207455	(1,114743 ; 1,307878)	0,000007
Média Coriza L1	1,192368	(1,092961 ; 1,300818)	0,000102
Soma Dor de Cabeça	0,988392	(0,969507 ; 1,007645)	0,236874
Média Dor de Cabeça	1,201151	(1,081156 ; 1,334464)	0,000771
Média Dor de Cabeça L1	1,246941	(1,139719 ; 1,364250)	0,000003
Soma Lacrimejamento	0,994913	(0,974087 ; 1,016185)	0,637088
Média Lacrimejamento	1,117603	(0,936394 ; 1,333880)	0,219380
Média Lacrimejamento L1	1,241477	(1,104024 ; 1,396044)	0,000379
Soma Remédio	0,991265	(0,978820 ; 1,003868)	0,174929
Média Remédio	1,151095	(1,007999 ; 1,314504)	0,038964
Média Remédio L1	1,082837	(0,917766 ; 1,277598)	0,346719
Média Temperatura Média	0,956488	(0,866037 ; 1,056384)	0,381087
Média Temperatura Média L1	1,006856	(0,905742 ; 1,119259)	0,899422
Média Temperatura Média L2	1,047393	(0,946361 ; 1,159212)	0,371967
Média Umidade Relativa	1,008028	(0,993606 ; 1,022659)	0,278032
Média Umidade Relativa L1	1,003834	(0,989539 ; 1,018336)	0,601576
Média Umidade Relativa L2	1,000756	(0,987214 ; 1,014483)	0,913579

* Os P-valores em vermelho são significantes a um nível de 20% para o teste de Wald.

Tabela 21: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor do Teste de Wald 2. Tangará da Serra, 2008.

Variáveis	\overline{RR}	IC95%	P-valor
Média PM10	0,999178	(0,991969 ; 1,006439)	0,824000
Média PM10 L1	1,003005	(0,995797 ; 1,010266)	0,415773
Média PM10 L2	1,001348	(0,995030 ; 1,007706)	0,676958
Média PM10 L3	1,003413	(0,996960 ; 1,009909)	0,301815
Média PM10 L4	1,003811	(0,996896 ; 1,010773)	0,282046
Média PM10 L5	1,003391	(0,995823 ; 1,011017)	0,381814
Média BC	0,864240	(0,512717 ; 1,456771)	0,584478
Média BC L1	1,032339	(0,548908 ; 1,941534)	0,921425
Média BC L2	1,018932	(0,571531 ; 1,816563)	0,949368
Média BC L3	1,028573	(0,559403 ; 1,891234)	0,927850
Média BC L4	1,041488	(0,604732 ; 1,793683)	0,883617
Média BC L5	1,163734	(0,708865 ; 1,910485)	0,549467
Média O3	0,984303	(0,961483 ; 1,007663)	0,187584
Média O3 L1	0,981100	(0,953073 ; 1,009952)	0,198363
Média O3 L2	0,990861	(0,959426 ; 1,023325)	0,577313
Média O3 L3	0,995423	(0,968365 ; 1,023236)	0,744515
Média O3 L4	1,006085	(0,975501 ; 1,037626)	0,700515
Média O3 L5	1,002259	(0,975498 ; 1,029753)	0,870374
FVC %	0,998182	(0,991378 ; 1,005033)	0,602578
FEV1 %	0,997152	(0,990626 ; 1,003721)	0,395591
FEV1FVC %	0,991788	(0,972187 ; 1,011784)	0,419050
PEFR %	0,996878	(0,990928 ; 1,002863)	0,307062
FEF2575 %	0,997828	(0,993405 ; 1,002271)	0,338454
Média do Pico de Fluxo	1,000891	(0,999452 ; 1,002333)	0,226484
Mediana do Pico de Fluxo	1,000860	(0,999437 ; 1,002285)	0,237638
Média do Pico de Fluxo 2	1,002051	(1,000991 ; 1,003112)	0,000195
Altura	1,001666	(0,993564 ; 1,009835)	0,688255
Peso	1,005181	(0,996535 ; 1,013902)	0,242348
Sexo	0,886113	(0,694725 ; 1,130226)	0,331216
Idade	1,025830	(0,977893 ; 1,076117)	0,297480
Idade da Mãe	0,991179	(0,976556 ; 1,006021)	0,243987
Idade de quem cuida	0,986552	(0,977475 ; 0,995713)	0,004516

* Os P-valores em vermelho são aqueles significantes a um nível de 20% para o teste de Wald.

Tabela 22: Estimativas dos Riscos Relativos (RR) e Intervalos de Confiança de 95% (IC95%) e P-valor do Teste de Wald 3. Tangará da Serra, 2008.

Variáveis		RR	IC95%	P-valor
Trabalha fora da escola	Sim	1	-	-
	Não	0,793915	(0,598637 ; 1,052892)	0,110626
Tipo de moradia	Madeira	1	-	-
	Alvenaria	0,840064	(0,653443 ; 1,079983)	0,175424
	Outros	0,467548	(0,117287 ; 1,863809)	0,282508
Tamanho da casa	Menor de 30m2	1	-	-
	De 30m2 a 80m2	0,748155	(0,577342 ; 0,969504)	0,029338
	Maior que 80m2	0,774876	(0,473996 ; 1,266749)	0,310307
Casa possui forro	Sim	1	-	-
	Não	1,245307	(0,934681 ; 1,659164)	0,135501
Numero de comodors	1 comodo e 1 banheiro	1	-	-
	Mais comodors	0,356918	(0,212754 ; 0,598770)	0,000128
Localização do Banheiro	Dentro de casa	1	-	-
	Fora de casa	1,062957	(0,815254 ; 1,385922)	0,652427
Quantas pessoas moram na residencia	Menos que 3	1	-	-
	De 3 a 5	0,777904	(0,426503 ; 1,418828)	0,413683
	Mais que 5	1,045248	(0,567411 ; 1,925488)	0,887235
Destino do lixo domestico	Queimado	1	-	-
	Enterrado	0,626423	(0,161969 ; 2,422712)	0,498680
	Coleta Pública	0,530956	(0,251989 ; 1,118758)	0,097449
	Jogado em terreno baldio	0,558802	(0,181440 ; 1,721005)	0,311756
Destino do lixo do quintal	Queimado	1	-	-
	Enterrado	0,905060	(0,545845 ; 1,500672)	0,699413
	Coleta Pública	1,114863	(0,727913 ; 1,707510)	0,617673
Destino do esgoto da casa	Céu aberto	1	-	-
	Fossa negra	0,407353	(0,199884 ; 0,830165)	0,014226
	Fossa séptica	0,425411	(0,241768 ; 0,748547)	0,003387
Possui energia elétrica	Sim	1	-	-
	Nao	1,374740	(0,645430 ; 2,928144)	0,410302
De onde vem a água para beber	Rede encanada	1	-	-
	Cisterna	0,931862	(0,147099 ; 5,903303)	0,940348
	Mineral	1,104102	(0,412986 ; 2,951776)	0,843722
	Poço individual	1,228654	(0,811694 ; 1,859804)	0,331415
	Outra	0,769915	(0,308555 ; 1,921112)	0,575767
A rua é asfaltada	Sim	1	-	-
	Nao	0,969182	(0,759363 ; 1,236976)	0,801692
Renda total da família	Menos de 1 salario min	1	-	-
	1 a 2 salarios min	1,390247	(0,650743 ; 2,970125)	0,395939
	3 a 4 salarios min	1,251538	(0,580489 ; 2,698326)	0,567667
	5 ou mais salarios min	0,960351	(0,389319 ; 2,368944)	0,930105
Possui alguma doença atualmente	Sim	1	-	-
	Nao	1,184847	(0,890367 ; 1,576724)	0,245968
Faz algum tratamento para a saúde	Sim	1	-	-
	Nao	0,601724	(0,376156 ; 0,962558)	0,035351
Membro da familia tem problema de saude	Sim	1	-	-
	Nao	1,025076	(0,803287 ; 1,308100)	0,842380
Nasceu com algum problema de saúde	Sim	1	-	-
	Nao	1,453153	(0,945963 ; 2,232278)	0,089454
Nasceu ou desenvolveu algum problema físico	Sim	1	-	-
	Nao	1,092893	(0,807847 ; 1,478517)	0,565169
Algum fumante em casa	Sim	1	-	-
	Nao	0,872148	(0,672750 ; 1,130645)	0,302859
Alergico a algum alimento	Sim	1	-	-
	Nao	0,881049	(0,484024 ; 1,603737)	0,679005
Alergico a algum produto de limpeza	Sim	1	-	-
	Nao	1,582120	(0,896664 ; 2,791573)	0,114817
Mora perto de uma plantação	Sim	1	-	-
	Não	0,741030	(0,569292 ; 0,964578)	0,026976
Ocorrem queimadas próximo a sua casa	Sim	1	-	-
	Nao	1,054367	(0,812283 ; 1,368598)	0,691192
Tem celular	Sim	1	-	-
	Nao	0,917666	(0,707611 ; 1,190076)	0,517784
Possui asma	Sim	1	-	-
	Nao	1,035075	(0,753815 ; 1,421277)	0,831461

* Os P-valores em vermelho são aqueles que deram significantes a um nível de 20% para o teste de Wald.

Tabela 23: Tabela de Correlação das Variáveis Quantitativas e Significantes no Teste de Wald. Tangará da Serra, 2008.

	PF3	SG1	TosT	Tos	TosL1	PMed	PMedL1	COuv	COuvL1	CGarT	CGar	CGarL1	CorT	Cor	CorL1	DCab	DCabL1	LacL1	RemT	Rem	O3	O3L1	
PF3	1																						
SG1	-0,334	1																					
TosT	0,360	0,791	1																				
Tos	-0,110	0,429	0,605	1																			
TosL1	-0,116	0,416	0,555	0,825	1																		
PMed	-0,100	0,118	0,088	0,161	0,051	1																	
PMedL1	0,024	0,034	0,066	0,165	0,207	0,207	1																
COuv	0,026	0,156	0,127	0,256	0,204	0,091	-0,045	1															
COuvL1	-0,003	0,221	0,139	0,244	0,251	0,064	-0,041	0,695	1														
CGarT	-0,126	0,582	0,506	0,296	0,205	0,135	0,116	0,197	0,261	1													
CGar	0,047	0,249	0,277	0,379	0,340	0,126	0,198	0,538	0,535	0,517	1												
CGarL1	0,033	0,220	0,241	0,363	0,422	0,062	0,152	0,519	0,549	0,442	0,838	1											
CorT	-0,336	0,802	0,757	0,427	0,450	0,092	-0,006	0,131	0,222	0,439	0,200	0,201	1										
Cor	-0,140	0,460	0,478	0,608	0,600	0,195	0,039	0,366	0,475	0,215	0,371	0,410	0,609	1									
CorL1	-0,123	0,459	0,485	0,594	0,674	0,181	0,061	0,306	0,426	0,198	0,357	0,415	0,635	0,873	1								
DCab	0,035	0,244	0,163	0,372	0,344	0,219	0,194	0,314	0,324	0,188	0,382	0,432	0,105	0,287	0,258	1							
DCabL1	0,040	0,240	0,168	0,214	0,309	0,105	0,051	0,318	0,389	0,236	0,379	0,517	0,178	0,379	0,385	0,591	1						
LacL1	0,135	0,085	0,046	0,281	0,337	-0,017	0,252	0,225	0,317	0,054	0,400	0,371	0,071	0,234	0,217	0,445	0,193	1					
RemT	-0,213	0,601	0,265	0,099	0,103	0,095	0,020	-0,023	0,063	0,303	0,053	0,051	0,257	0,101	0,084	0,204	0,167	0,030	1				
Rem	-0,050	0,338	0,188	0,373	0,311	0,223	0,228	0,159	0,227	0,269	0,307	0,308	0,165	0,293	0,221	0,549	0,355	0,289	0,522	1			
O3	0,037	-0,166	-0,177	-0,045	-0,058	-0,040	-0,022	-0,030	-0,047	-0,110	-0,066	-0,018	-0,141	-0,026	-0,059	-0,074	-0,069	0,019	-0,045	0,024	1		
O3L1	0,023	-0,131	-0,076	-0,019	-0,006	-0,014	0,008	0,003	-0,031	-0,076	-0,020	0,034	-0,129	-0,006	-0,026	-0,008	0,001	0,037	-0,044	0,030	0,765	1	

* Os valores das correlações em vermelho são aqueles que deram significantes a um nível de 5%.