Talita de Souza Costa Silva

MODELOS DE REGRESSÃO ADITIVOS DINÂMICOS APLICADOS A DADOS LOGITUDINAIS

BELO HORIZONTE-MG

25 DE MARÇO/2013

UNIVERSIDADE FEDERAL DE MINAS GERAIS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

MODELOS DE REGRESSÃO ADITIVOS DINÂMICOS APLICADOS A DADOS LOGITUDINAIS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Modelagem Estatística e Computacional

Orientador: Prof. Dra Lourdes Coral Contreras Montenegro
Co-orientador: Prof. Dra. Rosemeire Leovigildo Fiaccone

BELO HORIZONTE-MG
MARÇO/2013

UNIVERSIDADE FEDERAL DE MINAS GERAIS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

MODELOS DE REGRESSÃO ADITIVOS DINÂMICOS APLICADOS A DADOS LOGITUDINAIS

Talita de Souza Costa Silva

Dissertação julgada adequada para obtenção do título de mestre em Estatística, defendida e aprovada por unanimidade em 25/03/2013 pela Comissão Examinadora.

Orientador:

Prof. Dra. Lourdes Coral Contreras Montenegro
Universidade Federal de Minas Gerais

Banca Examinadora:

Prof. Dra. Lourdes Coral Contreras Montenegro
Universidade Federal de Minas Gerais
DEE-UFMG

Prof. Dr. Fábio Nogueira Demarqui
Universidade Federal de Minas Gerais
DEE-UFMG

Profª. Dra. Suely Ruiz Giolo
Universidade Federal do Paraná
DEE-UFPR

Dedico esta dissertação,
a minha mãe Rosa Costa Silva,
pelo apoio incondicional em todos os momentos.

Agradecimentos

Agradeço primeiramente a Deus, pela proteção constante e por todas as maravilhosas oportunidades concedidas, fundamentais para esta conquista.

Aos meus pais, Pedro e Rosa, pelo apoio, tornando este projeto possível.

Às minhas irmãs Jamile e Juliane, por serem tão especiais, apesar da distância.

Aos meus tios, Tonho e Glória, por acreditarem em mim enquanto tudo ainda era um sonho.

Ao meu namorado Alex, pelo carinho, cumplicidade e paciência.

À Silvana e Mariana, pelo companheirismo e momentos de distração e descontração, foi muito bom morar com vocês!

Aos meus amigos da Pós-Graduação em Estatística; Nívea por toda ajuda nos momentos difíceis, companheirismo e amizade, Gabriel, Paulo, Fabrícia, Wecsley e Ricardo, pela amizade, momentos gastronômicos e de distração.

Ao meu amigo Samuel, que apesar da distância mostrou-se sempre presente na minha vida pessoal e profissional.

À minha orientadora, Lourdes Montenegro, pelos ensinamentos, confiança e dedicação durante a realização deste trabalho.

À minha co-orientadora, Rosemeire Fiaccone, pelos primeiros conhecimentos e comprometimento.

Aos professores e funcionários do Departamento de Estatística da UFMG pela oportunidade e por terem contribuído na minha formação.

À Capes, pela bolsa concedida.

Enfim, agradeço a todos que me ajudaram e torceram por mim! Muito obrigada!

"*Quando estiver em dificuldade*
E pensar em desistir,
Lembre-se dos obstáculos que já superou
OLHE PARA TRÁS
Se tropeçar e cair, levante,
Não fique prostrado, esqueça o passado.
OLHE PARA FRENTE
Ao sentir-se orgulhoso, por alguma realização pessoal
Sonde suas motivações.
OLHE PARA DENTRO
Antes que o egoísmo o domine,
Enquanto seu coração é sensível
Socorra aos que o cercam.
OLHE PARA OS LADOS
Na escalada, rumo às altas posições,
No afã de concretizar seus sonhos,
Observe se não está pisando em alguém.
OLHE PARA BAIXO
Em todos os momentos da vida,
Seja qual for sua atividade,
Busque a aprovação de DEUS!
OLHE PARA CIMA
"

Charles Chaplin-Direcione seu olhar

Resumo

Este trabalho concentra-se nos modelos de regressão aditivos dinâmicos para dados longitudinais, nas versões sugeridas por Borgan et al. (2007) e Martinussen e Scheike (1999, 2000). A metodologia é destinada a dados de eventos recorrentes que são frequentemente de interesse em estudos longitudinais envolvendo múltiplas unidades de análise. Nas duas versões é modelada a média condicional dada a história prévia em função de covariáveis fixas e tempo dependentes. Com base nos resultados assintóticos são apresentados os intervalos de confiança, as bandas de confiança e os testes de hipóteses. A bondade do ajuste é considerada por meio do uso de resíduos martingais. Duas aplicações são apresentadas, cada uma referente as duas metodologias estudadas. Os dados provêm do estudo de Serrinha realizado pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia. Em cada aplicação partes diferentes dos dados são utilizados, mas com o mesmo objetivo de avaliar o efeito da suplementação periódica de vitamina A sobre a diarreia em crianças menores de 5 anos.

Palavras-chave: funções de regressão cumulativas, modelos aditivos dinâmicos, eventos recorrentes, dados longitudinais, martingais, modelos não-paramétricos, modelos semiparamétricos, coeficientes variáveis no tempo, incidência e prevalência de diarréia.

Abstract

This work focuses on the regression additive dynamic models for longitudinal data, suggested in the versions of Borgan et al. (2007) and Martinussen and Scheike (1999, 2000). The method is intended for recurrent event data of interest in longitudinal studies that involves multiple analysis units. In both versions is modeled conditional mean given the previous history in terms of fixed and time-dependent covariates. Based on the asymptotic results are presented the confidence intervals, the confidence bands and the hypothesis testing. The goodness of fit is considered through the use of martingales residuals.Two applications are presented, each one is referring to the methodologies studied. The data comes from the studies of Serrinha´s city conducted by the Public Health Institute, Federal University of Bahia. In each application different parts of the data are used, but with the same purpose of evaluating the effect of periodic supplementation from vitamin A about the diarrhea in children under 5 years old.

Keywords: regression functions cumulative, dynamic models additives, recurrent events, longitudinal data, martingales, non-parametric models, semiparametric models, time-varying coefficients, incidence and prevalence of diarrhea.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Neste capítulo será introduzido de forma breve o tipo de problema e o tipo de modelo que será considerado sem uma descrição completa dos detalhes técnicos associados às metodologias estatísticas a serem consideradas.

Em pesquisas biomédicas pode-se considerar duas estratégias para coletas de dados. Quando a coleta é realizada em um determinado instante envolvendo uma única observação para cada indivíduo, dizemos que o estudo tem um planejamento transversal. Quando a variável resposta para cada indivíduo envolve duas ou mais observações, realizadas em instantes diferentes, referimo-nos ao planejamento como longitudinal. Estudos longitudinais consistem em um caso especial daqueles conhecidos sob a denominação de medidas repetidas. Por exemplo, pode-se verificar ao longo do tempo a evolução da perda de peso entre pacientes sujeitos a diferentes tipos de dietas.

Com relação aos aspectos técnicos, em geral, análise de dados longitudinais é mais complicada que a análise de dados obtidos sob planejamentos transversais. Entretanto, vale ressaltar que em planejamentos longitudinais pode-se avaliar mudanças globais e/ou individuais ao longo do tempo. Existe uma variedade de procedimentos estatísticos desenvolvidos para a análise e quantificação dessas mudanças. As estratégias analíticas incluem modelos auto-regressivos, modelagem de equações estruturais, modelos de curvas de crescimento, modelos de regressão com coeficientes variando no tempo, entre outros. A escolha do modelo depende da natureza do fenômeno de estudo e da questão de pesquisa (DIGGLE, 2002)

Em muitas aplicações, as unidades de análise ou indivíduos podem experimentar o mesmo evento repetidamente ao longo do tempo, dando origem aos chamados eventos recorrentes. Métodos para lidar com esses tipos de dados têm sido estudados dentro da formulação de processos de contagem e têm sido utilizados em aplicações na análise de sobrevivência multivariada. Entretanto, esses eventos recorrentes são frequentemente de interesse em estudos longitudinais envolvendo múltiplas unidades de análise ou indivíduos

em diversos campos como na medicina, economia e ciências sociais, dentre outros. Por exemplo, um paciente pode experimentar a recorrência de um tumor após a remoção cirúrgica ou pode sofrer diversos ataques epilépticos ou ataques de asma ao longo de um estudo. Outros exemplos na área da saúde envolvem o uso de droga, a hospitalização de pacientes com doenças crônicas e a recorrência de cárie em estudos da saúde bucal. Pena, Slate e Gonzalez (2007) apresentam diversos exemplos em outras áreas.

Segundo Box-Steffensmeier e Boef (2006) existem duas fontes de variabilidade que devem ser consideradas nas análises de eventos recorrentes: heterogeneidade entre indivíduos e dependência intra indivíduos:

- **Heterogeneidade entre indivíduos**: É comum em certos estudos que algumas unidades de análise apresentem uma maior ou menor taxa de eventos recorrentes em relação à outras unidades. Esse fenômeno pode ocorrer devido a fatores desconhecidos ou não mensuráveis. Por exemplo, fatores ambientais ou traços genéticos ou mesmo estilo de vida podem afetar o evento de interesse ainda mais se os mesmos não puderam ser mensurados ou controlados. Além disso, em estudos com seres humanos alguns podem apresentar mais rapidamente o primeiro, segundo, terceiro e assim por diante, eventos recorrentes em relação a outros indivíduos sem necessariamente ser o único fator que aumenta o risco de eventos repetidos.

- **Dependência intra eventos**: A ocorrência de um evento pode influenciar o surgimento ou não de um outro evento para além da suscetibilidade (ou fragilidade) do indivíduo por exemplo. Esta dependência entre os eventos recorrentes pode ser produto da fraqueza ou força biológica do fenômeno estudado. Por exemplo, a manifestação de uma doença infecciosa tende a reduzir o risco de futuro eventos para um mesmo indivíduo devido à imunidade que se adquire à uma referida infecção.

A motivação desse trabalho baseia-se na modelagem de eventos recorrentes sob a ótica de dados longitudinais de um ensaio clínico com estrutura complexa cujo objetivo é avaliar em crianças menores de 5 anos o impacto da suplementação periódica de vitamina A na redução da morbidade por diarreia e infecção respiratória. Dessa forma, espera-se descrever melhor a susceptibilidade, heterogeneidade ou a fragilidade individual de cada criança através do uso de modelos aditivos de intensidade como forma de capturar efeitos variando no tempo de covariáveis tempo dependentes.

O objetivo deste trabalho é aplicar duas metodologias para lidar parcialmente com essas fontes de variabilidade. A primeira foi sugerida por Borgan et al. (2007), em que

a modelagem é aplicada para dados longitudinais binários na presença de dados omissos. A segunda foi proposta por Martinussen e Scheike (1999, 2000) em que os modelos de regressão para dados longitudinais são aplicados às observações censuradas. Em ambos é modelada a média condicional dada a história prévia em função de covariáveis fixas e tempo dependentes. A intenção aqui não é comparar metodologias e sim apresentar duas diferentes abordagens de análise de eventos recorrentes de forma abrangente e também contribuir para um maior subsídio aos pesquisadores da área de saúde, cujas respostas de interesse utilizadas nas diferentes análises são de carácter epidemiológico. Também neste capítulo é apresentada uma breve revisão bibliográfica sobre modelagem de eventos recorrentes e uma sinopse sobre os dados que serão analisados neste trabalho.

1.1 Modelagem de eventos recorrentes - Revisão da Literatura

Dados na análise de sobrevivência clássica usualmente concentram-se no tempo até a ocorrência de um determinado evento para cada indivíduo. Em contraste, muitas aplicações envolvem eventos que ocorrem repetidamente ao longo do tempo, em que indivíduos podem apresentar o mesmo evento diversas vezes num intervalo de tempo. Esses eventos fornecem um tipo especial de dados de sobrevivência multivariado os quais são denominados dados de eventos recorrentes ou eventos repetidos. Esse tipo de conjunto de dados surge naturalmente em estudos longitudinais envolvendo múltiplos indivíduos e consequentemente é preciso levar em consideração a dependência entre os eventos repetidos.

Procedimentos para análise de eventos recorrentes sob a ótica de dados longitudinais têm sido frequentes na literatura e como mencionado acima, incluem os métodos de processos de contagem baseado na intensidade, na análise do tempo entre a ocorrência de eventos, além dos modelos de fragilidade como alternativa para lidar com a heterogeneidade presente em eventos recorrentes. Cheng e Wei (2000) propuseram um modelo semiparamétrico, cuja função de estimação era a função escore de verossimilhança parcial do modelo de intensidade proporcional de Andersen-Gill no ajuste de dados longitudinais desbalanceados no tempo; Diggle, Farewell e Henderson (2007) demonstraram como um modelo geral para dados longitudinais contínuos poderia ser deduzido via análise de sobrevivência baseando-se na idéia do ajustamento de incrementos dado a história prévia da resposta de interesse. Ambas as propostas fundamentaram-se na teoria de processos de contagem.

Os modelos estatísticos baseados na teoria de processos de contagem para analisar eventos recorrentes foram originalmente introduzidos por Aalen (1980). Mais tarde, Martinussen e Scheike (2000), Scheike (2002), como também Lin e Ying (2001) propuseram um modelo de regressão com coeficientes variando no tempo para dados longitudinais. A proposta integrou de forma satisfatória as técnicas de processos de contagem na análise de dados longitudinais estabelecendo assim uma ponte entre análise de sobrevivência, eventos recorrentes e observações tempo-dependentes. Recentemente, Farewell e Henderson (2010) apontaram as conexões existentes entre essas duas áreas metodológicas através da teoria de modelos de eventos recorrentes com censura intervalar.

No modelo de intensidade multiplicativo de Aalen (1980, 1989) é assumindo que a intensidade, $\lambda_i(t)$ pode ser caracterizada por:

$$\lambda_i(t) = \alpha_i(t)Y_i(t)$$

em que $Y_i(t)$ é uma variável aleatória binária assumindo o valor 1 se o indivíduo i está em risco para a ocorrência do evento de interesse no tempo t e 0 em caso contrário e $\alpha(t)$ é uma função determinística não negativa associada ao processo de contagem $N(t)$.

Outra abordagem comum na análise de eventos recorrentes tem sido o uso de modelos de fragilidade (VAUPEL; MANTON; STALLARD, 1979; HOUGAARD, 2000). A idéia do efeito de fragilidade nesse contexto está ligada ao conceito de dependência, ou seja, espera-se que um número excessivo de eventos anteriores possa predizer uma maior intensidade de eventos no futuro. Os autores Kessing e Andersen (2005) argumentam que a fragilidade pode ser um processo estocástico que muda ao longo do tempo. Na verdade, qualquer tipo de dependência observada no processo pode, no sentido da matemática, ser interpretado como efeito do passado, isto é, pode ser dada uma interpretação dinâmica (FIACCONE, 2006).

Uma classe flexível de modelos de regressão capaz de incorporar covariáveis tempo dependentes através do uso da informação da resposta prévia na resposta atual (nomeada de covariável dinâmica) é o modelo de intensidade aditivo de Aalen (1989). Tais covariáveis podem ser definidas como o número de eventos prévios ou o tempo desde o ultimo evento e tem como objetivo melhorar o entendimento do fenômeno estudado. O referido modelo é completamente não-paramétrico uma vez que a estimação dos parâmetros usa apenas informação local além da base martingal associada para flexibilizar inferência de eventos recorrentes no contexto de dados longitudinais. Fiaccone (2006) ressalta que a inclusão da história passada no modelo da intensidade do processo é a forma mais natural de ca-

racterizar a susceptibilidade, heterogeneidade ou a fragilidade individual de desenvolver alguma doença por exemplo.

A idéia do modelo de Aalen (1989) é ajustar o efeito das p covariáveis $X_{i1}(t), X_{i2}(t), \ldots, X_{ip}(t)$ em relação a intensidade do processo, de forma aditiva, ou seja:

$$\alpha_i(t) = \beta_0(t) + \sum_{j=1}^{p} \beta_j(t) x_{ij}(t),$$

em que $\beta_0(t)$ é o risco basal e $\beta_j(t)$, $j = 1, \ldots, p$, são os coeficientes das covariáveis. O processo de estimação geralmente utiliza mínimos quadrados ou mínimos quadrados generalizados em cada momento do evento e inferência é realizada para os coeficientes cumulativos,

$$B_j(t) = \int_0^t \beta_j(s) ds.$$

Essas funções de regressão cumulativas são plotadas em relação ao tempo de forma a caracterizar se a covariável tem efeito constante ou não. Portanto, a mudança na função cumulativa é de interesse principal.

Além dos trabalhos de Martinussen e Scheike (2000), Scheike (2002), Lin e Ying (2001) também propuseram um modelo de regressão com coeficientes variando no tempo para dados longitudinais. Mais tarde, Borgan et al. (2007) sugeriram um modelo de regressão dinâmico aditivo para dados binários longitudinais na presença de dados omissos apoiando-se nos eventos recorrentes ocorridos em tempos discretos em que a média condicional dada a história prévia é ajustada em função de covariáveis tempo dependentes.

1.2 Estudo de Serrinha - Descrição dos dados

O conjunto de dados a ser utilizado nesse trabalho, como aplicação das metodologias a serem apresentadas nos capítulos posteriores, refere-se a um ensaio clínico aleatorizado, duplo-cego, placebo-controlado, realizado pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia, no período de dezembro de 1990 a dezembro de 1991, com o objetivo de avaliar o efeito da suplementação periódica de vitamina A sobre a morbidade e mortalidade em crianças menores de 5 anos - Estudo de Serrinha. O estudo foi realizado na cidade de Serrinha, a 170 Km noroeste de Salvador, Bahia. É uma cidade situada na zona do semi-árido, possuindo cerca de 30.000 habitantes e caracterizada por apresentar clima quente e seco, além de chuvas irregulares. Os serviços de saúde de Serrinha são

deficientes e aquém das necessidades de sua população.

Segundo Fiaccone (1998), o desenho do estudo é do tipo longitudinal formado por uma coorte fixa, com o acompanhamento de 1240 crianças de 6 a 48 meses, com o objetivo de testar o efeito da suplementação de vitamina A sobre a diarreia e a infecção respiratória aguda. As crianças foram aleatorizadas e receberam vitamina A ou placebo a cada 4 meses por um período de um ano. Elas foram visitadas três vezes por semana nos seus lares por entrevistadores que coletaram dados a respeito da ocorrência de diarreia, bem como o número de dejeções líquidas e amolecidas por períodos de 24 horas, bem como informações sobre infecção respiratória. No caso de haver 3 ou mais dejeções líquidas/amolecidas uma investigação mais detalhada acerca de sinais de vômitos, presença de muco ou sangue nas fezes, febre, uso de medicamento, uso de reidratação oral, internação hospitalar, foi conduzida (BARRETO et al., 1994). No caso de ter havido relato de tosse, a frequência respiratória foi medida duas vezes. Se a criança apresentava um número médio superior a 40 bat./min ou se fosse observado chiado no peito, o caso era relatado e o pediatra do projeto investigava o episódio mais profundamente.

Definiu-se como diarreia técnica o registro de três ou mais dejeções líquidas e/ou amolecidas em um período de 24 horas, e delimitou-se como um novo episódio de diarreia o intervalo de três ou mais dias sem diarreia. O intervalo de tempo estabelecido encaixa-se nas recomendações sugeridas em outros estudos, dentre eles, Morris et al., 1994 e Baqui et al. (1991)

As análises que serão apresentadas neste trabalho utilizarão somente uma parte dos dados coletados para este estudo.

O Capítulo 2 apresenta os principais conceitos básicos para melhor entendimento da dinâmica dos modelos. O Capítulo 3 apresenta as propriedades dos dois modelos que serão aplicados neste trabalho. No Capítulo 4 são aplicados os modelos a um conjunto de dados reais. E, finalmente, no Capítulo 5 são feitas as considerações finais do trabalho.

2 Concepções Metodológicas

Nas seções seguintes serão abordados os principais conceitos para melhor entendimento da metodologia empregada no Modelo Aditivo de Aalen.

2.1 Filtração

O termo *filtração* será considerado para representar a σ-álgebra para o i-ésimo indivíduo gerada pelas covariáveis do modelo, ou seja, a *filtração* F_{t^-} pode ser interpretada como a informação sobre o i-ésimo indivíduo obtida desde o início do estudo até o tempo imediatamente anterior a t.

Seja um espaço mensurável (Ω, F, P), em que Ω denota o espaço amostral, F a σ-álgebra e P a medida de probabilidade definida em F. Um processo estocástico é uma família de variáveis aleatórias indexadas no tempo $(X(t) : t \geq 0)$, ou seja,

$$t \to X(t, \omega), \text{ para } \omega \in \Omega.$$

O processo estocástico X induz a uma família crescente de sub-σ-álgebras

$$F_{t^-}^X = \sigma \{X(s) : 0 \leq s < t\}$$

chamada de história interna de X.

Em muitas aplicações será necessária mais informações do que a gerada por um único processo estocástico. Portanto, é definida a história de modo geral $(F_t : t \geq 0)$ como uma família de sub-σ-álgebras, de tal forma que, para todo $s \leq t$, $F_s \subset F_t$, o que significa se $A \in F_s$ implica $A \in F_t$. A *história* pode ser chamada de *filtração*. Às vezes as informações (filtrações) são combinadas e para duas filtrações F_t^1 e F_t^2, $(F_t^1 \vee F_t^2)$ é denotado como a menor filtração que contém F_t^1 e F_t^2. Um processo estocástico X é adaptado para a filtração F_t se, para cada $t \geq 0$, $X(t)$ é F_t-mensurável e neste caso $F_t^X \subset F_t$.

2.2 Conjunto sob Risco

Um outro importante conceito está ligado a ideia de um conjunto de elementos sob risco, ou seja, quando um certo número de indivíduos está em risco para a ocorrência de determinado evento. Esse conjunto pode ser restrito ou irrestrito. Um conjunto em risco é dito irrestrito quando todos os indivíduos contribuem igualmente para a ocorrência do evento independente do número de eventos já experimentado por cada indivíduo. Este tipo de conjunto será o estudado neste trabalho, uma vez que a função de intensidade *baseline* é comum.

Seguindo este conceito é definida a variável aleatória binária $Y_i(t)$, que assume o valor 1 se o indivíduo i está em risco para a ocorrência do evento de interesse no tempo t e 0 se o indivíduo i não está em risco.

2.3 Processo de Contagem e Martingais

Um processo de contagem é um processo estocástico no qual registra-se a ocorrência de eventos ao longo de um determinado período de tempo. Esses eventos podem ser considerados como um processo pontual, os quais têm sido estudados extensivamente em Andersen (1993).

Mais formalmente, um processo de contagem $\{N(t)\}_{t\geq 0}$ é um processo estocástico com as seguintes propriedades:

1. $N(0) = 0$.
2. $N(t) < \infty$ com probabilidade 1.
3. $N(t)$ é um processo contínuo à direita e cresce com saltos constantes de tamanho 1.

Aalen (1978) apresentou os elementos dessa teoria para audiência estatística demonstrando sua utilidade no corpo da inferência estatística, bem como seu papel fundamental na base matemática da análise de sobrevivência. Assim, define-se $\{N_i(t)\}_{t\geq 0}$, como sendo um processo que registra o número de ocorrências do evento de interesse até o tempo t para o indivíduo i. Portanto, $N(t)$ é uma função escada onde é constante entre os eventos e tem um salto de uma unidade quando se observa a ocorrência do evento. Para a modelagem do processo de contagem, um pequeno intervalo de tempo $[t, t+dt)$ é considerado e

consequentemente é feita a suposição de que no máximo um simples evento tenha ocorrido nesse intervalo com alta probabilidade.

A intensidade do processo, denotada por $\lambda_i(t)$, é proporcional a probabilidade condicional que um intervalo seja observado em um pequeno intervalo de tempo $[t, t+dt)$ dada a história do processo até o tempo anterior a t, o qual é chamado de filtração (AALEN et al., 2008). Ou seja, a filtração é toda a informação disponível ao pesquisador anteriormente a t. Dessa forma a intensidade pode ser definida como

$$\begin{aligned} \lambda_i(t)dt &= P(N_i(t+dt) - N_i(t) = 1|F_{t^-}) \\ &= P(dN_i(t) = 1|F_{t^-}), \end{aligned}$$

em que o tempo anterior a t é representado por t^-. $\mathbf{N}(t)$ será denotado como o vetor do processo de contagem para n indivíduos e $\lambda(t)$ o vetor de intensidades correspondente.

Desde que cada $dN_i(t)$ seja uma variável binária pode-se escrever

$$\boldsymbol{\lambda}(t)dt = E(d\mathbf{N}(t)|F_{t^-}),$$

ou equivalentemente,

$$E(d\mathbf{N}(t)|F_{t^-}) = d\boldsymbol{\Lambda(t)},$$

em que o processo $\boldsymbol{\Lambda(t)} = \int_0^t \boldsymbol{\lambda}(s)ds, t \geq 0$ representa o vetor de n intensidades cumulativas. Essa intensidade cumulativa representa o número esperado de saltos do vetor $\mathbf{N}$ sobre o intervalo $(0, t]$ e a média condicional de $d\mathbf{N}(t)$ fornece a história até o tempo imediatamente anterior a t, representando assim a média condicional de $d\Lambda(t)$, ou seja,

$$E(d\mathbf{N}(t)|F_{t^-}) = d\boldsymbol{\Lambda}(t) = E(d\boldsymbol{\Lambda}(t)|F_{t^-}).$$

É importante ressaltar que $d\boldsymbol{\Lambda}(t)$ é um processo estocástico pois depende dos elementos aleatórios no passado F_{t^-}, incluindo censura, no qual evita mudanças em $\mathbf{N}(t)$. Contudo assume-se que $d\boldsymbol{\Lambda}(t)$ é previsível, significando que seu valor é conhecido apenas antes do tempo t. Subtraindo-se o processo de intensidade cumulativo do processo de contagem tem-se um processo estocástico denominado *martingale*,

$$\mathbf{M}(t) = \mathbf{N}(t) - \boldsymbol{\Lambda}(t), \tag{2.1}$$

ou equivalentemente,

$$d\mathbf{M}(t) = d\mathbf{N}(t) - d\boldsymbol{\Lambda}(t).$$

Este processo tem a propriedade que $E[d\mathbf{M}(t)|F_{t^-}] = 0$, onde têm-se a definição de um martingal, uma vez que $E[d\mathbf{N}(t) - \lambda(t)dt|F_{t^-}] = 0$.

A partir de (2.1), o processo de contagem $\mathbf{N}(t)$ pode ser escrito como,

$$\mathbf{N}(t) = \mathbf{\Lambda}(t) + \mathbf{M}(t),$$

em que $\mathbf{M}(t)$ pode ser considerado um ruído com média 0 que surge quando subtraída a intensidade cumulativa $\mathbf{\Lambda}(t)$, em que é conhecido na decomposição de Doob-Meyer como o compensador do processo de contagem $\mathbf{N}(t)$.

Dentre as diferentes estratégias para modelar o processo de contagem pode-se citar o Modelo de Regressão de Cox semiparamétrico (ANDERSEN; GILL, 1982) e o Modelo Aditivo de Aalen (1980).

Uma vez que já foram supracitados os conceitos necessários para o entendimento do Modelo Aditivo de Aalen, a seção seguinte irá descrever este modelo que dará base aos dois modelos que serão apresentados no Capítulo 3.

2.4 Modelo Aditivo de Aalen

Aalen (1980) propôs um modelo de regressão aditivo não-paramétrico para análise de dados de sobrevivência baseando-se na estrutura de processos de contagem. De acordo com esse modelo, os parâmetros de regressão podem variar no tempo e o método dos mínimos quadrados é empregado para estimar os parâmetros do modelo em cada evento no tempo. O mesmo autor em 1989, propôs uma estatística de teste para inferência dos coeficientes de regressão. Entretanto, em 1993 Aalen desenvolveu a teoria de resíduos martingais para o modelo aditivo e considerou o mesmo como uma importante ferramenta para avaliar bondade do ajuste. O autor também utilizou o método *bootstrap* para investigar o efeito tempo dependente das covariáveis nos gráficos de regressão cumulativo.

Para ilustrar o modelo em questão, considere que n indivíduos são observados ao longo de um período de tempo, tal que os tempos de ocorrência dos eventos recorrentes sejam registrados. É assumido independência entre os diferentes indivíduos. O modelo pode ser escrito em notação matricial como:

$$\boldsymbol{\lambda}(t) = \mathbf{X}(t)\boldsymbol{\beta}(t),$$

em que $\boldsymbol{\lambda}(t)$ representa o vetor das intensidades do processo, $\mathbf{X}(t)$ é a matriz das co-

variáveis e $\boldsymbol{\beta}(t)$ é o vetor dos parâmetros desconhecidos.

Aplicações do modelo aditivo de Aalen podem ser vista em Mau (1986), Andersen e Vaeth (1989), McKeague (1987), Huffer e McKeague (1991) e Andersen (1993). No caso dos dados de corte transversal, os modelos com variáveis tempo dependentes foram descritos por vários autores na configuração de modelos de regressão. No entanto, as propriedades teóricas dos métodos de estimação não são bem conhecidas. Pensando nisso, Fan e Zhang (2000) e Hoover (1998) consideraram técnicas de regressão locais. Em Aalen (1989) foram apresentados gráficos empíricos para mostrar e testar o efeito das covariáveis sobre a resposta de interesse.

O modelo de Aalen assume que as covariáveis agem de forma aditiva sobre a taxa de risco basal. Seja $N_i(t)$ o processo de contagem que registra o número de eventos de interesse para o indivíduo i ao longo do tempo t. Assumimos que para cada indivíduo pode existir um conjunto de variáveis fixas e/ou tempo dependente, representado pelo vetor $X_i(t) = (1, X_{i1}(t), X_{i2}(t), \ldots, X_{ip}(t))^T$. A modelagem desse processo de contagem é realizada através da função de intensidade ou risco $\lambda_i(t)$ ligando de forma aditiva as covariáveis. Assim o modelo de risco aditivo (AALEN, 1980; AALEN, 1989) pode ser expresso da seguinte forma

$$\lambda_i(t) = Y_i(t)(\beta_0(t) + \beta_1(t)X_{i1}(t) + \ldots + \beta_p(t)X_{ip}(t)) \tag{2.2}$$

em que Y_i é uma variável aleatória binária assumindo o valor 1 se o indivíduo i está em risco para a ocorrência do evento de interesse no tempo t, $\beta_0(t)$ é o risco basal e $\beta_j(t), j = 1, \ldots, p$, são as as funções de regressão das covariáveis. Este modelo é considerado não-paramétrico pelo fato de que nenhuma forma paramétrica particular é assumida para as funções de regressão. Estas funções podem variar arbitrariamente com o tempo, revelando influência nas covariáveis. Esta é uma das suas grandes vantagens, bem como a não exigência de tamanho de amostra extremamente grande. Uma de suas desvantagens é a possibilidade de se estimar valores negativos para a função de risco.

Como apresentado anteriormente, o Modelo Aditivo de Aalen é completamente não-paramétrico e um bom resumo gráfico das estimativas dos coeficientes cumulativos permite também fazer inferência sobre os efeitos das covariáveis. Se forem observadas covariáveis constantes ao longo do tempo será de interesse considerar modelos semiparamétricos, em que o objetivo é a obtenção de um modelo parcimonioso.

Então a partir do modelo (2.2) pode ser obtido um sub-modelo chamado modelo semiparamétrico de riscos aditivos, sugerido por McKeague e Sasieni (1994). Neste caso

assume-se que a intensidade está na forma

$$\boldsymbol{\lambda}(t) = \mathbf{Y}(t)(\mathbf{X}^T(t)\boldsymbol{\beta}(t) + \mathbf{Z}^T(t)\boldsymbol{\gamma}),$$

em que $\boldsymbol{\gamma}$ é um vetor de dimensão q com coeficientes invariantes no tempo. Portanto, o efeito de algumas covariáveis pode mudar com o tempo, enquanto o efeito de outras é assumido como sendo constante.

2.4.1 Inferência

Existem métodos não-paramétricos para dados longitudinais que são normalmente baseados na suavização dos estimadores da função de regressão. Na abordagem em questão estima-se a função de regressão cumulativa $B_j(t) = \int_0^t \beta(s)ds$. Estes coeficientes são de fácil estimação e suas propriedades assintóticas são satisfatórias, o que não ocorre com $\beta(t)$ que em geral não são consistentes. A estimação dos coeficientes de regressão cumulativos são baseados no método dos mínimos quadrados. Assim, toda vez que um evento ocorre, um modelo linear é estimado pela regressão das observações $dN_i(t)$ sobre as covariáveis. Individualmente essas estimativas tem muito ruído e não são informativas. Porém, somando essas estimativas ao longo do tempo obtém-se algo mais sensato. A construção de gráficos dos elementos do vetor $B_j(t)$ versus o tempo permite identificar se determinada covariável tem efeito relevante no fenômeno em questão. Tais gráficos são conhecidos como gráfico de regressão cumulativa ou gráfico de Aalen (Mau, 1986). Dessa forma quanto maior a inclinação apresentada no referido gráfico mais evidência do efeito da covariável. Com esta abordagem é permitido o uso de martingais, que por sua vez irá permitir um comportamento assintótico dos estimadores de mínimos quadrados para as funções de regressão acumuladas. Assim, é possível o cálculo das bandas de confiança e fazer inferência sobre as funções de regressão acumuladas.

Nesta subseção serão apresentados os estimadores dos coeficientes de regressão, assim como também os intervalos de confiança e as bandas de confiança.

2.4.1.1 Estimação

A partir das equações (2.1) e (2.2), o processo de contagem pode ser decomposto em compensador e martingal

$$\mathbf{N}(t) = \int_0^t \lambda(s)ds + \mathbf{M}(t)$$

e consequentemente o compensador pode ser reescrito em termos do efeito das covariáveis

$$dN_i(t) = Y_i(t)dB_0(t) + \sum_{k=1}^{p} Y_i(t)dB_k(t)x_{ik} + dM_i(t). \tag{2.3}$$

Para melhor entendimento do processo de estimação é conveniente a utilização da notação matricial e vetorial:

$$\begin{aligned}
\mathbf{N}(t) &= (N_1(t), \cdots, N_n(t))^T \\
\mathbf{M}(t) &= (M_1(t), \cdots, M_n(t))^T \\
\mathbf{B}(t) &= (B_0(t), \cdots, B_p(t))^T \\
\mathbf{X}(t) &= \begin{pmatrix} Y_1(t) & Y_1(t)x_{11}(t) & \cdots & Y_1(t)x_{1p}(t) \\ \vdots & \vdots & \ddots & \vdots \\ Y_n(t) & Y_n(t)x_{n1}(t) & \cdots & Y_n(t)x_{np}(t) \end{pmatrix},
\end{aligned}$$

em que $\mathbf{N}$(t) é o vetor do processo de contagem, $\mathbf{M}$(t) é o vetor correspondente aos martingais e $\mathbf{B}$(t) é o vetor dos coeficientes de regressão acumulados. Pode-se definir a matrix $\mathbf{X}(t)$ de ordem $n \times (p+1)$ da seguinte maneira: se o evento considerado ainda não ocorreu para o i-ésimo indivíduo e ele não é censurado, então a i-ésima linha de $\mathbf{X}(t)$ é o vetor $\mathbf{x}_i(t) = (1, x_{i1}(t), x_{i2}(t), \cdots, x_{ip}(t))^T$. Caso contrário, ou seja, se o indivíduo não está sob risco no tempo t, então, a correspondente linha de $\mathbf{X}(t)$ contém apenas zeros. Com a notação matricial e vetorial, as equações (2.2) e (2.3) tornam-se

$$\boldsymbol{\lambda}(t) = \mathbf{X}(t)d\mathbf{B}(t)$$

$$d\mathbf{N}(t) = \underbrace{\mathbf{X}(t)d\mathbf{B}(t)}_{\text{Modelo}} + \underbrace{d\mathbf{M}(t)}_{\text{Resíduo}}, \tag{2.4}$$

em que a equação (2.4) tem a forma do modelo de regressão linear padrão. Isto leva naturalmente a estimação por mínimos quadrados:

$$d\widehat{\mathbf{B}}(t) = (\mathbf{X}(t)^T\mathbf{X}(t))^{-1}\mathbf{X}(t)^T d\mathbf{N}(t),$$

que é bem definida se $\mathbf{X}$(t) tem posto completo, ou seja, a estimação pára quando $\mathbf{X}$(t) deixa de ser uma matriz não-singular, que é uma consequência do princípio não-paramétrico. Então é introduzida J(t) como a indicadora de que $\mathbf{X}$(t) tem posto completo

$$d\widehat{\mathbf{B}}(t) = J(t)(\mathbf{X}(t)^T\mathbf{X}(t))^{-1}\mathbf{X}(t)^T d\mathbf{N}(t).$$

O estimador da função de regressão acumulada, utilizando a inversa generalizada $\mathbf{X}^{-}(t) = (\mathbf{X}(t)^T\mathbf{X}(t))^{-1}\mathbf{X}(t)^T$ é,

$$\begin{aligned}\widehat{\mathbf{B}}(t) &= \int_0^t J(u)(\mathbf{X}(u)^T\mathbf{X}(u))^{-1}\mathbf{X}(u)^T d\mathbf{N}(u) \\ &= \sum_{T_j \leq t} J(T_j)\mathbf{X}^{-}(j)\Delta\mathbf{N}(T_j),\end{aligned} \tag{2.5}$$

em que T_j são os tempos distintos dos eventos e $\Delta\mathbf{N}(T_j)$ é um vetor de um's para os indivíduos em que houve ocorrência do evento no tempo T_j e zeros caso contrário. O fato da estimação incidir sobre as funções de regressão acumulada $B_j(t) = \int_0^t \beta_j(s)ds$, se deve também a justificativa de que estimar a função de distribuição acumulada é mais fácil do que estimar a função de densidade de probabilidade.

As funções de regressão acumuladas são obtidas em cada ponto de tempo pela estimação da contribuição instantânea das covariáveis para o risco. A influência da j-ésima covariável pode ser verificada pela inclinação do gráfico da função de regressão acumulada contra o tempo. Por exemplo, se $B_j(t)$ é constante, então, o gráfico deve se aproximar de uma linha reta. Inclinações positivas ocorrem durante períodos em que aumentos dos valores das covariáveis são associados com aumentos na função risco. Por outro lado, inclinações positivas ocorrem em períodos em que crescimento nos valores das covariáveis estão associados com decréscimos na função risco. As funções de regressão acumulada têm inclinações aproximadamente iguais a zero em períodos em que as covariáveis não influenciam na função risco.

Os componentes de $\widehat{\mathbf{B}}(t)$ convergem assintoticamente, sob condições apropriadas, para um processo Gaussiano (AALEN, 1989). Então, o estimador da matriz de covariância é

$$\begin{aligned}\widehat{\Sigma}(t) &= \int_0^t J(u)\mathbf{X}^{-}(u)diag(d\mathbf{N}(u))\mathbf{X}^{-}(u)^T \\ &= \sum_{T_j \leq t} J(T_j)\mathbf{X}^{-}(T_j)diag(\Delta N(T_j))\mathbf{X}^{-}(T_j)^T.\end{aligned} \tag{2.6}$$

Como consequência dos resultados apresentados anteriormente, é possível estimar o risco acumulado $\widehat{\Lambda}(t)$:

$$\widehat{\Lambda}(t) = \mathbf{X}^T\widehat{B}(t),\ \mathrm{t}\ \leq\ \tau$$

Essas estimativas encontram-se somente disponíveis para $t \leq \tau$, sendo τ o valor maximal de t para o qual a matriz $\mathbf{X}$(t) é não singular.

Pelo Teorema 5.1.1 definido em Martinussen e Scheike (2006), Capítulo V, Seção 5.1,

p. 110, segue que

$$\sqrt{n}(\widehat{\mathbf{B}}(t) - \mathbf{B}(t)) \xrightarrow{D} \mathbf{U}(t), \tag{2.7}$$

em que $\xrightarrow{D}$ denota convergência em distribuição e $\mathbf{U}$(t) é um martingal contínuo Gaussiano de média zero com dimensão p. Este resultado será utilizado para a construção de intervalos e bandas de confiança, assim como para testes de hipóteses. A prova deste teorema pode ser encontrada em Andersen (1993), Capítulo VII, Seção 4.2, p. 576.

2.4.1.2 Intervalos de Confiança

Para a construção de intervalos de confiança para $B_j(t)$, a matriz de covariância de U(t), pode ser expressa como (AALEN et al., 2008)

$$\mathbf{A}(t) = var(\mathbf{U}(t)) = \int_0^t J(u)X^-(u)diag(\lambda(u)du)X^-(u)^T, \tag{2.8}$$

e pode ser estimada como

$$\widehat{var(U(t))} = n\widehat{\Sigma}(t) = n\sum_{T_j \leq t} J(T_j)\mathbf{X}^-(T_j)diag(\Delta N(T_j))\mathbf{X}^-(T_j)^T.$$

Pode-se mostrar (ANDERSEN, 1993), Capítulo VII, Seção 4.2, p. 576 que

$$n\widehat{\Sigma}(t) \to \mathbf{A}(t). \tag{2.9}$$

Por simplicidade de notação, $\mathbf{X}$(t) será substituído por $\mathbf{X}$. A partir de (2.7), (2.8) e (2.9) tem-se que

$$\begin{aligned} n\widehat{\Sigma}(t) &= n\int_0^t J(u)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T diag(d\mathbf{N}(u))\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \int_0^t J(u)n(\mathbf{X}^T\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}^T diag(d\mathbf{N}(u))\mathbf{X}n(\mathbf{X}^T\mathbf{X})^{-1} \\ &\xrightarrow{P} \mathbf{\Lambda}(t), \end{aligned}$$

em que $\xrightarrow{P}$ denota convergência em probabilidade e $\mathbf{\Lambda}(t)$ é a intensidade acumulada. Assim, $\sqrt{n}(\widehat{\mathbf{B}}(t) - \mathbf{B}(t))$ converge em distribuição para um martingal Gaussiano com média zero de dimensão p cuja matriz de covariância converge em probabilidade para $\mathbf{A}$(t). Portanto, o intervalo de confiança para $B_i(t)$ é dado por:

$$\left[\widehat{B}_i(t) - z_{1-\alpha/2}\sqrt{\widehat{\Sigma}_{ii}(t)}; \widehat{B}_i(t) + z_{1-\alpha/2}\sqrt{\widehat{\Sigma}_{ii}(t)}\right],$$

em que $z_{1-\alpha/2}$ é o quantil correspondente a distribuição normal padrão.

2.4.1.3 Bandas de Confiança

Nesta subseção serão construidas bandas de confiança para $B_i(t)$, que serão utilizadas na construção dos gráficos do Capítulo 4. As bandas de confiança não tem largura constante, mas permitem o uso de quantis conhecidos usando a tabela de Hall e Wellner (1980). Uma das vantagens da construção das bandas de confiança é que elas podem ser utilizadas para avaliar a hipótese $H_0 : B_i(t) = 0, \ \forall \mathrm{t} \ \in \ [0, \tau]$, ou seja, observa-se se a função zero está contida na banda.

A construção das bandas de confiança baseia-se no fato de $B^0(t)$ ser uma ponte Browniana (ou seja, $B^0(t) = W(t) - tW(t)$, onde $W(t)$ é um movimento Browniano padrão), mais detalhes (MARTINUSSEN; SCHEIKE, 2006), Seção 5.2, p. 117. Assim, as bandas de confiança de Hall-Wellner são representadas como:

$$\left(\widehat{\mathbf{B}}_i(t) - c_{1-\alpha} \sqrt{\frac{\widehat{\Sigma}_{ii}(\tau)}{n}} \left(1 + \frac{\widehat{\Sigma}_{ii}(t)}{\widehat{\Sigma}_{ii}(\tau)} \right) \leq B_i(t) \leq \widehat{\mathbf{B}}_i(t) + c_{1-\alpha} \sqrt{\frac{\widehat{\Sigma}_{ii}(\tau)}{n}} \left(1 + \frac{\widehat{\Sigma}_{ii}(t)}{\widehat{\Sigma}_{ii}(\tau)} \right) \right),$$

com cobertura assintótica $1 - \alpha$, em que $\widehat{\Sigma}_{ii}$ é o i-ésimo elemento da diagonal de $\widehat{\Sigma}$ e $c_{1-\alpha}$ são quantis que podem ser encontrados em Hall e Wellner (1980).

2.4.2 Testes de Hipóteses

Utilizando o resultado (2.7) pode-se obter testes de hipóteses. Neste trabalho serão apresentados dois testes. O primeiro verifica os efeitos das covariáveis e o segundo visa verificar se uma função de regressão tem seu efeito constante ao longo do tempo. Formalmente:

1. $H_0 : B_j(t) = 0, \ \forall \mathrm{t} \ \in \ [0, \tau] \quad \text{vs} \quad H_1 : B_j(t) \neq 0, \ \text{para efeitos das covariáveis.}$
2. $H_0 : B_j(t) = \gamma t, \ \forall \mathrm{t} \ \in \ [0, \tau] \quad \text{vs} \quad H_1 : B_j(t) \neq \gamma t, \ \text{para efeitos constantes,}$

em que $j = 1, \ldots, n$. Neste contexto $B_j(t)$ é uma função de regressão acumulada dada e γ é um parâmetro a ser estimado. Costuma-se testar em $[0, \tau]$, em que τ é o tempo final do estudo, de modo que as funções são consideradas em todo intervalo, mas qualquer intervalo de tempo menor $[0, t_0]$, com $t_0 \in [0, \tau]$, pode ser considerado.

2.4.2.1 Teste para os Efeitos das Covariáveis

Será considerado o primeiro teste, mas por questões práticas o teste irá incidir sobre $\beta_j(t)$, em que a hipótese nula para algum $j \geq 1$ é dada como:

$$H_0 : \beta_j(t) = 0, \ \forall \mathrm{t} \ \in \ [0, \tau].$$

Para o modelo aditivo de Aalen isto corresponde a testar a hipótese nula de que não existe efeito da covariável sobre a função risco.

Considere uma função peso previsível não negativa $L_j(t)$ que supostamente é nula sempre que $J(t) = 0$, ou seja, a seguinte hipótese nula pode apenas ser testada sobre intervalos de tempo em que $\mathbf{X}(\mathrm{t})$ tenha posto completo. Uma boa estatística para o teste pode ser dada por:

$$Z_j(\tau) = \int_0^\tau L_j(t) d\widehat{B}_j(t) = \sum_{T_j \leq \tau} L_j(T_j) \Delta \widehat{B}_j(T_j).$$

Esta estatística é eficiente para testar H_0 contra as alternativas da forma $\beta_j(t) < 0$ ou $\beta_j(t) > 0$ para todo t.

Aalen et al. (2008) consideraram a seguinte função peso:

$$L_j(t) = \frac{1}{(\mathbf{X}(t)^T \mathbf{X}(t))^{-1}_{jj}},$$

em que o denominador é a inversa de uma matriz diagonal tendo a mesma diagonal principal da matriz $(\mathbf{X}^T\mathbf{X})^{-1}$. A escolha do peso é diretamente inspirado na regressão por mínimos quadrados, em que as variâncias dos estimadores são proporcionais a $(\mathbf{X}^T\mathbf{X})^{-1}$. Em seguida, as estatísticas de teste torna-se uma soma ponderada das funções de regressão acumuladas.

Sob a hipótese nula, $Z_j(\tau)$ é assintoticamente normal com média zero e variância que pode ser estimada por:

$$V_{jj}(\tau) = \int_0^\tau L_j^2(t) d\widehat{\Sigma}_{jj}(t) = \sum_{T_j \leq \tau} L_j^2(T_j) \Delta \widehat{\Sigma}_{jj}(T_j).$$

Portanto, pelo Teorema Central do Limite, obtemos que, se H_0 é verdadeira,

$$\frac{Z_j(\tau)}{V_{jj}(\tau)} \xrightarrow{D} N(0, 1)$$

Então, a hipótese nula é rejeitada para valores grandes da estatística $Z_j(\tau)$.

2.4.2.2 Teste para Efeitos Constantes

O interesse agora é testar

$$H_0 : B_j(t) = \gamma t \ \forall \mathrm{t} \ \in \ [0, \tau] \quad \text{vs} \quad H_1 : B_j(t) \neq \gamma t.$$

Após o modelo ser ajustado em coeficientes paramétricos e não-paramétricos, aqui o interesse é verificar se os coeficientes paramétricos realmente são invariantes com o tempo. Sob a hipótese nula, deve ser estimado γ. Para isso, basta tomar

$$\widehat{\gamma} = \frac{\widehat{B}_j(\tau)}{\tau}.$$

Devido ao fato de $H = \sqrt{n}\left(\widehat{B}_j(t) - \frac{t}{\tau}\widehat{B}_j(\tau)\right)$ não ser um martingal, pois, $\widehat{B}_j(\tau)$ depende do futuro, o que não é compatível com a definição de um martingal, um método para a construção do teste de hipótese foi introduzido por **??**), que é chamado de procedimento multiplicador condicional. Baseia-se na reamostragem. Também foi previamente utilizado em Borgan et al. (2007), Elgmati et al. (2008) e Elgmati (2009). Primeiro é necessário mostrar que

$$\sqrt{n}\left(\widehat{\mathbf{B}}(t) - \mathbf{B}(t)\right) = \sqrt{n}\int_0^t \mathbf{X}^-(s)d\mathbf{M}(s).$$

Isso decorre de

$$\begin{aligned}
\sqrt{n}\left(\widehat{\mathbf{B}}(t) - \mathbf{B}(t)\right) &= \sqrt{n}\left(\int_0^t \mathbf{X}^-(s)d\mathbf{N}(s) - \mathbf{B}(t)\right) \\
&= \sqrt{n}\left(\int_0^t \mathbf{X}^-(s)d\mathbf{M}(s) - \int_0^t \underbrace{\mathbf{X}^-(s)\mathbf{X}(s)}_{=\mathbf{I}} d\mathbf{B}(s) - \mathbf{B}(t)\right) \\
&= \sqrt{n}\int_0^t \mathbf{X}^-(s)d\mathbf{M}(s).
\end{aligned}$$

Assim,

$$\sqrt{n}\left(\widehat{\mathbf{B}}(t) - \mathbf{B}(t)\right) = \sqrt{n}\sum_{i=1}^{n}\int_0^t (\mathbf{X}(s)^T\mathbf{X}(s))\mathbf{X}_i(s)^T dM_i(s),$$

onde $\mathbf{X}_i(s)$ é o vetor das covariáveis para cada indivíduo i.

Seja

$$\epsilon_i(t) = \int_0^t (\frac{1}{n}\mathbf{X}(s)^T\mathbf{X}(s))^{-1}\mathbf{X}_i(s)dM_i(s).$$

Então

$$\sqrt{n}\left(\widehat{\mathbf{B}}(t)-\mathbf{B}(t)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(t).$$

O estimador $\widehat{\epsilon}_i(t)$ de $\epsilon_i(t)$ é

$$\widehat{\epsilon}_i(t)=\int_0^t(\frac{1}{n}\mathbf{X}(s)^T\mathbf{X}(s))^{-1}\mathbf{X}_i(s)d\widehat{M}_i(s),$$

em que $\widehat{M}_i(t)$ é definido em (2.1).

O interesse é que $\sqrt{n}\left(\widehat{\mathbf{B}}(t)-\mathbf{B}(t)\right)$ seja uma soma de termos independentes e identicamente distribuídos, o que não é possível. No entanto, devido à correlação entre $d\widehat{M}_i$ ser de ordem $1/n$, $\sqrt{n}\left(\widehat{\mathbf{B}}(t)-\mathbf{B}(t)\right)$ se comporta como uma soma de termos independentes e identicamente distribuídos.

O teste para efeitos constantes é baseado no seguinte teorema:

Teorema 2.4.2.1. *Seja* $Z_1,\cdots,Z_n\overset{iid}{\sim}N(0,1)$. *Sob algumas condições técnicas, segue-se que* $\sqrt{n}\left(\widehat{\boldsymbol{B}}(t)-\boldsymbol{B}(t)\right)$ *tem a mesma distribuição com limite*

$$\Delta(t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\epsilon}_i(t)Z_i.$$

A prova deste teorema, bem como as condições nas quais o teorema está válido podem ser encontrados em **??**), Capítulo V, Seção 5.2. Como as propriedades para martingais não podem ser utilizadas para $H=\sqrt{n}\left(\widehat{B}_j(t)-\frac{t}{\tau}\widehat{B}_j(\tau)\right)$, não é possível estimar a sua variância, portanto, é difícil avaliar sua distribuição. O Teorema 2.4.2.1 fornece uma maneira simples para estimar a distribuição de H, em que são feitas simulações de Z_i.

Deve-se construir estatísticas de teste para testar a hipótese nula. Aqui são apresentadas duas:

$$\begin{aligned}T_{1,const}&=\sqrt{n}sup_{t\in[0,\tau]}\left|\widehat{B}_j(t)-\frac{t}{\tau}\widehat{B}_j(\tau)\right| \\ T_{2,const}&=n\int_0^\tau\left(\widehat{B}_j(t)-\frac{t}{\tau}\widehat{B}_j(\tau)\right)^2\end{aligned} \tag{2.10}$$

A primeira olha para a maior distância entre $\widehat{B}_j(t)$ e uma linha reta, enquanto a segunda

acumula o quadrado das diferências. Agora, pode-se notar que, sob a hipótese nula,

$$\begin{aligned}\sqrt{n}\left(\widehat{B}_j(t)-\frac{t}{\tau}\widehat{B}_j(\tau)\right) &= \sqrt{n}\left(\widehat{B}_j(t)-B_j(t)-\frac{t}{\tau}\widehat{B}_j(\tau)+\underbrace{B_j(t)}_{=\frac{\mathrm{t}}{\tau}\mathrm{B_j}(\tau)}\right)\\ &= \sqrt{n}\left(\widehat{B}_j(t)-B_j(t)\right)-\sqrt{n}\frac{t}{\tau}\left(\widehat{B}_j(\tau)-B_j(\tau)\right).\end{aligned}$$

Portanto, pelo Teorema 2.4.2.1, $\sqrt{n}\left(\widehat{B}_j(t)-\frac{t}{\tau}\widehat{B}_j(\tau)\right)$ tem, sob a hipótese nula, a mesma distribuição assintótica que $\Delta_j(t)-\frac{t}{\tau}\Delta_j(\tau)$, em que $\Delta(t)$ é definido no Teorema (1.3.1). Assim, quando se quer testar a hipótese $H_0 : B_j(t) = \gamma t$, é adotado o seguinte algoritmo:

Para $r = 1, \cdots, R$:

1. Calcule $\Delta^r(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\epsilon}_i(t)Z_i$, onde $Z_i \overset{iid}{\sim} N(0,1)$.
2. Calcule $T^r_{sim} = sup_{t\in[0,\tau]}|\Delta^r(t)-(t/\tau)\Delta^r(\tau)|$ ou $T^r_{sim} = \int_0^\tau(\Delta^r(t)-(t/\tau)\Delta^r(\tau))^2dt$.
3. Tem-se agora R replicações $T^1_{sim}, \cdots, T^R_{sim}$.
4. Calcular a estimativa $\widehat{p}$ do p-valor, $\widehat{p} = \dfrac{\#\left\{T^r_{sim} > T_{obs}\right\}}{R}$, r $= 1, \cdots,$ R,

em que T_{obs} é definida por uma das duas estatísticas em (2.10).

No algoritmo anterior, foram utilizados $Z_i \overset{iid}{\sim} N(0,1)$, mas também pode ser utilizado

$$Z_i = \begin{cases} 1 & \text{com probabilidade } 1/2 \\ -1 & \text{com probabilidade } 1/2, \end{cases}$$

e obter os mesmos resultados.

2.4.3 Resíduos Martingais e Bondade do Ajuste

Um passo importante para adequação do modelo é checar se ele é apropriado ou não. Uma ferramenta útil neste caso, são os resíduos martingais. Eles são definidos para cada indivíduo i em Aalen et al. (2008) como

$$\widehat{M_i}(t) = N_i(t) - \widehat{\Lambda}_i(t). \tag{2.11}$$

Ao escrever um pouco diferente, pode ser mostrado que (2.11) é um martingal

$$\begin{aligned}
\widehat{\mathbf{M}}(t) &= \int_0^t d\mathbf{N}(s) - \int_0^t \mathbf{X}(s) d\widehat{\mathbf{B}}(s) \\
&= \int_0^t d\mathbf{N}(s) - \int_0^t \underbrace{\mathbf{X}(s)\mathbf{X}^{-}(s)}_{\mathbf{H(s)}} d\mathbf{N}(s) \\
&= \int_0^t (\mathbf{I} - \mathbf{H}(s)) d\mathbf{N}(s) \\
&= \int_0^t \underbrace{(\mathbf{I} - \mathbf{H}(s))\mathbf{X}(s) d\mathbf{B}(s)}_{=0, \text{ pela definição de } \mathbf{H}} + \int_0^t (\mathbf{I} - \mathbf{H}(s)) d\mathbf{M}(s) \\
&= \int_0^t (\mathbf{I} - \mathbf{H}(s)) d\mathbf{M}(s).
\end{aligned}$$

Portanto, $\widehat{\mathbf{M}}(t)$ é um martingal, uma vez que é a integral do processo previsível com relação a um martingal ((AALEN et al., 2008), Seção 2.2.2). Pode também ser dado o estimador da matriz de variância e covariância de $\widehat{\mathbf{M}}(t)$:

$$\widehat{\Omega}(t) = \widehat{var(\widehat{\mathbf{M}}(t))} = \int_0^t (\mathbf{I} - \mathbf{H}(s)) diag(\lambda(s) ds)(\mathbf{I} - \mathbf{H}(s))^T ds.$$

Uma outra forma de construir os resíduos padronizados pode ser escrita como,

$$M_i^*(t) = \frac{\widehat{M}_i(t)}{\sqrt{\widehat{\Omega}_{ii}(t)}}, \quad \text{para i} = 1, \cdots, \text{p.}$$

Portanto, se o modelo for corretamente especificado, verifica-se se a variância dos resíduos padronizados deve está próxima de 1 em todos os tempos t, isto é, $var(M_i^*(t)) = 1$. A fim de verificar este modelo, pode-se representar graficamente o desvio-padrão dos resíduos padronizados observados e verificar se está próximo de 1 para todo t.

3 Modelos de Eventos Recorrentes no Contexto de Dados Longitudinais

Neste capítulo serão apresentados dois modelos de regressão baseados no Modelo Aditivo de Aalen definido no Capítulo 2. O primeiro foi sugerido por Borgan et al. (2007), em que são modelados dados longitudinais binários em tempos discretos onde é possível incluir covariáveis tempo dependentes por meio do uso de covariáveis dinâmicas bem como estimar e testar os seus efeitos tempo dependentes. O segundo foi sugerido por Martinussen e Scheike (2000), onde o modelo de regressão aditivo dinâmico também é aplicado em medidas longitudinais, porém em dados censurados em tempos contínuos.

3.1 Modelo de Regressão Aditivo Para Dados Longitudinais Binários

Nesta seção serão considerados modelos de regressão aditivos dimâmicos para dados de eventos recorrentes em tempos discretos em que a média condicional com base na história é modelada como uma função de possíveis covariáveis tempo dependentes. Este tipo de modelo tem sido estudado extensivamente ao longo dos anos em análise de sobrevivência e é conhecido como Modelo Aditivo de Aalen; mais detalhes em (AALEN, 1989; AALEN, 1993; HUFFER; MCKEAGUE, 1991; ANDERSEN, 1993).

A ideia aqui é apresentar e aplicar a metodologia utilizada em Fiaccone (2006) e Borgan et al. (2007) para análise de dados binários sobre diarreia infantil por meio da inclusão de variáveis tempo dependentes. Assim, através de um modelo aditivo dinâmico espera-se analisar eventos recorrentes no contexto de dados longitudinais considerando as inter-relações entre os mesmos. Os referidos autores acima salientam as vantagens dessa abordagem como a facilidade em capturar efeitos variando no tempo de covariáveis tempo dependentes, baixa carga computacional no procedimento de estimação e a inclusão de

covariáveis dinâmicas no modelo. Segundo Aalen et al. (2004), a inclusão de covariáveis dinâmicas como o número de eventos prévios ou o tempo desde o último evento pode melhorar o entendimento do fenômeno estudado. Vale ressaltar que a inclusão da história passada no modelo de intensidade do processo é a forma mais natural de caracterizar a susceptibilidade, heterogeneidade ou a fragilidade individual de desenvolver alguma doença por exemplo.

3.1.1 A Modelagem Considerada

Os dados longitudinais surgem sempre que observações repetidas da variável resposta são obtidas ao longo do tempo para cada indivíduo ou unidade de análise. Existem uma grande variedade de desafios na análise deste tipo de dados .

Por um lado, devido a sua natureza, as medições repetidas provenientes de estudos longitudinais são multivariadas e têm uma estrutura complexa de autocorrelação. Por outro, a natureza da variável resposta pode ser contínua ou discreta. Além disso, os estudos longitudinais permitem a introdução de covariáveis que variam ao longo do tempo, o que torna mais complexa a sua análise. Finalmente, neste tipo de estrutura a existência de dados omissos é um dos maiores desafios.

A modelagem incorreta quando existem dados omissos ou incompletos pode levar a três implicações gerais na análise: o conjunto de dados torna-se desbalanceado, a perda de informação leva a ineficiência do modelo e; este tipo de situação pode levar a estimativas viesadas dos parâmetros levando a inferências enganosas.

Nos dados que serão trabalhados nesta dissertação nem todas as crianças são observadas para o pleno estudo, sendo que cerca de 12% dos dados foram perdidos. Nesta seção será apresentada a modelagem considerada para lidar com este tipo de situação.

Para efeito de ilustração, $Y_i(t)$ representará um processo binário, denotando a ocorrência do evento de interesse para cada criança i até o tempo t. Em seguida o processo de contagem contará o número de ocorrências do evento. Como $Y_i(t)$ não será observado em todos os tempos, $R_i(t)$ indicará se a criança i está em risco no tempo t.

Considere uma situação em que não se tem observações faltantes. Logo podem ser definidos o modelo e os parâmetros de interesse. As observações para o i-ésimo indivíduo, $i = 1, \ldots, n$, irão formar um processo binário $\widetilde{Y}_i(1), \ldots, \widetilde{Y}_i(t)$, em que $\widetilde{Y}_i(t) = 1$ se há ocorrência do evento de interesse no tempo t e $\widetilde{Y}_i(t) = 0$ caso contrário. Lembrando que no tempo $t = 0$, tem-se $\widetilde{Y}_i(0) = 0$. O evento de interesse será o início do episódio

de diarreia (quando a incidência é estudada), ou se a criança sofre de diarreia (quando a prevalência é estudada). Para cada indivíduo, em cada tempo t, tem-se um vetor de covariáveis $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))$ de dimensão p. Estas podem ser fixas ou tempo dependentes.

Denotando por H_{i0} a σ-álgebra para o i-ésimo indivíduo gerada pelas covariáveis fixas e tempo dependentes, em que $H_{it} = H_{i0} \vee \sigma\left\{\widetilde{Y}_i(1), \widetilde{Y}_i(2), \ldots, \widetilde{Y}_i(t)\right\}$, a filtração H_{it} pode ser interpretada como a informação sobre o i-ésimo indivíduo obtida desde o início do estudo até o tempo t, se não houver observações em falta. A distribuição conjunta de $\widetilde{Y}_i(1), \ldots, \widetilde{Y}_i(t)$, pode ser obtida pela seguinte probabilidade condicional:

$$\alpha_i(t) = P(\tilde{Y}_i(t) = 1 | H_{it^-}). \tag{3.1}$$

O objetivo principal para a análise de dados longitudinais é estudar como estas probabilidades condicionais variam ao longo do tempo e como elas dependem das covariáveis.

Devido às observações faltantes, o estudo sobre $\alpha_i(t)$ é complicado. A fim de lidar com as faltas, será apresentado o "processo de falta"categórico $Z_i(t), \ldots, Z_i(T)$, em que $Z_i(t)$ indica se o resultado $\widetilde{Y}_i(t)$ para o i-ésimo indivíduo é observado, perdido devido a falta intermitente ou perdido devido ao abandono:

$$Z_i(t) = \begin{cases} 0, & observado \\ 1, & falta\ intermitente \\ 2, & abandono. \end{cases}$$

Mais uma vez, $Z_i(0) = 0$ a fim de se ter $\widetilde{Y}_i(t)$ definido para todo $t \in \tau$.

A introdução do "processo de falta"(geralmente) traz alguma variação aleatória adicional. Portanto, agora será de uso uma filtração maior (G_{it}) dada pela σ-álgebra,

$$G_{it} = G_{i0} \vee \sigma\left\{\widetilde{Y}_i(1), Z_i(1), \widetilde{Y}_i(2), Z_i(2), \ldots, \widetilde{Y}_i(t), Z_i(t)\right\}$$

representando os aspectos do "processo de falta"externo para o i-ésimo indivíduo, ou seja, quando um investigador perde uma visita domiciliar por razões que nada têm a ver com o estado de saúde de uma criança. Como consequência, a distribuição condicional de $\widetilde{Y}_i(t)$ pode ser alterada para:

$$P(\tilde{Y}_i(t) = 1 | G_{it^-}) = P(\tilde{Y}_i(t) = 1 | H_{it^-}), \tag{3.2}$$

para todo $t \in \tau$. Em (ANDERSEN, 1993), Capítulo III, Seção 2.2, encontra-se uma condição

semelhante no caso de censura independente na análise da história do evento .

Nenhuma das filtrações anteriores descrevem os dados disponíveis para este trabalho, sendo assim será necessário considerar a filtração (F_{it}) para o i-ésimo indivíduo, com $i = 1, \ldots, n$. Para isto é introduzido o indicador de risco $R_i(t) = I\{Z_i(t) = 0\}$ assumindo o valor 1 se o indivíduo i é observado no tempo t e o valor 0 caso contrário e o processo $Y_i(t) = R_i(t)\widetilde{Y}_i(t)$, registrando os eventos observados para cada indivíduo. Então,

$$F_{it} = G_{i0} \vee \sigma\{Y_i(1), Z_i(1), Y_i(2), Z_i(2), \ldots, Y_i(t), Z_i(t)\}, \tag{3.3}$$

assumindo que todas as covariáveis fixas e tempo dependentes são observáveis.

Acima assume-se que as filtrações (F_{it}) correspondem aos dados realmente disponíveis, mas há algumas situações em que filtrações maiores podem ser definidas para outros processos observados em paralelo com os dados longitudinais binários $Y_i(t)$. De acordo com Borgan et al. (2007) a extensão das filtrações não causam problema para os métodos estatísticos do modelo aditivo desde que a sua previsão não seja uma preocupação.

Assumindo que $\alpha_i(t)$ é (F_{it})-previsível, implica que as covariáveis dinâmicas tempo dependentes podem depender apenas de partes da informação de G_{it^-} que também estão contidas em F_{it^-}. Então por (3.1) e (3.2),

$$\begin{aligned}
\lambda_i(t) &= P(Y_i(t)|F_{it^-}) \\
&= E\left\{P(R_i(t) = 1, \widetilde{Y}_i(t) = 1|G_{it^-})|F_{it^-}\right\} \\
&= E\left\{P(\widetilde{Y}_i(t) = 1|G_{it^-})P(R_i(t) = 1|G_{it^-}, \widetilde{Y}_i(t) = 1)|F_{it^-}\right\} \\
&= \alpha_i(t)E\left\{P(R_i(t) = 1|G_{it^-}, \widetilde{Y}_i(t) = 1)|F_{it^-}\right\}.
\end{aligned} \tag{3.4}$$

Com o pressuposto de que a falta na distribuição não depende do resultado $\widetilde{Y}_i(t)$ (faltas aleatórias), pode-se escrever,

$$\lambda_i(t) = \alpha_i(t)E\{P(R_i(t) = 1|G_{it^-})|F_{it^-}\} = \alpha_i(t)\pi_{it}, \tag{3.5}$$

em que

$$\pi_{it} = P(R_i(t) = 1|F_{it^-}) \tag{3.6}$$

é a probabilidade condicional de observar $\widetilde{Y}_i(t)$ dado o "passado" F_{it^-}.

3.1.1.1 Inferência

De acordo com a Figura 2 apresentada no Capítulo 4, o padrão de falta intermitente parece essencialmente ser por motivos externos, portanto previsível. No entanto, não é evidente que o abandono tem o mesmo comportamento . Neste caso é assumido que o "processo de falta"é previsível, então:

$$\pi_{it} = P(R_i(t) = 1|F_{it^-}) = R_i(t). \tag{3.7}$$

Pelos resultados gerais, para dados longitudinais binários resumidos no Apêndice A, têm-se a decomposição $Y_i(t) = \lambda_i(t) + \epsilon_i(t)$ da observação $Y_i(t)$, em parte sistemática $\lambda_i(t)$ e um erro aleatório $\epsilon_i(t)$. Nesta abordagem, $\epsilon_i(t)$, são as diferenças martingais, isto é, o processo $M_i(t) = \sum_{s=0}^{t} \epsilon_i(s)$ é um martingal. Portanto, de (3.5) e (3.7) pode-se escrever

$$Y_i(t) = \beta_0(t)R_i(t) + \beta_1(t)X_{i1}(t)R_i(t) + \cdots + \beta_p(t)X_{ip}(t)R_i(t) + \epsilon_i(t), \tag{3.8}$$

que, para cada t, tem a forma de um modelo de regressão linear com erros não-correlacionados. Pode-se estimar $\beta_j(t)$ pela regressão das observações $Y_i(t)$ sobre as covariáveis $X_{ij}(t)R_i(t)$ pelo método de mínimos quadrados. Embora as estimativas em cada ponto de tempo esteja sujeita a grandes erros de amostragem, pode-se obter estimativas informativas e estáveis dos coeficientes de regressão cumulativos $B_j(t) = \sum_{s=0}^{t} \beta_j(s)$, em que são acumuladas as estimativas de $\beta_j(s)$ ao longo do tempo.

Para uma descrição detalhada do processo de estimação é conveniente uma introdução à notação vetorial e matricial. Visto que $t \in \tau$, tem-se que $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^T$ representa o vetor das observações, $\boldsymbol{\beta}(t) = (\beta_0(t), \beta_1(t), \ldots, \beta_p(t))^T$ é o vetor dos coeficientes de regressão, $\mathbf{X}(t)$ é a "matriz desenho"com linhas $\mathbf{x}_i(t)^T\mathbf{R}_i(t) = (1, x_{i1}(t), \ldots, x_{ip}(t))R_i(t)$. Desde que $\mathbf{X}(t)$ tenha posto completo, a estimação por mínimos quadrados de $\boldsymbol{\beta}(t)$ é dada por

$$\widehat{\boldsymbol{\beta}}(t) = (\mathbf{X}^T(t)\mathbf{X}(t))^{-1}\mathbf{X}^T(t)\mathbf{Y}(t). \tag{3.9}$$

Sendo $J(t)$ um indicador do processo que assume o valor 1 se $\mathbf{X}(t)$ tem posto completo e o valor 0 caso contrário. Ao acumular as estimativas de mínimos quadrados para todos os tempos, têm-se a estimativa do coeficiente de regressão acumulado.

$$\widehat{\mathbf{B}}(t) = \sum_{s=0}^{t} J(s)\widehat{\boldsymbol{\beta}}(s) = \sum_{s=0}^{t} J(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}\mathbf{X}^T(s)\mathbf{Y}(s) \tag{3.10}$$

em que $\widehat{\mathbf{B}}(t) = (B_0(t), B_1(t), \ldots, B_p(t))^T$ é o vetor das funções de regressão acumulada.

Para estudar as propriedades do estimador, é considerado $\mathbf{B}^*(t) = \sum_{s=0}^{t} J(s)\beta(s)$, que é fechado para $\mathbf{B}(t)$ quando existe uma pequena probabilidade de que $\mathbf{X}(s)$ não tenha posto completo para todo $s \leq t$ e seja $\boldsymbol{\epsilon}(t) = (\epsilon_1(t), \ldots, \epsilon_n(t))^T$ o vetor dos erros aleatórios definidos em (3.8). Então $\mathbf{Y}(s) = \mathbf{X}(s)\beta(s) + \epsilon(s)$ é substituido em (3.10) e é obtido

$$\widehat{\mathbf{B}}(t) - \mathbf{B}^*(t) = \sum_{s=0}^{t} J(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}\mathbf{X}^T(s)\boldsymbol{\epsilon}(s).$$

Em particular, $\widehat{B}(t) - \mathbf{B}^*(t)$ é uma transformação martingal, e portanto um martingal de média zero. Logo, $E(\widehat{B}(t)) = E(\mathbf{B}^*(t))$ para todo $t \in \tau$ de modo que (3.10) é um estimador não viciado. Por (A.10), a matriz de covariância de $\widehat{\mathbf{B}}(t)$ é estimada por

$$\widehat{cov}(\widehat{\mathbf{B}}(t)) = \sum_{s=0}^{t} J(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}\mathbf{X}^T(s)\widehat{\Sigma}(s)\mathbf{X}(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}, \tag{3.11}$$

em que $\widehat{\Sigma}(s) = diag\left\{\widehat{\lambda}_i(s)(1 - \widehat{\lambda}_i(s))\right\}$ é uma matriz diagonal de dimensão $n \times n$ com o i-ésimo elemento igual a $\widehat{\lambda}_i(s)(1 - \widehat{\lambda}_i(s))$ com

$$\widehat{\lambda}_i(s) = \mathbf{x}_i^T(t)\mathbf{R}_i(t)\widehat{\beta}(t) = \left\{\widehat{\beta}_0(s) + \widehat{\beta}_1(s)x_{i1}(s) + \cdots + \widehat{\beta}_p(s)x_{ip}(s)\right\} R_i(s) \tag{3.12}$$

sendo o modelo baseado na estimativa de $\lambda_i(s)$. Além disso, pelo teorema central do limite para martingais, a distribuição assintótica de (3.10) segue aproximadamente uma distribuição normal multivariada. As derivações acima são semelhantes aos do modelo aditivo de Aalen para tempos contínuos, ver Seção 2.4 do Capítulo 2.

Além da análise gráfica, uma outra forma de verificar se uma covariável específica tem algum efeito na função de risco total é testar a hipótese nula de que não existe efeito da covariável. A hipótese nula para algum $j \geq 1$ é definida como:

$$H_0(t) : \beta_{jt} = 0 \text{ para todo } t \in \tau,$$

com estatística de teste

$$U_j = \sum_{s \in \tau} L_j(s)\widehat{\beta}_j(s) \tag{3.13}$$

em que $L_j(s)$ é o peso predito do processo. De acordo com Aalen (1989), $L_j(s)$ será o $(j+1)$-ésimo elemento da diagonal da matriz $(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}$. Em H_{0j} a estatística de teste (3.13) é uma tranformação martingal da forma (A.8) com $\mathbf{K}(s) = J(s)\mathbf{L}^{(j)}(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}\mathbf{X}^T(s)$, em que $\mathbf{L}^{(j)}(s)$ é uma matriz de dimensão $(p+1) \times (p+1)$ com todas as entradas iguais a

zero, exceto o (j+1)st elemento da diagonal que é igual a $L_j(s)$. O estimador da variância, $\widehat{var}(U_j)$, de (3.13) é dado por (A.10) avaliada em $t = T$, com $\widehat{\lambda}_i(s)$ dado por (3.12). Pelo teorema central para martingais, pode-se chegar a estatística de teste $U_j \{\widehat{var}(U_j)\}^{-1/2}$. Esta estatística tem uma distribuição assintótica normal padrão sob a hipótese nula.

O estimador (3.11) da matriz de covariância de $\widehat{\mathbf{B}}(t)$ é válido quando o modelo $P(Y_i(t) = 1|F_{it^-})$ descreve adequadamente a sua dependência "do passado" F_{it^-}. Em particular, isto requer que as covariáveis dinâmicas utilizadas no modelo em questão capture (na maior parte) essa dependência. Como alternativa, pode-se recorrer a um modelo marginal, assumindo apenas

$$E(Y_i(t)|R_i(t), \mathbf{x}_i(t)) = \mathbf{x}_i^T(t)\mathbf{R}_i(t)\beta(t) = \{\beta_0(t) + \beta_1(t)x_{i1}(t) + \cdots + \beta_p(t)x_{ip}(t)\} R_i(t).$$

Então, se os indivíduos forem independentes, utiliza-se o argumento de Scheike (2002) para obter o estimador

$$\widetilde{cov}(\widehat{\mathbf{B}}(t)) = \sum_{i=1}^{n} \mathbf{Q}_{it}^{\otimes 2} \tag{3.14}$$

para a matriz de covariância de (3.10), sendo

$$\mathbf{Q}_{it} = \sum_{s=0}^{t} J(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}\mathbf{x}_i(s)(Y_i(s) - \widehat{\lambda}_i(s)),$$

em que, em um vetor $\mathbf{a}$, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$.

3.1.2 Covariáveis Dinâmicas

Em algumas situações é de interesse saber como o passado influencia o presente e o futuro ou simplesmente considerar covariáveis que variam com o tempo. Por exemplo, pode-se estar interessado em conhecer o comportamento de uma doença no paciente com o passar do tempo (idade varia com o tempo). É de interesse particular em dados de eventos recorrentes considerar como os eventos anteriores podem influenciar os próximos eventos. Isto deve ser tratado com o uso de modelos específicos. Duas soluções são geralmente propostas na literatura: fragilidade que introduz um componente aleatório diferente no modelo para cada indivíduo, que muitas vezes resulta de uma distribuição gama (AALEN et al., 2008) e os modelos dinâmicos que introduzem o número de eventos anteriores para um indivíduo como covariável para prever eventos futuros (FOSEN et al., 2006) e (AALEN et al., 2008). Esta seção irá se concentrar na segunda solução em que a construção de um modelo de regressão aditivo dinâmico é um simples modelo de regressão aditivo.

Para ilustração, considere a situação na qual há duas covariáveis, uma fixa X_{1t} e outra dinâmica Z_{1t}. Um modelo marginal somente com X_{1t} é ajustado. O diagrama do caminho deste modelo é dado na Figura 3.1. Um modelo com as duas covariáveis, fixa e dinâmica é ajustado. Este modelo é definido em Aalen et al. (2004) como um modelo dinâmico ingênuo. O coeficiente estimado, $\beta_{x.z_{1t}}$, para a covariável fixa será menor em comparação com o $\beta_{x_{1t}}$, do modelo marginal. Isto pode ser explicado pela definição dos efeitos diretos e indiretos. Os autores definiram o efeito total de X_{1t} em Y_t como uma soma de efeitos diretos e indiretos, ou seja,

$$\beta_{X_{1t}} = \beta_{x.z_{1t}} + \beta_{X_{1t}}.\beta_{Z_{1t}}.$$

Nesta situação, o efeito total tem a interpretação de um modelo marginal. Esta decomposição pode ser realizada apenas de forma aditiva ou numa estrutura linear.

Uma estratégia para a situação descrita acima é ajustar uma regressão linear por mínimos quadrados ordinários da covariável dinâmica Z_{1t} na covariável fixa X_{1t}. Uma vez ajustado o modelo, uma covariável dinâmica nova pode ser definida como o resíduo a partir deste modelo. Este argumento baseia-se, quando uma covariável ortogonal é adicionada no modelo de regressão linear ordinário, fazendo com que os coeficientes de todos os outros modelos sejam inalterados. Utilizando a mesma notação apresentada na subseção anterior, tem-se:

$$Z_{1t} = X_{1t}\Psi_t + \epsilon_t,$$

sendo um modelo de regressão da covariável dinâmica na covariável fixa. Desde que tenha as estimativas dos parâmetros, uma covariável nova pode ser definida como

$$W_{1t} = Z_{1t} - X_{1t}\widehat{\Psi}_t,$$

e, em seguida, um novo modelo dinâmico é ajustado

$$Y_{it} = R_{it}[\beta_{0t} + \beta_{1t}X_{i1t} + \beta_{2t}(W_{1t} + X_{1t}\widehat{\Psi}_t)], \tag{3.15}$$

assim são dadas estimativas corretas tanto para as covariáveis fixas quanto para as dinâmicas preservando a aditividade do modelo linear. Esta estratégia foi definida por Fosen et al. (2005) como análise de caminho dinâmico (*path analysis*). Os autores argumentam que esta estratégia permite a decomposição do efeito total em efeitos diretos e indiretos.

Portanto, um efeito direto é um efeito que é transmitido através de uma única aresta

de um grafo, enquanto um efeito indireto é um efeito que é mediado através de uma ou várias covariáveis. Esta decomposição permite não apenas para descrever como o efeito de uma covariável fixa está trabalhando indiretamente. Por exemplo, se X_{i1t} tem um efeito negativo sob Z_{1t} e Z_{1t} tem um efeito positivo sob Y_t, então X_{i1t} tem um efeito negativo indireto sob Y_t. De acordo com os autores esse tipo de estratégia é de difícil execução no modelo de Cox.

3.1.3 Resíduo Martingal

Uma ferramenta importante para avaliar o ajuste de um modelo aditivo, é o resíduo martingal. Esta ferramenta foi introduzida por Aalen (1993), no contexto do seu modelo aditivo para sobrevivência e os dados históricos dos eventos (AALEN, 1980; AALEN, 1989), e a sua utilização em tempo contínuo para dados de eventos recorrentes foi ilustrado por Aalen et al. (2004). Neste trabalho será considerado os resíduos martingais para dados longitudinais binários em tempo discreto.

Para cada indivíduo i, tem-se o processo $N_i(t) = \sum_{s=0}^{t} Y_i(s)$ que conta o número de eventos observados para o indivíduo até o tempo t (incluso), e $\Lambda_i(t) = \sum_{s=0}^{t} \lambda_i(s)$. Então

$$M_i(t) = N_i(t) - \Lambda_i(t) \tag{3.16}$$

é um martingal. A idéia é agora substituir $\Lambda_i(t)$ em (3.16) por sua estimativa $\widehat{\Lambda}_i(t) = \sum_{s=0}^{t} \widehat{\lambda}_i(s)$ sob o modelo (3.12) para obter o resíduo martingal $\widehat{M_i}(t)$. Se o modelo é corretamente especificado, cada um dos n resíduos individuais devem se comportar como um martingal.

Mais especificamente, seja o vetor $\widehat{\mathbf{\Lambda}}(t) = (\widehat{\Lambda}_1(t), \ldots, \widehat{\Lambda}_n(t))^T$, e pela notação em (3.9) e (3.12), este pode ser dado por

$$\widehat{\mathbf{\Lambda}}(t) = \sum_{s=0}^{t} J(s)\mathbf{X}(s)\widehat{\beta}(s) = \sum_{s=0}^{t} J(s)\mathbf{H}(s)\mathbf{Y}(s),$$

em que

$$\mathbf{H}(s) = \mathbf{X}(s)(\mathbf{X}^T(s)\mathbf{X}(s))^{-1}\mathbf{X}(s)^T,$$

é a matriz predita. Então, introduzindo o vetor $\mathbf{N}(t) = \sum_{s=0}^{t} J(s)\mathbf{Y}(s)$ do processo de contagem, restrito a pontos de tempo, onde é possível a estimação. O vetor $\widehat{\mathbf{M}}(t) =$

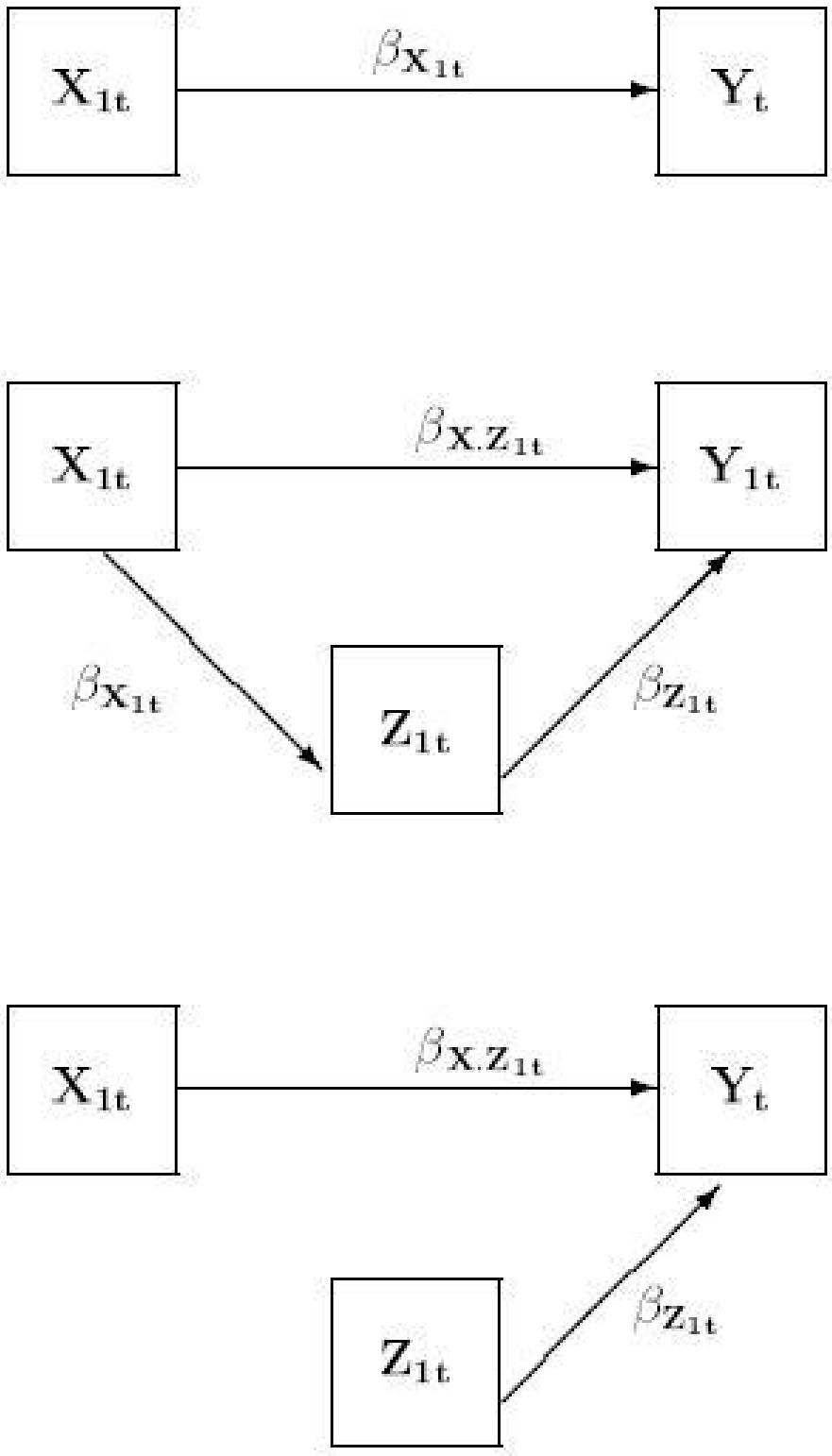

Figura 3.1: Diagrama do modelo marginal, modelo dinâmico ingênuo e dinâmico com ortogonalização. Fonte: Fiaccone (2006).

$(\widehat{M}_1(t), \ldots, \widehat{M}_n(t))^T$ são os resíduos martingais que podem ser escritos como

$$\widehat{\mathbf{M}}(t) = \mathbf{N}(t) - \widehat{\mathbf{\Lambda}}(t) = \sum_{s=1}^{t} J(s)(\mathbf{I} - \mathbf{H}(s))\mathbf{Y}(s). \tag{3.17}$$

Quando o modelo aditivo é corretamente especificado, $\mathbf{Y}(s) = \mathbf{X}(s)\beta(s) + \epsilon(s)$, e o vetor dos resíduos martingais torna-se

$$\widehat{\mathbf{M}}(t) = \sum_{s=1}^{t} J(s)(\mathbf{I} - \mathbf{H}(s))\epsilon(s)$$

ou seja, é uma transformação martingal e portanto terá média zero (A.8). Por (A.10) a matriz de covariância do vetor dos martingais residuais pode ser estimada por

$$\widehat{cov}(\widehat{\mathbf{M}}(t)) = \sum_{s=0}^{t} J(s)(\mathbf{I} - \mathbf{H}(s))\widehat{\Sigma}(s)(\mathbf{I} - \mathbf{H}(s))^T, \tag{3.18}$$

onde $\widehat{\Sigma}(s) = diag\left\{\widehat{\lambda}_i(s)(1 - \widehat{\lambda}_i(s))\right\}$.

Para o vetor de resíduos martingais e para as matrizes de covariância estimadas, pode-se derivar resíduos martingais padronizados a partir da divisão de cada entrada de (3.17) pela raiz quadrada do elemento correspondente a diagonal de (3.18). Se o modelo estiver bem especificado, estes devem ter média zero e variância um.

3.2 Modelos de Regressão Aditivos Para Dados Longitudinais Contínuos

Nesta seção serão considerados modelos de regressão aditivos dinâmicos para dados longitudinais em tempos contínuos, em que será estudada a modelagem sugerida por Martinussen e Scheike (1999, 2000). Estes modelos permitem a investigação da dinâmica temporal da variável resposta sobre as covariáveis e a média condicional da resposta dado a história é especificada como uma função linear dessas covariáveis.

Inicialmente os autores apresentaram a versão não-paramétrica, em que todos os coeficientes de regressão são não-paramétricos, permitindo uma dependência variável no tempo entre respostas e covariáveis. Além disso, as covariáveis são permitidas variar com o tempo. Aqui também serão consideradas inferências sobre os coeficientes de regressão cumulativos. A função de regressão cumulativa foi estudada da forma não-paramétrica no contexto de regressão em Scheike e Zhang (1998) e Scheike (2000). E os coeficientes variando no tempo foram estudados por Hastie & Tibshirani (1993) e Fahrmeir e Klinger (1998).

Às vezes a versão não-paramétrica apresenta estimativas não confiáveis para dados de amostras moderadas ou pequenas. Além disso, é sempre desejável reduzir os mode-

los para o mais simples possível, deixando-os robustos e com resultados fáceis de serem interpretados. Portanto, Martinussen e Scheike (1999) consideraram um submodelo do modelo não-paramétrico, em que a influência de algumas covariáveis variam com o tempo de forma não-paramétrica e o efeito das outras é constante. Este modelo é chamado de Modelo Aditivo Semiparamétrico. Um modelo semelhante foi considerado por Speckman (1988) para regressão de dados independentes e McKeague & Sasieni (1994) apresentaram o modelo de intensidade aditivo semiparamétrico para dados de sobrevivência.

3.2.1 A Modelagem Considerada

Considere que dados longitudinais podem ser descritos como um processo pontual marcado, ou seja, tem-se para o i-éssimo indivíduo a tríplice $(T_i^k, Z_i^k(T_i^k), X_i(t))$, em que T_i^k é o tempo da k-ésima mensuração, $Z_i^k(T_i^k)$ é a variável resposta longitudinal e $X_i(t)$ é uma possível covariável tempo dependente com vetor de dimensão *(p × 1)*. Considerando que são observados n indivíduos independentes no intervalo de tempo $[0, \tau]$. A dupla $(T_i^k, Z_i^k(T_i^k))$ constitue um processo pontual marcado, em que no tempo T^k ocorrem eventos específicos com (Z^k) sendo as marcas associadas. As marcas (Z^k) estão associadas ao espaço mensurável (E,ξ), denominado espaço de marcas. Para cada $A_i \in \xi$ é associado um processo de contagem

$$N_i(t)(A_i) = \sum_{k \geq 1} I(T_i^k \leq t) I(Z_i^k(T_i^k) \in A_i),$$

que conta o número de saltos anterior ao tempo t com marcas em A_i. Pode-se denotar $N_i(t) = N_i(t)(E)$, $i = 1, \ldots, n$, como um processo de contagem. Logo, um processo pontual marcado, assim como um processo de contagem, acumula informações ao longo do tempo.

Os processos pontuais marcados também podem ser identificados pela sua medida de contagem induzida $p_i(ds \times dz_i)$ definida por

$$p_i((0,t] \times A_i) = N_i(t)(A_i), \ \mathrm{A_i} \in \xi.$$

Como já definida no Capítulo 2, Seção 2.1, a filtração F_{t^-} representa a história de n processos. Esta história contém informações prévias sobre as respostas, os tempos de mensuração bem como as covariáveis. Com base no vetor básico das covariáveis $\mathbf{X}_i(t)$, são definidos dois vetores $\mathbf{X}_{\beta i}(t)$ e $\mathbf{X}_{\gamma i}(t)$ que são covariáveis que refletem diferentes aspectos do modelo. Ou seja, o primeiro representa o vetor das covariáveis em que seus efeitos são variáveis no tempo e o segundo o vetor das covariáveis com efeitos invariantes no tempo.

3.2.2 O Modelo Aditivo Não-Paramétrico

Assim como o modelo anterior, a variável resposta é modelada através da sua esperança condicional dada as covariáveis relevantes que podem incluir informações prévias de cada processo. A diferença nesse modelo é que é modelada a média de ocorrência de certo evento. Então o modelo médio aditivo não-paramétrico é dado por

$$m_i(t) = \int_E z_i \Phi_t^i(dz_i) = \beta_0(t) + \beta_1(t)X_{i1} + ... + \beta_p(t)X_{ip}, \tag{3.19}$$

em que E é o espaço onde as respostas assumem seus valores (o espaço de marcas) e Φ_t^i é a distribuição condicional da marca dada a informação prévia até o ponto de tempo em que a marca é obtida, isto é,

$$\Phi_{T_i^k}^i(A_i) = P(Z_i^k \in A_i | F_{T_i^k-}^i)$$

com $F_{T_i^k-}^i$ representado a história prévia até o tempo T_i^k e $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ sendo o vetor das funções de regressão tempo dependentes. A função de variância $\int_E (z_i - m_i(t))^2 \Phi_t^i(dz_i)$ é assumida independente para cada i e denotada por $\sigma^2(t)$. Outros modelos podem ser relevantes, ver Cheng e Wei (2000).

Uma forma alternativa de representar o modelo (3.19) é

$$Z_i^k = \beta_0(T_i^k) + \beta_1(T_i^k)X_{i1}(T_i^k) + \cdots + \beta_p(T_i^k)X_{ip}(T_i^k) + \epsilon_i^k,$$

com $E(\epsilon_i^k | F_{T_i^k-}^i) = 0$ e $V(\epsilon_i^k | F_{T_i^k-}^i) = \sigma^2(T_i^k)$. Pode-se observar que este modelo é semelhante ao Modelo Aditivo de Aalen. O interessante desses dois modelos é que os efeitos das covariáveis podem mudar com o tempo. Como no modelo aditivo estima-se as funções de regressão acumuladas $B(t) = \int_0^t \beta(s)ds$.

3.2.3 O Modelo Aditivo Semiparamétrico

O modelo de coeficientes variando no tempo descrito na subseção anterior fornece um bom resumo gráfico dos tempos das covariáveis dinâmicas e ainda permite a inferência sobre os efeitos das covariáveis. No entanto, muitas vezes é de interesse considerar modelos semiparamétricos.

Uma hipótese relevante sobre o efeito de cada covariável, é se de fato o seu efeito sofre alterações com o tempo ou se é constante. Portanto, pode ser considerado o seguinte

submodelo ((MARTINUSSEN; SCHEIKE, 1999))

$$m_i(t) = \boldsymbol{\beta}^T(t)\mathbf{X}_{\beta i}(t) + \boldsymbol{\gamma}^T\mathbf{X}_{\gamma i}(t) \tag{3.20}$$

onde $\mathbf{X}_{\beta i}(t)$ e $\mathbf{X}_{\gamma i}(t)$ são vetores de dimensão p e q respectivamente, $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ é o vetor das funções de regressão dependentes do tempo localmente integráveis e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$ é o vetor dos parâmetros desconhecidos. Portanto, o efeito de algumas covariáveis são considerados variáveis no tempo e o efeito das outras são assumidos constantes.

Uma forma alternativa de representar o modelo (3.20) é

$$Z_i^k = \boldsymbol{\beta}(T_i^k)^T\mathbf{X}_{\beta_i}(T_i^k) + \boldsymbol{\gamma}^T\mathbf{X}_{\gamma_i}(T_i^k) + \epsilon_i^k,$$

com $E(\epsilon_i^k|F^i_{T_i^k\_}) = 0$ e $V(\epsilon_i^k|F^i_{T_i^k\_}) = \sigma^2(T_i^k)$.

3.2.4 Inferência

Assim como o Modelo Aditivo de Aalen, o tempo de mensuração das respostas para o i-ésimo indivíduo são modeladas por processos de contagem $N_i(t)$ com intensidade $\lambda_i(t)$. E por simplicidade Martinussen e Scheike (1999), Martinussen e Scheike (2000) também consideraram o modelo multiplicativo de Aalen:

$$\lambda_i(t) = Y_i(t)\alpha(t)$$

em que $\alpha(t)$ é uma função determinística desconhecida e $Y_i(t)$ é uma variável indicadora do risco no tempo anterior a t.

A grande vantagem de uma abordagem com base em processos pontuais marcados é que a acumulação das respostas (e geralmente funções da resposta) ao longo do tempo dá origem a uma decomposição em martingal e compensador. Em que, o compensador é a distribuição da reposta dado a informação acumulada até um certo ponto de tempo, ou seja, representa em média o número de saltos do processo.

Os estimadores propostos por Martinussen e Scheike (1999, 2000) serão as integrais do processo pontual marcado básico e aqui será explicado brevemente a notação do processo pontual marcado. Para $H_i(t, z)$ uma função do tempo e respostas ($t \in [0, \infty[, z \in E$) que depende da informação anterior a t, é dada a seguinte notação

$$\int_0^t \int_E H_i(s, z_i)p_i(ds \times dz_i) = \sum_{k=1}^{\infty} H_i(T_i^k, Z_i(T_i^k))I(T_i^k \leq t),$$

referindo-se a integral do processo pontual marcado. Considere n processos pontuais marcados $p(dt \times dz) = (p_1(dt \times dz_1), \ldots, p_n(dt \times dz_n))^T$ e $H(s,z) = (H_{ij}(s,z))$ e $K(s,z) = (K_{ij}(s,z))$ matrizes de dimensão $n \times r$ com $z = (z_1, \ldots, z_n)$, e definindo como a j-ésima entrada da matriz de dimensão $r \times 1$

$$\int_0^t \int_E H(s,z)^T p(ds \times dz)$$

pode ser escrita por

$$\sum_{i=1}^{n} \sum_{k=1}^{\infty} H_{ij}(T_i^k, Z_i(T_i^k)) I(T_i^k \leq t)$$

e (i,j) são elementos da matriz de dimensão $r \times r$

$$\int_0^t \int_E H(s,z)^T diag(p(ds \times dz)) K(s,z)$$

pode ser expressa por

$$\sum_{i=1}^{n} \sum_{k=1}^{\infty} H_{ij}(T_i^k, Z_i(T_i^k)) K_{il}(T_i^k, Z_i(T_i^k)) I(T_i^k \leq t).$$

3.2.4.1 Estimação

Para os modelos não-paramétrico e semiparamétrico, será de interesse os estimadores das funções de regressão cumulativa, bem como o estimador paramétrico da parte paramétrica do modelo semiparamétrico. Todos os estimadores serão apresentados com suas variâncias e suas distribuições assintóticas, que serão válidos sob condições de regularidade. Os estimadores dos componentes não-paramétricos são martingais conjuntamente Gaussianos e os estimadores paramétricos são normalmente distribuidos. A inferência de todos os componentes pode ser efetuada sob as distribuições assintóticas.

3.2.4.2 Modelo Não-Paramétrico

Considerando o processo $H_i(t, z_i) = z_i$ e definindo $B_j(t) = \int_0^t \beta(s)ds$ e $\mathbf{B}(t) = (B_1(t), \ldots, B_p(t))^T$, observam-se as somas das respostas anterior ao tempo t para o i-ésimo indivíduo (integrando $H_i(t, z_i)$ com respeito a $p_i(dt \times dz_i)$) em que dá-se a seguinte

decomposição

$$\begin{aligned}\sum_k Z_i^k I(T_i^k \leq t) &= \int_0^t \int_E z_i p_i(ds \times dz_i) \\ &= \int_0^t \alpha(s) Y_i(s) m_i(s) ds + M_i(z_i)(t) \\ &= \int_0^t \alpha(s) Y_i(s) X_i^T(s) dB(s) + M_i(z_i)(t),\end{aligned}$$

para $i = 1, \ldots, n$, em que $M_i(z_i)(t)$ é o martingal do processo pontual marcado. Coletando n dessas equações uma única equação vetorial, e escrevê-la na forma diferencial, obtem-se

$$\int_0^t \int_E D(z) p(ds \times dz) = \int_0^t \alpha(s) Y(s) dB(s) + M(z)(t), \tag{3.21}$$

em que $z = (z_1, \ldots, z_n)$, $D(z) = diag(z)$, $M(z)(t) = (M_1(z_1)(t), \ldots, M_n(z_n)(t))^T$ e $Y(t) = (Y_{ij}(t))$ é uma matriz de dimensão $n \times (p+1)$ com a i-ésima linha, $i = 1, \ldots, n$, representada por

$$Y_i(t)(1, X_{i1}(t), \ldots, X_{ip}(t)).$$

Escrevendo (3.21) na forma diferencial,

$$\int_E D(z) p(dt \times dz) = \alpha(t) Y(t) dB(t) + dM(z)(t), \tag{3.22}$$

desde que $E(dM(z)(t)|F_{t-}) = 0$, são obtidos os estimadores de mínimos quadrados de $B(t)$ da forma

$$\widehat{B}(t) = \int_0^t \int_E \frac{J(s)}{\widehat{\alpha}(s)} Y^-(s) D(z) p(ds \times dz), \tag{3.23}$$

onde $\widehat{\alpha}(t)$ é o estimador de $\alpha(t)$, $Y^-(t)$ é a inversa generalizada de $Y(t)$, ou seja, uma matriz de dimensão $(p+1) \times n$ satisfazendo

$$Y^-(t) Y(t) = I_{(p+1)},$$

uma matriz identidade de dimensão $(p+1) \times (p+1)$ e $J(t)$=I($Y(t)$ tem posto completo e $\widehat{\alpha}(t) > 0$). Uma escolha da inversa generalizada é

$$(Y(t)^T Y(t))^{-1} Y(t)^T,$$

com base nos mínimos quadrados. Segundo Martinussen e Scheike (1999, 2000), esta escolha é ineficiente, mas de simples implementação.

Utilizando (3.23) é preciso especificar uma estimativa de $\alpha(t)$. Por simplicidade Martinussen e Scheike (1999, 2000), sugerem uma estimativa baseada na suavização *Kernel*, isto é,

$$\widehat{\alpha}(t) = \frac{1}{b_n} \int K\left(\frac{t-s}{b_n}\right) d\widehat{A}(s)$$

em que $\widehat{A}(t) = \int_0^t \frac{1}{Y.(s)} dN.(s)$, $Y.(t) = \sum_i Y_i(t)$, $N.(t) = \sum_i N_i(t)$, K é uma função *kernel* limitada com integração 1 e suporte $[-1, 1]$ e b_n é o parâmetro janela. Este é o estimador *kernel* de Ramlau-Hansen (1983).

O estimador para a variância é dado por

$$\widehat{\Sigma}(t) = n \int_0^t \int_E \frac{J(s)}{\widehat{\alpha}^2(s)} H(z,s) diag(p(ds \times dz)) H(z,s)^T$$

em que

$$H(z,t) = Y^-(t) \left(D(z) - \frac{1}{Y.(t)} Y(t) \widehat{\beta}(t) a \right)$$

com a sendo o vetor de 1's de dimensão $1 \times n$.

Utilizando o resultado (2.7) pode-se testar a hipótese de que o coeficiente de regressão cumulativo $B_j(t)$ é igual a zero. Uma boa estatística de teste é dada por

$$\sqrt{n} \frac{\widehat{B}_j(S)}{\sqrt{\widehat{\Sigma}_{jj}(S)}}, \tag{3.24}$$

que é assintoticamente normal padrão e $\widehat{\Sigma}_j j(t)$ denota o elemento (j,j) de $\widehat{\Sigma}(t)$. Esta estatística de teste tem sido discutida na configuração de regressão não-paramétrica por Scheike e Zhang (1998) e Scheike (2000).

3.2.4.3 Modelo semiparamétrico

Considerando agora que o modelo possui coeficientes de regressão não-paramétricos e paramétricos, a equação (3.22) pode ser expressa da seguinte forma:

$$\int_E D(z) p(dt \times dz) = \alpha(t) Y_\beta(t) dB(t) + \alpha(t) Y_\gamma(t) \gamma dt + dM(z)(t) \tag{3.25}$$

em que $\mathbf{Y}_\beta(t) = (Y_1(t)X_{\beta 1}(t), \ldots, Y_n(t)X_{\beta n}(t))^T$ e $\mathbf{Y}_\gamma = (Y_1(t)X_{\gamma 1}(t), \ldots, Y_n(t)X_{\gamma n}(t))^T$.

Suponha que γ seja conhecido. Outra forma de representar o estimador (3.23) é:

$$\widehat{B}(\gamma)(t) = \int_0^t \int_E \frac{1}{\widehat{\alpha}(s)} Y_\beta^-(s) D(z) p(ds \times dz) - \int_0^t Y_\beta^-(s) Y_\gamma(s) \gamma ds. \tag{3.26}$$

A estimação de γ pode ser realizada através da equação a partir de (3.25) com B(t) substituído por $\widehat{B}(\gamma)(t)$. O que resulta

$$\alpha(t) G(t) Y_\gamma(t) \gamma dt = \int_E G(t) D(z) p(dt \times dz), \tag{3.27}$$

em que $G(t) = \left\{ I - Y_\beta(t) Y_\beta^-(t) \right\}$.

Agora a estimação pode ser feita por dois caminhos. Pode-se resolver (3.27) localmente, obtendo-se a estimativa de γ a cada ponto de tempo e, em seguida, tomando-se uma estimativa final sob a forma de uma média ao longo do tempo. Outra possibilidade é multiplicar (3.27) por $Y_\gamma^T(t)$, antes de resolver para γ. Este estimador é similar ao estimador obtido por McKeague e Sasieni (1994) na versão semiparamétrica do Modelo Aditivo de Aalen. Logo, Martinussen e Scheike (1999) sugeriram o seguinte estimador para γ,

$$\widehat{\gamma} = \left\{ \int_0^\tau Y_\gamma^T(s) G(s) Y_\gamma(s) \widehat{\alpha}(s) ds \right\}^{-1} \int_0^\tau \int_E Y_\gamma^T(s) G(s) D(z) p(ds \times dz), \tag{3.28}$$

em que τ representa o tempo do estudo. Quando o modelo não contém quaisquer componentes não-paramétricos, γ pode ser interpretado como estimador de mínimos quadrados. Para mais detalhes ver Scheike (1994) em que os estimadores de mínimos quadrados do modelo paramétrico são baseados em processos pontuais marcados. Finalmente, pode-se estimar $B(t)$ por $\widehat{B}(\gamma)(t)$.

Considere a hipótese nula $H_0 : \beta_p(t) = \gamma_{q+1}$ para todo t. Note que, para $q = 0$, a hipótese é a de que a última função de regressão do modelo totalmente paramétrico é igual a uma constante. Para $q > 0$, compara-se dois modelos semiparamétricos. Em ambas as situações, pode-se utilizar a seguinte estatística de teste do desvio máximo

$$TST = sup_{t \in [0,\tau]} \left| \sqrt{n} \left\{ \widehat{B}_p(t) - \widehat{\gamma}_{q+1} t \right\} \right|.$$

Para mais detalhes ver equação (2.10) ou **??**), Capítulo V, Seção 5.4.

4 Aplicação

Neste capítulo serão apresentadas duas aplicações referentes aos dados de Serrinha, em que cada análise utilizará partes diferentes dos dados coletados. Na primeira aplicação a variável resposta é binária representando a ocorrência ou não de um episódio ou dia com diarréia (Seção 3.1). Na segunda estratégia modela-se o número médio de episódios de diarreia (Seção 3.2). As duas aplicações têm o mesmo objetivo de avaliar o efeito da suplementação periódica de vitamina A sobre a diarreia em crianças menores de 5 anos.

4.1 Aplicação 1

Nesta seção será aplicada a metodologia utilizada em Fiaccone (2006) e Borgan et al. (2007), em que a ênfase é ajustar um modelo para a ocorrência de diarreia ou episódio de diarreia, mensurada pela prevalência e incidência respectivamente, ao longo de um período de um ano de seguimento de um estudo longitudinal. Vale ressaltar que essa estratégia permite a inclusão de covariáveis tempo dependentes, bem como avaliar se o efeito da mesma também varia no tempo.

Segundo Fiaccone (1998, 2006), o desenho do estudo é do tipo longitudinal formado por uma coorte fixa, com o acompanhamento de 1240 crianças de 6 a 48 meses, com o objetivo de testar o efeito da suplementação de vitamina A sobre a diarreia, bem como a influência de outras covariáveis através da análise de prevalência e de incidência. As crianças com menos de 90 dias de acompanhamento foram excluídas da análise, ficando 1092 crianças.

Várias características sociais, demográficas e econômicas foram colhidas no início do estudo, as quais podem influenciar no resultado. Todas as covariáveis fixas foram recodificadas como binárias. Um grupo de três categorizações foi criado para a covariável idade cujo efeito é avaliado ao longo do tempo. O grupo de $18-36$ meses foi utilizado como categoria de referência. A Tabela 4.1 resume as covariáveis fixas e idade.

Tabela 4.1: Resumo das covariáveis fixas e idade utilizadas nos dados de Serrinha 1.

Variáveis	Descrição	Frequência
Sexo	Masculino (Male)	47%
	Feminino	57%
Idade no início do estudo (meses)	≤ 18	26%
	$18 - 36$	43%
	> 36	31%
Grupo de tratamento	Vit A	50%
	Placebo	50%
Tipo de piso	Terra (Land)	10%
	Outros	90%
Fonte de água	Canalizada	96%
	Outra (Contsrce)	4%
Água potável	Sim	72%
	Não (Notrt)	28%
Sistema de esgoto	Sim	35%
	Não (Nosew)	65%
Banheiro dentro de casa	Sim	73%
	Não (Abstoil)	27%
Escolaridade da mãe	Com estudo	66%
	Sem estudo (Illit)	24%
Outras crianças ≤ 5 anos	Sim (Othchld)	59%
	Não	41%

Algumas covariáveis dinâmicas foram definidas nesta análise: taxa de episódios em cada criança (Epsrate), dias com diarreia (Dayrate), interação entre os espisódios e dias com diarreia (Epsdays), dias com febre (Feverrate), dias com tosse (Tosrate) e dias com alguma doença (Doerate).

Nesses dados serão utilizados dias como a unidade de tempo, e tempo de calendário como a escala de tempo. A Figura 4.1 mostra a prevalência e incidência durante o período de estudo que representam, em um determinado dia, respectivamente, as proporções de crianças com diarreia e de um novo episódio de diarreia. No início, a prevalência é de cerca de 8%, tendo uma queda acentuada entre o ducentésimo e tricentésimo dia e passando dos 8% iniciais no final do estudo. A incidência é de aproximadamente 8% no início do estudo, tendo uma queda no decorrer dos dias e passando do valor inicial no final do acompanhamento. A queda em ambas as parcelas pode refletir a melhoria da saúde sobre o período de estudo, ou o envelhecimento da coorte. Deixando em alerta o comportamento atípico no final do estudo.

A Figura 4.2 mostra 10 crianças selecionadas aleatoriamente. As linhas acinzentadas

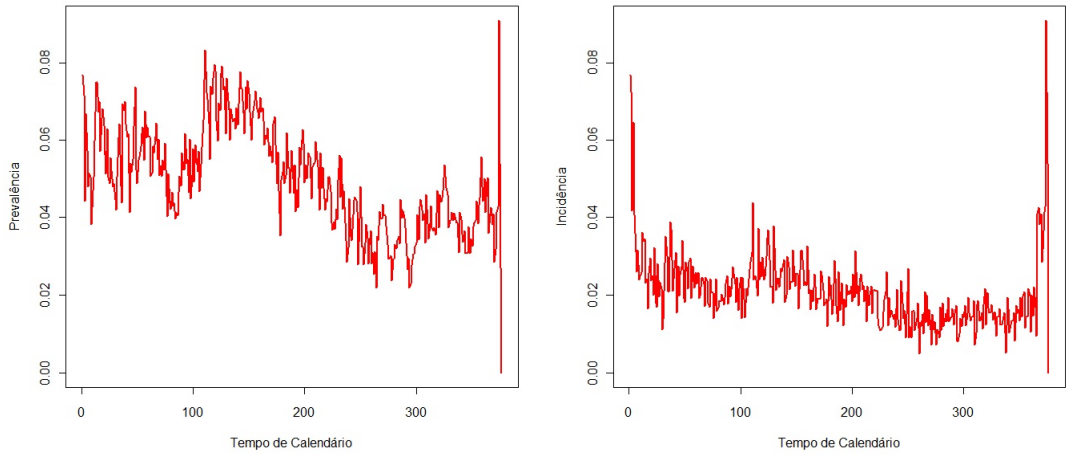

Figura 4.1: Prevalência e incidência diária de diarreia após início do estudo.

indicam os dados disponíveis; as cruzes marcam os dias consecutivos com diarreia; os círculos indicam abandono das crianças do estudo. Pode-se observar que algumas crianças são mais suscetíveis a diarreia do que outras, com algumas tendo episódios relativamente longos.

4.1.1 Modelagem

Considerando que a ocorrência ou não de diarreia em um determinado dia é a variável resposta, caracteriza-se, assim, um dado longitudinal binário para cada criança. Logo, a resposta longitudinal binária para cada criança é um processo de contagem que indica se um dia com diarreia ou episódios repetidos de diarreia aconteceram até o tempo t. A ideia é investigar como as covariáveis influenciam na mudança instantânea do processo de diarreia em cada ponto de tempo.

A variável Y_{it}, é um processo binário, denotando um dia com diarreia ou episódios de diarreia para cada criança i no tempo t, em que $t = 0$ é o tempo de início do estudo. Em seguida, o processo de contagem de interesse conta o número de dias/episódios de diarreia. O processo Y_{it} não pode ser observado em todos os tempos. Assim, a função de risco R_{it} é usada para indicar se a criança i estava em risco no tempo t. Isso significa que a criança estava ou não estava presente no dia da visita domiciliar. Logo, todas as covariáveis incluídas no modelo foram incorporadas nos estudos de prevalência e incidência.

Os modelos dinâmicos introduzem o número de eventos anteriores para um indivíduo como covariável que prevê eventos futuros (FOSEN et al., 2006) e (AALEN et al., 2008). Mas a inclusão de covariáveis dinâmicas distorce a estimação dos efeitos das covariáveis

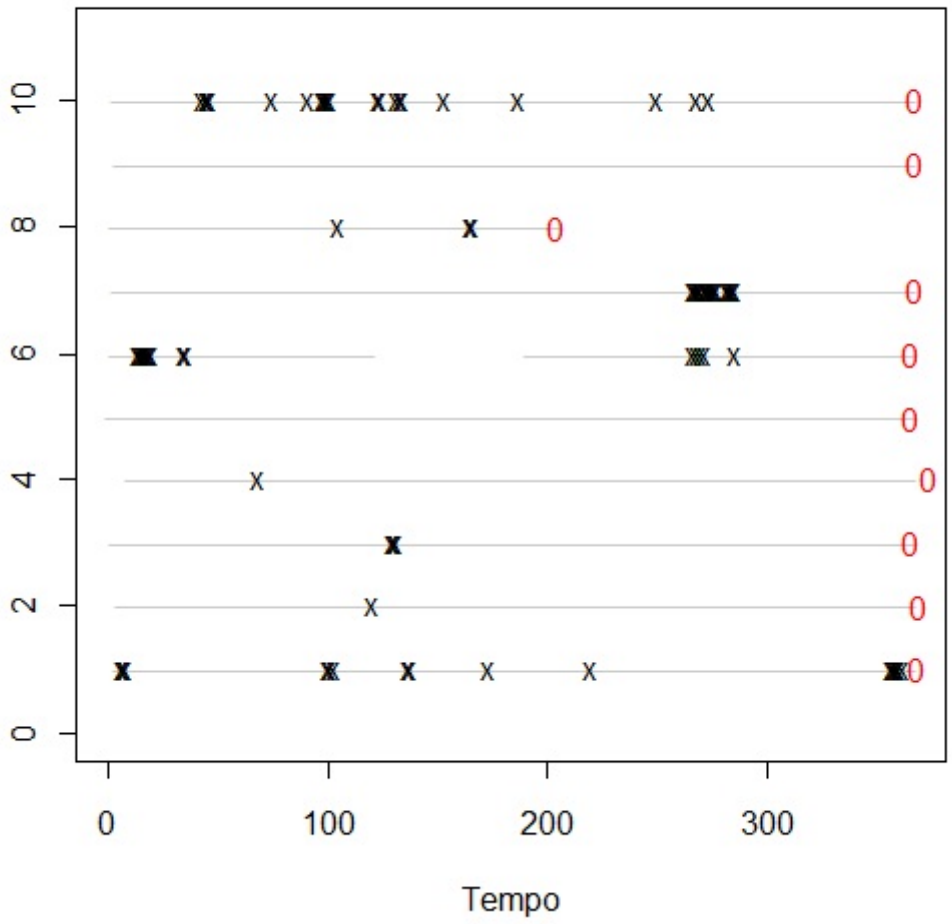

Figura 4.2: Dados da amostra. Cada linha corresponde a uma das 10 crianças selecionadas aleatoriamente.

fixas. A fim de evitar isso, Fiaccone (2006) propôs que em cada covariável dinâmica fosse determinada a covariável dinâmica X^*_{ijt} para cada indivíduo i como:

$$\begin{aligned} X^*_{ijt} &= \frac{\sum_{s=0}^{t-1} w_s R_{is} \tilde{Y}_{is}}{\sum_{s=0}^{t-1} w_s R_{is}} \\ &= \frac{\sum_{s=0}^{t-1} w_s Y_{is}}{\sum_{s=0}^{t-1} w_s R_{is}}, \end{aligned}$$

em que $\tilde{Y}_{is}$ é o evento do processo, R_{is} é o indicador de risco e w_s é uma função peso.

Esta função dá o mesmo peso aos eventos anteriores e é representada por:

$$w_s = \begin{cases} 1, & (t-s) \leq \tau \\ \\ \varrho^{-\rho(t-s-\tau)}, & (t-s) > \tau, \end{cases}$$

ou seja, é dado o mesmo peso a todos os eventos prévios ao dia τ. O peso decresce ao longo do tempo. Depois de consideráveis experimentações através da soma dos quadrados foram escolhidos $\tau = 120$ e $\rho = 0,01$ para a análise de incidência e prevalência. O uso dos resíduos a partir destas covariáveis será aplicado como no modelo de regressão aditivo (FOSEN et al., 2006) discutido anteriormente. Por este procedimento, os efeitos estimados das covariáveis fixas são os mesmos em um modelo com covariáveis dinâmicas como no modelo onde apenas covariáveis fixas estão incluídas. Portanto, as covariáveis dinâmicas poderiam considerar situações em que uma criança não tenha apresentado o evento diarréia enquanto ele/ela apresentou pelo menos um evento.

Para a análise de prevalência também foram incluídas covariáveis dinâmicas binárias, onde descreveu se a criança teve diarréia em cada um dos quatro dias anteriores, ou seja, *lags* 1-4. Essas covariáveis medem o efeito extra sobre o estado de diarréia em dias anteriores.

4.1.2 Resultados

Para ajustar o modelo multivariado para a ocorrência de diarréia, os primeiros 20 dias foram excluídos da estimação quando se presume que haviam ocorridos poucos eventos. Nas análises de incidência e prevalência, o método de seleção de modelos utilizado foi o *backward*.

4.1.2.1 Análise de Incidência

A Tabela 4.2 fornece as estatísticas de teste para as covariáveis fixas selecionadas. Basicamente, as crianças que vivem nas piores condições têm maior incidência de diarreia. Como esperado, as crianças que fizeram tratamento com vitamina A tiveram uma incidência mais baixa da doença. Pode-se observar que episódios de diarreia são mais comuns em crianças muito jovens.

As Figuras 4.3 e 4.4 mostram respectivamente os gráficos das funções de regressão acumuladas para as covariáveis fixas e dinâmicas. É observada uma acentuada diminuição

Tabela 4.2: Estatísticas de teste para efeitos das covariáveis fixas e idade nos modelos de regressão aditivos.

Covariáveis	Teste T	
	Prevalência	Incidência
Sexo masculino	8,23	2,69
Grupo de tratamento (VitA)	-10,51	-3,61
Casa sem água canalizada	-5,86	-2,29
Casa sem banheiro	26,27	8,48
Casa sem tratamento de água	6,97	3,53
Casa sem sistema de esgoto	12,13	2,64
Outras crianças ≤ 5 anos	15,70	3,96
Mãe sem estudo	2,87	1,01
Crianças menores que 18 meses	29,62	12,24
Crianças maiores que 36 meses	-57,96	-23,74

da incidência de diarréia em crianças que receberam a suplementação de vitamina A. Entretanto, depois de 200 dias há um leve aumento do seu efeito. Esse efeito acumulativo para o modelo de incidência atinge aproximadamente o valor $-0,6$ no final dos 365 dias. Esse valor é significativo, mas ainda pequeno, uma vez que representa, em média, 0,002 episódios por criança tratada. A ausência de banheiro em casa tem um aumento acentuado na incidência de diarréia. Há uma flutuação estável para casas sem o sistema de esgotos. Crianças que moram em casas onde não há água potável tem um acentuado aumento na incidência de diarréia. A intensidade *baseline* acumulativa parece ser linear. Além disso, o efeito da idade é forte, com um grande aumento na incidência de diarréia quando as crianças com idade inferior a 18 meses. Por outro lado, as crianças acima de 36 meses têm uma tendência de queda na incidência de diarréia.

A primeira covariável dinâmica conta o número médio de episódios anteriores por dias em risco (epsrate). Isso é altamente significativo, fornecendo evidências do efeito de fragilidade, ou seja, algumas crianças são mais suscetíveis do que outras, mesmo conhecendo os fatores de risco. A segunda covariável dinâmica (dayrate) mede a proporção de dias anteriores em que a criança tem diarréia, e por isso leva em conta comprimento dos episódios. Novamente, há uma associação positiva, embora não tão forte como a taxa de episódio. Ademais, história de tosse (tosrate) e de qualquer outra doença (doerate) contribui para o aparecimento de futuros episódios.

A Figura 4.5 mostra uma amostra de 10 valores do processo residual padronizado, escolhidos aleatoriamente, para o modelo de incidência. O efeito pode ser observado com a inclusão de covariáveis dinâmicas. Vale salientar a vantagem da inclusão de covariáveis

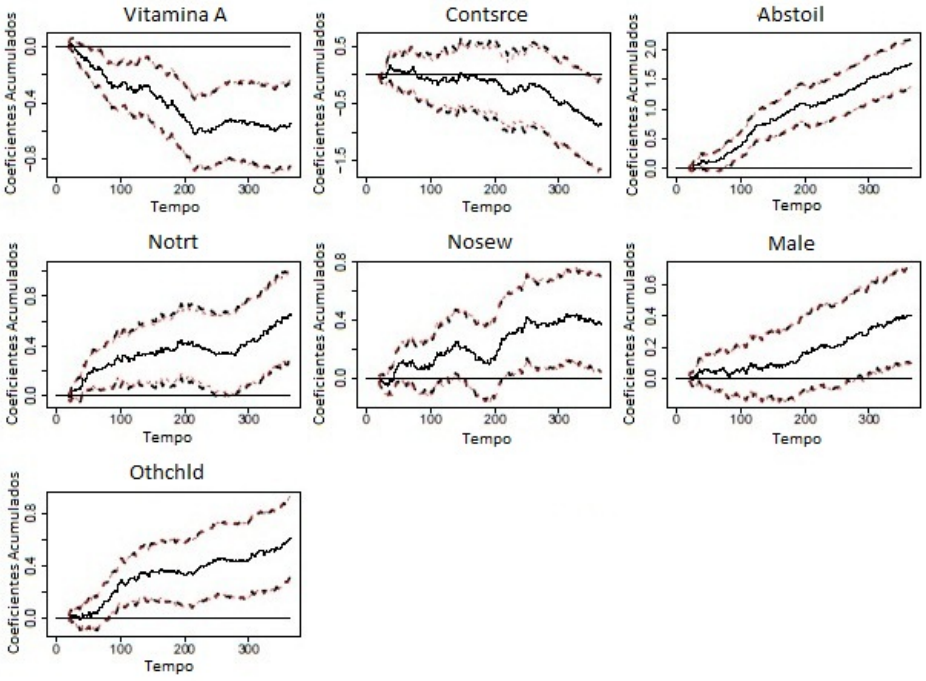

Figura 4.3: Estimativas das funções de regressão acumuladas e seus respectivos intervalos de confiança de 95% para as covariáveis fixas para a análise de incidência.

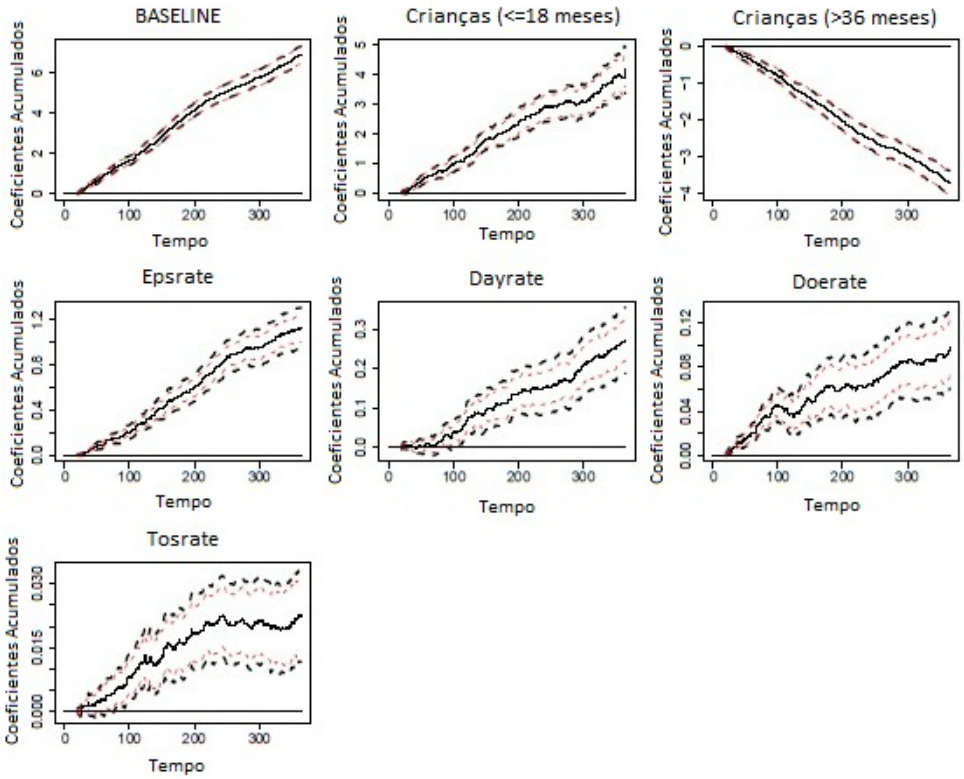

Figura 4.4: Estimativas das funções de regressão acumuladas e seus respectivos intervalos de confiança de 95% para idade e covariáveis dinâmicas para a análise de incidência.

dinâmicas no modelo, em um outro gráfico de desvios padrão empíricos dos resíduos martingais padronizados, apresentado na Figura 4.6. Observa-se que o modelo com covariáveis dinâmicas dá resultados muito razoáveis.

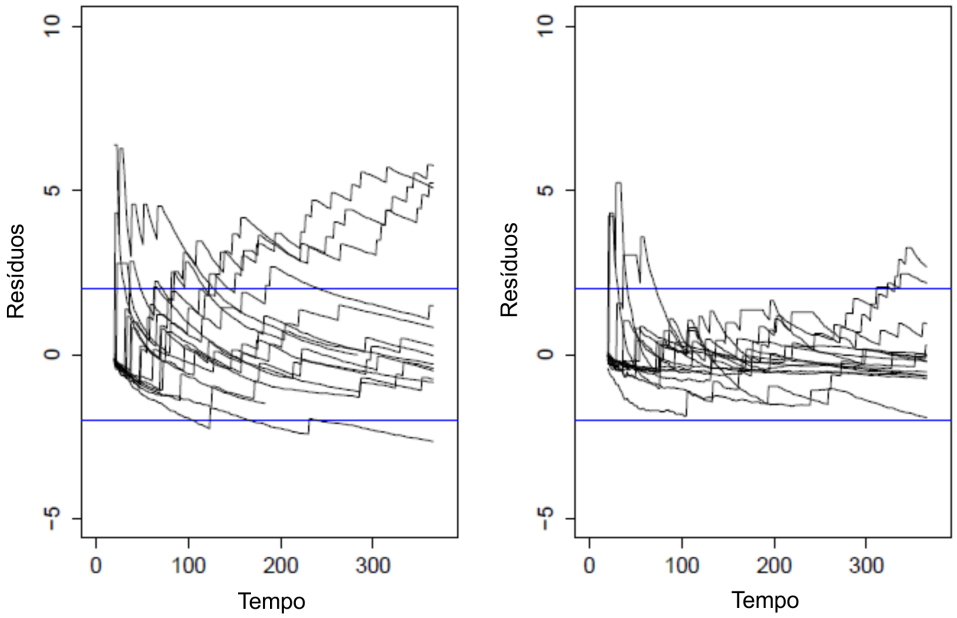

Figura 4.5: Resíduos martingais padronizados para o modelo de incidência: sem covariáveis dinâmicas (à esquerda) e com covariáveis dinâmicas (à direita).

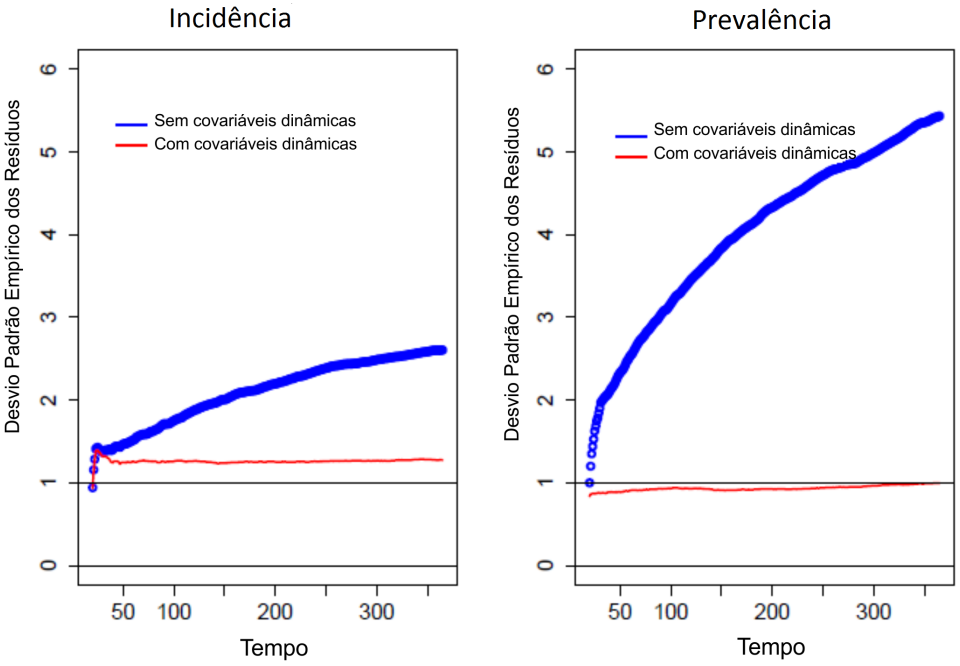

Figura 4.6: Desvio padrão empírico para os resíduos padronizados para o modelo de incidência e o modelo de prevalência.

4.1.2.2 Análise de Prevalência

A Tabela 4.2 resume alguns dos resultados da análise de prevalência. Como na incidência percebe-se que as crianças que vivem nas piores condições têm maior prevalência de diarreia. Também tendo exceção para a fonte de água. E as crianças mais velhas e que receberam tratamento com vitamina A têm efeitos negativos na prevalencia de diarreia.

A Figura 4.7 fornece o coeficiente *baseline* cumulativo, o efeito de grupo e de seis covariáveis dinâmicas importantes para a análise de prevalência. Mais uma vez, a vitamina A tem um efeito relevante, no sentido de diminuir a prevalência da diarreia em comparação com as crianças que receberam placebo. As covariáveis dinâmicas são a proporção de dias em que a criança teve diarreia (dayrate), a proporção de dias com alguma outra doença (doerate) e as variáveis binárias que descrevem se uma criança teve diarreia em cada um dos quatro dias anteriores, isto é, *lags* 1-4. Nota-se que o efeito de *lag* reduz em magnitude e significância com d aumentos de dias. As conclusões para as covariáveis fixas foram omitidas por serem as mesmas para as covariáveis fixas do modelo de incidência.

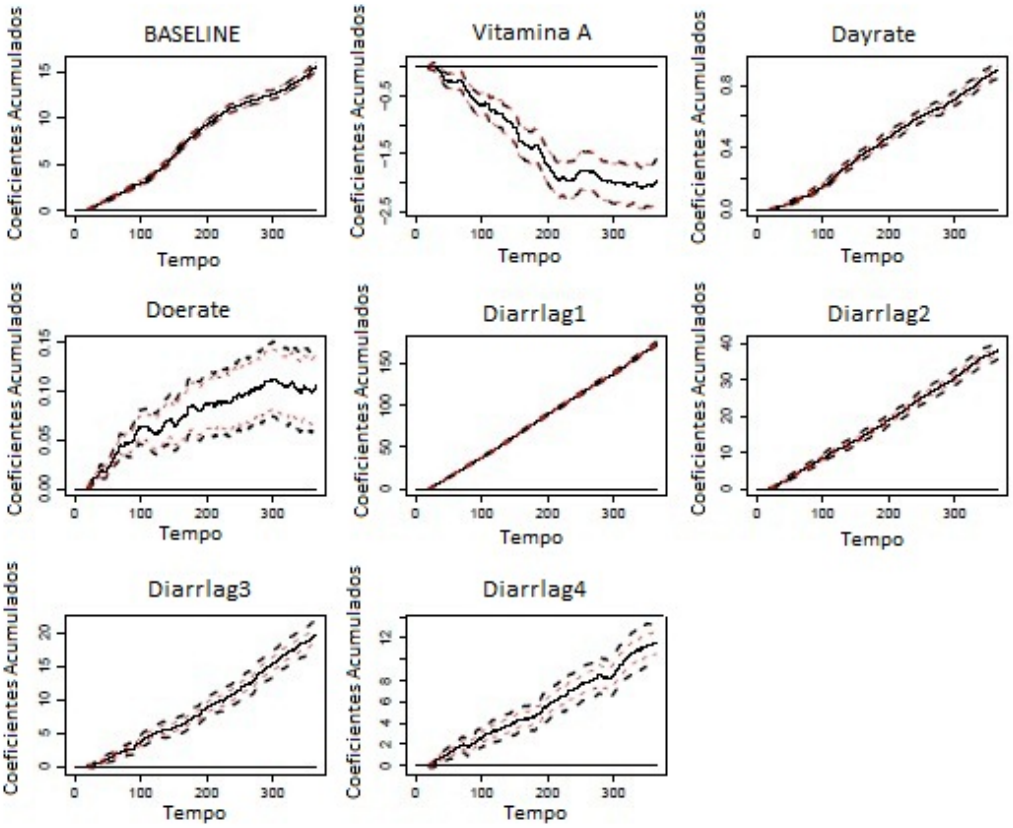

Figura 4.7: Estimativas das funções de regressão acumuladas e seus respectivos intervalos de confiança de 95% para idade, covariáveis fixas e dinâmicas para a análise de prevalência.

A Tabela 4.3 mostra os efeitos estimados destas covariáveis sobre a probabilidade de

ocorrência de diarreia. Sabendo que a criança teve diarreia no dia anterior aumenta a probabilidade em cerca de 45% de ocorrência de diarreia no dia posterior. Bem como, o fato de que a criança teve diarreia por dois dias anteriores aumenta a probabilidade de diarreia em 11%.

Tabela 4.3: Probabilidade estimada e observada de diarreia.

Dias prévios com diarreia				Prevalência	
4	3	2	1	Observada	Modelo
				5%	5%
			x	55%	45%
		x	x	63%	56%
	x	x	x	68%	61%
x	x	x	x	71%	64%

A partir dos vários fatores de risco, juntamente com a idade das crianças, estima-se um efeito benéfico da suplementação de vitamina A em ambos, incidência e prevalência de diarreia. A partir da Figura 4.8 tem-se indicação de uma diminuição da prevalência diária ao longo do tempo em ambos os grupos. Também, é de se esperar que haja um efeito negativo da idade sobre a incidência e prevalência de diarreia, isto é, a chance de uma criança experimentar um episódio de diarreia diminui à medida que a criança se torna mais velha.

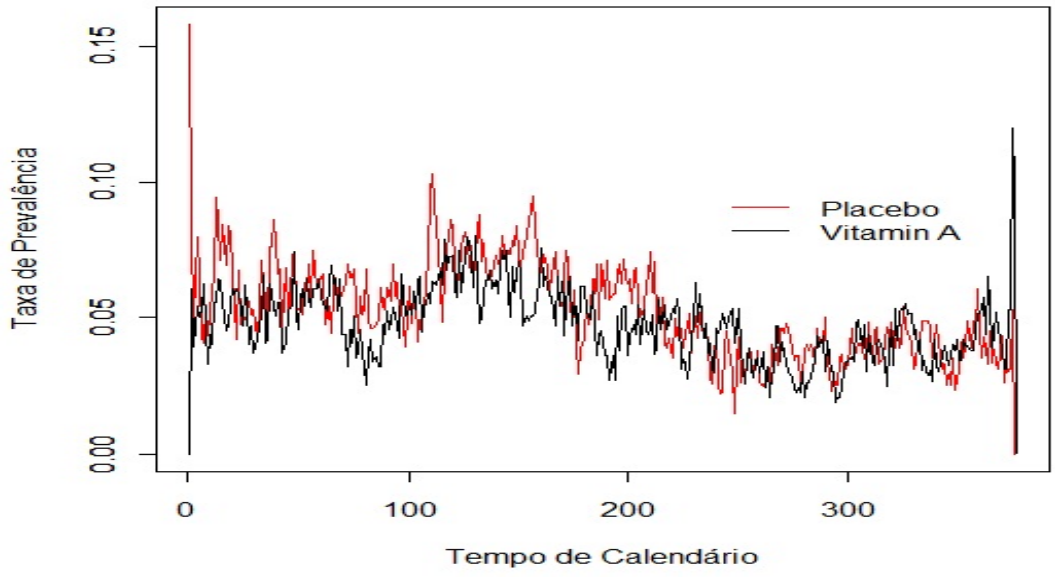

Figura 4.8: Prevalência diária dos grupos placebo e vitamina A.

A Figura 4.6 apresenta o gráfico dos desvios padrão empíricos dos resíduos martingais padronizados, pode-se perceber que o modelo com covariáveis dinâmicas fornece um melhor ajuste.

4.2 Aplicação 2

Nesta seção será apresentada uma aplicação dos modelos aditivos não-paramétrico e semiparamétrico propostos por Martinussen e Scheike (1999, 2000).

Como mencionado anteriormente, cada análise utilizará partes diferentes dos dados coletados e aqui o desenho de estudo também é do tipo longitudinal formado por uma coorte fixa, porém com o acompanhamento de 860 crianças (426 no grupo placebo e 434 no grupo vitamina A) de 6 a 48 meses analisadas no mesmo período, também com o mesmo objetivo de testar o efeito da suplementação de vitamina A sobre a diarreia.

No banco de dados cada linha referente ao mesmo indivíduo representa um tempo entre episódios de diarreia. As covariáveis fixas (**grupo, sexo e idade**) têm seus valores repetidos em todas as linhas referentes a essa criança. As covariáveis que variam de episódio para episódio são **diasant, mediadej** e a variável reposta **enum** representa o número de episódios de diarreia para cada criança. A Tabela 4.4 apresenta a descrição das covariáveis fixas e dinâmicas utilizadas neste estudo que podem ou não serem consideradas como fator de risco para ocorrência de diarreia.

Tabela 4.4: Descrição das covariáveis utilizadas nos dados de Serrinha 2.

Variáveis	Descrição
Grupo	Vit = receberam vitamina A, Pla = placebo
Sexo	Fem = feminino, Masc = masculino
Idade	Idade em meses no início do estudo
Diasant	Duração em dias do episódio de diarréia anterior
Mediadej	Média de dejeções líquidas ou semilíquidas do episódio de diarréia anterior

Na Tabela 4.5 encontram-se as estatísticas descritivas para algumas covariáveis. Note que o número mínimo de observações para cada criança foi de 1 e máximo de 27 e, em média, cada uma apresentou 5,86 episódios. A duração média em dias do episódio de diarreia anterior para cada criança foi de 2,83. Quanto à média de dejeções líquidas ou semilíquidas do episódio de diarreia anterior, o máximo foi de 18,05 e metade delas tiveram uma média de 3 dejeções líquidas e semilíquidas. Os coeficientes de variação mostram o

quanto os dados são heterogêneos, de forma especial na covariável **diasant**.

Tabela 4.5: Resultado da análise descritiva dos dados de Serrinha 2.

Variáveis	Mín	Mediana	Média	Máximo	CV(%)
Idade	6,00	25,00	25,71	48,00	46,48
Diasant	0,00	2,00	2,83	65,00	133,38
Mediadej	0,00	3,00	3,07	18,05	54,29
Enum	1,00	5,00	5,86	27,00	79,07

A Figura 4.9 apresenta os *boxplots* correspondentes ao número de episódios de diarreia para cada criança por grupo e por sexo. Observa-se que o número eventos teve menor concentração nas crianças do grupo vitamina A e maior concentração nas crianças do sexo masculino.

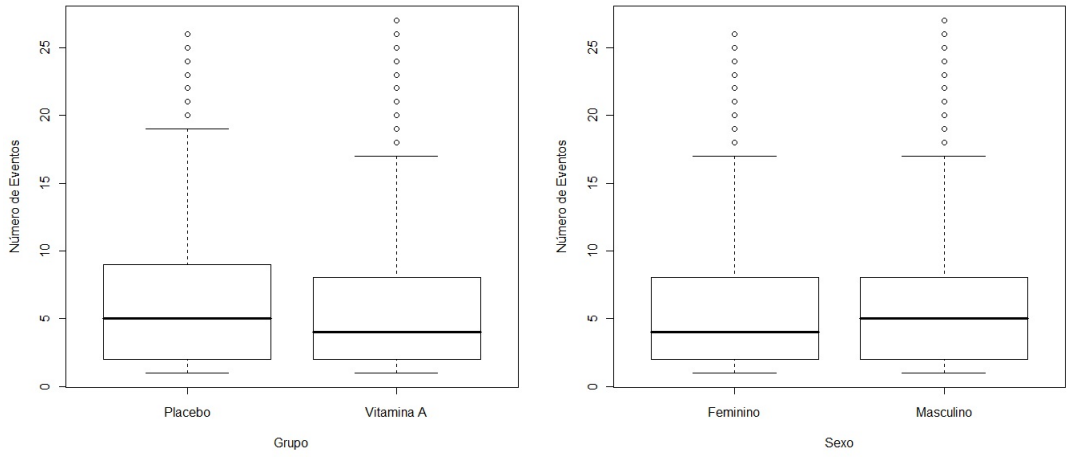

Figura 4.9: Boxplots do número de episódios de diarreia por grupo e por sexo.

4.2.1 Modelagem

Considere que o tempo de mensuração das respostas para a i-ésima criança são modeladas pelo processo de contagem $N_i(t)$, que conta o número de dias/episódios de diarreia, com intensidade $\lambda_i(t)$, que representa a taxa média de episódios de diarreia da criança i no tempo anterior a t. A variável indicadora do risco no tempo anterior a t, R_{it}, indica se a criança i estava ou não em risco.

O objetivo da análise aqui proposta é verificar se o número médio de episódios de diarreia para cada criança mudou ao longo do tempo. Portanto, a modelagem é inicial-

mente realizada pelo Modelo Aditivo de Aalen, em que todos os parâmetros do modelo têm seus efeitos variando no tempo. Depois o modelo anterior é reduzido só com as constantes para a constatação de suas significâncias. Em seguida, é dado continuidade, com o modelo para a média baseado em processos pontuais marcados com a taxa dependendo das covariáveis significativas do modelo anterior, ou seja, as covariáveis que realmente são constantes. E, finalmente, é ajustado o modelo semiparamétrico final.

4.2.2 Resultados

Nesta subseção são aplicados quatro modelos aditivos para respostas longitudinais. O primeiro é o Modelo Aditivo de Aalen (1989) que é baseado em processos de contagem, sendo um modelo totalmente não-paramétrico. O segundo é sua versão semiparamétrica. O terceiro é o modelo de regressão baseado em processos pontuais marcados e o quarto sua versão semiparamétrica, ambos sugeridos por Martinussen e Scheike (1999, 2000).

A partir dos resultados dos testes e dos gráficos das funções de regressão é possível detectar as covariáveis que têm seus efeitos variáveis no tempo, bem como extrair outras informações adicionais úteis. Por exemplo, na Figura 4.10 é possível verificar se existe influência positiva ou negativa do uso de suplementos de vitamina A sobre a taxa de diarreia.

Para o ajuste dos modelos foram utilizadas as funções *Aalen* e *Dynreg* da biblioteca *Timereg* do pacote *R*. Em todos os modelos foram consideradas 1000 simulações de reamostragem, em que essas reamostragens são utilizadas para calcular os p-valores dos testes para verificação do efeito constante de cada covariável. De início foi considerado o Modelo Aditivo de Aalen, representado por

$$\begin{aligned} \lambda_i(t) &= \beta_0(t) + \beta_1(t)(grupo)_{i1} + \beta_2(t)(sexo)_{i2} + \beta_3(t)(idade)_{i3} \\ &\quad +\beta_4(t)(diasant)_{i4} + \beta_5(t)(mediadj)_{i5}. \end{aligned} \tag{4.1}$$

A importância de considerar primeiramente o modelo (4.1) é de conhecer a influência de cada covariável sobre a taxa $\lambda_i(t)$. Na Tabela 4.6 é apresentado o resumo das estatísticas. Pode-se verificar através do Teste do Supremo (considerando o supremo da estatística (3.24)) que a covariável **sexo** (p-valor=0,53) não é significativa no modelo. Os efeitos de **sexo** (p-valor=0,43), **idade** (p-valor=0,05) e **diasant** (p-valor=0,05) parecem ser constantes ou insignificantes. A Figura 4.10 mostra as estimativas dos coeficientes de regressão acumulados com seus respectivos intervalos de confiança de 95% e bandas de confiança de *Hall-Wellner* (visto na Subseção 2.4.1.3). Observando este gráfico é possível

confirmar que os efeitos de **sexo**, **idade** e **diasant** são constantes ao longo do tempo, pois os coeficientes acumulativos são aproximadamente linhas retas. Nota-se que há uma diminuição na taxa de diarréia em crianças que receberam suplementos de vitamina A. No entanto, depois de 200 dias de acompanhamento há um leve aumento neste efeito. As crianças do sexo masculino tiveram um aumento estável na taxa de diarreia. Com o aumento da idade da criança há uma diminuição na taxa de diarreia.

Tabela 4.6: Teste do Supremo e Cramer Von Mises para testar a significância das covariáveis e efeito tempo invariante (função *Aalen*).

Teste do Supremo		Cramer von Mises ($T_{2,const}$)
Covariável	(Valor p)	(Valor p)
Grupo:Vit	3,91 (0,00)	14,50(0,00)
Sexo:Masc	2,01 (0,53)	2,41 (0,43)
Idade	10,80(0,00)	0,01 (0,05)
Diasant	3,00 (0,03)	2,59 (0,05)
Mediadej	7,78 (0,00)	3,32 (0,02)

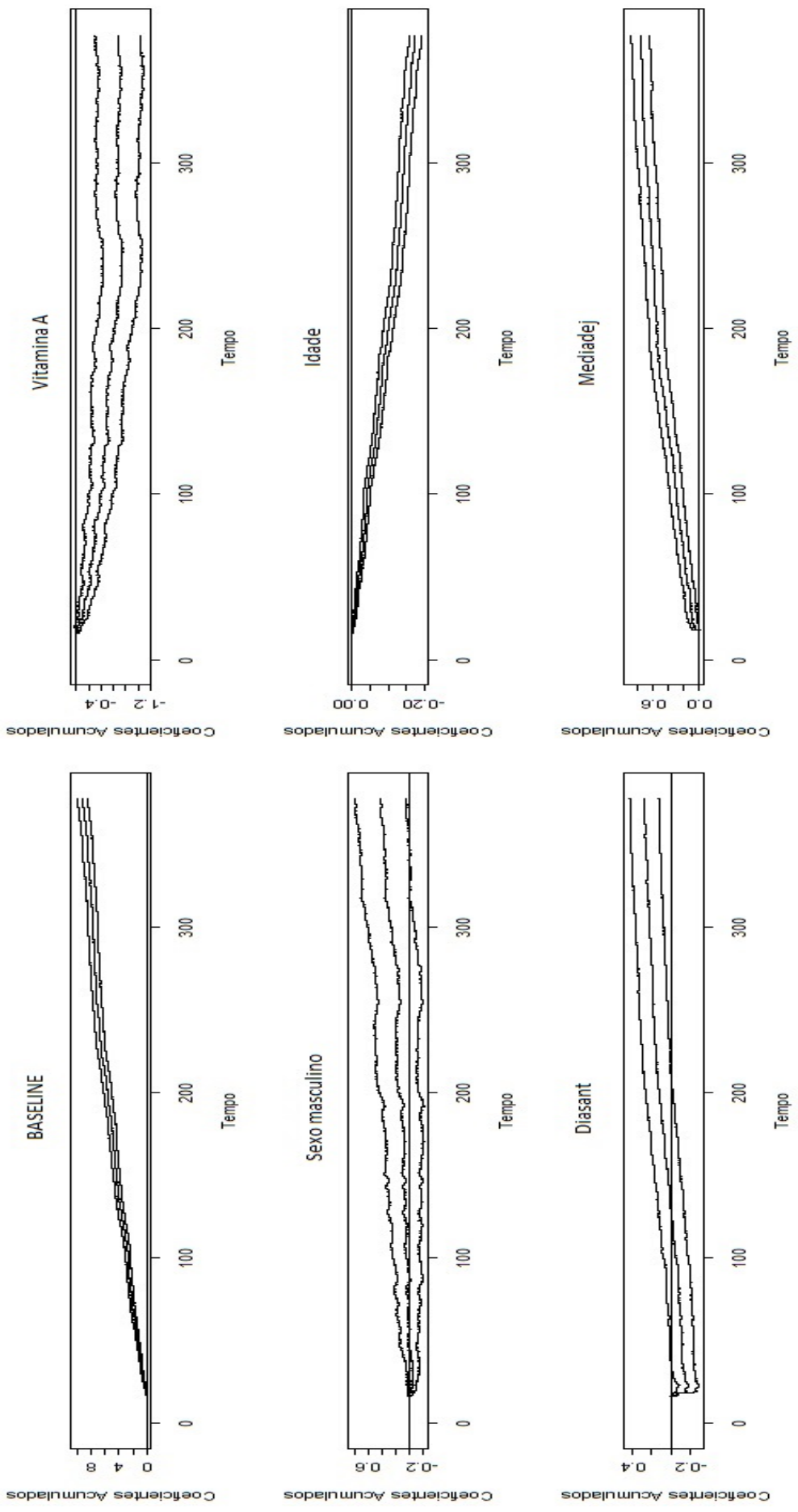

Figura 4.10: Estimativas das funções de regressão acumuladas e seus respectivos intervalos de confiança de 95%.

Na Figura 4.11 são apresentados os gráficos correspondentes ao Teste do Processo (algoritmo da Subseção 2.4.2.2) onde são realizadas 50 simulações de cada processo sob a hipótese nula que os efeitos de cada covariável é constante. Os gráficos confirmam que as covariáveis (**grupo** e **mediadej**) tem seus efeitos variando no tempo.

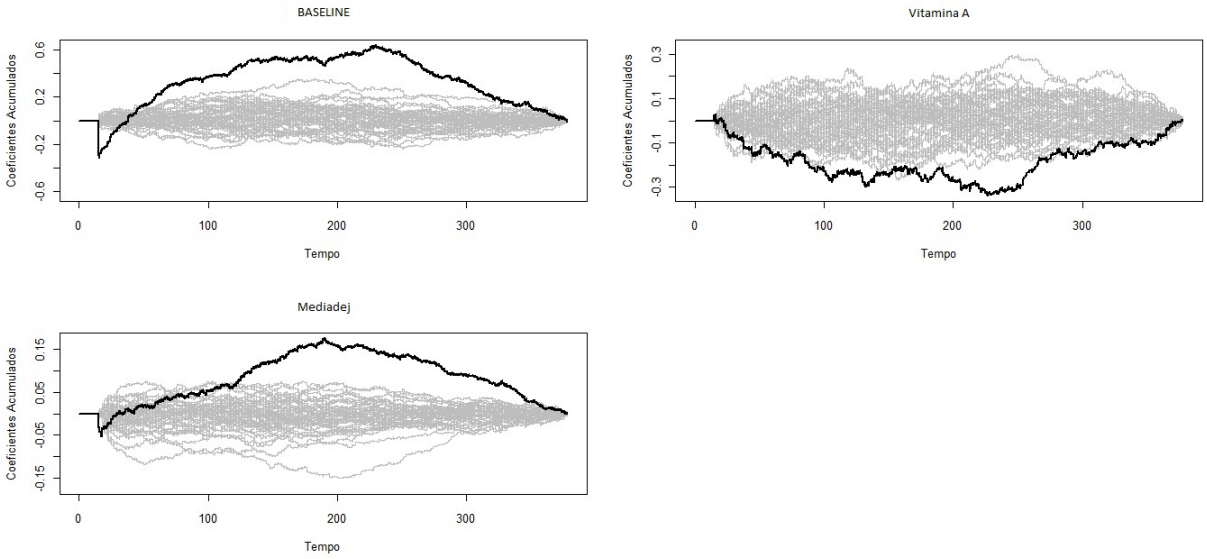

Figura 4.11: Teste para verificação do efeito constante para cada covariável.

Em seguida será apresentado o modelo (4.1) na sua forma semiparamétrica, ou seja, com suas covaráveis fixas e variáveis no tempo. O modelo é representado por:

$$\begin{aligned} \lambda_i(t) &= \beta_0(t) + \beta_1(t)(grupo)_{i1} + \beta_2(t)(mediadj)_{i2} + \gamma_1(sexo)_{i1} \\ &\quad +\gamma_2(idade)_{i2} + \gamma_3(diasant)_{i3}. \end{aligned} \tag{4.2}$$

em que $\beta(t) = (\beta_1(t), \beta_2(t))^T$ é o vetor das funções de regressão dependentes do tempo e $\gamma = (\gamma_1, ..., \gamma_3)^T$ é o vetor dos parâmetros em que os efeitos das covariáveis são constantes ao longo do tempo. O objetivo principal do modelo (4.2) é verificar a significância das covariáveis consideradas constantes. A Tabela 4.7 apresenta os resultados da parte paramétrica (que não depende do tempo). Assim, pode-se notar que os efeitos das três covariáveis consideradas com efeitos constantes são significativos. Para a utilização da função *Dynreg* é de grande importância primeiramente a aplicação do Modelo Aditivo de Aalen para verificar quais as covariáveis serão consideradas com efeitos constantes. Pois nesta função a taxa do modelo de Aalen irá depender das covariáveis constantes. Depois, é dada continuidade com a variável resposta sendo modelada através da função *Dynreg* e

Tabela 4.7: Estimativas dos parâmetros das covariáveis com efeitos invariantes no tempo (função *Aalen*).

Covariáveis	Estimativas	Erros Padrão	Valor P
Sexo:Masc	0,000791	0,001040	4,47e-01
Idade	-0,000518	0,000044	0,00e+00
Diasant	-0,001600	0,000210	2,89e-14

este modelo será representado por:

$$\begin{aligned} m_i(t) &= \beta_0(t) + \beta_1(t)(grupo)_{i1} + \beta_2(t)(sexo)_{i2} + \beta_3(t)(idade)_{i3} \\ &\quad + \beta_4(t)(diasant)_{i4} + \beta_5(t)(mediadj)_{i5}. \end{aligned} \tag{4.3}$$

A Tabela 4.8 e a Figura 4.12 indicam que o efeito das covariáveis **grupo** (p-valor=0,84), **sexo** (p-valor=0,24) e **mediadej** (p-valor=0,11) são constantes ao longo do tempo.

Tabela 4.8: Teste do Supremo e Cramer Von Mises para testar a significância das covariáveis e efeito tempo invariante (função *Dynreg*).

Teste do Supremo		Cramer von Mises ($T_{2,const}$)
Covariável	(Valor p)	(Valor p)
Grupo:Vit	4,70 (0,00)	95800 (0,84)
Sexo:Masc	1,45 (0,92)	736000 (0,24)
Idade	6,52 (0,00)	33900 (0,00)
Diasant	2,45 (0,18)	552000 (0,00)
Mediadej	5,16 (0.00)	62500 (0,11)

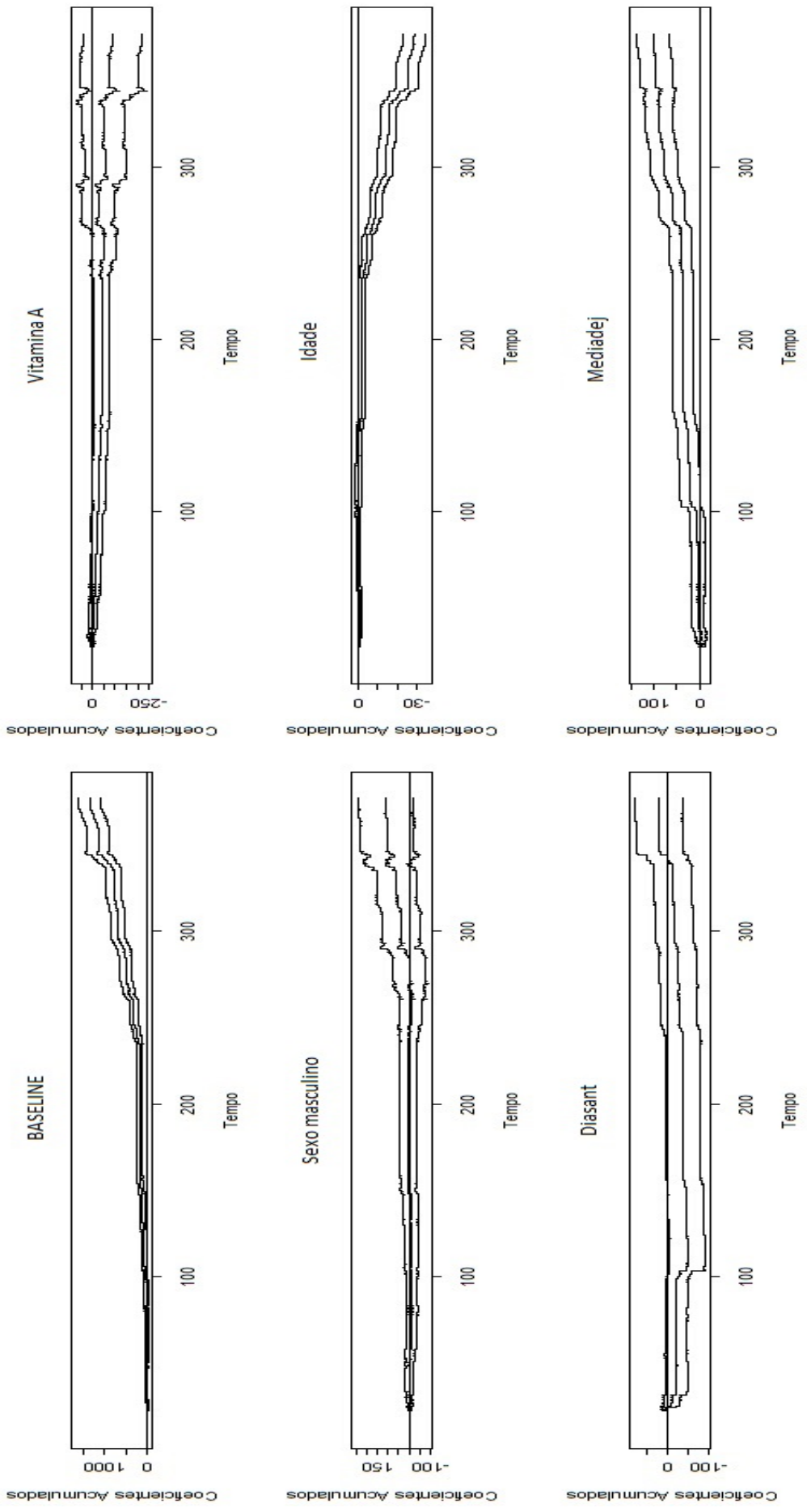

Figura 4.12: Estimativas das funções de regressão acumuladas e seus respectivos intervalos de confiança de 95%.

Finalmente é considerado o modelo semiparamétrico seguinte, em que as estimativas parecem razoáveis,

$$\begin{aligned} m_i(t) &= \beta_0(t) + \gamma_1(grupo)_{i1} + \beta_1(t)(idade)_{i1} + \gamma_2(Sexo)_{i2} \\ &\quad +\beta_2(t)(diasant)_{i2} + \gamma_3(mediadj)_{i3}. \end{aligned} \tag{4.4}$$

A partir das Tabelas 4.9 e 4.10 pode-se concluir que as covariáveis **idade** (p-valor=0,00), **diasant** (p-valor=0,00), **sexo** (p-valor=0,00), **mediadej** (p-valor=0,00), tem seus efeitos significativos, ou seja, as duas primeiras têm seus efeitos variáveis no tempo e as duas últimas têm seus efeitos constantes ao longo do tempo. O efeito da covariavel **grupo** (p-valor=0,85) é considerada insignificante para o modelo. O aumento em um dia na duração em dias do episódio de diarreia anterior(**diasant**), por exemplo, leva a um aumento na média estimada de 7,04. Os meninos têm o índice de diarreia de 1,68, com erro padrão de 0,51.

Tabela 4.9: Teste do Supremo e Cramer Von Mises para testar a significância das covariáveis e efeito tempo invariante (função *Dynreg*).

Teste do Supremo		Cramer von Mises ($T_{2,const}$)
Covariável	(Valor p)	(Valor p)
Diasant	7,04 (0,00)	305000 (0,00)
Idade	7,19 (0,00)	29800 (0,00)

Tabela 4.10: Estimativas dos parâmetros das covariáveis invariantes no tempo (função *Dynreg*).

Covariáveis	Estimativas	Erros Padrão	Valor P
Grupo:Vit A	0,0936	0,501	0,85
Sexo:Masc	1,6800	0,518	0,00
Mediadej	1,6000	0,124	0,00

Partindo da hipótese nula que os efeitos são invariantes no tempo, a Figura 4.13 mostra o Teste do Processo com 50 reamostragens. O comportamento das covariáveis **idade** e **diasant** nas figuras revela que seus efeitos são variáveis no tempo.

A qualidade do ajuste do modelo 4.4 foi verificada pela inspeção dos residuos martingais, como descrito na subseção 2.4.3. Na Figura 4.14 foram consideradas parcelas dos resíduos martingais para três grupos definidos pela covariável idade. , os resíduos

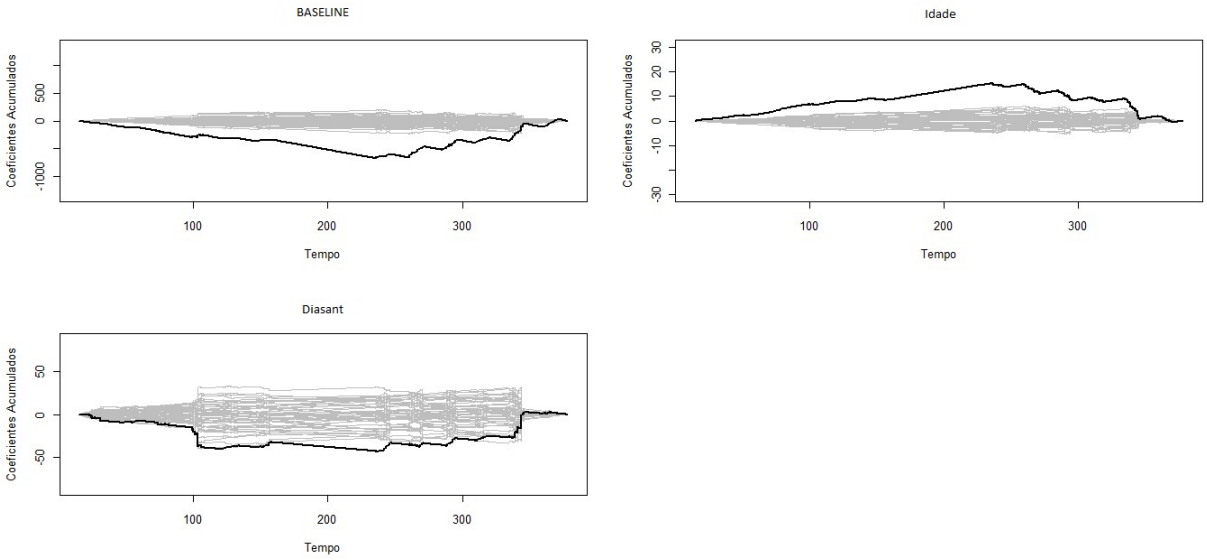

Figura 4.13: Teste para verificação do efeito constante para cada covariável.

juntamente com bandas de 95% de confiança. As parcelas para o grupo 1 e grupo 3 foram semelhantes, ou seja, deram indícios de um modelo mal ajustado. Porém o grupo 2 mostra um modelo bem ajustado. Mas considera-se que não há violação do modelo pois a maioria das crianças estão inseridas no grupo 2. A análise dos resíduos é omitida para a covariável diasant pois os resultados são parecidos com a anterior.

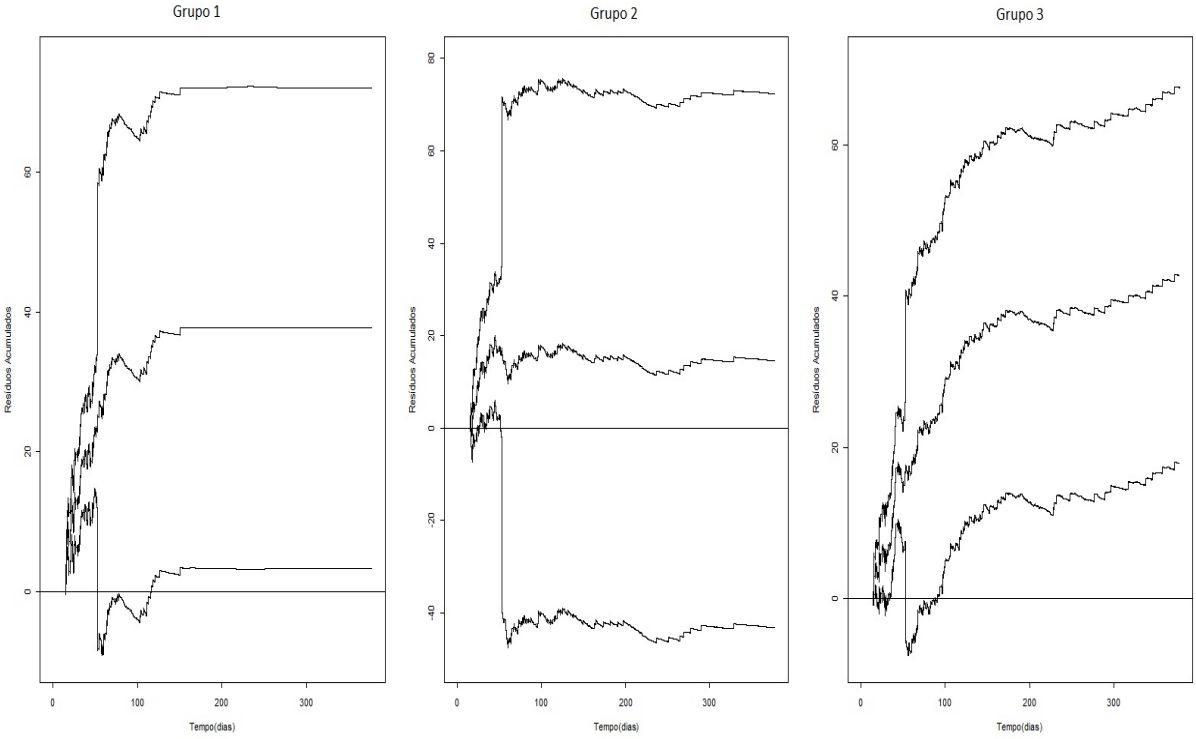

Figura 4.14: Resíduos martingais para a covariável idade.

5 Conclusões

Métodos para lidar com dados de eventos recorrentes têm sido estudados na formulação de processos de contagem e frequentemente utilizados em aplicações em dados longitudinais. As metodologias aqui empregadas podem ser vistas como uma alternativa flexível em que combinam os efeitos das covariáveis fixas e dinâmicas.

Foram apresentados modelos de regressão capazes de incorporar covariáveis tempo dependentes através do uso da informação da resposta prévia na resposta atual (nomeada covariável dinâmica). Esta abordagem permite o entendimento do fenômeno estudado com base na história dos eventos. Portanto, é permitido caracterizar a susceptibilidade, heterogeneidade ou a fragilidade individual baseada na história passada (FIACCONE, 2006).

Uma das desvantagens dos modelos é a possibilidade de se estimar valores negativos para a função de risco. Esta particularidade é, por vezes, utilizada contra a utilização do Modelo Aditivo de Aalen. No entanto, uma variedade de razões superam esta deficiência. A primeira, é que o interesse deste trabalho é na função de regressão cumulativa, que são estimadas de forma consistente nas abordagens. A segunda, é a permissão do uso da teoria martingal, que por sua vez irá permitir um comportamento assintótico dos estimadores de mínimos quadrados para as funções de regressão cumulativas. A terceira, é que se houver interesse na previsão individual, então faz sentido em qualquer caso, aplicar alisamento local para função risco reduzir o ruído, e isso deve trazer estimativas dentro dos limites. A quarta, e mais importante, é que as estimativas são de rápida estimação.

Na primeira aplicação foi considerada a variável resposta binária representando a ocorrência ou não de um episódio ou dia com diarreia e, na segunda, o número médio de episódios de diarréia. As duas aplicações tiveram o mesmo objetivo de avaliar o efeito da suplementação periódica de vitamina A sobre a diarreia em crianças menores de 5 anos.

Os resultados obtidos apontam que a partir dos vários fatores de risco, juntamente com a idade das crianças, estima-se um efeito benéfico significativo da suplementação de vitamina A na incidência e prevalência de diarreia. Também foi visto que há um efeito

negativo da idade sobre a incidência e prevalência de diarreia, isto é, a chance de uma criança experimentar um episódio de diarreia diminui à medida que a criança se torna mais velha e que as condições de moradia influenciam no aparecimento da doença.

Outra abordagem comum para a análise de eventos recorrentes tem sido o uso de modelos de fragilidade. Então, em termos de pesquisa futura pretendemos incluir um termo de fragilidade variando no tempo na estrutura do Modelo Aditivo de Aalen de forma a se ter uma interpretabilidade dinâmica.

APÊNDICE A – Resultados Referentes aos Dados Longitudinais Binários

Este apêndice irá resumir alguns resultados para dados longitudinais binários em tempo discreto que são necessários na Seção 3.1 do Capítulo 3. Serão considerados dados longitudinais binários para um indivíduo genérico.

Seja um processo estocástico binário $Y = Y(t)$; $t \in \tau$; com $Y(0) = 0$, adaptado a uma filtração (F_t) em que N é o processo $N(t) = \sum_{s=0}^{t} Y(s)$ que conta o número de $1's$ em Y até o tempo t (incluso). Então o processo

$$\lambda(t) = P(Y(t) = 1|F_{t^-}) = E(Y(t)|F_{t^-}) \tag{A.1}$$

com sua parte cumulativa $\Lambda(t) = \sum_{s=0}^{t} \lambda(s)$, em que os processos λ e Λ são previsíveis (isto é, $\lambda(t)$ e $\Lambda(t)$ são F_{t^-}-mensuráveis em cada tempo $t \in \tau$). Em seguida será considerado o processo $M = N - \Lambda$. Note que $M(0) = 0$, e que

$$M(t) = N(t) - \Lambda(t) = \sum_{s=0}^{t} \epsilon(s)$$

onde $\epsilon(t) = Y(t) - \lambda(t)$. De (A.1) têm-se que $E(\epsilon(t)|F_{t^-}) = 0$. Assim

$$E(M(t)|F_{t^-}) = E(M(t^-) + \epsilon(t)|F_{t^-}) = M(t^-) + E(\epsilon(t)|F_{t^-}) = M(t^-),$$

o que mostra que M é um martingal. Em particular $E(M(t)) = E(E(M(t)|F_0)) = M(0) = 0$, para todo $t \in \tau$, assim M tem média zero. O processo de variação previsível $\langle M \rangle$ de M é dado por

$$\langle M \rangle_t = \sum_{s=0}^{t} Var(\epsilon(s)|F_{s^-}) = \sum_{s=0}^{t} \lambda(s)(1 - \lambda(s)). \tag{A.2}$$

O processo $M^2 - \langle M \rangle$ é um martingal de média zero. Em particular, $Var(M(t)) =$

$E(M^2(t)) = E\langle M\rangle_t$.

Se $K = \{K_t\}$ é um processo previsível, então têm-se o processo KoM definido por

$$(KoM)_t = \sum_{s=0}^{t} K_s(M(s) - M_{s^-}) = \sum_{s=0}^{t} K_s\epsilon(s) \tag{A.3}$$

Este processo é uma versão discreta de uma integral estocástica, e denotado pela *transformação* de M por K. É de fácil verificação, utilizando a previsibilidade de K, que KoM é um martingal. E o seu processo de variação previsível é dada por

$$\langle KoM\rangle_t = \sum_{s=0}^{t} Var(K_s\epsilon(s)|F_{s^-}) = \sum_{s=0}^{t} K_s^2\lambda(s)(1-\lambda(s)) \tag{A.4}$$

Suponha n séries binárias $Y_i; i = 1, \ldots, n$; considerando N_i, λ_i, M_i e ϵ_i os processos derivados desta. Pressupondo que os processos sejam adaptados ou previsiveis, tal como descrito acima com respeito a uma filtração comum (F_t). Para tempos discretos, situação considerada aqui, pode-se ter $Y_i(t) = 1$ para dois ou mais índices i com probabilidade positiva. Assim, os processos de contagem N_i podem ter saltos comuns, e portanto, o processo de n-variáveis $(N_1, \ldots, N_n)$ não é um processo de contagem multivariado. No entanto, muitos dos resultados da teoria de processos de contagem em tempos contínuos transitam para a situação atual de assumir que os erros individuais $\epsilon_i(t)$ são condicionalmente não-correlacionados, especificamente, que para todo t e todo $i \neq j$

$$Cov(\epsilon_i(t), \epsilon_j(t)|F_{t^-}) = 0. \tag{A.5}$$

Para os processos de covariância previsíveis $\langle M_i, M_j\rangle$ torna-se

$$\langle M_i, M_j\rangle_t = \sum_{s=0}^{t} = Cov(\epsilon_i(s), \epsilon_j(s)|F_{s^-}) = \delta_{ij}\langle M_i\rangle_t, \tag{A.6}$$

com δ_{ij} um delta de *Kronecker*. Assim os martingais M_i são ortogonais; uma propriedade chave para a teoria clássica de processo de contagem em tempos contínuos. Nota-se que o pressuposto (A.5) é cumprido quando os processos estocásticos binários $Y_i; i = 1, \ldots, n$; são condicionalmente independentes dado F_0, como é o caso neste trabalho.

Se $K^{(1)} = \left\{K_t^{(1)}\right\}$ e se $K^{(2)} = \left\{K_t^{(2)}\right\}$ são processos previsíveis, seguindo por (A.2) e

(A.5) têm-se

$$\begin{aligned}\left\langle K^{(1)}oM_i, K^{(2)}oM_j\right\rangle_t &= \sum_{s=0}^{t} Cov(K_s^{(1)}\epsilon_i(s), K_s^{(2)}\epsilon_j(s)|F_{s^-}) \\ &= \sum_{s=0}^{t} \delta_{ij}K_s^{(1)}K_s^{(2)}\lambda_i(s)(1-\lambda_i(s))\end{aligned} \tag{A.7}$$

Considerando $\mathbf{M}(t) = (M_1(t), \ldots, M_n(t))^T$ e $\epsilon(t) = (\epsilon_1(t), \ldots, \epsilon_n(t))^T$, os vetores dos martingais e seus incrementos, e nota-se que o processo de variação previsível $\langle \mathbf{M} \rangle$ de $\mathbf{M}$ é a matriz avaliada com a (i,j)-ésima entrada igual a $\langle \mathbf{M}_i, \mathbf{M}_j \rangle$. Além disso, $\mathbf{K} = \{\mathbf{K}_t\}$ é uma matriz $p \times n$ de processos previsíveis. Em seguida, a transformação de $\mathbf{M}$ por $\mathbf{K}$ é um vetor martingal KoM de média zero com dimensão p dado por

$$(\mathbf{K}o\mathbf{M})_t = \sum_{s=0}^{t} \mathbf{K}_s\epsilon(s). \tag{A.8}$$

De acordo (A.4), (A.6) e (A.7) têm-se a matriz de variação previsível do processo $\langle \mathbf{K}o\mathbf{M} \rangle$ de dimensão $p \times p$ de $\mathbf{K}o\mathbf{M}$ que é dada por

$$\langle \mathbf{K}o\mathbf{M} \rangle_t = \sum_{s=0}^{t} \mathbf{K}_s\Sigma(s)\mathbf{K}_s^T. \tag{A.9}$$

onde $\Sigma(s) = diag\{\lambda_i(s)(1-\lambda_i(s))\}$ é uma matriz diagonal com o i-ésimo elemento diagonal igual a $\lambda_i(s)(1-\lambda_i(s))$. Em particular a matriz de covariância de $(\mathbf{H}o\mathbf{M})_t$ assume a forma $Cov(\mathbf{K}o\mathbf{M})_t = \sum_{s=0}^{t} E(\mathbf{K}_s\Sigma(s)\mathbf{K}_s^T)$, e pode ser estimada por

$$\widehat{Cov}(\mathbf{K}o\mathbf{M})_t = \sum_{s=0}^{t} \mathbf{K}_s\widehat{\Sigma}(s)\mathbf{K}_s^T \tag{A.10}$$

onde $\hat{\Sigma}(s) = diag\left\{\hat{\lambda}_i(s)(1-\hat{\lambda}_i(s))\right\}$ com $\hat{\lambda}_i(s)$ o estimador de (A.1).

Referências

AALEN, O. Nonparametric inference for a family of counting processes. **The Annals of Statistics**, v. 6, n. 4, p. 701–726, 1978.

AALEN, O. et al. **Survival and Event History Analysis: A Process Point of View**. [S.l.]: Springer-Verlag New York, 2008. (Statistics for Biology and Health). ISBN 9780387685601.

AALEN, O. O. A model for non-parametric regression analysis of counting processes. **Lecture Notes in Statistics**, McGraw-Hill, v. 2, p. 1–25, 1980.

AALEN, O. O. A linear regression model for the analysis of life times. **Statistics in Medicine**, v. 8, n. 8, p. 907–925, 1989.

AALEN, O. O. Further results on the non-parametric linear regression model in survival analysis. **Statistics in Medicine**, v. 12, n. 17, p. 1569–1588, 1993.

AALEN, O. O. et al. Dynamic analysis of multivariate failure time data. **Biometrics**, Wiley Online Library, v. 60, n. 3, p. 764–773, 2004.

ANDERSEN, P. **Statistical Models Based on Counting Processes**. [S.l.]: Springer, 1993. (Springer Series in Statistics). ISBN 9780387945194.

ANDERSEN, P. K.; GILL, R. D. Cox$\hat{\text{E}}\frac{1}{4}$s regression model for counting processes: A large sample study. **Annals of Statistics**, JSTOR, v. 10, n. 4, p. 1100–1120, 1982.

ANDERSEN, P. K.; VAETH, M. Simple parametric and nonparametric models for excess and relative mortality. **Biometrics**, v. 45, n. 2, p. 523–535, 1989.

BAQUI, A. H. et al. Methodological issues in diarrhoeal diseases epidemiology: definition of diarrhoeal episodes. **International Journal of Epidemiology**, AMER SOC CLINICAL PATHOLOGY, v. 20, n. 4, p. 1057–1063, 1991.

BARRETO, M. L. et al. Effect of vitamin a supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in brazil. **Lance**, v. 344, n. 8917, p. 228–231, 1994.

BORGAN, R. et al. Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in brazil. **Scandinavian Journal of Statistics**, Blackwell Publishing Ltd, v. 34, n. 1, p. 53–69, 2007. ISSN 1467-9469.

BOX-STEFFENSMEIER, J. M.; BOEF, S. D. Repeated events survival models: the conditional frailty model. **Statistics in Medicine**, John Wiley & Sons, Ltd., v. 25, n. 20, p. 3518–3533, 2006.

DIGGLE, P. **The Analysis of Longitudinal Data**. [S.l.]: Clarendon Press, 2002. (Oxford Statistical Science Series). ISBN 9780198524847.

DIGGLE, P.; FAREWELL, D.; HENDERSON, R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. **Writing**, Wiley Online Library, v. 56, n. 5, p. 499–550, 2007.

ELGMATI, E. et al. Frailty modelling for clustered recurrent incidence of diarrhoea. **Statistics in Medicine**, v. 27, n. 30, p. 6489–6504, 2008.

FAHRMEIR, L.; KLINGER, A. A Nonparametric Multiplicative Hazard Model for Event History Analysis. **Biometrika**, v. 85, p. 581–592, 1998.

FAN, J.; ZHANG, W. Statistical estimation in varying-coefficient models. **The Annals of Statistics**, v. 27, p. 1491–1518, 2000.

FAREWELL, D.; HENDERSON, R. Longitudinal perspectives on event history analysis. **Lifetime Data Analysis**, Springer, v. 16, n. 1, p. 102–117, 2010.

FIACCONE, R. L. **Métodos Estatísticos para Análise de Dados Categorizados com Estruturas Complexas**. Dissertação (Mestrado) — Universidade Estadual de Campinas — UNICAMP, 1998.

FIACCONE, R. L. **Modelling Multivariate Binary and Count Data, with Application to Infant Diarrhoea in Brazil**. Tese (Doutorado) — Lancaster University, United Kingdom, 2006.

FOSEN, J. et al. Dynamic analysis of recurrent event data using the additive hazard model. **Biometrical journal Biometrische Zeitschrift**, Wiley Online Library, v. 48, p. 381–398, 2006.

HALL, W. J.; WELLNER, J. A. Confidence bands for a survival curve from censored data. **Biometrika**, v. 67, n. 1, p. 133–143, 1980.

HOOVER. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. **Biometrika**, Biometrika Trust, v. 85, n. 4, p. 809–822, 1998.

HOUGAARD, P. **Analysis of Multivariate Survival Data**. [S.l.]: Springer-Verlag GmbH, 2000. (Statistics for Biology and Health Series). ISBN 9780387988733.

HUFFER, F. W.; MCKEAGUE, I. W. Weighted least squares estimation for aalen's additive risk model. **Journal of the American Statistical Association**, v. 86, n. 413, p. 114–129, 1991.

KESSING, L. V.; ANDERSEN, P. K. Predictive effects of previous episodes on the risk of recurrence in depressive and bipolar disorders. **Current Psychiatry Reports**, v. 7, n. 6, p. 413–420, 2005.

LIN, Y. D.; YING, Z. Semiparametric and nonparametric regression analysis of longitudinal data. **Journal of the American Statistical Association**, v. 96, p. 103–126, 2001.

MARTINUSSEN, T.; SCHEIKE, T. H. A semiparametric additive regression model for longitudinal data. **Biometrika**, v. 86, n. 3, p. 691–702, 1999.

MARTINUSSEN, T.; SCHEIKE, T. H. A non-parametric dynamic additive regression model for longitudinal data. **Annals of Statistics**, Dept. of Biostatistics, University of Copenhagen, v. 98/6, n. 4, p. 1000–1025, 2000.

MARTINUSSEN, T.; SCHEIKE, T. H. **Dynamic Regression Models for Survival Data (Statistics for Biology and Health)**. 1. ed. [S.l.]: Springer, 2006. ISBN 0387202749.

MCKEAGUE, I. **Asymptotic Theory for Weighted Least Squares Estimators in Aalen's Additive Risk Model**. FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS: Defense Technical Information Center, 1987.

MCKEAGUE, I. W.; SASIENI, P. D. A partly parametric additive risk model. **Biometrika**, v. 81, n. 3, p. 501–514, 1994.

PENA, E. A.; SLATE, E. H.; GONZALEZ, J. R. Semiparametric inference for a general class of models for recurrent events. **Journal of Statistical Planning and Inference**, v. 137, n. 6, p. 1727–1747, 2007.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2005. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

SCHEIKE, T. H. Comparison of non-parametric regression functions through their cumulatives. **Statistics & Probability Letters**, v. 46, n. 1, p. 21–32, 2000.

SCHEIKE, T. H. The additive non-parametric and semiparametric aalen model as the rate function for a counting process. **Lifetime Data Analysis**, v. 8, n. 3, p. 247–262, 2002.

SCHEIKE, T. H.; ZHANG, M. Cumulative regression function tests for longitudinal data. **Annals of Statistics**, n. 26, p. 1328–1355, 1998.

SPECKMAN, P. Kernel Smoothing in Partial Linear Models. **Journal of the Royal Statistical Society. Series B (Methodological)**, Blackwell Publishing for the Royal Statistical Society, v. 50, n. 3, p. 413–436, 1988. ISSN 00359246.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, Springer, v. 16, n. 3, p. 439–454, 1979.

?