ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Dynamic Additive Regression Models Applied to the Study of Recurrent Event Data: Childhood Diarrhoea in Salvador, Brazil

Alix Leboucq

Master Project

Carried out at Newcastle University, UK,

under the supervision of Pr. R. Henderson

February 2010
Section of Mathematics, Master Project

Contents

Introduction

In many different areas, one may be interested in the time to some events: death or cancer in medicine, wedding or divorce in sociology, etc... These data are called survival data. Several events or multiple occurences of a single event lead to event histories or recurrent event data respectively and appear just as often. Even if those data appear so frequently and that they do not seem to be so different from any other classical data, they cannot be analysed using the usual statistical tools. Indeed they have particularities that make their analysis special. First, the strong time dependence requires the use of objects such as counting processes and other stochastic processes that take time into account. Furthermore, partly also due to the dependence in time, the data are often incompletely observed and we can be in presence of censoring and truncation. Censoring occurs when the recorded time does not correspond to an event. This can happen when an individual drops out of the study before it ends or simply when the study ends itself. Truncation, on the other hand, corresponds to when the starting time is not the same for all the individuals involved in the study. Having a common starting point is impossible in many studies: consider for example survival after a myocardial infarction; it is impossible to have all patients having a myocardial infarction the same day to enroll them in the study. All the particularities described above are evidence for the fact that survival data cannot be dealt with using usual statistical tools. In particular classical linear regression cannot be used. Nevertheless, it could be of interest to study the effect of some covariates on the occurence of the event of interest. In that case several models have been proposed but we will focus on the additive regression model which was introduced by Aalen (1980). A full theory about the construction of such models as well as inference, testing and model checking will be presented in Chapter 1. We choose to consider this model because it has really nice and useful properties among which is the easy and straightforward interpretation of the effect of each single covariate on the event of interest. The use of this model will be illustrated by an analysis of recurrent event data: the Blue Bay data (Strina *et al.*, 2005), concerning the occurence of diarrhoea in young children in Salvador, Brazil. In 1997 in Salvador, a city-wide sanitation programme was started in order to reduce the risk of morbidity due to diarrhoea in children aged less than three years old. Three studies were then carried out in order to assess the effectiveness of such measures. The first one was done in 1997-1998, just before the intervention, the second one was done in 2000-2002 and finally the last one, which we will focus on, was done in 2003-2004. In this last phase, a total number of 1127 children were enrolled and each day the occurence of diarrhoea as well as other symptoms were recorded over 231 days. We will be interested in prevalence and incidence of diarrhoea. Prevalence is defined as the probability that a child has diarrhoea on a given day, whereas incidence is the probability that a child starts a new episode of diarrhoea (an episode is a sequence of days with diarrhoea until there have been at least three consecutive clear days). A complete analysis of these data will be carried out in Chapter 2. Then, as we are also interested in seeing if the measures taken against childhood diarrhoea were effective, we will compare Phase II and Phase III of the Blue Bay data in Chapter 3. Finally, we will study the effect of clustering in Chapter 4.

Throughout the report, all the computationnal results were obtained using the `R` software and writting personal code instead of using the available packages for survival analysis.

Chapter 1

Additive Regression Models

We want to study the situation where we have several individuals that may experience either a single event (such as death) or several events (such as seizures or hospitalisations) and we are interested in the time to, or between, those events. This situation arises in many fields such as medicine, biology, demography ...
In order to study this type of data, we need to introduce an object that can record the number of events experienced by an individual. The use of counting processes seems therefore to be appropriate as it gives at each time t the number of events that an individual has experienced up to time t. We denote it $N(t)$.
We first recall some basic definitions, which are generally taken from Dalang (2008), see also Steele (2003) and Kuo (2006). For a complete theory about relevant stochastic processes see Andersen *et al.* (1993).
We start by defining a filtration $(\mathcal{F}_t)$, which is a family of sub-algebras of $\mathcal{F}$, an algebra, such that for all $s \leq t$, $\mathcal{F}_s \subseteq \mathcal{F}_t$. We say that a family of random variables $(X(t))$ is adapted to $(\mathcal{F}_t)$ if for all t, $X(t)$ is $\mathcal{F}_t$-measurable. A particularly interesting class of processes are the predictable processes, $H(t)$, which are adapted to the filtration $\mathcal{F}_t$ and whose sample paths (realisations of H as functions of t) are left continuous. Finally, we define a martingale $M(t)$ as being a stochastic process with the following properties:

- $\mathbb{E}\left(M(t)\right) < \infty$,
- $(M(t))$ is adapted to $(\mathcal{F}_t)$,
- $\mathbb{E}(M(t)|\mathcal{F}_s) = M(s)$, $\forall s < t$.

Note that if in the last condition we replace the equality by an inequality, then $(M(t))$ is called a super-martingale (for $\leq$) or a submartingale (for $\geq$).

Suppose now we have a counting process $N(t)$ which is assumed to be adapted to a filtration $(\mathcal{F}_t)$. It is associated with the *intensity* $\lambda(t)$, which is a predictable process defined by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(N(t + \Delta t) - N(t) = 1 | \mathcal{F}_{t^-})}{\Delta t}.$$

Less formally, following the definition given by Aalen *et al.* (2008) we can write:

$$\lambda(t) = P(dN(t) = 1 | \mathcal{F}_{t-}),$$

where $\mathcal{F}_{t-}$ is the history up to a time just before t and $dN(t)$ is the instantaneous change at t of $N(t)$. Another way to define it is to notice that $N(t)$ is a submartingale and use the Doob-Meyer decomposition (Martinussen and Scheike, 2006). Then we can write

$$N(t) = \int_0^t \lambda(s) ds + M(t), \tag{1.1}$$

where $M(t)$ is a martingale and $\lambda(t)$ is the intensity.

We now consider n individuals, each having a counting process $N_i^\star(t)$ which we assume to be fully observed and adapted to the same filtration $(\mathcal{F}_t)$. We also have a predictable observation indicator $Y_i(t)$ for each individual, defined as:

$$Y_i(t) = \begin{cases} 1, & \text{if individual } i \text{ is at risk at time } t, \\ 0, & \text{otherwise,} \end{cases}$$

where an individual is said to be *at risk* on day t if they are not censored nor have missing information on day t. We denote the observed counting process as $N_i(t)$ and its respective intensity process as $\lambda_i(t)$ which is often assumed to have the multiplicative form

$$\lambda_i(t) = Y_i(t)\alpha(t), \tag{1.2}$$

where $\alpha(t)$ is the intensity associated with $N^\star(t)$.
We want to model the effect of p covariates $x_{i1}(t), \ldots, x_{ip}(t)$ on the intensity process. In order to do this, under the *additive model*, one can write

$$\alpha(t) = \beta_0(t) + \beta_1(t)x_{i1}(t) + \cdots + \beta_p(t)x_{ip}(t). \tag{1.3}$$

This model was first proposed by Aalen (1980). The functions $\beta_j(t)$ are called the regression functions and are to be estimated. However, their estimation cannot be obtained in a straightforward way. This motivates the introduction of the cumulative regression functions

$$B_j(t) = \int_0^t \beta_j(u) du,$$

which can be estimated consistently, whereas $\beta_j(t)$ cannot.
We recall, from (1.1) and (1.2), that

$$N(t) = \int_0^t \lambda(s)ds + M(t) \quad \text{and} \quad \lambda_i(t) = \alpha(t)Y_i(t),$$

which means we can write

$$\begin{aligned} \lambda_i(t) &= Y_i(t)\left(\beta_0(t) + \beta_1(t)x_{i1}(t) + \cdots + \beta_p(t)x_{ip}(t)\right) \quad \text{and, loosely,} \\ dN_i(t) &= Y_i(t)dB_0(t) + \sum_{k=1}^{p} Y_i(t)dB_k(t)x_{ik}(t) + dM_i(t). \end{aligned} \tag{1.4}$$

It is convenient to rewrite all this using matrix and vector notation:

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

With this notation, equations (1.4) become

$$\begin{aligned} \lambda(t) &= \mathbf{X}(t)d\mathbf{B}(t) \\ d\mathbf{N}(t) &= \underbrace{\mathbf{X}(t)d\mathbf{B}(t)}_{\text{Model}} + \underbrace{d\mathbf{M}(t)}_{\text{Noise}}, \end{aligned}$$

where the last equation has the form of the standard linear regression model. This leads naturally to least squares estimation:

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

which is well defined if $\mathbf{X}(t)$ is full rank. We introduce $J(t)$ as an indicator that $\mathbf{X}(t)$ is full rank to obtain

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

The estimator of the cumulative regression functions is, using $\mathbf{X}^-(t) = (\mathbf{X}(t)^T\mathbf{X}(t))^{-1}\mathbf{X}(t)^T$,

$$\begin{aligned} \widehat{\mathbf{B}}(t) &= \int_0^t J(u)(\mathbf{X}(u)^T\mathbf{X}(u))^{-1}\mathbf{X}(u)^T d\mathbf{N}(u) \\ &= \sum_{T_j \le t} J(T_j)\mathbf{X}^-(T_j)\Delta\mathbf{N}(T_j), \end{aligned} \tag{1.5}$$

where the T_j are the distinct event times and $\Delta\mathbf{N}(T_j)$ is a vector with zeros except for the individuals who experienced the event at time T_j. An estimator of the variance covariance matrix is

$$\begin{aligned}\widehat{\mathbf{\Sigma}}(t) &= \int_0^t J(u)\mathbf{X}^{-}(u)\text{diag}\left(d\mathbf{N}(u)\right)\mathbf{X}^{-}(u)^T \\ &= \sum_{T_j \leq t} J(T_j)\mathbf{X}^{-}(T_j)\text{diag}\left(\Delta\mathbf{N}(T_j)\right)\mathbf{X}^{-}(T_j)^T.\end{aligned} \tag{1.6}$$

An important result is

$$\sqrt{n}\left(\widehat{\mathbf{B}}(t) - \mathbf{B}(t)\right) \xrightarrow{D} U(t), \tag{1.7}$$

where $U(t)$ is a continuous mean zero p-dimensional Gaussian martingale, and $\xrightarrow{D}$ means convergence in distribution. This result will be useful for constructing confidence intervals and confidence bands as well as for testing. It will also play an essential role in the assessment of the model for the construction of martingale residuals. The proof of this theorem can be found in Andersen *et al.* (1993), Chapter VII, Section 4.2, p.575.

1.1 Confidence intervals and confidence bands

1.1.1 Confidence intervals

In order to construct confidence intervals for $B_j(t)$, we need to find the variance covariance matrix of $U(t)$, which is given in Aalen *et al.* (2008) as

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

and can be estimated by

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

We can show (Andersen *et al.*, 1993, Section VII.4.2) that

$$n\widehat{\Sigma}(t) \rightarrow A(t). \tag{1.8}$$

Indeed,

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

where $\xrightarrow{P}$ means convergence in probability and for simplicity, we have written $\mathbf{X}$ instead of $\mathbf{X}(t)$. Thus, $\sqrt{n}(\widehat{\mathbf{B}}(t) - \mathbf{B}(t))$ converges in distribution to a mean zero p-dimentional Gaussian martingale whose covariance matrix converges in probability to $\mathbf{A}(t)$. We can therefore construct a $(1-\alpha)$ confidence interval for $B_i(t)$ as:

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

where $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal law.

1.1.2 Confidence bands

Confidence intervals give for each value of t the region in which the true value is susceptible to fall in. Confidence bands, by contrast, give the region for the entire function. Thus, a confidence interval is pointwise whereas confidence bands are uniform. We will here construct confidence bands for $B_i(t)$ which will be used in the plots in the next sections. They are non constant width confidence bands, but allow use of known quantiles by using tables in Hall and Wellner (1980).

We base our approach on the following fact: suppose that $B^0(t)$ is a Brownian bridge (i.e. $B^0(t) = W(t) - tW(1)$, where $W(t)$ is a standard Brownian motion), then Doob's transformation (Hall and Wellner, 1980) gives that

$$\left\{B^0\left(\frac{t}{1+t}\right), \quad t \in [0,1]\right\} \overset{D}{=} \left\{\frac{1}{1+t}W(t), \quad t \in [0,1]\right\}. \tag{1.9}$$

We know (Aalen *et al.*, 2008, Section 2.3.1) that for a mean zero Gaussian martingale $U(t)$ with predictable variation process $V(t)$, we have

$$U(t) = W(V(t)),$$

where $W(t)$ is a standard Brownian motion.

Therefore, as $\mathbf{A}(t)$ is the predictable variation process of $U(t)$,

$$\sqrt{n}\left(\widehat{B}_j(t) - B_j(t)\right) \xrightarrow{D} U_j(t) = W(A_{jj}(t)), \tag{1.10}$$

with $W(t)$ a standard Brownian motion. Then, defining

$$G(t) = \frac{\left(\widehat{B}_j(t) - B_j(t)\right)}{\sqrt{\widehat{\Sigma}_{jj}(\tau)}} \frac{1}{1 + \frac{\widehat{\Sigma}_{jj}(t)}{\widehat{\Sigma}_{jj}(\tau)}},$$

where τ is the end of the time interval, and using (1.8) and (1.9), we obtain

$$G(t) \xrightarrow{D} B^0\left(\frac{\frac{A(t)}{A(\tau)}}{1 + \frac{A(t)}{A(\tau)}}\right).$$

Taking the supremum of the absolute value of $G(t)$ leads to

$$\begin{aligned}
\lim_{n\to\infty} P\left(\sup_{t\in[0,\tau]} |G(t)| > c_{1-\alpha}\right) &= 1-\alpha \\
&= P\left(\sup_{t\in[0,\tau]} \left| B^0\left(\frac{\frac{A(t)}{A(\tau)}}{1+\frac{A(t)}{A(\tau)}}\right)\right| > c_{1-\alpha}\right) \\
&= P\left(\sup_{t\in[0,1]} \left| B^0\left(\frac{t}{1+t}\right)\right| > c_{1-\alpha}\right) \\
&= P\left(\sup_{t\in[0,1/2]} |B^0(t)| > c_{1-\alpha}\right).
\end{aligned}$$

Tables of values of $c_{1-\alpha}$ can be found in Hall and Wellner (1980), where also a formula for the density function of $\sup_{t\in[0,\frac{1}{2}]}|B^0(t)|$ is given:

$$\begin{aligned}
P\left(\sup_{t\in[0,\frac{1}{2}]} |B^0(t)| \le \lambda\right) =& 1-2\Phi\left(\frac{\lambda}{\sqrt{a(1-a)}}\right) \\
&+2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\lambda^2}\left(\Phi(r(2k-d))-\Phi(r(2k+d))\right),
\end{aligned}$$

where Φ is the standard normal distribution function, $r=\lambda\sqrt{\frac{1-a}{a}}$ and $d=\frac{1}{1-a}$.

Thus, using

$$1-\alpha = P\left(\sup_{t\in[0,\tau]}\left|\frac{\sqrt{n}(\widehat{B}_i(t)-B_i(t))}{\sqrt{\widehat{\Sigma}_{ii}(\tau)}}\frac{1}{1+\frac{\widehat{\Sigma}_{ii}(t)}{\widehat{\Sigma}_{ii}(\tau)}}\right| > c_{1-\alpha}\right).$$

we obtain the confidence bands:

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

with asymptotic coverage $1-\alpha$.

1.2 Martingale residuals and model checking

An important step after the construction of a model is to check whether it is appropriate or not, i.e. whether it fits the data well. A useful tool to look at, in that case, are the martingale residuals processes. They are defined for each individual i in Aalen *et al.* (2008) as

$$\widehat{M}_i(t) = N_i(t) - \widehat{\Lambda}_i(t). \tag{1.11}$$

By writing it a little differently, we can show it is a martingale:

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

Therefore, $\widehat{\mathbf{M}}(t)$ is a martingale since it is the integral of a predictable process with respect to a martingale (Aalen *et al.*, 2008, Section 2.2.2). We can also give an estimator of the variance covariance matrix of $\widehat{\mathbf{M}}(t)$:

$$\widehat{\mathbf{\Omega}}(t) = \widehat{\mathrm{var}\left(\widehat{\mathbf{M}}(t)\right)} = \int_0^t (\mathbf{I} - \mathbf{H}(s))\mathrm{diag}(\lambda(s)ds)(\mathbf{I} - \mathbf{H}(s))^T ds. \tag{1.12}$$

We can now construct the standardized residuals

$$\mathbf{M}_i^*(t) = \frac{\widehat{\mathbf{M}}_i(t)}{\sqrt{\widehat{\mathbf{\Omega}}_{ii}(t)}}, \text{ for } i = 1, \dots p.$$

Then, if the model is correctly specified, the variance of the above standardized residuals should be close to one at all times t, i.e. $\mathrm{var}(\mathbf{M}^*(t)) = 1$. Therefore, in order to check the fit of the model, one can plot the standard deviation of the observed standardized residuals and see if it is close to one for all t.

1.3 Tests

Result (1.7) is not only useful for the construction of confidence intervals and confidence bands as we saw in Section 1.1, but also for testing.

Here we will focus on three tests. First in testing if a regression function is equal to a given function and second to test for constancy, which aims to see if a regression function has a constant effect over time. Finally, we will introduce a version of the log-rank test for testing the equality of several functions. More formally, we want to test either:

$$\begin{aligned}
&(1) H_0 : B_j(t) = B_j^0(t) \quad \forall t \in [0,\tau] \quad \text{vs} \quad H_1 : B_j(t) \neq B_j^0(t), \text{ for significance}\\
&(2) H_0 : B_j(t) = \gamma t \quad \forall t \in [0,\tau] \quad \text{vs} \quad H_1 : B_j(t) \neq \gamma t, \text{ for constancy.}\\
&(3) H_0 : \alpha_1(t) = \cdots = \alpha_K(t) \quad \forall t \in [0,\tau]\\
&\text{vs } H_1 : \text{There is at least one difference for some } t, \text{for the log-rank test.}
\end{aligned}$$

Here $B_j^0(t)$ is a given function and γ is a parameter to be estimated. We usually test on $[0,\tau]$, where τ is the end of the study, so we consider the functions on the whole study interval, but any smaller time interval $[0,t_0]$, with $t_0 \in [0,\tau]$ can be considered.

1.3.1 Significance test

We focus on the first test

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

We recall from (1.10) that

$$\sqrt{n}\left(\widehat{B}_j(t) - B_j(t)\right) \xrightarrow{D} W(A_{jj}(t)).$$

Then, by the continuous mapping theorem (Panaretos (2008) and Knight (2000)),

$$\sup_{t\in[0,\tau]} \sqrt{n}\left|\widehat{B}_i(t) - B_i(t)\right| \xrightarrow{D} \sup_{t\in[0,\tau]} U_i(t).$$

By Slutsky's theorem (Panaretos (2008) and Knight (2000)),

$$\begin{aligned}
\sup_{t\in[0,\tau]} \frac{\sqrt{n}\left|\widehat{B}_i(t) - B_i(t)\right|}{\sqrt{n\widehat{\Sigma}_{ii}(\tau)}} \xrightarrow{D} &\sup_{t\in[0,\tau]} \frac{|U_i(t)|}{A_{ii}(\tau)}\\
= &\sup_{t\in[0,\tau]} \left|\frac{W(A_{ii}(t))}{A_{ii}(\tau)}\right|.
\end{aligned}$$

By the scaling property of Brownian motion (Dalang, 2008),

$$\begin{aligned}
&= \sup_{t\in[0,\tau]} \left|W\left(\frac{A_{ii}(t)}{A_{ii}(\tau)}\right)\right|\\
&= \sup_{t\in[0,1]} |W(t)|,
\end{aligned}$$

where the last equality is due to the fact that $A_{ii}(t)$ is a continuous nondecreasing mapping of $[0,\tau]$ on $[0, A_{ii}(\tau)]$.

Thus, if we define T as

$$T = \sup_{t \in [0,\tau]} \frac{\sqrt{n}\left|\widehat{B}_j(t) - B_j(t)\right|}{\sqrt{n\widehat{\Sigma}_{jj}(\tau)}},$$

then

$$T \xrightarrow{D} \sup_{t\in[0,1]} |W(t)| .$$

As $\sup_{t\in[0,1]} |W(t)|$ has a known distribution (Billingsley (1968)), given by

$$P\left(\sup_{t\in[0,1]} |W(t)| \leq b\right) = \sum_{k=-\infty}^{\infty} (-1)^k \Phi((2k+1)b) - \Phi((2k-1)b), \qquad (1.13)$$

where Φ is the standard normal distribution function, we can obtain the quantile $c_{1-\alpha}$, where $1-\alpha$ is the significance level of the test.
Therefore, we will reject H_0 if $T > c_{1-\alpha}$. The p-value of the test can also be computed using the formula of the density function of $\sup_{t\in[0,1]} |W(t)|$.
If we want to test that the regression function is null, we can simply take $B_j^0(t) = 0$.
Notice that the fact that $\sqrt{n}\left(\widehat{B}_j(t) - B_j(t)\right)$ is a martingale was essential in the construction of this test and made it quite simple.

We can also apply another method for testing. We consider a nonnegative predictable weight process $L_j(t)$ which is supposed to be null whenever $J(t) = 0$.
A good statistic for testing

$$H_0 : \beta_j(t) = 0 \quad \forall t \in [0,\tau]$$

is

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

This statistic is good for testing H_0 against alternatives of the form $\beta_j(t) < 0$ or $\beta_j(t) > 0$ for all t. It will be more difficult to detect crossing effects using this test statistic. In the particular case of crossing effect, it would be better to use the first test. However, when working with real data later in this work, we will use the second test for significance.
Under the null hypothesis, $Z_j(\tau)$ is asymptotically a mean zero Gaussian variable with variance which can be estimated by

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

Therefore, by the central limit theorem, we obtain that, if H_0 is true,

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

Then, the null hypothesis is rejected for significantly large values of this statistic. This last significance test will be applied when working with real data, later in this report, using

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

as suggested by Aalen *et al.* (2008) (in Section 4.2.1) or by Elgmati (2009). This choice of weight process is directly inspired by ordinary least squares regression where the variances of the estimators are proportional to $(\mathbf{X}^T\mathbf{X})^{-1}$, $\mathbf{X}$ being the design matrix. Then the test statistics become simply a weighted sum of the cumulative regression functions.

1.3.2 Constancy test

We now want to test

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

Under the null hypothesis, we must estimate γ. We simply take

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

The first method used in the previous section for testing for significance is not applicable here because $H = \sqrt{n}\left(\widehat{B}_j(t) - \frac{t}{\tau}\widehat{B}_j(\tau)\right)$ is not a martingale. Indeed, $\widehat{B}_j(\tau)$ depends on the future which is not compatible with the definition of a martingale.

One method, which was introduced by Martinussen and Scheike (2006), and which is called a conditional multiplier procedure, can be used. It is based on a resampling approach. It was also previously used in Borgan *et al.* (2007), Elgmati (2009) and Elgmati *et al.* (2008).

First of all, we need to show that

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

This follows from

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

Thus,

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

where $\mathbf{X}_i(s)$ is the vector of covariates for the individual i.
Now let

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

Then,

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

An estimator $\widehat{\epsilon_i}(t)$ of $\epsilon_i(t)$ is

$$\widehat{\epsilon_i}(t)=\int_0^t\left(\frac{1}{n}\mathbf{X}^T(s)\mathbf{X}(s)\right)^{-1}\mathbf{X}_i(s)d\widehat{M_i}(s),$$

where $\widehat{M_i}(t)$ is defined by (1.11).
What we would like is writting $\sqrt{n}\left(\widehat{\mathbf{B}}(t)-\mathbf{B}(t)\right)$ as a sum of iid terms, which is not possible. However, because correlation between the $d\widehat{M_i}$ is of order $1/n$, $\sqrt{n}\left(\widehat{\mathbf{B}}(t)-\mathbf{B}(t)\right)$ behaves like a sum of iid terms.
The constancy test we are going to use is based on the following theorem:

Theorem 1.3.1. *Let $Z_1,\dots,Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Under some technical conditions, it follows that $\sqrt{n}\left(\widehat{\mathbf{B}}(t)-\mathbf{B}(t)\right)$ has the same limit distribution as*

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

The proof of this theorem as well as the conditions under which the theorem is valid can be found in Martinussen and Scheike (2006) (Section 5.2). As the martingale property cannot be used for $H = \sqrt{n}\left(\widehat{B}_j(t) - \frac{t}{\tau}\widehat{B}_j(\tau)\right)$, it is not possible to estimate its variance and therefore its distribution is hard to evaluate. Theorem (1.3.1) gives us a simple way to estimate the distribution of H by simulating Z_i as we will see.

We must now construct test statistics for testing the null hypothesis. We present two here:

$$\begin{aligned} T_{1,const} &= \sqrt{n} \sup_{t\in[0,\tau]} \left| \widehat{B}_j(t) - \frac{t}{\tau}\widehat{B}_j(\tau) \right|. \\ T_{2,const} &= n \int_0^\tau \left(\widehat{B}_j(t) - \frac{t}{\tau}\widehat{B}_j(\tau) \right)^2. \end{aligned} \tag{1.14}$$

The first looks at the largest difference between $\widehat{B}_j(t)$ and a straight line, whereas the second accumulates the squared differences. Now, we can notice that under the null hypothesis,

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

Therefore, by Theorem (1.3.1), $\sqrt{n}\left(\widehat{B}_j(t) - \frac{t}{\tau}\widehat{B}_j(\tau)\right)$ has, under the null hypothesis, the same asymptotic distribution as $\Delta_j(t) - \frac{t}{\tau}\Delta_j(\tau)$, where $\Delta(t)$ is defined in Theorem (1.3.1). Thus, when wanting to test our hypothesis $H_0 : B_j(t) = \gamma t$, we will adopt the following algorithm:
For $r = 1, \ldots, R$:

1. Compute $\Delta^r(t) = n^{-1/2}\sum_{i=1}^n \widehat{\epsilon}_i(t) Z_i$, where $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
2. Compute either $T^r_{sim} = \sup_{t\in[0,\tau]} |\Delta^r(t) - (t/\tau)\Delta^r(\tau)|$
   or $T^r_{sim} = \int_0^\tau \left(\Delta^r(t) - (t/\tau)\Delta^r(\tau)\right)^2 dt$.
3. We now have R replications $T^1_{sim}, \ldots, T^R_{sim}$.
4. Compute an estimate $\widehat{p}$ of the p-value as

$$\widehat{p} = \frac{\sharp\{T^r_{sim} \geq T_{obs}\}}{R}, \quad r = 1, \ldots, R,$$

where T_{obs} is defined by one of the two statistics in (1.14).
In the previous algorithm we used $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ but notice that we can also use

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

and obtain the same results.

1.3.3 Simulation study

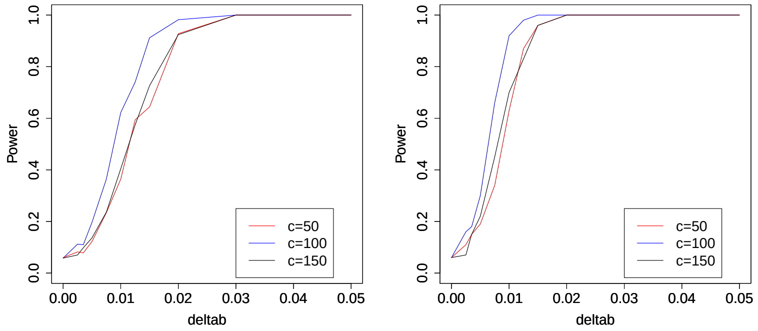
In order to see if the previous test gives satisfactory results, i.e. if the test size is correct and the power is high, we carry out a simulation study. We consider n individuals followed to time τ. The additive model we consider is

$$\alpha(t|x) = \beta_0 + \beta_1 x(t) + \Delta\beta x(t) I_{t \geq c},$$

where c is a changepoint and $\Delta\beta$ is the amplitude of the change in the slope. We consider $R = 100$ replications for the conditional multipliers and we choose the test statistic $T_{1,const}$. Note that if we want a constant slope for the regression function β_1 we simply set $\Delta\beta = 0$.
Several simulations were carried out in order to see the influence of c, $\Delta\beta$ and n on the power of the test. We recall that the power of a test is the proportion of rejected tests. In order for the test to be effective, we should obtain 5% rejections when $\Delta\beta = 0$ (constant slope) and we expect to have a high power when the slope changes ($\Delta\beta \neq 0$).
We base the estimation of the power on a sample of 500 replications of the test when $n = 500$ and 100 replications when $n = 1000$. This difference in the number of replications used is due to the time taken by the simulations which is considerably longer when we pass from 500 to 1000 individuals.
The results obtained are presented in Figures 1.1 and 1.2.

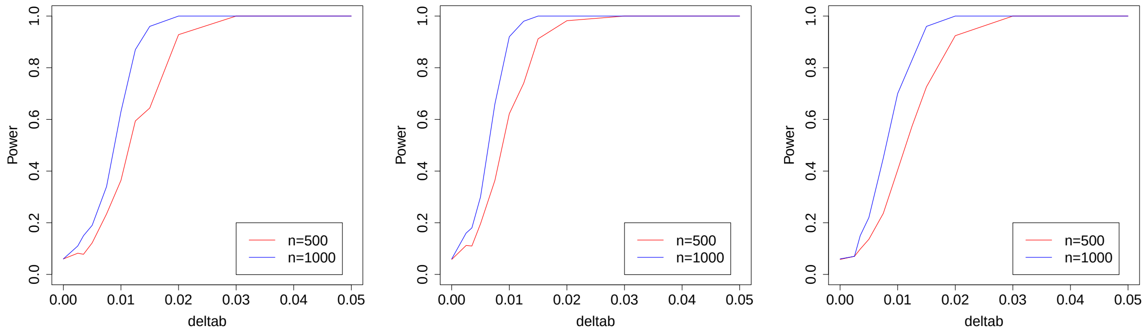
The test size is the proportion of rejection under the null $\Delta\beta = 0$. We use a nominal 5% test and found that for all combinations, the empirical rejection rates are within simulation noise of the expected value. Turning to power, we notice that it is improved when n or $\Delta\beta$ is larger as expected. Indeed, by increasing the size of n or $\Delta\beta$, the difference between the cumulative regression function and the straight line becomes clearer and it is therefore more easily detectable using $T_{1,const}$. The test detects differences more easily if the change of the slope happens near the center at $\tau/2$ instead of near the extremes. This again is what we were expecting as if the slope changes close to the extremes, the difference between the cumulative regression function and the straight line will appear less neatly compared to when the slope change in the middle, as illustrated in Figure 1.3.

Figure 1.1: Estimated power of the constancy test for different values of $\Delta\beta$, based on 500 simulations for $n = 500$ and 100 simulations when $n = 1000$.



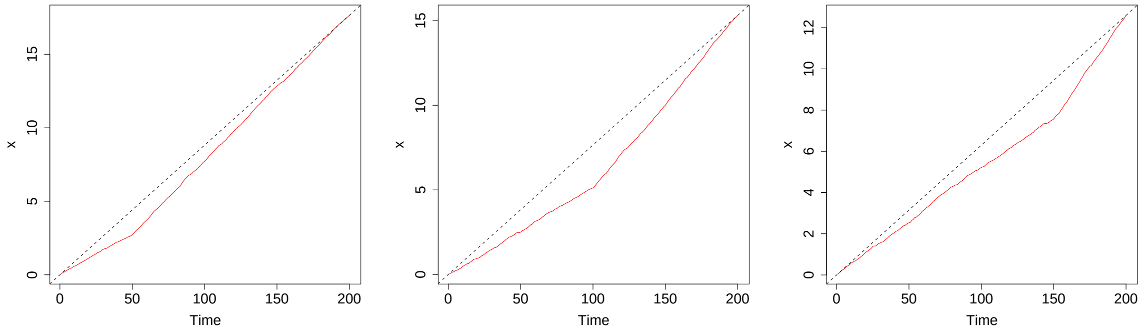
(a) n=500 (b) n=1000

Figure 1.2: Estimated power of the constancy test for different values of the changepoint c, based on 500 simulations when $n = 500$ and 100 simulations when $n = 1000$.



(a) c=50 (b) c=100 (c) c=150

Figure 1.3: Example of simulated cumulative functions for different values of the changepoint.



(a) c=50

(b) c=100

(c) c=150

When $\Delta\beta \neq 0$, the power reaches 1 from $\Delta\beta = 0.03$ for $n = 500$ and from $\Delta\beta = 0.02$ when $n = 1000$, which means that the test can detect small changes in the slope quite easily with a really good power.

1.3.4 Log-rank test

The previous tests applied to only one function. However, sometimes it can be of interest to compare several functions. This situation can occur if we want to compare the effect of a variable, or several, on the survival function between different groups of individuals, for example. In other words, we are interested in testing if there is a difference in the intensity between groups. In order to do that, we will introduce a test which is presented in Andersen *et al.* (1993) (Section V.2.1) and in Martinussen and Scheike (2006) (Section 4.2.1) and is of log-rank form.

Suppose we have K groups and denote by α_k the intensity function for the kth group of individuals, $k = 1, \dots, K$. We want to test

$$\begin{aligned} H_0 &: \alpha_1(t) = \cdots = \alpha_K(t) \text{ for all } t \in [0, \tau], \\ \text{vs } H_1 &: \text{there is at least one difference for some } t. \end{aligned}$$

Suppose there are n_k individuals in group k and we denote by N_{ik} the counting process of individual i in group k and by Y_{ik} the "at risk" indicator, $i = 1, \dots, n_k$, $k = 1, \dots, K$.

A test statistic for each group k is

$$Z_k(t) = \int_0^t w(s) \left(dN_k(s) - \frac{Y_k(s)}{Y_\bullet(s)} dN_\bullet(s) \right),$$

where $w(s)$ is a weight function taken as $w(t) = I_{Y_\bullet(t)>0}$ in the log-rank test, $N_k(t) = \sum_{i=1}^{n_k} N_{ik}$ and $N_\bullet(t) = \sum_{k=1}^{K} N_k(t)$. The test effectively contrasts the number of events in group k at each time s with the expected number under the null, conditional on the total number of events at s in the combined sample.
We can notice that

$$\sum_{k=1}^{K} Z_k(t) = 0.$$

By denoting $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_{K-1}(t))$, it can be shown (Martinussen and Scheike (2006)) that $\sqrt{n}\mathbf{Z}(t)$ is asymptotically normally distributed around zero with variance covariance matrix $\mathbf{\Gamma}(t)$ which may be estimated by

$$\widehat{\mathbf{\Gamma}}_{kl} = \int_0^t w(s)^2 \frac{Y_k(s)}{Y_\bullet(s)} \left(\delta_{kl} - \frac{Y_l(s)}{Y_\bullet(s)} \right) dN_\bullet(s), \quad k, l = 1, \ldots, K-1,$$

where, $Y_\bullet(t) = \sum_{k=1}^{K} Y_k(t)$ and $\delta_{kl} = 1$ if $k = l$ and zero otherwise.
Thus the test statistic we will be using is

$$Q(t) = \mathbf{Z}(t)^T \widehat{\mathbf{\Gamma}}^{-1}(t) \mathbf{Z}(t),$$

which under the null hypothesis, follows a χ^2_{K-1} distribution.

1.3.5 Simulation study

In order to assess the quality of the test, we conduct a simulation study. We consider 1006 individuals followed up to time $\tau = 200$. Those numbers were chosen based on the data to be studied in the next chapter. We choose an additive model with constant intensity $\alpha = 0.02$ and then we cluster the individuals in 24 different groups. As the intensity is the same for all groups, the log-rank test should not reject the null hypothesis. On 1000 simulations, the test rejected 32 times giving an approximate power of $\widehat{p} = 3.2\%$ which is within simulation noise of the expected value of 5%.

1.4 Dynamic models

In some situations, we may be interested in seeing how the past influences the present and the future or simply to consider covariates that vary with time. For example, we may be interested in seeing the behaviour of a patient's illness as he grows up (age varies with time). It is particularly of interest in the case of recurrent event data to consider how the previous events can influence the upcoming events. This must be dealt with using particular models. Two solutions are usually proposed in the literature: frailty which introduces a random component in the model different for each individual and which often follows a gamma distribution (Aalen *et al.*, 2008) and dynamic models which introduce the number of previous events for an individual as a covariate for predicting future events (Fosen *et al.* (2006) and Aalen *et al.* (2008)). In this section we will focus on the second solution constructing a dynamic additive regression model which is a simple additive regression model but where we allow the covariates to be time varying and depend upon the individual past (Aalen *et al.*, 2008). We refer to functions of the individual-specific histories as dynamic covariates (they are also often called endogeneous or internal covariates) whereas the other covariates are called fixed covariates. Later in this work, we will be considering variables that depend upon the individual past events and that may therefore be defined as (Borgan *et al.*, 2007)

$$x(t) = \frac{N(t^-)}{t},$$

to include the previous events in the model.

However, the use of dynamic covariates brings several problems. First, as dynamic covariates depend on the past, the model cannot be fitted from the beginning, as we will not be able to construct them. The solution is therefore to start the estimation after a few events happened. Another problem is that the fixed covariates and the dynamic ones are not independent (Elgmati, 2009). The influence of the fixed covariates on the event of interest may be hidden by the dynamic ones as illustrated in Figure 1.4 (left). We say that the dynamic covariates lie in the causal pathway between the fixed covariates and the response. Therefore the estimates for the effects of the fixed covariates may be biased. This will be illustrated in Section 2.3.1 when considering the Blue Bay data to be described in Section 2.1. We will here illustrate a solution to this problem by using path analysis which models the relations between the different variables of a model, as suggested in Fosen *et al.* (2006). Here, we will denote by Z_i the fixed covariates for individual i and by $D_i(t)$ the dynamic ones. In the simple additive regression model, we have the following marginal model:

$$dN_i(t) = Y_i(t)\Big(\beta_0(t) + \beta_1(t)Z_i\Big) + dM_i(t).$$

We can then simply introduce the dynamic covariates in a naive way

$$dN_i(t) = Y_i(t)\Big(\gamma_0(t) + \gamma_1(t)Z_i + \gamma_2 D_i(t)\Big) + dM_i(t).$$

The problem with this naive model, as it was explained previously, is that the estimate $\widehat{\gamma}_1(t)$ in the naive model will be underestimated compared to the value of $\widehat{\beta}_1(t)$ in the marginal model because of the introduction of the dynamic covariates. Indeed some of the effect of the fixed covariates will be accounted to the dynamic covariates leading to an underestimation of $\widehat{\gamma}_1(t)$.

The solution to this problem is inspired by linear regression, where it is well known that the estimated regression coefficients remain unchanged when adding an orthogonal covariate to the model. Therefore, when wanting to preserve the effect of the fixed covariates when adding the dynamic covariates, one can simply add an orthogonal covariate. Instead of using the dynamic covariates, one can use the residuals of a linear model that regresses the dynamic covariates on the other covariates (Fosen *et al.*, 2006). In other words, we are trying to explain as much as possible of the dynamic covariate by the past. The residual is what is left over and we investigate how this affects the current events. The key is that the residuals are orthogonal to covariates under least squares estimation of a linear model. Therefore, when fitting the additive regression model, the effect of the fixed covariates will not be affected by the effect of the residuals, as illustrated in Figure 1.4 (right). In other words, we fit the linear model

$$D(t) = \psi(t)Z + \epsilon(t),$$

obtain the residuals, $R(t)$ which are orthogonal to Z,

$$R(t) = D(t) - \widehat{\psi}(t)Z.$$

The dynamic covariates can therefore be expressed as

$$D(t) = R(t) + \widehat{\psi}(t)Z.$$

To obtain the final model, we simply replace the dynamic covariates by the residuals

$$dN_i(t) = Y_i(t)\Big(\beta_0(t) + \beta_1(t)Z_i + \beta_2(t)R_i(t)\Big) + dM_i(t).$$

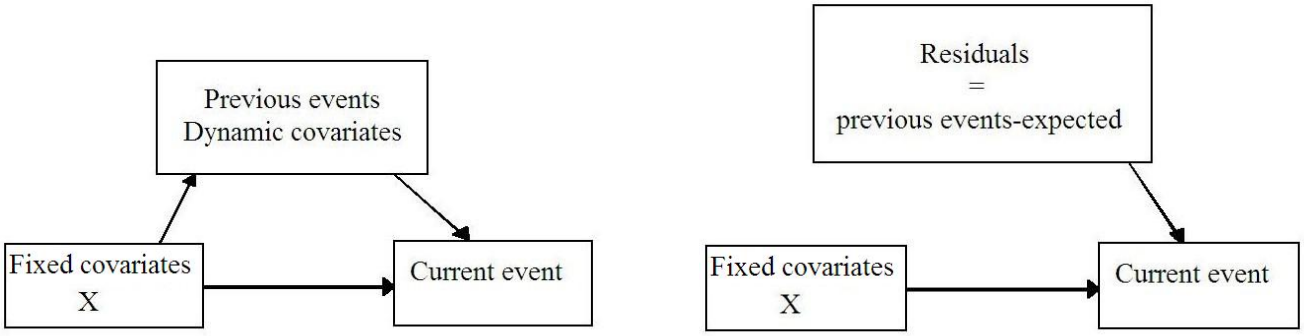
We can also compute the quantity of which we underestimate the effect of the fixed covariates in the naive model by noticing that

$$D_i(t) = R_i(t) + \widehat{\psi}(t)Z_i$$

and replacing $D_i(t)$ by the above formula in the naive model, we get

$$\begin{aligned} dN_i(t) &= Y_i(t)\Big(\gamma_0(t) + \gamma_1(t)Z_i + \gamma_2(t)\big(R_i(t) + \widehat{\psi}(t)Z_i\big)\Big) + dM_i(t) \\ &= Y_i(t)\Big(\gamma_0(t) + \big(\underbrace{\gamma_1(t) + \gamma_2(t)\widehat{\psi}(t)}_{=\beta_1(t)}\big)Z_i + \gamma_2(t)R_i(t)\Big) + dM_i(t). \end{aligned}$$

Figure 1.4: Path diagram of the dynamic model.



Thus in the naive model, the effect of the fixed covariates is underestimated by $\gamma_2(t)\widehat{\psi}(t)$.

1.5 Discrete time

In the previous sections, the theory was presented for the continuous time case. However, when working with real data in the remainder of this work, we will have to consider discrete time. All the previous results still apply in the case of discrete time and some formulas become even simpler. All the integrals over time can be replaced by sums. Only the formulas for variance estimation need adjusting. For example, we can write formula (1.12) as

$$\widehat{\boldsymbol{\Omega}}(t) = \sum_{s=0}^{t} (\mathbf{I} - \mathbf{H}(s)) \mathrm{diag}\left(\lambda(s)(1-\lambda(s))\right) (\mathbf{I} - \mathbf{H}(s))^T.$$

Indeed, in discrete time, ΔN is a binomial and therefore its variance may be estimated by $\lambda(1-\lambda)$. This is properly taken into account in our code for analyses to come.

Chapter 2

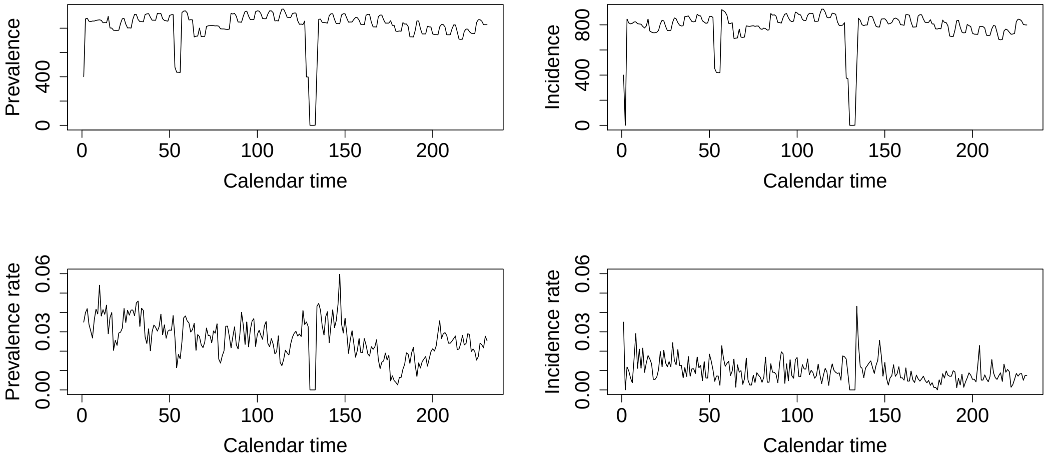
Data

2.1 Exploratory Analysis

According to the World Health Organization (WHO, 2009), diarrhoea occurs world-wide and causes 4% of all deaths in the world. Mostly children in developing countries are concerned. It has been long known that adequate water supplies, promotion of sanitation plus hygiene are important in prevention of diarrhoea and related problems. In this regard, a city-wide sanitation programme was started in Salvador, Brazil, in 1997, in order to improve sewerage coverage and thus, reduce diarrhoea morbidity in children less than 3 years old (Barreto *et al.*, 2007). The Institute of Public Health of the Federal University of Bahia undertook three studies in 1997-1998, 2000-2002 and 2003-2004, together called Blue Bay, which meant to evaluate the impact of the measures. The first phase was done before the intervention and the second phase was studied by Borgan *et al.* (2007). We will focus on the third study which started on the 13th October 2003 and ended on the 30th May 2004, for a maximum follow up of 231 days. A total of 1127 children aged between 0 and 36 months were followed each day and days with diarrhoea were registered. In this paper, we will concentrate on the 1006 children who had at least 90 days of follow up as in the previous study (Borgan *et al.*, 2007). The children came from 24 different districts. The events of interest will be prevalence and incidence of diarrhoea. Prevalence is the probability that a child has diarrhoea on a given day whereas incidence is the probability that a child starts a new episode of diarrhoea, where an episode is a sequence of days with diarrhoea until there have been at least three consecutive clear days.

In Figure 2.1, we first show the number of children at risk for both incidence and prevalence (upper plots) along with the daily prevalence and incidence rate through the study period (lower plots). The rates are computed as the number of children experiencing the event of interest divided by the total number of children at risk that day. We can notice that prevalence starts around 4% at the beginning

Figure 2.1: Empirical prevalence and incidence of diarrhoea. Top line: Number of children at risk for prevalence (left) and incidence (right). Bottom line: daily prevalence (left) and incidence (right).

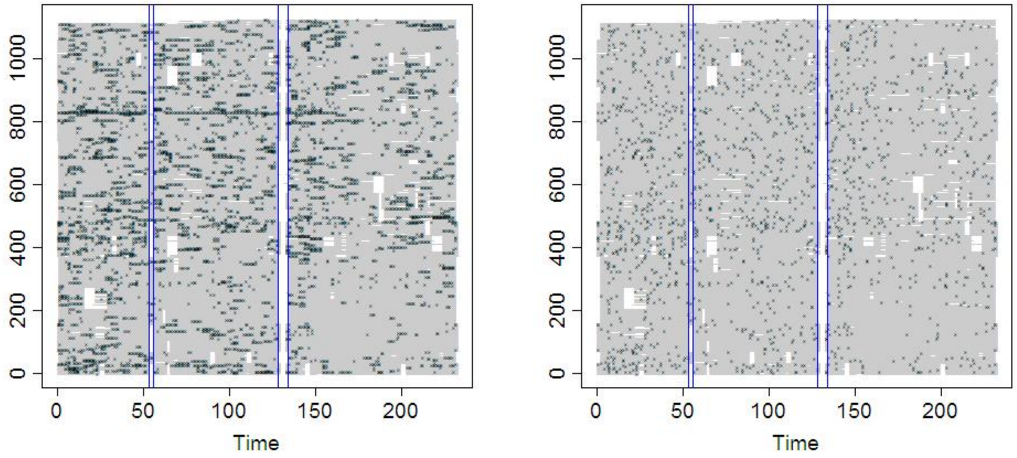


of the study and finishes at approximatively 3% at the end, whereas the rate of incidence is much lower (around 1%). We can also notice two periods were the number of children at risk for prevalence and incidence is really low. We will see in the next figure that this is due to a large amount of missingness.

In Figure 2.2, we present the data for each child and for all the study period, for both prevalence and incidence. Grey lines indicate that the child is at risk, blank spaces correspond to missing information, crosses correspond to a day with diarrhoea in the left plot and the beginning of a new period of diarrhoea in the right one. Blue lines highlight periods when at least half of the children have missing information.

The analysis of these data is complicated by the fact that we do not have complete information for all children for all days. First of all, some of the children entered late (4.1% of the children) and some of them dropped out of the study before the final ending date (about 9.2%). Moreover, there are two periods when data are missing for at least half of the children involved in the experiment. On the one hand, data are missing for days 53 to 56. These correspond to dates between the 4th and the 7th December, which was the Festa de Santa Barbara, holidays in Salvador. On the other hand, there are missing data on days 128 to 134 which correspond to dates between 17th and 23rd February 2004, which was

Figure 2.2: All data for prevalence (left) and incidence (right), the blue lines highlight periods when more than half the children are missing and the crosses represent events of interest.



the Carnival for the city. These two periods contain really few observations and therefore we will not consider them in our analysis.

In Figure 2.3, we present two samples of 20 individuals. For each child, a gray line represents the days when data were collected while blank spaces stand for missing information. The red points show when the child entered the study and the green points represent the leaving date. We represent a day with diarrhoea and start of an episode by × and | respectively. The length of the follow up varies considerably with children entering late in the study (individual 14 in the right plot for instance) or on the contrary dropping out of the study (see for example individual 19 on the left plot). Also, the number of events per child is very variable with some children experiencing none or very few events and some having plenty. This last fact can be observed in Table 2.1 and Figure 2.4, where summary and plots of the number of events per child and per day are presented, an event being a day with diarrhoea. In particular, child 748 experienced a particularly large number of events (146 events in 231 days).

For each child in the study, information was collected concerning basic neighbourhood and household sanitation conditions. These covariates are presented below.

- `MA` : micro areas, 24 different. They are the districts of the child's residence.

Figure 2.3: Two examples of a sample of 20 individuals. The gray lines represents the days where the children were at risk, red and green circles are for the beginning and the end of the study respectively, crosses are days with diarrhoea and bars are the start of an episode.

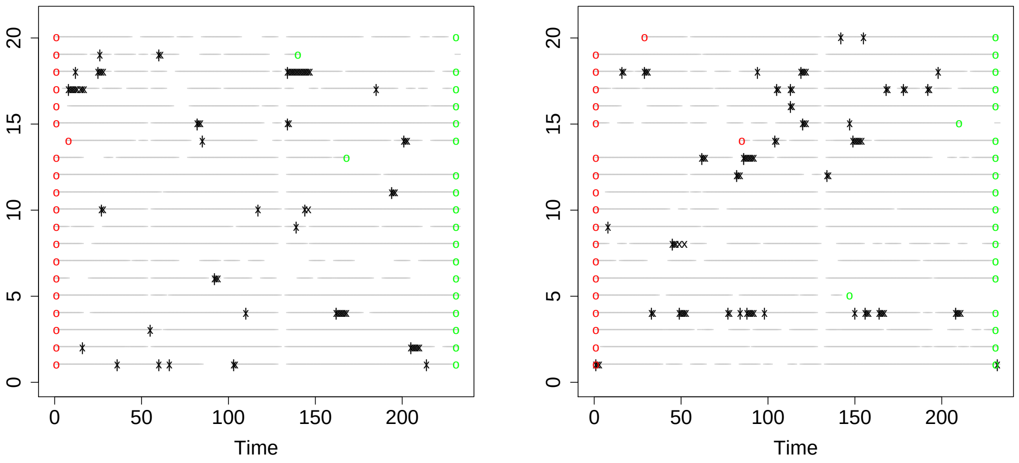


Table 2.1: Summary of the number of events per child and per day.

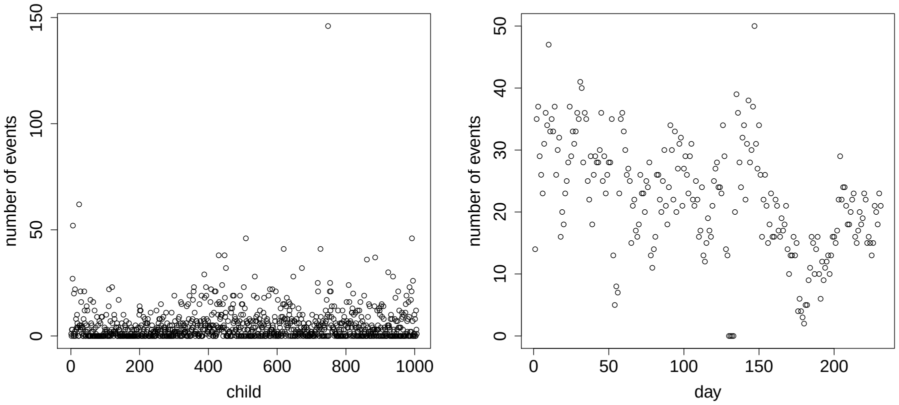
(a) number of events per child

Min	Median	Mean	Max
0	3	5.8	146

(b) number of events per day

Min	Median	Mean	Max
0	22	22.12	50

Figure 2.4: Plot of the number of events per child and per day respectively.



- `age` : age of the child in months.
- `num5y` : number of children under 5 years old in the house (0: one child or less, 1: at least 2 children).
- `motherage` : mother's age.
- `mothercatage` : mother's age categorized (0: 25 years old or more, 1: less than 25 years old).
- `streetqual` : quality of street (0: good, 1: bad).
- `habqual` : habitation type (0: good, 1: bad).
- `dens` : number of people per bedroom (0: one person per bedroom, 1: two people per bedroom, 2: more than 2 people per bedroom).
- `water_origin` : where water comes from (0: goverment link, 1: other).
- `resagcat` : type of drinking water reserve (0: good, 1: bad).
- `waterqual` : quality of drinking water (0: good, 1: bad).
- `toilets` : presence of toilet (0: inside the house, 1: not inside the house).
- `exc_disp` : excretal disposal (0: appropriate. 1: not appropriate).
- `dirtrivers`: presence of small rivers with dirty water (0: No, 1: yes).

Table 2.2: Summary of the covariates.

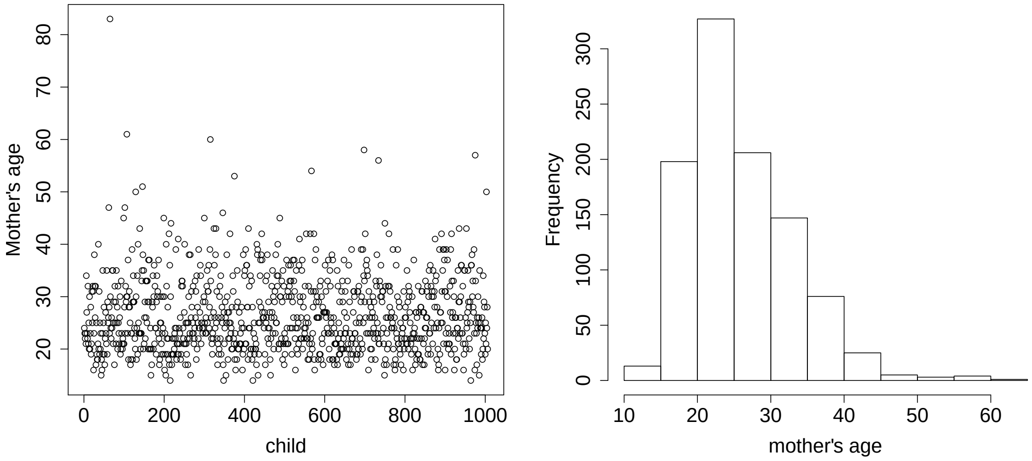
covariate	value	percentage
age (months)	≤ 12	37%
	> 12 and < 24	35%
	≥ 24	28%
num5y	0	92%
	1	8%
motherage	0	52%
	1	48%
streetqual	0	61%
	1	39%
habqual	0	98%
	1	2%
water_origin	0	88%
	1	12%
dirtrivers	0	79%
	1	21%
exc_disp	0	91%
	1	9%

covariate	value	percentage
toilets	0	87%
	1	13%
dens	0	34%
	1	44%
	2	22%
garbage	0	97%
	1	3%
flooding	0	70%
	1	30%
mother_education	0	25%
	1	61%
	2	14%
sex	0	48%
	1	52%
waterqual	0	85%
	1	15%

- `garbage` : destination of the garbage of the house (0: appropriate, 1: non appropriate).
- `flooding` : flooding in the house during rain (0: no, 1: yes).
- `mother_education` : mother's education categorized (from 0 to 2).
- `sex` : sex of the child (0: female, 1: male).

The Micro Areas will not be considered first. We will devote Chapter 4 to the study of this variable. We found that 316 values of the variable `resagcat` were missing which represents around 31% of missingness. Due to this, we choose not to consider that variable. Moreover, the variables `waterqual` and `exc_disp` present a particularity. These variables also have missing values but only rather few (14 and 38 out of 1006 are missing respectively). Thus, we choose to impute the value of the low risk category instead of missing value, which will make the analysis conservative (i.e in these cases, we will replace the missing values by a zero).

Figure 2.5: Histogram and plot of the mother's age.



In Table 2.2, a summary of the covariates is presented. All these covariates are considered as binary or categorized. The variable `age` will be considered as a time varying covariate whereas all the other covariates are considered fixed.
In Figure 2.5, plots of the age of the mother are presented. In the first plot, we can see the mother's age of each child. The mothers are mainly aged between 15 and 40 years old. We can notice that child 65 has an 83 years old mother. This suggests that either it is not his mother but the person who looks after him (such as his grandmother) or that there is an error in the data collection. However, we can also remark that some of the other mothers are aged between 50 and 70 years old, which suggest we should prefer the first hypothesis. The second picture is an histogram showing the frequency of the children given the age of their mother. Notice that in the histogram, we deleted the mother of 83 years in order to make it clearer.

2.2 Additive regression model without dynamic covariates

As described in Section 1, we will fit an additive regression model to the data presented in the previous section using all the covariates introduced previously except the micro areas, which will be studied later. First, we do not consider dynamic covariates.

Table 2.3: Significance and constancy test for the prevalence model: all covariates.

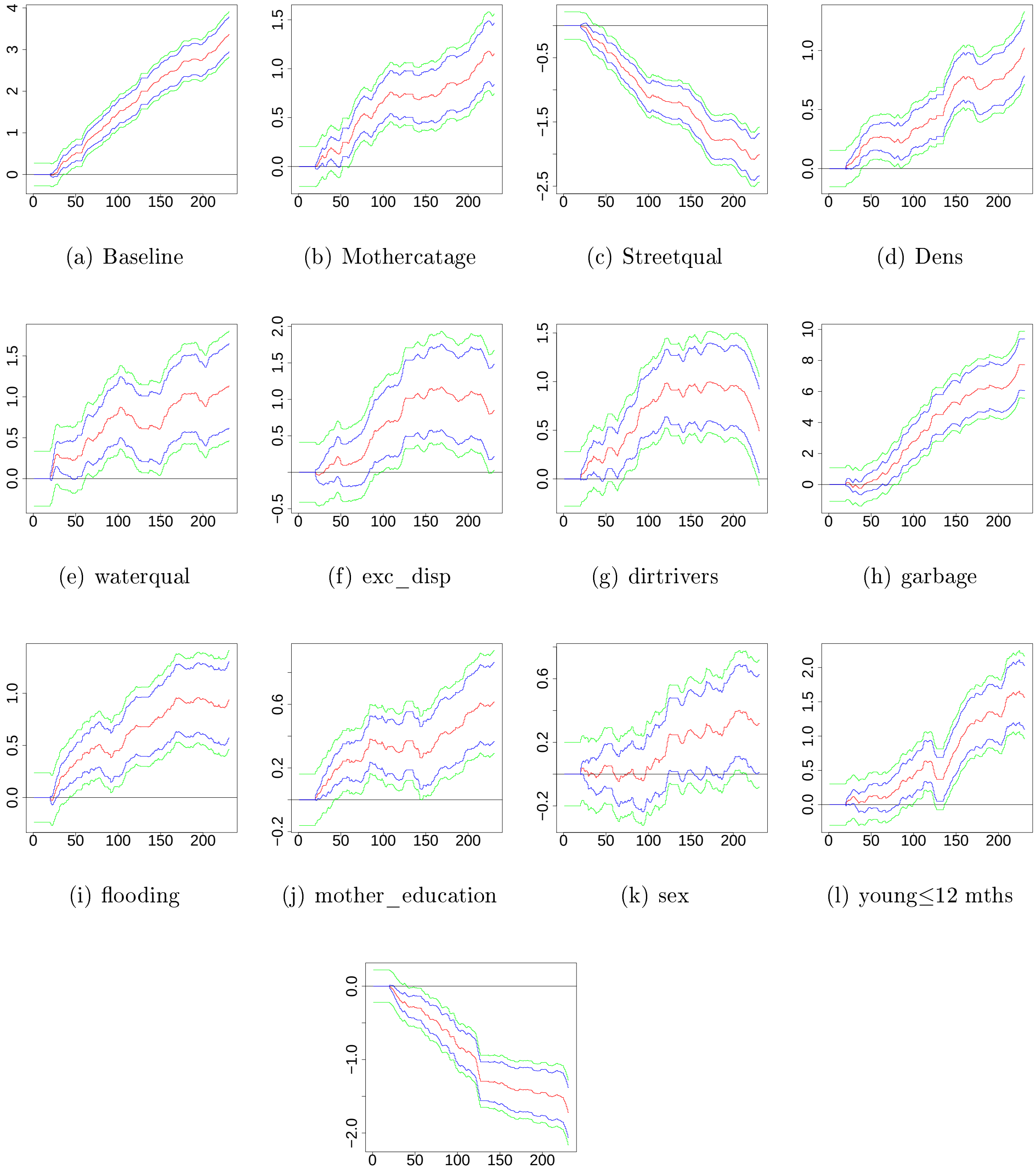
	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	15.526	$<$0.001	10.536	0.530
num5y	0.689	0.490	38.580	$<$0.001
mothercatage	7.107	$<$0.001	7.763	0.440
streetqual	-11.795	$<$0.001	11.164	0.210
habqual	-1.476	0.140	22.996	0.890
dens	8.162	$<$0.001	6.159	0.520
water_origin	-0.130	0.900	14.299	0.740
waterqual	4.253	$<$0.001	13.800	0.470
toilets	-0.469	0.640	15.488	0.670
exc_disp	2.753	0.010	20.730	0.450
dirtrivers	2.436	0.010	22.524	0.060
garbage	9.567	$<$0.001	35.906	0.670
flooding	5.454	$<$0.001	9.117	0.550
mother_education	4.758	$<$0.001	6.322	0.490
sex	2.060	0.040	5.017	0.850
young$\leq$12 mths	6.629	$<$0.001	14.438	0.310
old $>$24 mths	-9.888	$<$0.001	13.394	0.130

2.2.1 Study of prevalence

We start with the study of prevalence. The results for the first model are presented in Table 2.3, where T_{sig} is the test statistic for testing $B_j(t) = B_j^0(t) = 0$. Given Table 2.3, we can see that some of the covariates are not significant. This means that they do not have any great effect on the prevalence of diarrhoea. Therefore, we choose to remove them from the model. The final model is composed by the covariates that were significant in Table 2.3 and the obtained results are presented in Table 2.4, for the significance and constancy test, and in Figure 2.6 for the plots of the cumulative regression functions of each of the covariates. In this figure, we can notice that confidence bands (in green) are wider than confidence intervals (in blue), as expected, due to the fact that confidence intervals are pointwise whereas confidence bands are uniform. However this difference is not great.

The advantage of the additive regression model, as it was outlined in Section 1, is that the plots of the cumulative regression functions give a direct interpretation of their effect on the event of interest. A decreasing cumulative regression function for instance implies that the covariate reduces the risk of the event happening

Figure 2.6: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the final prevalence study without dynamic covariates. The x-axis is in days.



(a) Baseline

(b) Mothercatage

(c) Streetqual

(d) Dens

(e) waterqual

(f) exc_disp

(g) dirtrivers

(h) garbage

(i) flooding

(j) mother_education

(k) sex

(l) young≤12 mths

(m) old>24 mths

Table 2.4: Significance and constancy test for the prevalence final model: reduced covariate set.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	15.821	<0.001	10.204	0.480
mothercatage	7.015	<0.001	8.684	0.410
streetqual	-11.809	<0.001	11.541	0.220
dens	8.550	<0.001	4.328	0.790
waterqual	4.238	<0.001	13.790	0.410
exc_disp	2.560	0.010	19.825	0.390
dirtrivers	2.176	0.030	23.802	<0.001
garbage	9.564	<0.001	33.001	0.880
flooding	5.432	<0.001	9.171	0.520
mother_education	4.741	<0.001	6.045	0.700
sex	2.036	0.040	4.855	0.930
young≤12 mths	6.518	<0.001	14.007	0.280
old >24 mths	-10.129	<0.001	12.323	0.170

and an increasing cumulative regression function implies the opposite. Take the variable `mothercatage` for instance (second plot in the top row of Figure 2.6). Its cumulative regression function is increasing with time, seems to be approximately linear and reaches 1 at the end of the study. Moreover, recall that this variable takes the value 1 if the mother is younger than 25 years old. Therefore, the interpretation would be that children with younger mothers are more at risk to have diarrhoea on a given day than those who have an older mother, with on average one more day with diarrhoea per child. This effect is roughly constant with time, as we cannot reject the null hypothesis in the constancy test (p-value>0.05). The variable `dirtrivers` (plot 3, row 2) is quite interesting because it is the only one with a definitely non constant effect over time. Indeed, we can notice that the null hypothesis is rejected in Table 2.4 (p-value<0.05). The shape of the cumulative regression function gives evidence in favor of this result: we can see that it is increasing up to time 100 and then it is roughly flat up to the end. Therefore, the presence of dirty rivers would increase the risk of having diarrhoea but only during the first 100 days or so and then it seems to have no influence.

All the other variables can be interpreted in the same way. Notice that young children are more susceptible to have diarrhoea than older ones as we would have expected. All the results seem to be close to what we could have expected, i.e. poor life and hygiene conditions lead to a higher risk. This corresponds to increasing

Table 2.5: Significance and constancy test for the incidence model: all covariates.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	12.154	<0.001	5.159	0.060
num5y	1.296	0.190	5.231	0.540
mothercatage	2.877	<0.001	2.568	0.360
streetqual	-4.539	<0.001	2.682	0.470
habqual	1.185	0.240	15.942	0.120
dens	3.476	<0.001	1.929	0.480
water_origin	-0.160	0.870	5.439	0.470
waterqual	2.377	0.020	4.714	0.350
toilets	1.031	0.300	4.306	0.790
exc_disp	-1.227	0.220	5.023	0.370
dirtrivers	-0.818	0.410	5.475	0.080
garbage	4.036	<0.001	18.149	0.080
flooding	1.726	0.080	2.330	0.700
mother_education	1.516	0.130	1.444	0.870
sex	0.298	0.770	2.141	0.510
young≤12 mths	1.233	0.220	4.927	0.220
old >24 mths	-6.168	<0.001	4.684	0.060

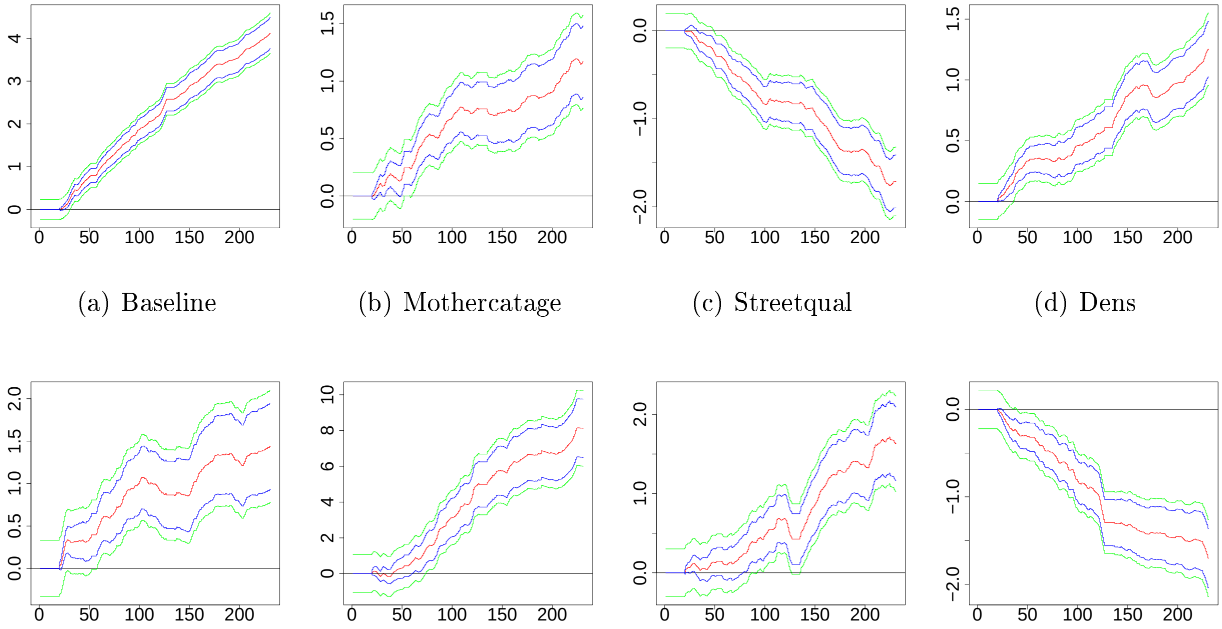
cumulative regression functions over time as poor hygiene conditions were coded as 1 when constructing the covariates. However the variable `streetqual` stands out. Its cumulative regression function (third plot of the upper row of Figure 2.6) is decreasing which means that a child who lives in a street of bad quality has less chance to start an episode of diarrhoea, which is not intuitive. This same contradiction will be observed when studying Phase II later on in this work.

2.2.2 Study of incidence

In this section we apply the same methods as in the previous section but this time, our event of interest is incidence. Again, we do not consider dynamic covariates. The first model was fitted by introducing all fixed covariates and results of both significance and constancy tests are presented in Table 2.5. All the non-significant variables were then removed to construct the final model. Results are in Table 2.6 and Figure 2.7.

In the final model, we only included significant covariates. The only exception is variable `young` that we have to include even though it is not significant because

Figure 2.7: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the incidence study without dynamic covariates. The x axis is in days.



(a) Baseline

(b) Mothercatage

(c) Streetqual

(d) Dens

(e) Waterqual

(f) Garbage

(g) Young≤12 mths

(h) Old>24 mths

Table 2.6: Significance and constancy tests for the incidence model: reduced covariate set.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	15.086	<0.001	5.734	0.010
mothercatage	2.805	0.010	2.464	0.450
streetqual	-5.019	<0.001	2.861	0.350
dens	5.374	<0.001	2.238	0.260
waterqual	2.894	<0.001	5.172	0.170
garbage	4.610	<0.001	17.511	0.030
young$\leq$12 mths	1.136	0.260	4.961	0.220
old >24 mths	-6.314	<0.001	4.769	0.050

it is linked with the variable `old` which is significant. All the results seems to agree with what we would have expected, except for the variable `streetqual` as it was already noticed when studying prevalence.

2.3 Additive regression model with dynamic covariates

As seen in Section 2.1, some children experienced many events whereas others experienced hardly any. Therefore it seems not realistic to consider that all the children have the same probability to have diarrhoea or to start an episode of diarrhoea on a given day, even given the same covariates. Some children are more at risk than others and this is called frailty in the event history literature (Hougaard, 2000). If we are in the presence of frailty, the results of the previous section may not be correct. To deal with that problem, we can construct dynamic models as described in Section 1.4. We saw in this section that any function of the past $\mathcal{F}_{t-}$ can be incorporated into a model for $\alpha(t)$ without changing the properties.
Thus, we are now going to look at a model where we will include some dynamic covariates. In what follows, we will construct the dynamic covariate j for each child i in the following way, according to Borgan *et al.* (2007):

$$x_{ij}(t) = \frac{\sum_{s=0}^{t} w(s)R_i(s)Y_i(s)}{\sum_{s=0}^{t-1} w(s)R_i(s)},$$

where $Y_i(s)$ is the at risk indicator for child i, $R_i(s)$ is the event process of interest and

$$w(s) = \begin{cases} 1 & \text{if } t-s \leq \tau, \\ e^{-\rho(t-s-\tau)} & \text{if } t-s > \tau. \end{cases} \tag{2.1}$$

In what follows, we will take $\tau = 30$ and $\rho = 0.01$ as in Borgan *et al.* (2007). We will see in Section 2.5.2 that this arbitrary choice of the value of this two variables does not affect the estimation to any extent.
We will consider the following dynamic covariates:

- days rate: previous days with diarrhoea scaled by days at risk.
- episodes rate: days with episodes of diarrhoea scaled by days at risk.
- fever rate: days with fever scaled by days at risk.
- sick rate: days when the child was sick scaled by days at risk.
- cough rate: days when the child had cough scaled by days at risk.
- can rate: days when the child had shortness of breath scaled by days at risk.

In the case where we want to study prevalence, we will also include the lags (from 1 to 4 days) which will be denoted by lag1, ..., lag4 and which describe whether a child had diarrhoea over the previous four days. By definition of incidence, these dynamic covariates cannot be included in the analysis of incidence.

2.3.1 Study of prevalence

We then first fit an additive regression model for the prevalence analysis. The results of the significance test are presented in the left part of Table 2.7. We can notice that once we introduce the dynamic covariates, none of the fixed covariates, except `streetqual`, are significant. This is evidence for the problem that was introduced in Section 1.4, which was that the fixed covariates also influence the dynamic covariates and therefore their effect is hidden when introducing the dynamic covariates in the model. Thus, we fit a new dynamic model by replacing the dynamic covariates by the residuals as described in Section 1.4. The results obtained for the significance test are presented in the right part of Table 2.7.

We can see that, by fitting the model using the residuals, there are now some fixed significant covariates. For the final dynamic model, we choose to keep in the analysis only significant covariates. The results of both significant and constancy tests are presented in Table 2.8 and the cumulative regression functions for each of the covariates included in the final model are presented in Figures 2.8 and 2.9.

Table 2.7: Significance test for the study of prevalence with dynamic covariates: original (left) and residuals (right).

(a) Dynamic covariates

Covariate	value	p-value
Baseline	0.14	0.89
num5y	0.554	0.58
mothercatage	1.982	0.05
streetqual	-2.371	0.02
habqual	-0.782	0.43
dens	0.731	0.46
water_origin	-0.004	1
waterqual	0.425	0.67
toilets	0.314	0.75
exc_disp	-0.582	0.56
dirtrivers	-0.099	0.92
garbage	1.866	0.06
flooding	0.715	0.47
mother_education	1.286	0.20
sex	1.179	0.24
young≤12 mths	2.28	0.02
old >24 mths	-0.869	0.39
days_rate	**3.017**	**<0.01**
episodes_rate	**1.273**	**0.2**
sick_rate	**0.642**	**0.52**
fever_rate	**1.266**	**0.21**
cough_rate	**2.264**	**0.02**
can_rate	**-0.177**	**0.86**
lag1	**39.656**	**<0.01**
lag2	**4.404**	**<0.01**
lag3	**0.201**	**0.84**
lag4	**1.988**	**0.05**

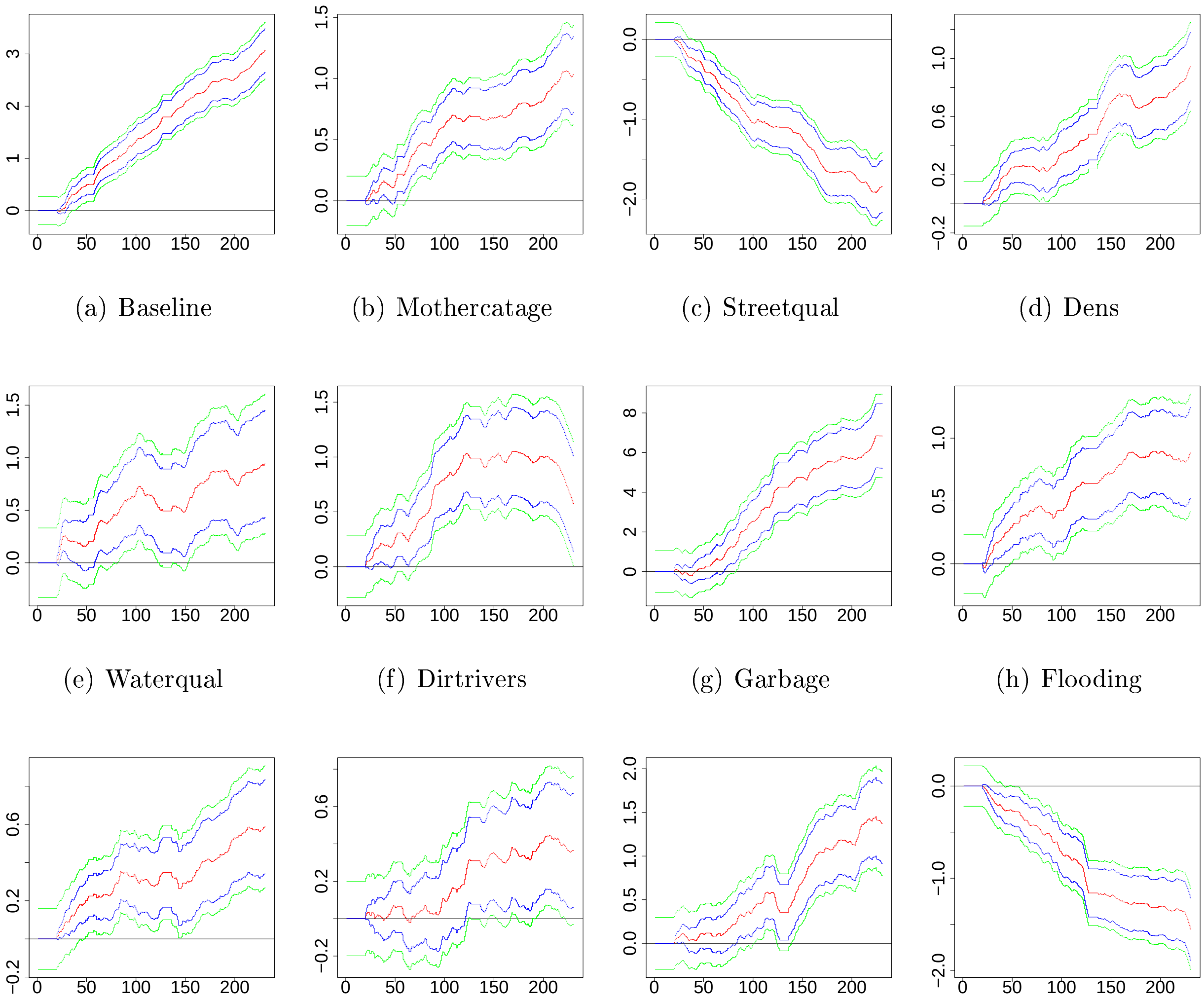
(b) Dynamic residuals

Covariate	value	p-value
Baseline	14.169	<0.001
num5y	1.035	0.300
mothercatage	6.486	<0.001
streetqual	-10.981	<0.001
habqual	-0.720	0.470
dens	6.999	<0.001
water_origin	-0.409	0.680
waterqual	3.407	<0.001
toilets	0.149	0.880
exc_disp	1.795	0.070
dirtrivers	2.561	0.010
garbage	8.360	<0.001
flooding	5.240	<0.001
mother_education	4.740	<0.001
sex	2.497	0.010
young≤12 mths	5.672	<0.001
old >24 mths	-9.007	<0.001
days_rate	**27.693**	**<0.001**
episodes_rate	**5.682**	**<0.001**
sick_rate	**3.103**	**<0.001**
fever_rate	**3.711**	**<0.001**
cough_rate	**2.455**	**0.010**
can_rate	**-0.196**	**0.840**
lag1	**49.468**	**<0.001**
lag2	**5.292**	**<0.001**
lag3	**1.092**	**0.270**
lag4	**1.986**	**0.050**

Table 2.8: Significance and constancy tests for the prevalence final model with dynamic covariates.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	14.526	<0.001	8.360	0.010
mothercatage	6.387	<0.001	8.350	<0.001
streetqual	-11.009	<0.001	11.065	<0.001
dens	8.006	<0.001	4.888	<0.001
waterqual	3.576	<0.001	11.314	<0.001
dirtrivers	2.584	0.010	23.786	<0.001
garbage	8.775	<0.001	29.722	0.060
flooding	5.186	<0.001	8.304	<0.001
mother_education	4.565	<0.001	5.371	<0.001
sex	2.344	0.020	4.384	0.110
young≤12 mths	5.806	<0.001	11.432	<0.001
old >24 mths	-9.228	<0.001	10.741	<0.001
days_rate	27.733	<0.001	4.352	<0.001
episodes_rate	5.622	<0.001	6.746	<0.001
sick_rate	2.555	0.010	1.157	<0.001
fever_rate	3.635	<0.001	0.827	0.020
cough_rate	2.471	0.010	0.224	0.010
lag1	49.886	<0.001	92.900	<0.001
lag2	5.204	<0.001	86.342	0.010

Figure 2.8: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the prevalence study with dynamic covariates. The x axis is in days.

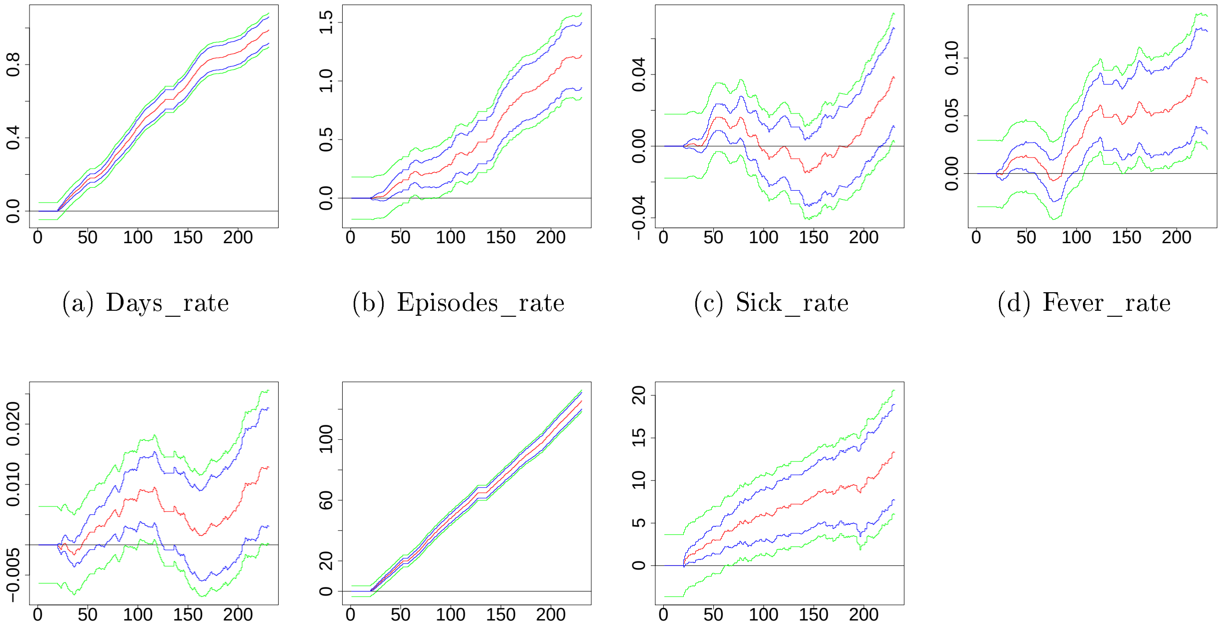


(a) Baseline (b) Mothercatage (c) Streetqual (d) Dens

(e) Waterqual (f) Dirtrivers (g) Garbage (h) Flooding

(i) Mother_education (j) Sex (k) Young≤12 mths (l) Old>24 mths

Figure 2.9: Cumulative regression functions (continued) with confidence intervals (blue) and confidence bands (green) for the prevalence study with dynamic covariates. The x axis is in days.



(a) Days_rate

(b) Episodes_rate

(c) Sick_rate

(d) Fever_rate

(e) Cough_rate

(f) Lag 1

(g) Lag 2

By comparing Tables 2.3 and 2.7, we can notice that exc_disp and sex become non-significant when we add dynamic covariates. The other covariates are still significant. In Figures 2.8 and 2.9 cumulative regression functions for each covariates are presented. We can remark that the plots for the fixed covariates in this figure are similar to those in Figure 2.6. When now looking at the dynamic covariates, we see that they are highly significant. We notice that days_rate and lag1 have a really small variance and the cumulative regression function of lag1 reaches 120 at the end of the time study, which is huge compared to the value of the other covariates. This means that a child who had diarrhoea the day before has a really strong risk of having diarrhoea the next day. Looking carefully at Table 2.8, we notice that the constancy test is rejected for every covariate included in the model. However in Figures 2.8 and 2.9 there seem to be some empirically constant covariates which contradicts the results obtained by computation. A reason for this would be that introducing the lags explains a lot and there is not much variability left. Therefore a really small variation in the slope of a covariate is detected by the test as being highly significant whereas it does not seem to be when looking at the plots. Thus, this model with the lags in is maybe a little over-fitted and may not give the correct results for the constancy test.

2.3.2 Study of incidence

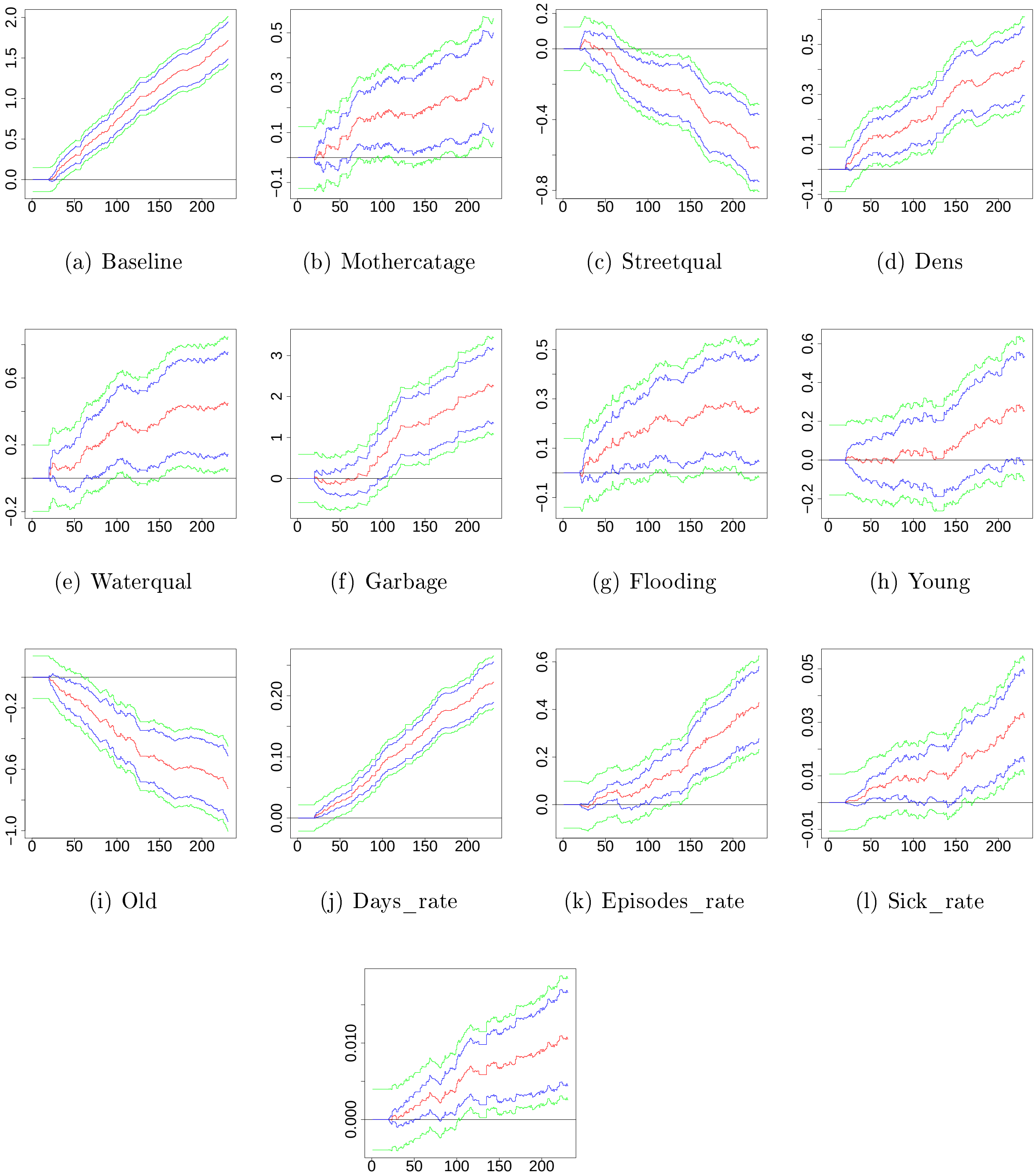
We now study incidence and this time we include dynamic covariates as described in the previous section. The only difference here is that we do not add the lags as they are not appropriate when studying incidence. We carry out the model selection in the exact same way as we did in the previous sections: we remove all the non-significant covariates from the model to keep only the significant covariates. The results of both significance and constancy tests for the final model are presented in Table 2.10 while the plots of the cumulative regression functions of the covariates are in Figure 2.10.

Here again, we notice that there is almost no difference between the plots of the cumulative regression functions of the fixed covariates in the model without dynamics (Figure 2.7) and in the model with dynamics (Figure 2.10). Only flooding becomes not significant between the two models. The variable days_rate also has a small variance. However, contrary to the prevalence model with dynamic covariates, the dynamic covariates do not take huge values. Moreover, results of the constancy test of Table 2.10 seem to conform to what we can observe in Figure 2.10, which is that the covariates almost all have a constant effect over time. It seems therefore that we encounter less problems and obtain more appropriate results when using the model for incidence with dynamic covariates. This conclusion will later be reinforced by the study of the residuals in Section 2.5.

Table 2.9: Significance and constancy tests for the incidence final model with dynamic covariates.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	12.006	<0.001	4.429	0.210
num5y	1.041	0.300	5.886	0.320
mothercatage	3.033	<0.001	2.810	0.380
streetqual	-5.413	<0.001	2.431	0.630
habqual	1.386	0.170	16.768	0.110
dens	4.410	<0.001	2.214	0.510
water_origin	-0.303	0.760	4.995	0.650
waterqual	2.480	0.010	4.690	0.420
toilets	0.863	0.390	4.640	0.740
exc_disp	-0.677	0.500	5.061	0.730
dirtrivers	0.137	0.890	6.005	0.120
garbage	4.617	<0.001	19.772	0.040
flooding	2.217	0.030	2.461	0.710
mother_education	1.725	0.080	1.539	0.770
sex	0.988	0.320	2.544	0.480
young<=12 mths	1.680	0.090	4.390	0.220
old >24mths	-6.631	<0.001	4.271	0.030
days_rate	13.471	<0.001	0.596	0.590
episodes_rate	4.109	<0.001	3.110	0.590
sick_rate	3.864	<0.001	0.286	0.270
fever_rate	0.831	0.410	0.221	0.880
cough_rate	3.729	<0.001	0.070	0.640
can_rate	0.284	0.780	0.294	0.850

Figure 2.10: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the incidence study with dynamic covariates. The x axis is in days.



(a) Baseline

(b) Mothercatage

(c) Streetqual

(d) Dens

(e) Waterqual

(f) Garbage

(g) Flooding

(h) Young

(i) Old

(j) Days_rate

(k) Episodes_rate

(l) Sick_rate

(m) Cough_rate

Table 2.10: Significance and constancy tests for the incidence final model with dynamic covariates.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	14.791	<0.001	4.667	0.090
mothercatage	2.950	<0.001	2.789	0.400
streetqual	-5.697	<0.001	2.487	0.540
dens	6.239	<0.001	2.517	0.200
waterqual	2.882	<0.001	5.111	0.290
garbage	5.237	<0.001	19.113	0.060
flooding	2.582	0.010	3.499	0.430
young<=12 mths	1.617	0.110	4.178	0.410
old >24mths	-6.714	<0.001	4.244	0.070
days_rate	13.388	<0.001	0.585	0.620
episodes_rate	4.062	<0.001	3.255	0.480
sick_rate	3.814	<0.001	0.286	0.270
cough_rate	3.744	<0.001	0.069	0.740

2.4 Dropout

As we saw in Section 2.1, the data we are studying present missingness. And more particularly, we noticed that some children dropped out of the study. We cannot test for missing not at random. However, we can suppose missing at random (MAR) as we condition on the past. In that case,

$$P(dropout|past, future) = P(dropout|past).$$

One can be interested in seeing whether the dropout effect is dependent on the covariates or if it depends on unknown factors. This means investigate MAR by looking at covariate effects on dropout. Thus, we fitted an additive regression model using all fixed and dynamic covariates with the dropout as the response variable. We found that none of the covariates were significant apart from `num5y`[1], which means that children with `num5y`=1 are more likely to drop out. Our analysis of prevalence and incidence is valid under MAR dropout as by definition our models condition on the past, and hence how that past affects dropout does not bias our conclusions.

[1] The complete results of this analysis are available in Section A of the Appendix.

2.5 Model checking and study of the martingale residuals.

We have fitted several models to the data but we have not so far considered model adequacy. For this, we need to develop a diagnostic procedure. First we will study residuals with or without dynamic covariates. Then, we will study the influence of the weight function, defined by equation (2.1), on our estimates.

2.5.1 Study of the residuals

Study of prevalence

In this section, we will study the goodness of fit of the two models we fitted for the prevalence analysis. As suggested in Section 1.2, we will construct the martingale residuals and if the model is correctly specified, the standard deviation of the standardized martingale residuals should be close to one at all time.

In what follows, we use the final models described in the previous sections, where we took $\tau = 30$ and $\rho = 0.01$ for the weights of the dynamic covariates.

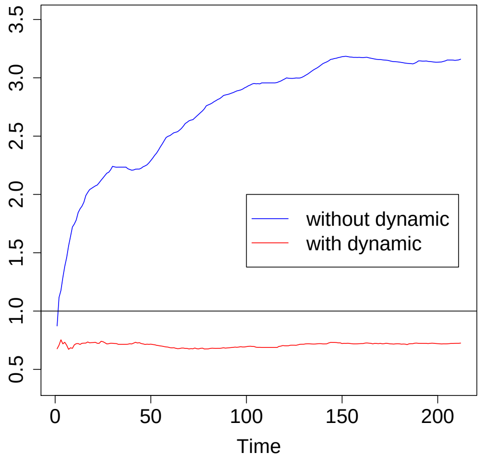
The standard deviations of the standardized martingale residuals for prevalence for both the models with and without dynamic covariates are presented in Figure 2.11. We notice that including the dynamic covariates provides a huge improvement as the standard deviation is much closer to one. It seems therefore more appropriate to use the model that includes dynamic covariates when wanting to study prevalence of diarrhoea. However, we can also notice that in the case when we introduce the dynamic covariates, we obtain residuals with variance below one which suggests that there may be some over-fitting. This is consistent with our concerns over inclusion of the lags effect.

Study of incidence

For the two models fitted for the incidence analysis, we computed the martingale residuals. The standard deviations obtained are plotted in Figure 2.12. This time, opposite to the prevalence case, it seems that the model without dynamic has residuals that are acceptable as their value at around 1.2, is reasonably close to one. Of course, introducing the dynamic covariates leads to an improvement of the model as the standard deviation of the residuals is closer to one, though there is some indication of overfitting. However the difference between the models with and without dynamic covariates is not as large as in the prevalence study.

We see that the standard deviation of the martingale residuals is closer to one when we study incidence than when we study prevalence. Moreover, we saw there was a problem with the constancy test when introducing the lags in the prevalence

Figure 2.11: Standard deviation of the standardized martingale residuals for the study of prevalence for models without dynamic covariates (blue) and with dynamic covariates (red).



model (see Section 2.3.1). For these reasons, we choose to keep the dynamic model for incidence as the best model for studying diarrhoea. However, we also keep in mind that the model for incidence without dynamic covariates is also quite good.

2.5.2 Study of the weight function

So far, every time we fitted a model including dynamic covariates, we chose $\tau = 30$ and $\rho = 0.01$. We are now interested in studying the role played by the weight function on the fit of the model. This is to see if other values of τ and ρ could produce better models, i.e. models with better residuals.
First, recall that the weights were defined in the following way

$$w(s) = \begin{cases} 1 & \text{if } t - s \leq \tau, \\ e^{-\rho(t-s-\tau)} & \text{if } t - s > \tau, \end{cases}$$

where τ and ρ must be chosen.

In Figure 2.13, we show the weight function $w(t)$ for different values of the parameters τ and ρ. We can notice that the weights are equal for the events that are in the most recent τ days and reduce the importance of the earlier events.

In Figure 2.13, we show the residuals obtained when fitting the model for incidence with dynamic covariates using the four weight functions described previously. We can notice that changing the value of τ or ρ does not affect the results.

Figure 2.12: Standard deviation of the martingale residuals for the study of incidence for models without dynamic covariates (blue) and with dynamic covariates (red).

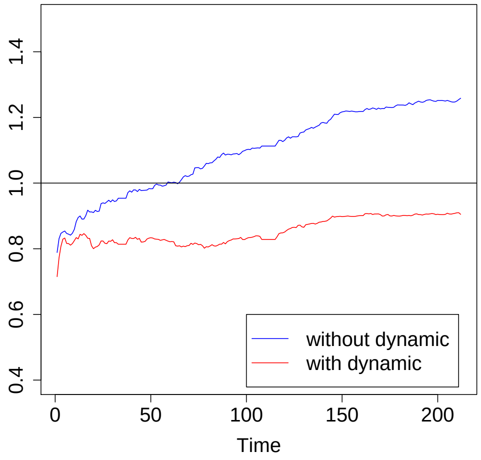
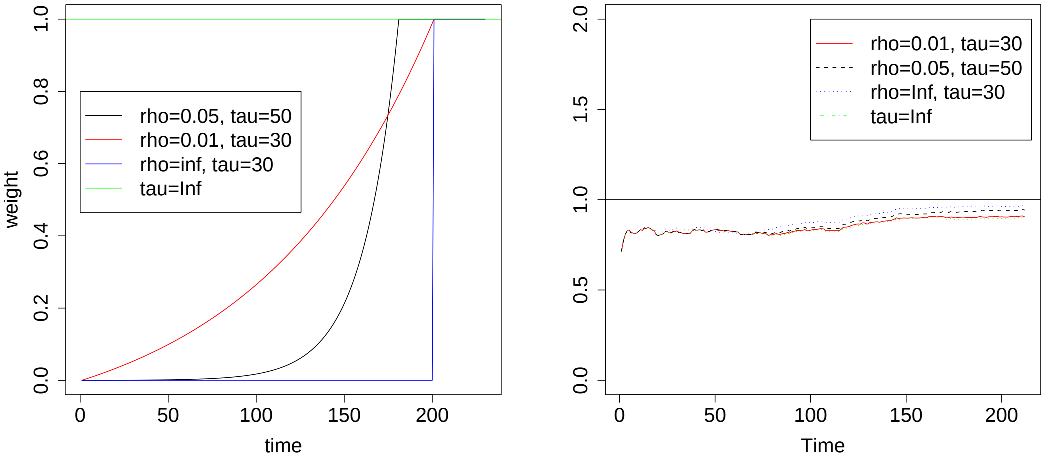


Figure 2.13: Weight function for different values of τ and ρ (left) and the impact of different weight functions on the martingale residuals (right).



It seems we have constant time frailty, so the dynamic covariates are fairly stable over time. Therefore, if some reasonable history is included, changing the weights will not affect the model.

2.6 Logistic regression

The additive regression model presented above has lots of advantages. Indeed, it produces results that are directly interpretable and all the theory of martingales can be applied for inference. However it presents also an important disadvantage which is that the estimated intensity $\widehat{\alpha}(t)$ defined by (1.3) can be negative. This happens when for example one of the covariates has a really strong negative effect. The estimates are not constrained to be non-negative which leads to the possibility of having a negative $\widehat{\alpha}(t)$.
In order to deal with that problem, one can consider a logistic regression approach. For each time point t, we fit a logistic regression for all individuals at risk. We replace our model (1.3) by:

$$\alpha(t) = Y(t)\frac{e^{X(t)\beta(t)}}{1+e^{X(t)\beta(t)}}.$$

However this method loses some of the advantages of the additive regression model. First, the plot of the output, which is the estimates $\widehat{\beta}(t)$ is not informative and does not give any direct information on its influence on $\alpha(t)$ (we cannot obtain the same kind of plots as in the previous sections). Moreover, the cumulative coefficients $\widehat{B}(t)$ do not have the martingale property needed for inference. Further, the logistic approach needs sufficient events and sufficient variability in the covariates for the iterative estimation routine to converge. For example, it is not possible to fit the logistic regression model if all the individuals who experienced an event have a covariate equal to 0. This is illustrated in Table 2.11, where we can see that on day 25, the individuals who experienced an event all had `habqual`=0 and as we fit the model each day, we will not be able to estimate the probability for day 25.

We first fitted the logistic model by omitting days when there was not enough events (≤ 10) or when the covariates were not well distributed. When using all fixed covariates for the prevalence analysis, we were only able to fit the model for 14 of the 231 days. This is clearly not enough to do a meaningful analysis. Therefore, even if we can have negative estimate of the intensity, we prefer the additive regression model for rare recurrent events data. Moreover, provided we do not over-interpret the intensity estimates at particular times, inferences from the additive model based on cumulative coefficients or intensities are usually valid.

Table 2.11: Illustration of a covariate which is not well distributed. Each cell contains the number of individuals in each category.

		Event (day 25)	
		0	1
habqual	0	954	23
	1	29	0

Chapter 3

Comparison of Phase II and Phase III

Three longitudinal studies composed the investigation of the sanitation intervention in Salvador, Brazil. The first was done in 1997-1998, before the intervention. The second one, which we will call Phase II was conducted between October 2000 and January 2002. A complete analysis of these data can be found in Borgan *et al.* (2007) and Elgmati (2009). The last study, Phase III, was the one studied in the previous chapter and was conducted in 2003-2004 after the intervention.

In this chapter, we are interested in comparing Phase II and Phase III in order to see whether a real improvement can be observed concerning the occurence of childhood diarrhoea. We will first conduct an exploratory analysis of Phase II and then we will construct two models in order to be able to compare the two phases in the last section.

3.1 Study of Phase II

3.1.1 Exploratory analysis of Phase II

A total of 926 children were involved in the second phase of the Blue Bay project. they were followed at home twice a week from October 2000 to January 2002 for a maximum number of 455 days of follow-up (Borgan *et al.*, 2007). However, as only 231 days were considered in the next phase, we will only consider the first 230 days of the study. As well as recording each day if the child had diarrhoea, fever or was sick, some information about the environnement in which the child lived were collected. As our goal is to compare Phase II with Phase III, we will only consider the covariates that are common to both studies. They are: `num5y`, `dens`, `streetqual`, `waterqual`, `dirtrivers`, `flooding`, `mothercatage`, `sex` and `age`: these variables were defined in Section 2.1. As we are trying to see if the

Table 3.1: Proportion of children for each common covariate in each phase.

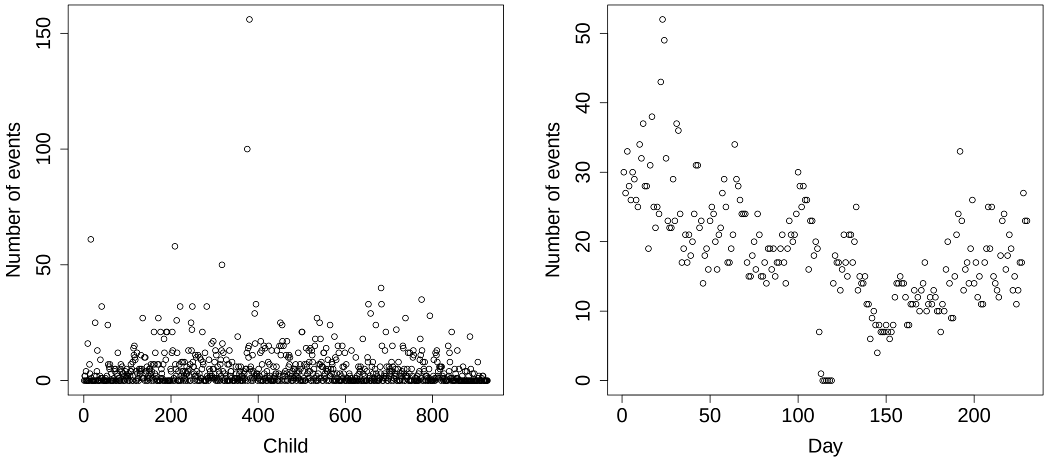
Covariate	value	Phase II	Phase III
Num5y	0	0.56	0.92
	1	0.44	0.08
Dens	0	0.82	0.78
	1	0.18	0.22
Streetqual	0	0.43	0.61
	1	0.57	0.39
waterqual	0	0.78	0.85
	1	0.22	0.15
dirtrivers	0	0.84	0.79
	1	0.16	0.21
flooding	0	0.71	0.70
	1	0.29	0.30
mothercatage	0	0.54	0.52
	1	0.46	0.48
sex	0	0.47	0.48
	1	0.53	0.52
age	$\leq$12	0.28	0.37
	$\geq$ 24	0.37	0.28

sanitation conditions were improved between both phases, we will not consider any dynamic covariates as they depend mostly on the children involved in the experiment rather than on the conditions in which they live. In the same way as for the study of Phase III, the micro areas will be considered and studied later in Chapter 4.

First, a table summarizing the proportion of individuals in each category of the covariates for both phases is presented in Table 3.1. With one exception, the distribution of covariates seems to be very similar in the two phases. The exception is `num5y`, the indicator of there being at least two children under five years old in the household. In Phase II about half the households had this property but by Phase III the proportion dropped to 8%. It is possible that these was a change in definition by Phase III but in the absence of other information we must accept the data as provided.

Figure 3.1 presents the number of events per child and per day respectively. As in Phase III, we can observe that some children experienced a lot more events than others.

Figure 3.1: Number of events per child (left) and per day (right) for Phase II.



3.1.2 Additive regression model for phase II

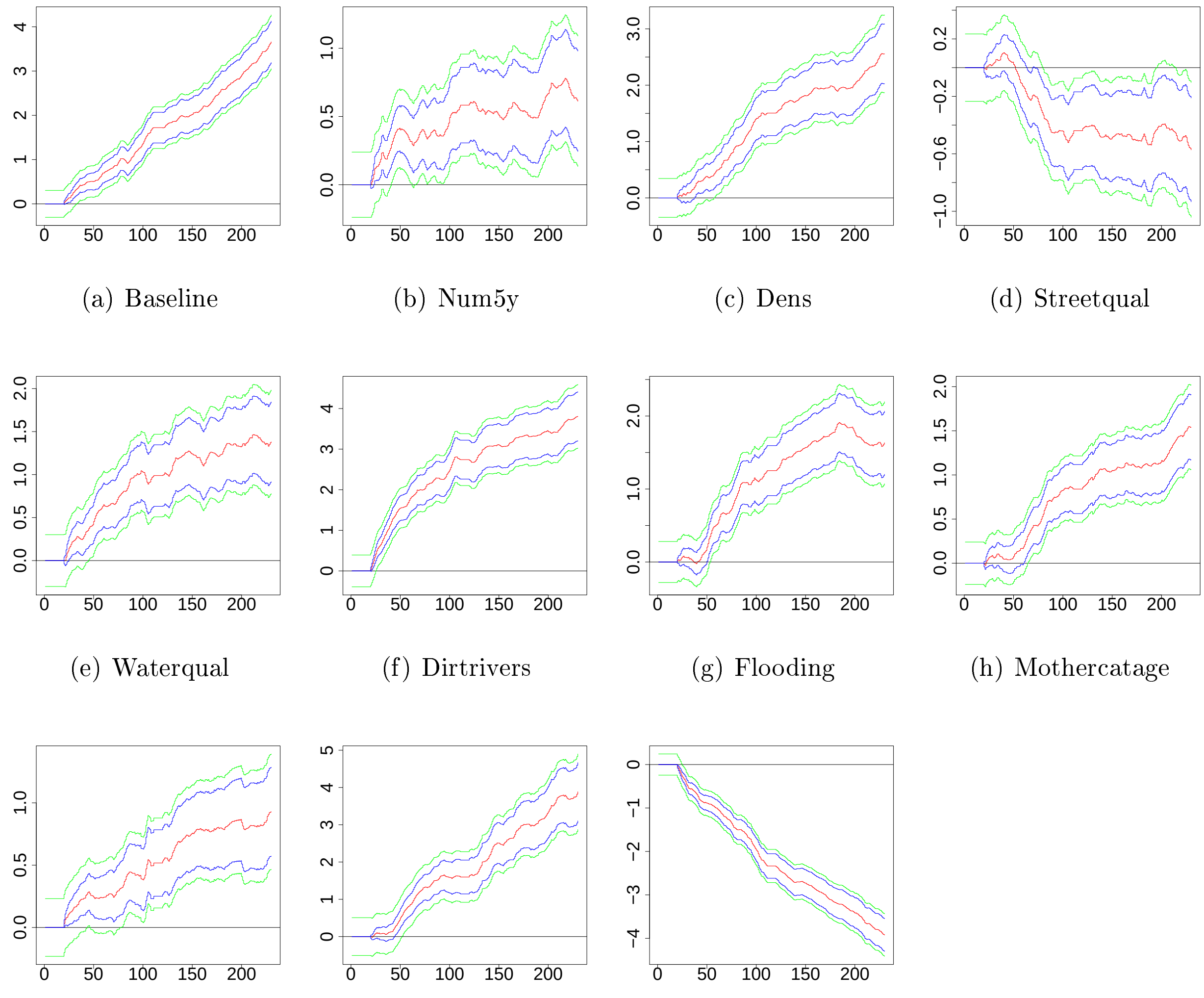
In this section we will only present the final models, as the model selection was done in a similar way as for Phase III. We include in the models only covariates that are common to both phases in order to be able to make a comparison. Again, in order to be consistent with Phase III, we only fit the models up to day 230. For each model, we present a table summarizing the significance and the constancy tests and we show figures of the cumulative regression functions for each covariate. Recall that we do not consider the dynamic covariates in this section as they depend on the children involved in the experiment and therefore are not suitable for a comparison.

Study of Prevalence

We start the construction of models by considering prevalence.

From Table 3.2, we can notice that all the covariates involved in both phases are significant here. The only covariate that was not significant in Phase III but is in Phase II is `num5y`. This is consistent with what was observed in the previous section where we noticed that the proportion of households with `num5y` taking the value one had dropped between both phases. As in the study of Phase III, the results we obtained in Figure 3.2 are consistent with what we would expect: bad hygiene and environment tend to increase the risk of having diarrhoea on a given day. However, note that again the same problem appears with the vari-

Figure 3.2: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the study of prevalence for Phase II. The x axis is in days.



(a) Baseline

(b) Num5y

(c) Dens

(d) Streetqual

(e) Waterqual

(f) Dirtrivers

(g) Flooding

(h) Mothercatage

(i) Sex

(j) Young

(k) Old

Table 3.2: Significance and constancy tests for the final model for prevalence for Phase II.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	15.461	<0.001	8.844	0.740
num5y	3.369	<0.001	9.845	0.410
dens2	9.515	<0.001	12.797	0.500
streetqual	-3.091	<0.001	8.494	0.530
waterqual	5.873	<0.001	15.421	0.250
dirtrivers	12.542	<0.001	36.035	0.010
flooding	7.124	<0.001	19.676	0.120
mothercatage	8.263	<0.001	6.924	0.740
sex	5.229	<0.001	6.269	0.740
young<=12 mths	10.339	<0.001	8.980	0.970
old >24mths	-20.364	<0.001	17.374	0.020

able `streetqual` which has a cumulative regression function in a counterintuitive direction.

Study of incidence

We now focus on the study of incidence. The results for significance and constancy tests are presented in Table 3.3 whereas the plots of the cumulative regression functions are in Figure 3.3.

The results seem to correspond once again with what we expected. However, this time it seems that the significant covariates involved to explain the incidence of diarrhoea are different between Phase II and Phase III (compare Tables 2.6 and 3.3). Variable `age` is common but now `dirtrivers`, `flooding` and `sex` appear to be important. In both cases the selected covariates can generally be considered as proxies for "bad environment".

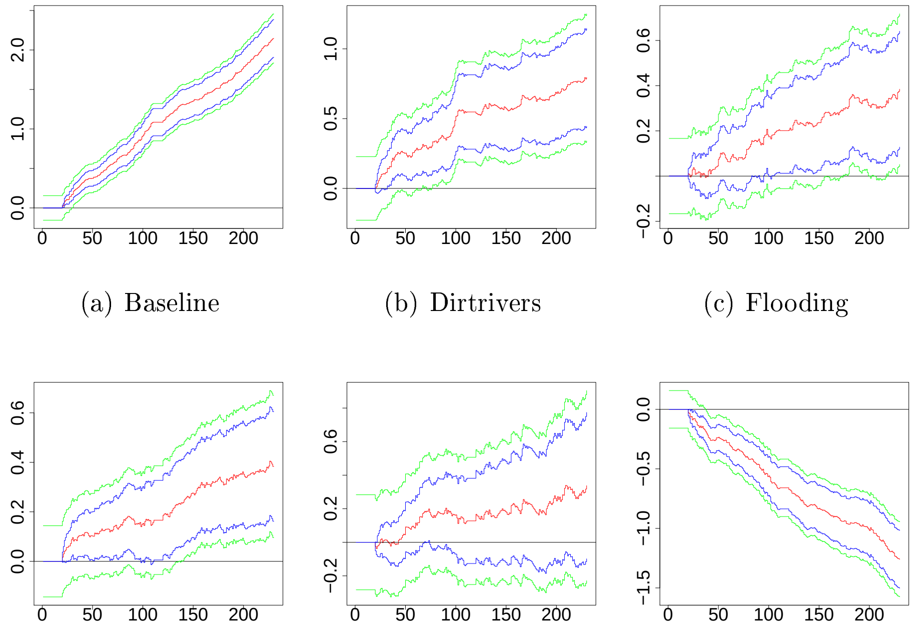
Martingale residuals

We can compare the standard deviation of the martingale residuals for both models. The result is presented in Figure 3.4. We notice that the model for incidence is much better in term of the residuals. Recall that when studying Phase III, the model for incidence without dynamics was also found to be quite good. Therefore, in what follows, we will consider the model for incidence without dynamics for

Table 3.3: Significance and constancy tests for the incidence final model for Phase II.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	17.848	<0.001	4.257	0.120
dirtrivers	4.373	<0.001	7.117	0.090
flooding	2.849	<0.001	1.991	0.850
sex	3.491	<0.001	2.281	0.740
young<=12 mths	1.875	0.060	3.309	0.940
old >24mths	-10.257	<0.001	3.914	0.300

Figure 3.3: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the incidence final model for Phase II.



(a) Baseline

(b) Dirtrivers

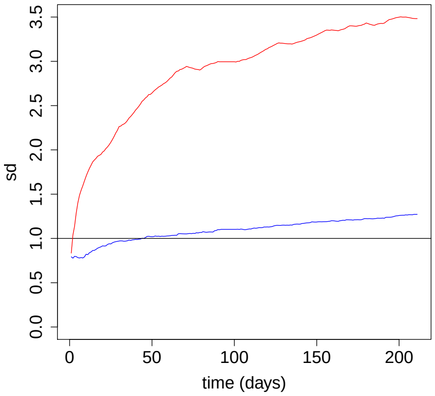
(c) Flooding

(d) Sex

(e) Young

(f) Old

Figure 3.4: Standard deviation of the martingale residuals for prevalence (red) and incidence (blue) for Phase II.



comparison and will not pursue prevalence further.

3.2 Comparison Phase II and Phase III

Now that we studied Phase II and that we found a model that is acceptable to explain incidence of diarrhoea, we aim to compare Phase II and Phase III. A model for incidence was constructed for Phase III using only the covariates that were in common and using only the 230 first days. As the results are similar to those presented in Section 2.2.2, we choose not to show them here. We begin our comparison with Figure 3.5, which shows for each covariate the final cumulative regression effect $B_j(\tau)$ for each phase. The colours indicate whether the covariate was significant or constant in each phase. We learn from the figure that few covariates appear in both phases. It seems that incidence of diarrhoea is explained by different covariates depending on the phase under study.

We now want to see if there is a difference in the intensity processes of the two phases. In Figure 3.6, we present the predicted intensity process for the incidence model for Phase II and Phase III. In order to do a meaningful comparison we consider two cases: first, children with "all the best", i.e. with all covariates equal to zero, and then children with "all the worst", i.e. all covariates equal to one. We also split into old and young children as we saw that age influences the incidence of diarrhoea.

From Figure 3.6, we see that the intensity in Phase III is lower than in Phase II, for young children at least. As expected young children are more at risk to

Figure 3.5: Value of the cumulative regression coefficients for incidence at time $\tau = 230$ for Phase II and Phase III.

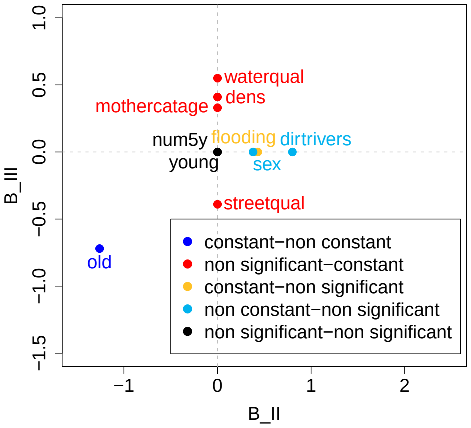
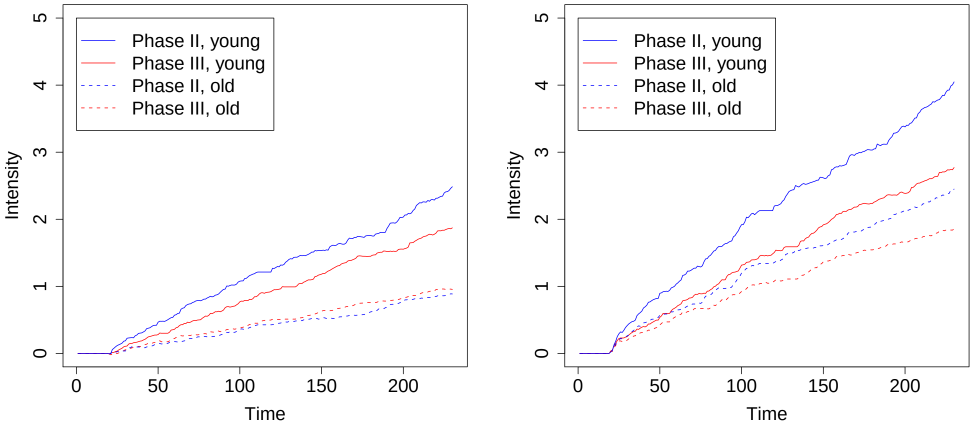


Figure 3.6: Predicted intensity process for incidence for children with all covariates=0, young and old (left) and with all covariates=1, young and old (right) for Phase II and Phase III.



have an episode of diarrhoea than older ones and of course having all "the worst" covariates increases considerably the intensity. Even if it seems that there is a difference between Phase II and Phase III, we need to see if it is significant. In order to do that, one can compute the variance of the intensity process, which we will denote $\widehat{\Gamma}(t)$. The intensity process is

$$\widehat{N}(t) = \int_0^t \widehat{\beta}(s) x_0 ds,$$

where x_0 is a vector of zeros and ones depending if the corresponding covariates take the value zero or one respectively. The computation of the variance is then inspired by the computation of the variance in simple linear regression.

$$\begin{aligned}\widehat{\Gamma}(t) = \widehat{\mathrm{var}(\widehat{N}(t))} &= \int_0^t x_0^T \mathrm{var}(\widehat{\beta}(s)) x_0 \; ds \\ &= x_0^T \int_0^t \mathrm{var}(\widehat{\beta}(s)) ds \; x_0 \\ &= x_0^T \widehat{\Sigma}(t) x_0,\end{aligned}$$

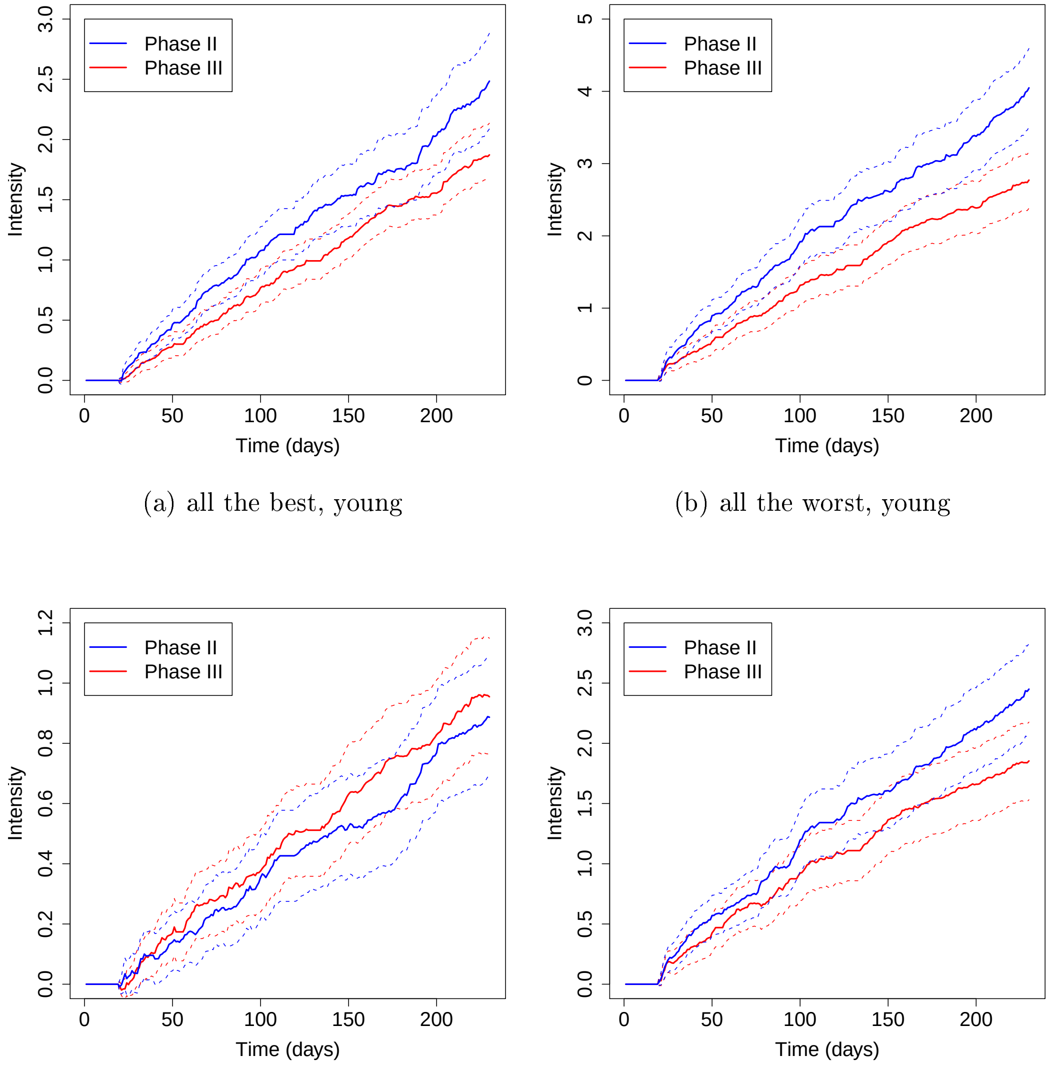
where $\widehat{\Sigma}(t)$ is the variance of $\widehat{\beta}(t)$ and was defined by Formula (1.6). Now that we have the variance, we can construct a 95% confidence interval for each of the functions plotted in Figure 3.6. The results are presented in Figure 3.7. Note that we change the scale and we separate each of the previous plots in two for clarity. By looking at this figure it is now easy to see if the difference observed in Figure 3.6 is significant or not. It suffices to see if the two confidence intervals include both functions for the last time point $\tau = 230$ in order for the difference to be non significant. This situation only happens once: for old children with "all the best" meaning that there was no improvement between Phase II and Phase III for this category of children. However for all other children, there is a significant difference between Phase II and Phase III and as the intensity for Phase III is always lower than the one for Phase II, this means that there was a real improvement in reducing the incidence of diarrhoea.

These results can be found again using a formal normal test. We denote by $\widehat{A}_{II}(\tau)$ and $\widehat{A}_{III}(\tau)$ the intensity at time τ in Phase II and Phase III respectively. The variance of $\widehat{A}_i$, $i \in \{II, III\}$ is denoted V_i. Under the null hypothesis that the two intensities are equal, we have

$$Z = \frac{\widehat{A}_{II}(\tau) - \widehat{A}_{III}(\tau)}{\sqrt{V_{II} + V_{III}}} \sim \mathcal{N}(0,1)$$

We apply the test to the four categories of children described above and obtain the following results:

Figure 3.7: Intensity processes for children with all best covariates (left) and all worst (right), young (upper) and old (lower), with 95% pointwise confidence intervals as dashed lines.



(a) all the best, young

(b) all the worst, young

(c) all the best, old

(d) all the worst, old

- Young children with all the best: $Z = 2.52$, leading to a p-value of 0.012.
- Old children with all the best: $Z = -0.47$ leading to a p-value of 0.63.
- Young children with all the worst: $Z = 3.73$ leading to a p-value$<$0.001.
- Old children with all the worst: $Z = 2.35$, leading to a p-value of 0.018.

The results of the test give evidence to support what was concluded from Figure 3.7. It seems that there was a real improvement between both phases concerning the incidence of diarrhoea, except for old children with good living condition. However, in this category, the incidence was already very low in Phase II and hence there is reduced scope for improvement by Phase III. As we saw earlier in this section, the observed difference cannot be due to the covariates. Therefore, it seems that the difference in the intensity that was observed may be due to the actions on hygiene and behaviour that was part of the measures taken by the government and not by an improvement of the living conditions of the children.

Chapter 4

Study of the micro areas

One of the covariates, the micro areas (MA), was left apart when constructing models because it was not binary and needed a special study. Considering that diarrhoea is maybe due to bacteria, it makes sense to suppose that children living in the same area as a child who had an episode of diarrhoea have more risk to have one themselves. In this chapter, we will first study the effect of the micro areas on the incidence of diarrhoea in Phase III. Then we will compare the effects of the micro areas in Phase II and Phase III.

4.1 Study of the micro areas in Phase III

In this section we will study if there is a significant difference between the areas by clustering the children depending on the area they live in and testing whether we obtain significant differences between the clusters. We have 24 different micro areas and the number of children in each varies from 22 to 61 (Figure 4.1). In all this section we concentrate on incidence only.

We first fitted models for each cluster including only a baseline. Figure 4.2 shows the cumulative regression functions for the baseline in each cluster (gray lines) and for all the children without clustering (black line). The big spread between the functions indicates there may be some differences between the micro areas. Three cumulative functions are highlighted in this plots. They are the furthest (MA 1 and 20) and the closest (MA 10) cumulative baselines to the general cumulative baseline.
The second plot of the same figure presents the standard deviation of the martingale residuals for each of the micro areas. The differences between the different clusters is not striking, and they all seem quite close to the standard deviation of the martingale residuals of the general model without clusters.

A more formal way to see if there is a difference between the clusters is to study the difference between their intensity. This means that if we denote α_k the

Figure 4.1: Number of children per micro area.

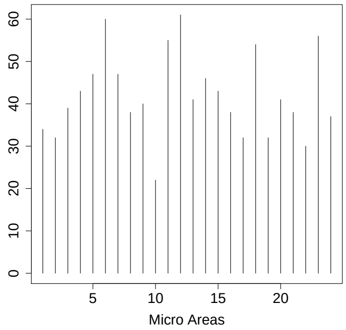


Figure 4.2: Left: cumulative baseline functions for the 24 MA (gray lines) and for all children (black line), for the study of incidence, with the furthest MA (blue) and the closest (red) to the general baseline. Right: standard deviation of the standardized martingale residuals for each cluster (gray lines) and for all children (black line) for the study of incidence.

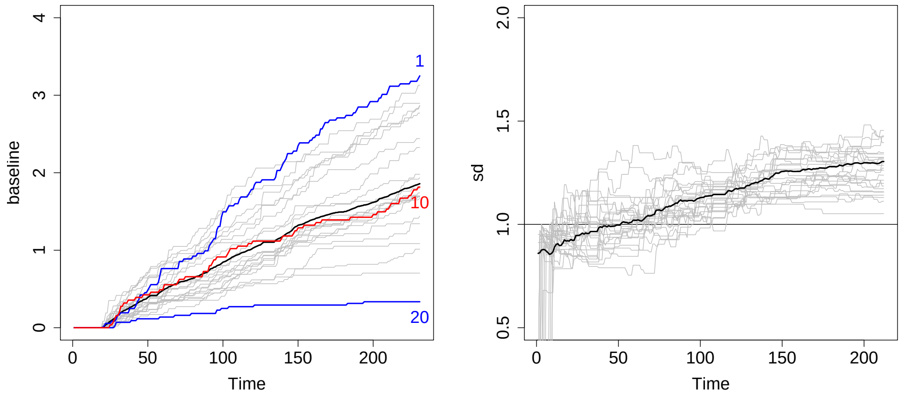
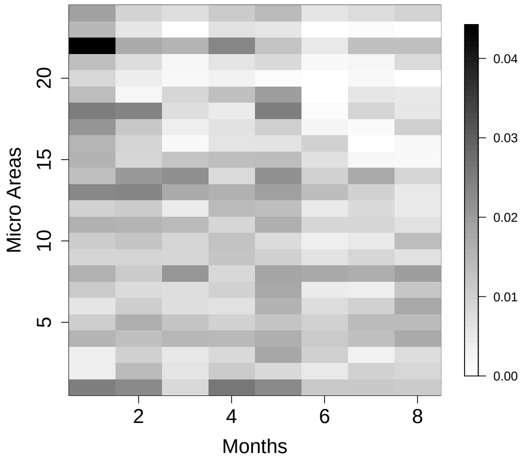


Figure 4.3: Space time plot for incidence. Each cell represents the average number of events in the MA for the given month. The darker the color, the higher the incidence.



intensity of cluster k, we want to test H_0: $\alpha_1(t) = \cdots = \alpha_{24}(t)$, for all $t \in [0, \tau]$. For this, we use the log-rank test which was introduced in Section 1.3.4. We obtained a value of the test statistic of 250.89, which under the null hypothesis, should come from a χ^2 distribution with 23 degrees of freedom. The corresponding p-value was below 0.001 which is strong evidence against the null hypothesis that all the clusters have the same intensity.

Another way to see the difference between clusters and if they are time varying, is to produce a space time plot where we can observe month to month differences. It is presented in Figure 4.3. In each cell, the average number of events per month, computed as the number of events in the month divided by the number of children at risk in the same month, is represented. The darker the colour of the cell, the more events happened in that cluster, that month. Differences between clusters are highlighted as some lines are darker than others (compare clusters 20 and 13 for instance). If we now look for differences between months for a given cluster, we can see that some of them, like clusters 1 or 8, seem to vary with time whereas cluster 9 is pretty much constant.

To see whether the differences between clusters, that we noticed previously, were by chance, we permute all children and we distribute them in 24 clusters of the same size as the original micro areas. Then, as before, the average number of events per month is computed. We did those permutations 1000 times. A space

Figure 4.4: Space time plots for 12 permutations.

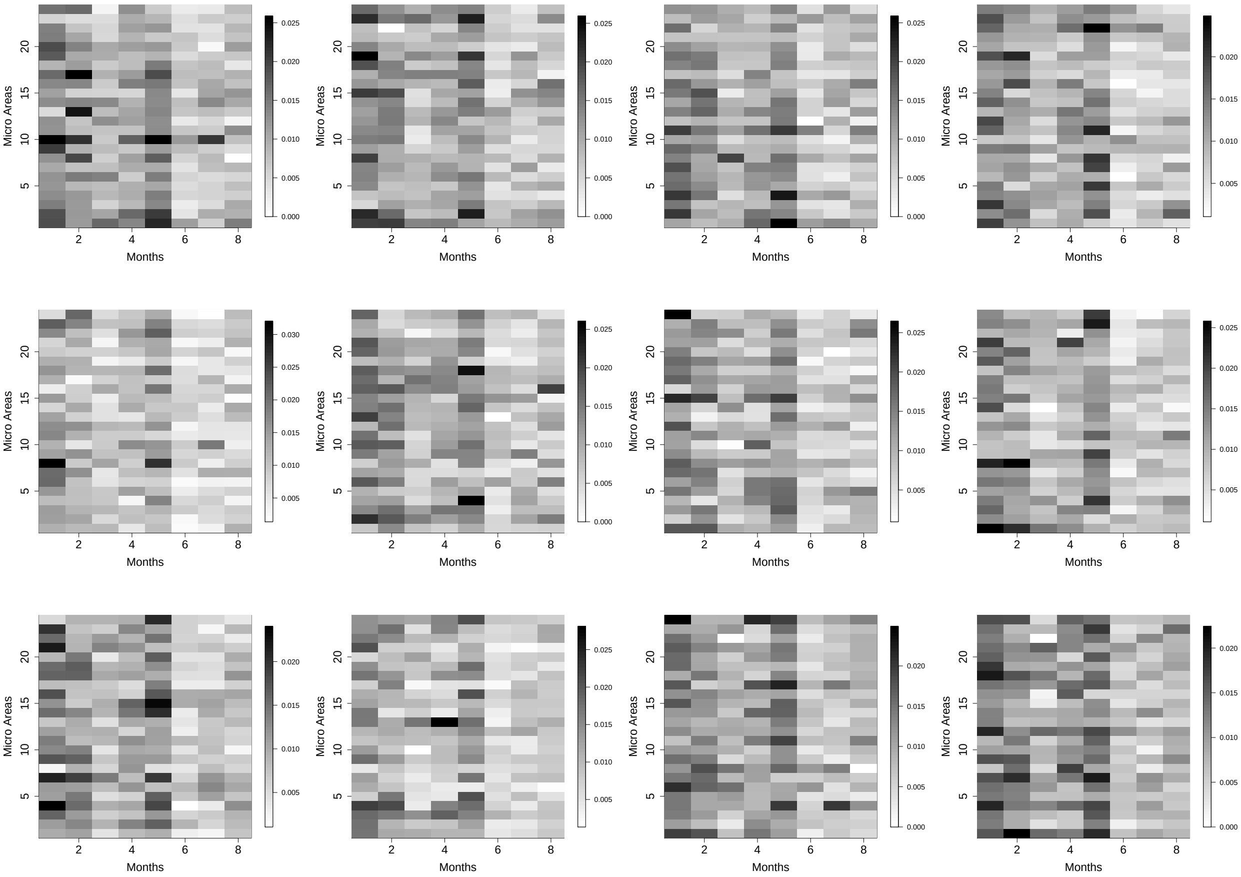
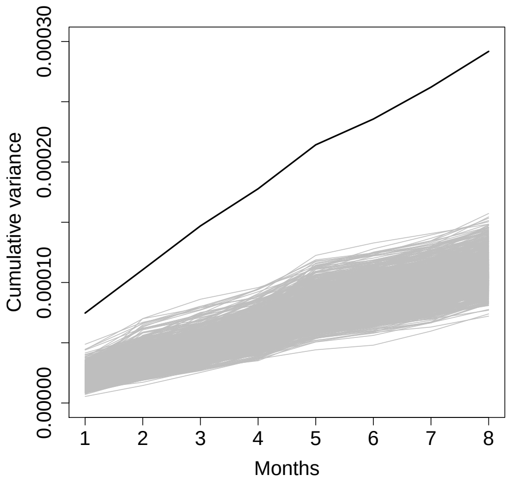


Figure 4.5: Cumulative variance of 1000 permutations (gray lines) and the cumulative variance of the real clusters (black line) over 8 months.

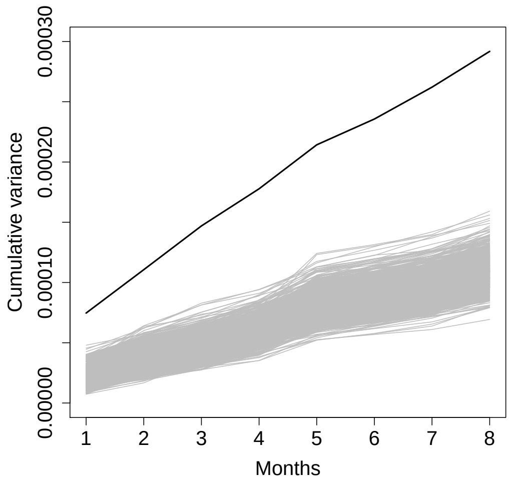


time plot for some permutations is presented in Figure 4.4. To see if the general data are different from the permutations, we compute for each the between cluster variance each month. The plot of the cumulative between cluster variance is in Figure 4.5. The variance of the original data is presented in black whereas the variances for each of the 1000 permutations are plotted in gray. We can notice a huge difference between the general variance and the variances of the permutations giving strong evidence for the fact that there is significant differences between the micro areas.

Even if Figure 4.5 clearly shows that we have a difference between clusters, we have to moderate this conclusion. Indeed, suppose there is a child in our sample who has a really large number of events. Then when doing permutations, the cluster containing that child will probably always be detected as different than the others. A solution to this would be to apply a permutation of the children each month. Then we apply the exact same process we used before. The plot of the cumulative variance for the 1000 temporal permutations is presented in Figure 4.6. We can see that changing the kind of permutation does not affect the results and the conclusion still apply: there is a big difference between clusters in the real data.

We now want to see what is the effect of each cluster on the general intensity and if this effect is constant or not. In order to analyze this, we have to fit an additive regression model including the MA as factor covariates. However,

Figure 4.6: Cumulative variances of 1000 temporal permutations (gray lines) and the cumulative variance of the real clusters (black line) over 8 months.



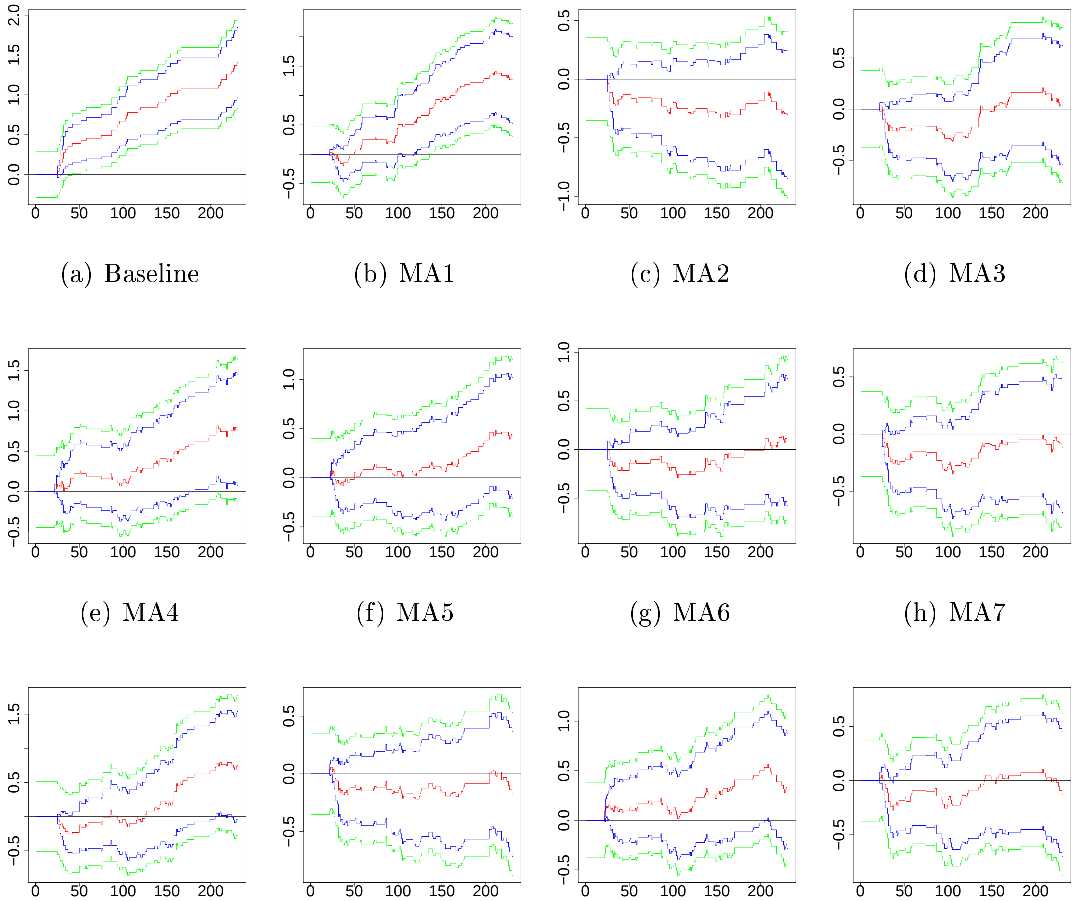
we cannot introduce all the MA, as the design matrix will not have full rank in that case and we will not be able to estimate the regression functions. Then, the estimates we will obtain from fitting the model will be the effect of a given MA compared to the one we removed. As we want to be able to compare the effect of each micro area with the general effect, it seems therefore appropriate to remove the MA that has the closest cumulative baseline to the general baseline. By doing this, we will approximately compare the effect of each MA with the general baseline. We recall from Figure 4.2 that MA10 was detected as being the closest to the general baseline. Thus, we fitted an additive regression model for the study of incidence, including all the MA but one (MA 10) in the same model. The results of the significance and the constancy tests are presented in Table 4.1 and Figures 4.7 and 4.8 present the plots of the cumulative regression functions of each micro area.

We can notice from Table 4.1 that micro areas 1, 4, 21, 22 and 23 are significant whereas the others are not and they have a constant effect. However, we have to remember that the effect of the micro areas that is shown by this model corresponds to the effect of the MA compared to the general effect without clustering (i.e. the baseline). Therefore, when we see a significant effect of one of the MA it means that this MA is significantly different from the baseline. On the other hand, when a micro area is detected as being constant it means that it varies in the same way as the cumulative baseline (first plot top row in Figure 4.7). The

Table 4.1: Significance and constancy tests for the study of incidence with all the micro areas but one included.

	Significance test		Constancy test	
Covariate	value of T_{sig}	p-value	value of $T_{1,const}$	p-value
Baseline	6.245	<0.001	7.305	0.110
MA1	3.537	<0.001	9.797	0.390
MA2	-1.052	0.290	6.540	0.500
MA3	0.137	0.890	10.475	0.100
MA4	2.508	0.010	7.224	0.660
MA5	1.508	0.130	5.935	0.760
MA6	0.138	0.890	10.282	0.200
MA7	-0.264	0.790	9.619	0.250
MA8	1.953	0.050	14.641	0.060
MA9	-0.477	0.630	5.343	0.700
MA11	1.187	0.240	9.567	0.240
MA12	-0.234	0.820	8.336	0.290
MA13	1.873	0.060	9.389	0.390
MA14	1.475	0.140	7.386	0.860
MA15	-1.516	0.130	7.613	0.320
MA16	-1.502	0.130	7.414	0.440
MA17	-0.847	0.400	6.059	0.660
MA18	0.393	0.690	9.608	0.150
MA19	-1.200	0.230	9.027	0.160
MA20	-4.786	<0.001	5.998	0.410
MA21	-1.886	0.060	3.997	0.910
MA22	2.251	0.020	11.674	0.210
MA23	-3.282	<0.001	7.078	0.340
MA24	0.336	0.740	7.720	0.480

Figure 4.7: Cumulative regression functions for incidence, showing the effect of each cluster compared with the general effect. The x axis is in days.



(a) Baseline

(b) MA1

(c) MA2

(d) MA3

(e) MA4

(f) MA5

(g) MA6

(h) MA7

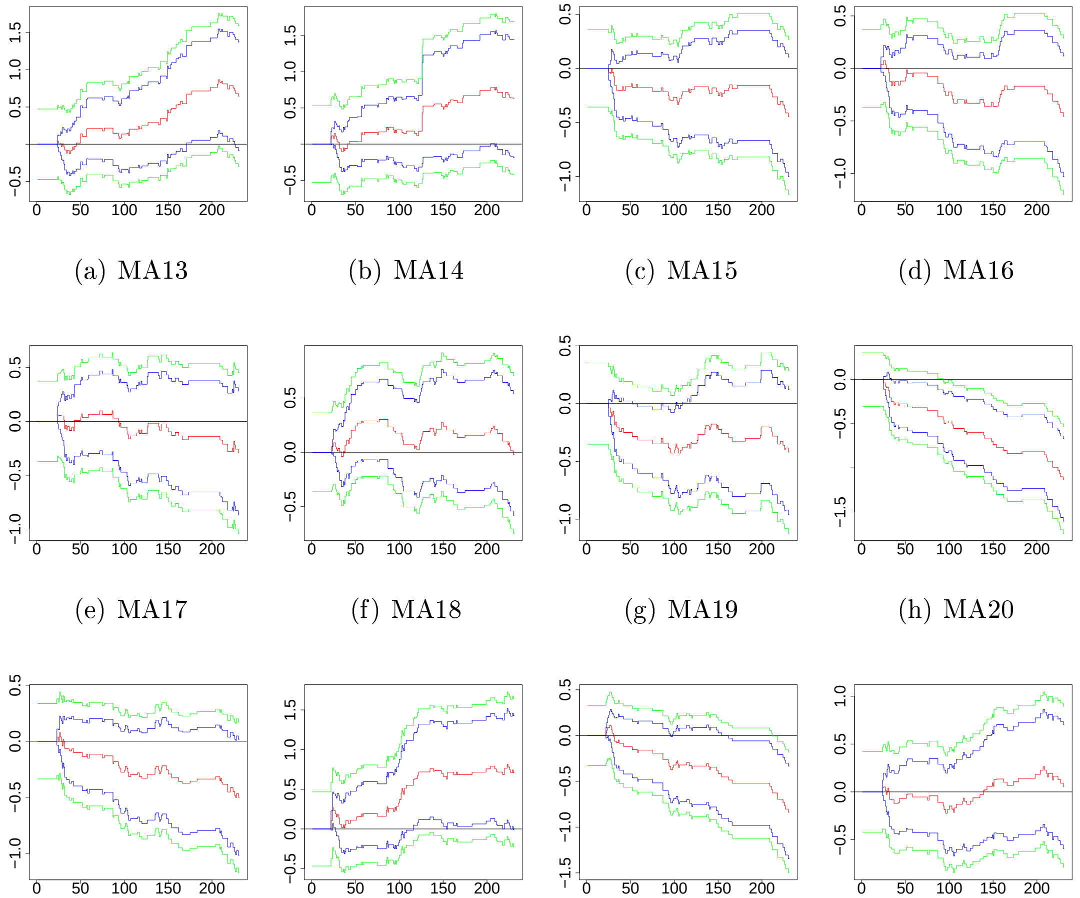
(i) MA8

(j) MA8

(k) MA11

(l) MA12

Figure 4.8: Cumulative regression functions (continued) for incidence, showing the effect of each cluster compared with the general effect. The x axis is in days.



(a) MA13

(b) MA14

(c) MA15

(d) MA16

(e) MA17

(f) MA18

(g) MA19

(h) MA20

(i) MA21

(j) MA22

(k) MA23

(l) MA24

cumulative regression functions of MAs 1, 4 and 22 are increasing as can be seen in Figures 4.7 and 4.8. This means that children living in those areas tend to have more risk to have an episode of diarrhoea on a given day than other children. On the contrary, MA 20 and 23 have decreasing cumulative regression functions indicating that children in those areas are less at risk than other children.

In this section we applied different methods in order to see if there was a significant difference between the micro areas and if it was the case, identify the effect of the clusters. We first looked at the differences in the cumulative baselines when fitting a model for each cluster. We could assess graphically and using a log-rank test that the difference was significant. Then, we constructed a space time plot and noticed again that it seemed there was an important difference between the Micro Areas. The permutation study gave evidence for this conclusion no matter if time constant or temporal permutations were applied. Finally the additive regression model that considered only the micro areas as covariates showed that in some of them the risk of having an episode of diarrhoea was increased whereas it was decreased in others. Therefore, the micro areas are an important factor for the risk of having diarrhoea. It seems that depending on where they live, children have not equal risk of having an episode of diarrhoea.

4.1.1 Comparison with Phase II

We will now focus on a comparison of the micro areas between Phase II and Phase III. It could be interesting to see if there is a difference between the micro areas, i.e. if the improvement that was observed in Chapter 3 was general or if some areas are more affected than others. Recall that in the previous section, we considered 24 different micro areas. However only 21 of them are present in both studies. In Figure 4.9, a plot of the number of children in each MA is presented for Phase II. It seems that children were taken approximately equally in all areas.

In Table 4.2, the proportion of children for each covariate and each MA in both phases is presented. The idea behind this table is to see if there is an area where living conditions were significantly improved or made worse between phases. It seems that the proportion is nearly the same between the phases for all the covariates and the MA. Again, as it was noticed in Table 3.1, the variable `num5y` stands out due to the big difference in the proportion of children for this variable between both phases. Nevertheless, this variable is the only one with such a characteristic and this difference applies to all MA.

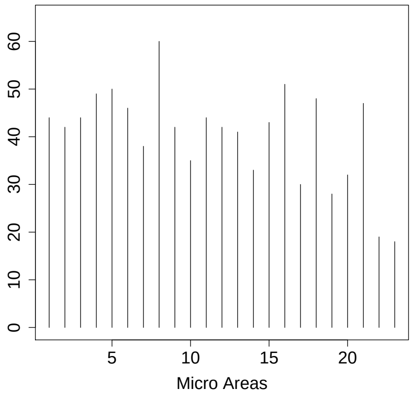
For each micro area we construct a residual between observed and expected counts as follows: for each phase,

- fit a model with baseline and covariates but no area effect.

Table 4.2: Percentage of children for each covariate (the percentage is given for the value 1 of the covariate) in each micro area for Phase II (left in each column) and Phase III (right in each column).

MA	Num5y		dens		streetqual		waterqual		dirtrivers		flooding		mothercatage		sex		young		old	
1	53	7	52	49	79	63	38	24	45	27	53	34	55	56	45	56	76	59	55	78
2	53	13	14	19	62	38	14	19	5	9	12	21	38	49	50	60	71	64	69	72
3	49	9	20	20	27	7	25	15	6	0	6	7	45	47	59	62	63	56	59	64
4	46	5	26	35	37	25	3	22	11	22	31	22	39	30	51	40	74	65	49	72
5	50	11	16	30	45	11	30	9	0	13	27	28	44	43	55	57	70	65	66	57
7	50	2	13	23	30	0	20	7	20	0	7	21	37	40	47	56	67	63	60	72
10	41	12	6	12	44	0	28	16	0	3	3	19	50	41	50	53	72	69	75	62
11	46	7	23	17	31	11	21	11	2	15	54	33	48	61	54	61	75	54	56	78
12	32	10	13	24	26	2	15	12	0	7	21	24	45	37	47	39	70	59	68	68
13	50	3	11	23	18	7	7	7	14	7	14	30	32	47	46	53	79	67	64	77
14	56	3	29	24	80	42	22	26	44	37	59	37	39	42	51	61	78	53	56	79
15	47	5	19	14	93	41	19	11	33	54	51	41	58	46	53	51	84	65	72	73
16	43	9	19	9	64	56	12	9	15	50	39	44	30	34	36	47	70	44	64	69
17	42	5	26	5	71	84	21	5	8	8	0	21	55	34	68	53	66	71	71	71
18	40	11	5	18	70	70	32	15	13	30	15	31	47	52	50	51	62	59	63	77
19	33	2	16	16	94	91	18	9	24	51	24	33	51	49	39	44	84	63	51	63
20	39	2	11	11	57	64	26	9	15	26	48	36	35	40	61	47	59	57	63	70
21	50	8	24	30	94	82	16	18	14	37	32	40	70	67	54	50	70	70	60	67
22	43	6	20	35	59	85	32	26	41	44	25	38	52	50	48	56	75	65	73	79
23	43	10	14	3	95	90	23	5	34	54	66	54	41	36	55	59	75	74	66	67
24	60	19	26	25	34	19	26	28	2	0	21	19	64	69	62	50	76	59	67	56

Figure 4.9: Number of children per micro area.



- Calculate the expected number of events for each MA for each phase:

$$\begin{aligned} E_k &= \sum_{i \in \mathrm{MA}_k} \int_0^\tau Y_i(t) \left(\widehat{\beta}_0(t) + \beta_1(t) x_{1i}(t) + \cdots + \beta_p(t) x_{pi}(t) \right) dt \\ &= \sum_{i \in \mathrm{MA}_k} \int_0^\tau \widehat{\alpha}_i(t) dt \\ &= \sum_{i \in \mathrm{MA}_k} \widehat{A}_i(\tau) \end{aligned}$$

- Calculate the observed number of events for each MA

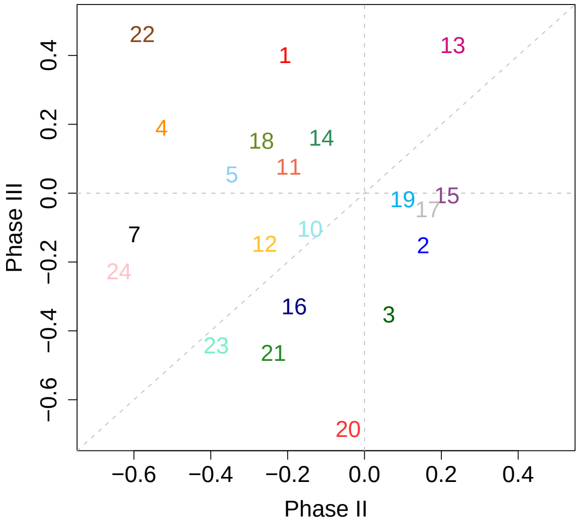
$$O_k = \sum_{i \in \mathrm{MA}_k} N_i(\tau)$$

- Construct R_k the relative residuals for each MA

$$R_k = \frac{O_k - E_k}{E_k}$$

The obtained residuals are then plotted for both phases in Figure 4.10. We can see that no systematic pattern appear. There is no area which stands out in both phases. Hence we have little evidence of residual area effects being related to Phase II to Phase III differences and no area stands out as being convincingly consistently good or consistently bad in terms of incidence.

Figure 4.10: Relative residuals R for Phase II and Phase III.



Conclusion

In this report, we focused on additive regression models to analyse survival data and more specifically recurrent events data. This method presents lots of advantages when dealing with those kind of data. First, the model is very simple as it is of linear form and therefore the computations are simplified, as we can use least squares estimation. Moreover, when wanting to assess the effect of covariates on the event of interest, the additive regression model produces results that give direct interpretation even for complex time varying effects of time varying covariates. Indeed it suffices to look at the plot of the cumulative regression function for each covariate to give a straightforward interpretation of its effect on the event of interest. Moreover, the martingale property gives very useful results regarding inference testing and model checking.

One problem is that the conditional probabilities can take negative values as there is no constraint to lie between zero and one. This situation occurs for example when a covariate has a strong negative effect. Due to the linear form of the model, it can happen that the resulting intensity is negative. In order to get round this problem, one can try to adopt a logistic regression approach which, due to the exponential form of the model, constrains the estimated intensity to lie between zero and one. This method was tried in Section 2.6, where we saw that unfortunately, due to the very few number of events, it was impossible to fit such a model for a sufficiently large amount of time points. Indeed the method used in the logistic modelling approach failed to converge for time points with low numbers of events.

After introducing the Blue Bay data, we applied several additive regression models. We applied each time the same model selection procedure, which consists in constructing a model with all covariates included and then removing the covariates that are not significant. First, models without dynamic covariates were fitted for both incidence and prevalence. Then after noticing that there was probably a frailty effect in our data, given the large difference in the number of events experienced by each child, we introduced the dynamic covariates in the model. All models were then compared using martingale residuals and it appears that the model for the study of incidence with dynamic covariates was the most appropriate. We learned from the models that bad hygiene and living conditions lead to a higher risk in having an episode of diarrhoea, as expected.

In the previous models, we always supposed that all the children were independent. However, given the fact that diarrhoea is maybe due to a bacteria, it seems to be sensible to think that children that live close to a child which is having an episode will tend to have one themselves too. We were therefore interested in seeing if there was some area effect. By applying various methods, such as testing (log-rank test), using permutation methods or even through additive regression models, the study of the clustering effect leads to the conclusion that there exists an area effect. Thus depending on where a child lives, he can be more at risk to have an episode of diarrhoea on a given day.

Finally, a comparison between the two last phases of the Blue Bay project was conducted. It revealed that there was a significant improvement between both phases, especially for young children. However, this improvement cannot be attributed to an improvement in the sanitation conditions as they were similar between both phases. We think that the improvement is therefore mainly due to prevention which was part of the measures taken by the Salvador governement. Moreover, this measure seems to have been applied uniformly over all areas as no difference was detected between the areas concerning the improvement in childhood diarrhoea.

As further work, one could consider doing a deeper analysis of the conditional multipliers that was studied when introducing the constancy test in Section 1.3.2. A comparison of the two terms that are often used in the literature: "conditional multipliers" and "wild bootstrap" and which seem to be similar could also be conducted. Some power analysis could be done in the log-rank test in order to assess which of the simulation or the permutation study gives the best power. Concerning the data and more specifically clustering, one information was not used when analyzing them. It is the geographics areas. There are eight of them and each is composed of three micro areas. An interesting question would be to assess if there is still a difference between them and if this time we can see some geographic difference between the improvements. Again concerning the data, we saw that problems occurred with the constancy test when introducing the lags in the model for prevalence. We could also notice that we were probably in the presence of over-fitting. It may therefore be interesting to investigate why the constancy test failed or construct an additive regression model with dynamics but without the lags and assess the goodness of fit of this model.

Finally, in this report, we focused on the additive regression models but there exist other models such as the well known Cox model. It could be of interest to fit such a model to the data and see if we obtain similar results.

Acknowledgements

I would like to thank the support, the helpfulness, the presence and the enthusiasm of Pr. Robin Henderson as well as the careful advice of Pr. Anthony C. Davison. My thanks also go to Rosemeire Fiaccone for providing the data studied in this report.

I also thank many people within the School of Mathematics and Statistics of Newcastle University, and especially my officemates, that make Newcastle such a nice place to live and study in.

Appendix A

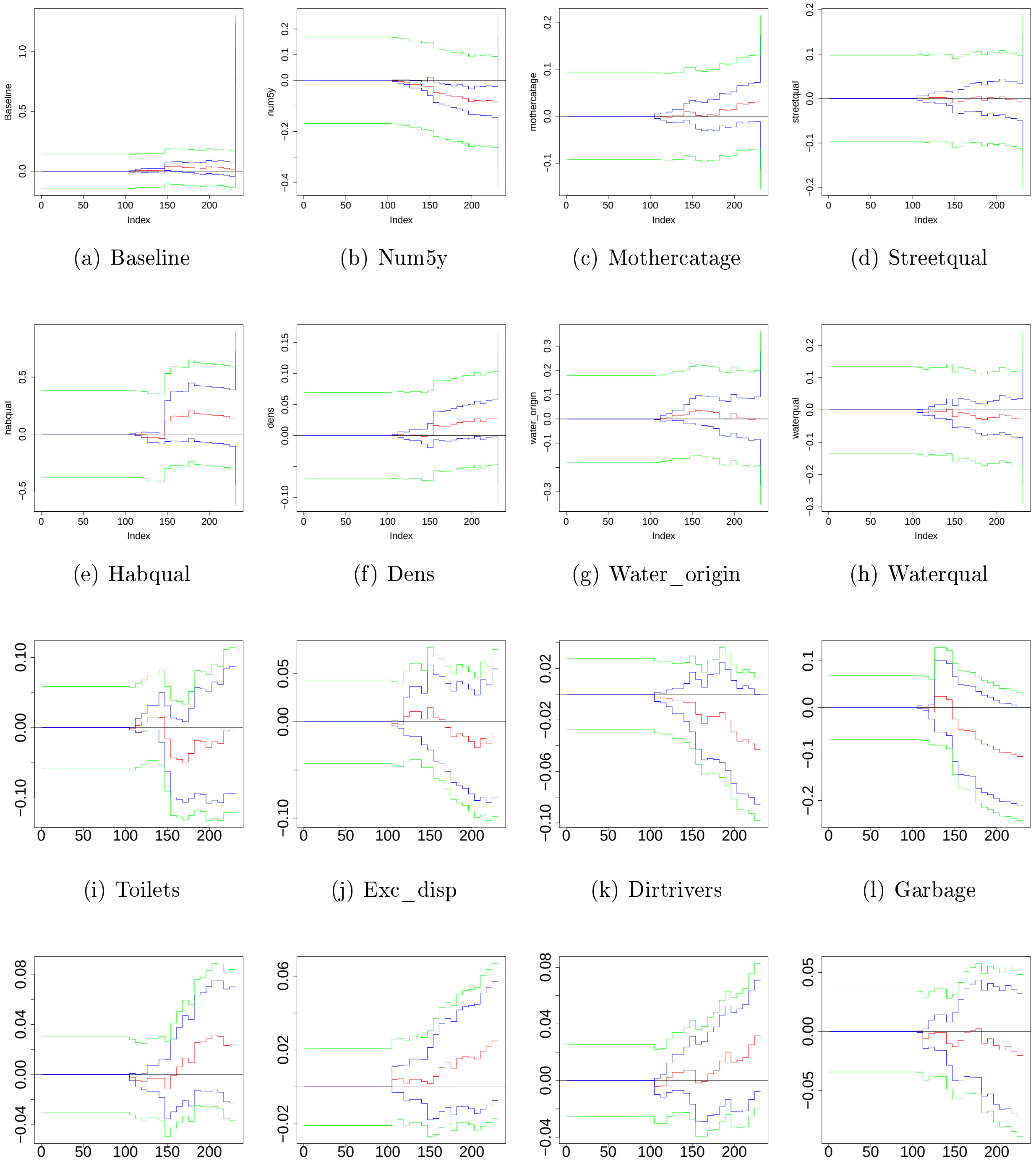
Results of the dropout study

Table A.1: Results of the significance test for the study of the dropout.

Covariates	Value of T_{sig}	p-value
Baseline	1.318	0.190
num5y	-2.375	0.020
mothercatage	1.386	0.170
streetqual	-0.388	0.700
habqual	1.290	0.200
dens	1.957	0.050
water_origin	0.134	0.890
waterqual	-0.851	0.390
toilets	-0.119	0.910
exc_disp	-0.330	0.740
dirtrivers	-1.943	0.050
garbage	-1.871	0.060
flooding	0.993	0.320
mother_education	1.491	0.140

Covariates	Value of T_{sig}	p-value
sex	1.536	0.120
young<=12 mths	-0.713	0.480
old >24mths	0.256	0.800
days_rate	-0.150	0.880
episodes_rate	2.015	0.040
sick_rate	1.449	0.150
fever_rate	0.228	0.820
caugh_rate	-0.819	0.410
can_rate	0.612	0.540
lag1	-0.086	0.930
lag2	-1.673	0.090
lag3	-0.248	0.800
lag4	1.402	0.160

Figure A.1: Cumulative regression functions with confidence intervals (blue) and confidence bands (green) for the dropout study. The x axis is in days.



(a) Baseline

(b) Num5y

(c) Mothercatage

(d) Streetqual

(e) Habqual

(f) Dens

(g) Water_origin

(h) Waterqual

(i) Toilets

(j) Exc_disp

(k) Dirtrivers

(l) Garbage

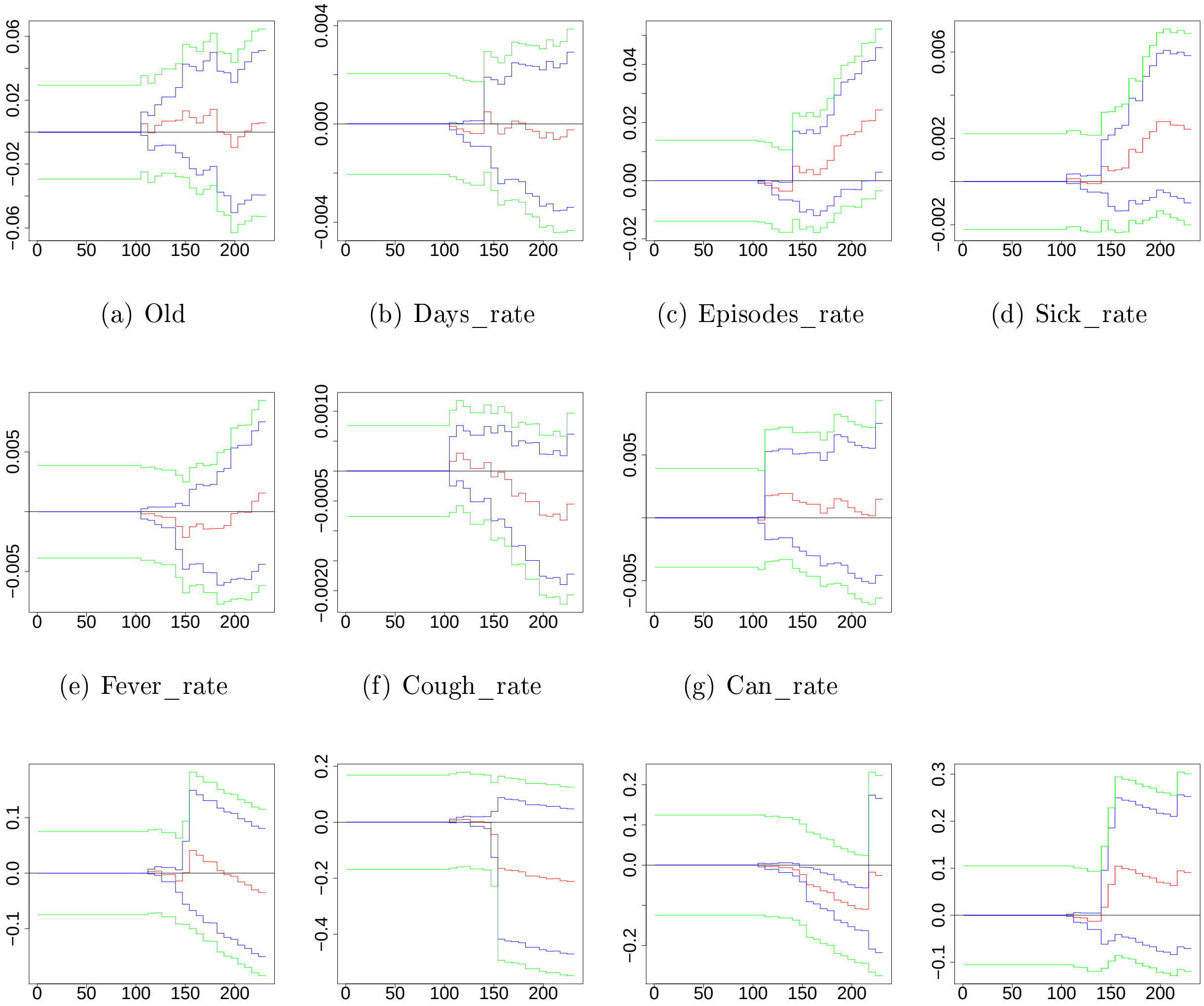
(m) Flooding

(n) Mother_education

(o) Sex

(p) Young

Figure A.2: Cumulative regression functions for the dropout study (continued). The x axis is in days.



(a) Old

(b) Days_rate

(c) Episodes_rate

(d) Sick_rate

(e) Fever_rate

(f) Cough_rate

(g) Can_rate

(h) Lag 1

(i) Lag 2

(j) Lag 3

(k) Lag 4

Bibliography

Aalen, O. O. (1980) A model for nonparametric regression analysis of counting process. *Lecture Notes in Statistics* **2**, 1–25.

Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008) *Survival and Event History Analysis, A Process Point of View*. Springer.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical models based on Counting Processes*. Springer.

Barreto, M. L., Genser, B., Strina, A., Teixeira, M., Assis, A. M. O., Rego, R. F., Teles, C. A., M.S.Prado, Matos, S. M. A., Santos, D. N., dos Santos, L. A. and Cairncross, S. (2007) Effect of city-wide sanitation programme on reduction in rate of chidhood diarrhoea in northeast brazil: assessment by two cohort studies. *Lancet* **370**, 1622–1628.

Billingsley, P. (1968) *Convergence of Probability Measures*. Wiley.

Borgan, Ø., Fiaccone, R. L., Henderson, R. and Barreto, M. (2007) Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in brazil. *Board of the Scandinavian Journal of Statistics* **34**, 53–69.

Dalang, R. (2008) Stochastic processes course. `http://ima.epfl.ch/prob/enseignement/calcul_sto/index.html`.

Elgmati, E. (2009) *Additive Intensity Models for Discrete Time Recurrent Event Data*. Ph.D. thesis, Newcastle University, United Kingdom.

Elgmati, E., Farewell, D. and Henderson, R. (2009) A martingale residual diagnostic for longitudinal and recurrent event data. *Lifetime Data Analysis* .

Elgmati, E., Fiaccone, R., Henderson, R. and Mohammadi, M. (2008) Frailty modelling for clustered recurrent incidence of diarrhoea. *Statistics in Medicine* **27**, 6489–6504.

Fosen, J., Borgan, Ø., Weedon-Fekjaer, H. and Aalen, O. O. (2006) Dynamic analysis of recurrent event data using the additive hazard model. *Biometrical Journal* **48**, 381–398.

Hall, W. J. and Wellner, J. A. (1980) Confidence bands for a survival curve from censored data. *Biometrika* **67**(1), 133–143.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data.* Springer.

Knight, K. (2000) *Mathematical Statistics.* Chapman and Hall.

Kuo, H. (2006) *Introduction to Stochastic Integration.* Springer.

Martinussen, T. and Scheike, T. H. (2006) *Dynamic Regression Models for Survival Data.* Springer.

Panaretos, V. (2008) Theoretical statistics course. `http://smat.epfl.ch/courses/theory.html`.

Steele, J. M. (2003) *Stochastic Calculus and Financial Applications.* Springer.

Strina, A., Cairncross, S., M, S. P., Teles, C. A. S. and Barreto, M. L. (2005) Childhood diarrhoea symptoms, management and duration: observation from a longitudinal cummunity study. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **99**, 407–416.

WHO (2009) World health organization. `http://www.who.int/water_sanitation_health/diseases/diarrhoea/en/`.