Stochastic Process Based Regression Modeling of Time-to-event Data

Application to Phenological Data

by

Song Cai

B.Sc., Peking University, 1999
M.Sc., Peking University, 2002
M.Sc., The University of British Columbia, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2010

Abstract

In agricultural study, the timings of *phenological events*, such as bud-bursting, blooming and fruiting, are considered to be mainly influenced by climate variables, especially accumulative daily average temperatures. We developed a stochastic process-based regression model to study the complicated relationship between phenological events and climate variables, and to predict the future phenological events. Compared with the traditional Cox model, the newly developed model is more efficient by using all available time-dependent covariate information, and is suitable for making predictions. Compared with parametric proportional hazards model, this model is less restrictive on assumptions, and fitting of this model to data is computationally straightforward. Also, this model may be easily extended to incorporate sequential events as responses. It may also be useful for a broad range of survival data in medical study.

This model was applied to the bloom-date data of six high-valued, woody perennial crops in the Okanagan Valley, BC Canada. Simulation results showed that the model provides a sensible way to estimate an important parameter, T_{base}, controlling phenological forcing events. Also, our statistical findings support Scientists' previous experimental findings that the temperature influence blooming events via accumulation of growing degree days (GDDs). Furthermore, a cross-validation procedure showed that this model can provide accurate predictions for future blooming events.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I owe my deepest gratitude to my supervisor, Dr. James V. Zidek, who has guided me and inspired me during the course of my graduate study in statistics, and has given me great help in my thesis writing process.

I am also deeply indebted to my co-supervisor, Dr. Nathaniel Newlands, who has supported and helped me throughout the project.

In addition, I would like to thanks Dr. Denise Neilsen, who has provided data and valuable inputs for this project.

Without them, this thesis would not have been possible.

Chapter 1

Introduction

The main purpose of this thesis is to investigate the dependence of timings of development stages (e.g. bloom dates) of six high-valued, woody perennial crops on local daily average temperature in the Okanagan Valley, British Columbia, Canada, and to forecast the timings of future development events of the crops. The data are *time-to-event data*, but with some special features: *time-dependent covariates* and *sequential multiple events*. Using standard techniques in survival analysis, e.g. *Cox model* and *Accelerated Failure Time model*, we found difficulties dealing with time-dependent covariates, both when estimating parameters and when predicting future events. We then developed a method capable of handling time-dependent covariates using a stochastic process based regression model. This model can be easily extended to incorporate sequential multi-state responses. Although it is originally designed for phenological data, we believe it will also be useful for a broad range of survival data in medical research.

In this chapter, we first introduce the basic concept and features of time-to-event data. Then we give two motivating examples, for analyzing *phenological data* and medical *survival data*. Phenological data are rarely discussed in statistical literature. We identify problems encountered in applying the Cox and parametric proportional hazards models to such data.

1.1 General features of time-to-event data

Time-to-event data are frequently encountered in medical and agriculture science. In cancer studies, one may want to know for how long that a patient is likely to survive after entering the study, and maybe also the association between the survival time and the age of the patient. When studying the development of a perennial plant, one is interested in the dates at which the plant reaches certain development stages in a year, e.g. blooming and fruiting, and how such timings relate to climatic conditions. The survival time in the first example is called survival data in medical research, and in the second example, the times to yearly periodic events of a plant along with related climatic covariates are known as phenology.

Both data types similarly characterize the duration of the time starting from a time point to the occurrences of response events of interest. Time is distinct from other variables. Other

variables are measured almost instantaneously, and usually independent of the response size, while in time-to-event data, the largest observations require more time to observe than other observations; also, time is observed sequentially (Hougaard, 2000). These characteristics make time hard to analysis. However, time-to-event data are critically important to many fields, so have been given increasing attention in the last few decades and may analysis techniques for them have been invented. In this section, we introduce some basic concepts and special features of time-to-event data.

1.1.1 The conditioning concept

Time perpetually moves forward. Thus inferences about the future are always changing, because the information we have up to the current time changes.

Consider the following two situations. In a cancer study, the estimated median survival time of patients is 98 days after entering the study in control group. Now if a patient in control group has survived for 90 days, should he expect that just eight days remain in his life, based on the estimated median? In an agricultural study, a scientist analyzed bloom-date data, and estimated the mean bloom date of apple trees as April 16th. Now, it is April 12th, and an apple tree hasn't bloom yet. Does it mean we could reasonably expect it to bloom after four days according to the estimated mean?

The answers to both above questions are "No"! The key is the concept of *conditioning*. The variables of interest in the above examples are times to some events that start from well-defined time origins of reference. In the cancer study example, the time origin is the moment that a patient enters the study, and in the agricultural study, the first day of a year. We denote the random variable (r.v.) associated with the time to event as variable T^*, and the probability distribution function of T as $\mathrm{F}(t) = \mathrm{P}(T \leq t)$. The distribution functions of times to events are estimated, and the median/mean of the distributions are calculated. However, as time passes by and we get more information about T, for example the event has not occurred up to the current time t_0, what shall we say about the distribution of T now, based on this new information? The answer is the conditional distribution of T given $T > t_0$:

$$\mathrm{F}(t \,|T > t_0) = \mathrm{P}(T \leq t \,|T > t_0)\,, \tag{1.1}$$

which is different from $\mathrm{F}(t)$.

Since in the analysis of time-to-event data, these kind of questions are far more common

*In this thesis, the names of (random or non-random) variables will be denoted as capital letters and their values will be denoted as lower-case letters.

than simply asking for the unconditional distribution of time, people have defined the *survivor function* and the *hazard function* to characterize this conditioning concept.

1.1 Definition. The *survivor function* of a random variable T, denoted as $S(t)$, is the probability that T exceeds a real number t:

$$\mathrm{S}(t) = \mathrm{P}(T > t),\ t \in \mathbb{R}\,. \tag{1.2}$$

Clearly, survivor function is right-continuous and decreasing.

1.2 Definition. The *hazard function* of a random variable T is defined as follows:

$$h(t) = \lim_{\delta \to 0} \frac{\mathrm{P}(t \le T < t + \delta \,|\, T \ge t)}{\delta}\,. \tag{1.3}$$

For an absolutely continuous random variable T, the hazard function is understood as the "instantaneous probability" of an event occurring within a short time interval after t, given that it hasn't occurred prior to the beginning of the interval.

The following properties of the hazard function and the survivor function hold for a continuous random variable T:

$$h(t) = \frac{f(t)}{\mathrm{S}(t)} = -\frac{\mathrm{d}\log \mathrm{S}(t)}{\mathrm{d}t}\,; \tag{1.4}$$

$$\mathrm{S}(t) = \exp\left[-\int_0^t h(u)\,\mathrm{d}u\right]\,; \tag{1.5}$$

$$f(t) = h(t)\exp\left[-\int_0^t h(u)\,\mathrm{d}u\right]\,, \tag{1.6}$$

where $f(t)$ is the probability density function (PDF) of T. For convenience, we define the *cumulative hazard function* of T as:

$$\mathrm{H}(t) = \int_0^t h(u)\,\mathrm{d}u\,. \tag{1.7}$$

Substituting (1.7) into (1.5) and (1.6), we have

$$\mathrm{S}(t) = \exp(-\mathrm{H}(t))\,, \tag{1.8}$$

and

$$f(t) = h(t)\exp(-\mathrm{H}(t))\,. \tag{1.9}$$

An advantage of the hazard function over the density function is that it does not change form under conditioning. This is illustrated by Table 1.1 (Hougaard, 2000).

Table 1.1: The effects of conditioning on $T > t_0$ on distributional quantities.

Quantity	Under full distribution	Under conditional distribution given $T > t_0$
Density function	$f(t)$	$f(t)/\mathrm{S}(t_0),\ t > t_0$
Survivor function	$\mathrm{S}(t)$	$\mathrm{S}(t)/\mathrm{S}(t_0),\ t > t_0$
Hazard function	$h(t)$	$h(t),\ t > t_0$

For a discrete random variable, according to Equation 1.3, the hazard function simplifies to:

$$h(t_k) = \mathrm{P}(T = t_k | T \geq t_k) = \frac{\mathrm{P}(T = t_k)}{\mathrm{S}(t_k^-)}\ , \tag{1.10}$$

and properties (1.5) and (1.6) become:

$$\mathrm{S}(t) = \prod_{k \in \{k: t_k \leq t\}} (1 - h(t_k))\ , \tag{1.11}$$

$$\mathrm{P}(T = t_k) = h(t_k) \prod_{i=1}^{k-1} (1 - h(t_i))\ . \tag{1.12}$$

Since the hazard function and the survivor function provide a convenient way to express conditioning, nowadays many popular data analysis techniques in survival analysis are based on modeling them.

1.1.2 Censoring

Clearly, time to an event can only be known after the event has occurred. Monitoring the status of the event is therefore necessary. Unfortunately, a subject can rarely be constantly monitored due to the limited resources and other reasons beyond the investigator's control, e.g. a patient intentionally leaves a medical study early. Typically, a subject is checked at regular intervals for the whole period of study. An event, then, may occur between two check points, or may not occur before the end of the study. In both cases, if we consider time as a continuous random variable, we do not observe the exact time to the event, but we get partial information about it. In these situations, we say that the data are *censored*.

We call an observation *interval censored* if the event occurs between two check points, and *right censored* if it has not occurred by the end of the study. We emphasize that censoring, as a special type of missing data, is an important feature we need to pay attention to when analyzing time-to-event data.

1.1.3 Distributions of the time responses

Recall that when we measure the time to an event, we usually have a well-defined time origin. Without loss of generality, we may fix this time origin as zero and assume as time passes by, its value becomes larger. Then in such a measurement system, the evaluated time T is non-negative and non-symmetric. Mathematically, T cannot be normally distributed. Nevertheless, under certain circumstances, T's distribution might be reasonably well modelled as normal.

In medical research, survival data distributions are not usually considered as normal. In classical parametric models, the exponential, the Weibull and the Gompertz distribution are often preferred choices (Hougaard, 2000). On the other hand, non-parametric models are more popular now since they allow a more modeling flexibility by avoiding distributional assumptions, although at the cost of losing efficiency.

1.2 Two examples of time-to-event data

In this section we give two examples of time-to-event data with special features, for phenological data and survival data.

1.2.1 Example one: Development stages of apple trees

Scientists in agricultural science are interested in the timings of annual biological events of perennial crops, such as the timings of bud-bursting, flowering, fruiting, etc., in relation to climate variables, such as daily average temperature. The occurrences of these events indicate the development stages of crops. Studying the relationships between them and climate variables help scientists to understand how the crops respond to future climate change so that they can take actions now to increase crop production. We will give a detailed introduction to phenological data in Chapter 3. Now let's look at an example of this type of data.

Representative bloom dates of apple trees from 1937 to 1964 in the Okanagan area, British Columbia, Canada are recorded, as shown in Table 1.2. The dates are counted as the number of days from the first day of a year (t_0 = January 1^{st}) to a representative bloom time of all the apple trees in the area in that year. Experimental phenologists consider that there is little relationship between the bloom dates of any two years, and the main factor that influences the bloom date of a year is a quantity called *growing degree day* (GDD) (D. Neilsen, Agriculture and Agri-Food Canada (AAFC), Personal Communication; Murray et al., 1989), which is a

Table 1.2: Representative bloom dates of apple trees from 1937 to 1964 in the Okanagan area, BC, Canada

Year	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946
Bloom Date	136	127	122	121	115	126	133	132	132	127
Year	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956
Bloom Date	122	142	128	139	132	131	127	137	146	132
Year	1957	1958	1959	1960	1961	1962	1963	1964		
Bloom Date	127	125	132	126	129	131	132	134		

function of time defined as follows:

$$GDD(t) = \begin{cases} \frac{T_{min}(t)+T_{max}(t)}{2} - T_{base} & \text{if } \frac{T_{min}(t)+T_{max}(t)}{2} > T_{base} \\ 0 & \text{Otherwise} \end{cases}, \text{ for } t \geq t_0 , \tag{1.13}$$

where t stands for discrete time with the unit of day, $T_{min}(t)$ and $T_{max}(t)$ are daily minimum and maximum temperatures, and T_{base} is a thresholding constant temperature. In this thesis, we will refer to $\frac{T_{min}(t)+T_{max}(t)}{2}$ as *daily average temperature*. GDD, defined in this way, is a discrete function of time measured on daily basis. Loosely speaking, GDD is a measure of energy that is available to contribute to the occurrence of the bloom event in each day. Only when daily average temperature is above some thresholding temperature, will it influence the time of blooming. The time origin t_0 is usually chosen to be the starting date of a development stage Chuine (2000). As we can see from (1.13), this date is actually controlled by T_{base}. If we choose some date earlier than the time origin, the daily average temperatures in those earlier days will be smaller than T_{base}, and the corresponding GDDs will be 0. Therefore, that choice will not impact the results of analysis. For bloom date, the choice of January 1^st may be far earlier than the starting date of the blooming stage, and so it's impact on results may be negligible. In the analysis, we may verify this by comparing the estimated T_{base} to observed daily temperatures in January.

Scientists are interested in answering the follow questions (N. Newlands and D. Neilsen, AAFC, Personal Communication):

(1) Is there any statistical relationship between GDDs and bloom date?

(2) If the answer to the first question is yes, how does GDDs relate to bloom dates? Is the GDD evaluated only on date of bloom related to the response, are the GDDs evaluated on several days prior to bloom date all related to the bloom date, or are accumulated effect of a much longer history of GDDs before blooming?

(3) What is the value of the thresholding temperature T_{base}?

(4) Based on the answers to the above questions, is GDD useful for predicting future bloom date, and how to predict it?

For statisticians, one possible way to answers these questions is to build a regression model with bloom date as the response and a function of GDD as the covariate. In this setting, the first problem is a hypothesis testing problem for regression coefficients, the second one may be solved by model selection, the third one is related to parameter estimation, and the last one is a prediction problem.

Such a regression model has the special feature that time-to-bloom event is the response. But it has one more special feature as well: the covariate is a function of time since GDD is a function of time. As we will see later, this will have consequences when formulating likelihood functions and estimating parameters in many existing models. It also make prediction more difficult. Suppose that we have built a regression model and all model parameters have been estimated. We now are on the position to predict the times for future events. Imagine that we are on the last day of a year, and want to predict the bloom date for next year using the fitted regression model. Here, we must know the temperatures of next year first, since the covariate in the regression model depends on the future time, and yet this is not enough. Even if we could predict the daily average temperature of the whole year, we might still not be able to directly apply the regression model, because the covariate may (and most likely will) depend on the GDD on the bloom date, the variable we are trying to predict!

In the above example, we have a single event, the bloom date, as response. Real situations are usually much more complicated. Each year, apple trees go through a sequence of biological events, e.g. bud-bursting, leafing, blooming, fruiting, etc. And blooming is just one of them. These events are all important indicators of the development stages of these trees. Moreover, they occur sequentially. That is, an event will not occur unless the event just before it in the list has occurred, e.g. an apple tree will not bear a fruit if it doesn't bloom. This feature presents additional challenges, and we seek fa regression model to deal with this and the time related covariate simultaneously.

1.2.2 Example two: Lung cancer and smoking

This example is hypothetical. A researcher wants to study the life lengths of patients who have lung cancer and smoke (or have a history of smoking), and their relation to the intensity of smoking, which is measured by the number of cigarettes each patient has per day. The data are

collected under the supervision of statisticians to validate the assumption of independence of responses. Now we are at the stage of analyzing the data.

One way of analyzing such data is to build a regression model with time to death as response and intensity of smoking as covariate. Again, the covariate is a function of time. We have the same problems about estimation and prediction as mentioned in the bloom-date data. However, in this example, the issue about how the covariate should be included is clearer. We know that it is unreasonable to assume the risk of death at time t is only related to the intensity of smoking at that time. A reasonable assumption should be that the response is related to the smoking history of the patient. Imagine that a patient had smoked for 100 days after entering the study, and in the 101st day, he quited smoking. Does this imply that the risk of death in 101st day goes to the level of a patient who never smokes? Obviously, the answer is "No". In fact, the effect of smoking may last for quite a long time. Nevertheless, this reasonable assumption – the risk of death relates to the covariate evaluated at current and previous times, will introduce more difficulty in modeling.

Like the example of development stages of apple trees, the responses in this example could be the times to the occurrences of several sequential events. For example, they could be times to four stages of lung cancer, which are defined according to how far the cancer has spread, or a sequence of declining health conditions.

1.3 Summary of problems and review of some existing models

In both of the above examples, covariates are functions of time and refer to as *time-dependent covariates*. According to the context, they could have different impacts in a statistical model when the response is time. Various aspects of time-dependent covariates has been considered (Collett 2003; Hougaard 2000; Cox and Oakes 1984; Kalbfleisch and Prentice 2002). Here focus on the type associated with our examples – covariates that vary with time (may or may not depend on the individual or unit we are studying) and whose values are not known in advance at any future time. We will discuss the difficulties this introduces to parameter estimation and prediction. Also, we will confine our discussion to the Cox model and the parametric proportional hazards models, as they are widely-applicable models for dealing with time-dependent covariates.

1.3.1 Difficulties in estimation

First we will introduce some basics about the *proportional hazards model*. Suppose we are to investigate the time to an event T in relation to the change in a covariate X which is not a function of time. Assume T is an absolutely continuous random variable. The proportional hazards model is based on modeling the hazard function T defined in Equation (1.3). It assumes that different covariate values correspond to different hazard functions, and the ratio of any two survivor functions evaluated at different covariate values is an explicit function only of the two covariates, but not of time:

$$\frac{h(t, x_1)}{h(t, x_2)} = \Psi(x_1, x_2) \ , \tag{1.14}$$

where x_1 and x_2 are any two covariate values, $h(\cdot)$ is the hazard function, and $\Psi(\cdot)$ is a function of the covariate with range $[0, \infty)$.

The denominator on the left-hand side (LHS) of Equation (1.14) is called the *baseline hazard function*, and could be (in terms of x_2) an arbitrary unspecified hazard function of T, since the above assumption holds for any two covariate values. Practically, it is usually chosen to be the hazard function that corresponds to a zero covariate value. With such a choice, the argument x_2 disappears in the baseline hazard, and Ψ becomes a function of only x_1. We can then omit the subscript of x_1 in Equation (1.14) without any confusion and denote the baseline hazard as $h_0(t)$, and the model is now:

$$\frac{h(t,x)}{h_0(t)} = \Psi(x) \ . \tag{1.15}$$

If we assume Ψ to be an exponential function of a linear combination of the covariate and a set of parameters, we get the popular proportional hazards model:

$$\frac{h(t,x)}{h_0(t)} = \exp(\beta_0 + \beta_1 x) \ , \tag{1.16}$$

β_0 and β_1 being fixed but unknown parameters.

The covariate in the above settings can be replaced by a vector of covariates with $x = (x_1, x_2, \cdots, x_k)^T$, k being a positive integer.

When the covariate in above model is *not* time dependent, the ratio $h(t,x)/h_0(t)$ would be constant over time. However, when it is time-dependent, we make a further assumption that *the response T only relates to the value of covariates evaluated at the current time*. We can

then directly adapt Equation (1.15) as,

$$\frac{h(t,x(t))}{h_0(t)} = \Psi(x(t)) \ . \tag{1.17}$$

In this case, the hazard ratio is no longer a constant with respect to time, and thus the hazard is no longer "proportional". Therefore some authors (e.g. Kalbfleisch and Prentice 2002) prefer calling this type of model a *relative risk model* to encompass both types of covariates.

Parametric proportional hazards model

Now we assume that at baseline, the random variable T comes from a known family of distributions but with unknown distributional parameters. We also keep all the assumptions made above about the proportional hazards model.

When the covariate is not a function of time, Equations (1.15), (1.6) and (1.7) imply that we can deduce the density function of T when the covariate takes value x as follows (assuming time origin is 0):

$$f(t \mid x) = h_0(t)\Psi(x)\exp\left[-\int_0^t h_0(u)\Psi(x)\,\mathrm{d}u\right] \tag{1.18}$$

$$= h_0(t)\Psi(x)\exp\left[-\Psi(x)\mathrm{H}_0(t)\right] \ . \tag{1.19}$$

We see that since $\Psi(x)$ is not a function of time, it can be pulled out of the integral in Equation (1.18). Then the whole integral in the exponential part becomes the product of Ψ and cumulative hazard at baseline H_0. Since the form of h_0 and H_0 are usually known by distributional assumptions, the expression of f is also known. The likelihood function is therefore easy to calculate.

However, if the covariate is time-dependent, the density function becomes

$$f(t \mid x) = h_0(t)\Psi(x(t))\exp\left[-\int_0^t h_0(u)\Psi(x(u))\,\mathrm{d}u\right] \ . \tag{1.20}$$

The expression does not have a simple form if the covariate is a not a simple function of time. If the response is related to several covariate values evaluated at previous time points, the expression becomes even more complicated. Usually, numerical integration is needed for calculating the likelihood function.

Parametric proportional hazards models may fit the data poorly if the distribution of the response is mis-specified. Non-parametric models, on the other hand, do not make any distri-

butional assumptions on the response distribution, thus applying to a broader range of data than parametric models. Next we discuss the famous Cox model.

The Cox model

Cox (1972) developed an estimation method for non-parametric proportional hazards models which soon became very popular. It is known as the *Cox model*. That model also assumes the proportional hazards described by (1.15) and the independence of the observations. When there are no ties and no censorings in the observations, it can be shown by successive conditioning that the likelihood function is

$$\mathrm{L} = \prod_{i=1}^{N} \frac{h(t, x_i)}{\sum_{k \in \mathscr{R}(\tau_i)} h(t, x_k)} \overset{(1.15)}{=} \prod_{i=1}^{N} \frac{h_0(t)\Psi(x_i)}{\sum_{k \in \mathscr{R}(\tau_i)} (h_0(t)\Psi(x_k))} = \prod_{i=1}^{N} \frac{\Psi(x_i)}{\sum_{k \in \mathscr{R}(\tau_i)} \Psi(x_k)}, \quad (1.21)$$

where τ_i is the i^{th} value of ordered event times of the N observations ($\tau_1 < \tau_2 < \cdots < \tau_N$), and $\mathscr{R}(\tau_i)$ is the so-called *risk set* which consists of the observations whose death times are equal to or larger than τ_i.

We see that the baseline hazard cancels out, and so does not appear in the final expression of the likelihood function. This is a major contribution by Cox: obtaining consistent estimators of regression parameters, while allowing the baseline hazard to be a nuisance parameter and entirely unspecified. Later, this technique was generalized and the likelihood called the *partial likelihood* (Cox and Oakes, 1984).

An advantage of the Cox model is that when the covariate is time-dependent, the form of the likelihood does not change, i.e. one just needs to replace all "x_i" terms in Equation (1.21) to "$x_i(t)$", if it is assumed that only the covariate value evaluated at the current time relates to the response. The likelihood does not need integration of hazard or covariate over time. Nonetheless, in doing so, for each observed individual response, we are throwing away all the information contained in the corresponding covariate values observed before the event time of this individual but not at the event times of other individuals (if the observation of these covariate values are available). Thus, the efficiency is lower than it is for the parametric models.

If the response relates to several historical values of the covariate, we could just write them as a vector and plug the result into (1.21).

For the above reasons, the Cox model is preferred for estimation especially when time-dependent covariates are present, although it is less efficient than parametric models.

1.3.2 Problems in prediction

Although the Cox model has certain advantages in estimation, it is not suitable for prediction, even when covariates are not time-dependent. This is because it does not estimate the baseline hazard function, so the distribution of the response cannot be calculated. Breslow (1972) provided the maximum likelihood estimator of the baseline hazard, and Kalbfleisch and Prentice (1973) gave an alternative estimator for the baseline hazard based on the marginal likelihood obtained from the marginal distribution of the ranks of the observations. However, both methods only allow the hazard function of the response within the range of observed event times, i.e. for $t \leq \max\{t_1, \cdots, t_N\}$ where $t_1, \cdots, t_N$ are observed event times, to be estimated. In other words, they do not extrapolate beyond the last observation. Therefore, for prediction, parametric proportional hazards models are usually the choice.

In a parametric proportional hazards model, when time-dependent covariates are present, to predict the future events using a fitted model, we need to known the future values of these covariates in advance. Here we assume that they are either given as a function of time or predicted by other models. Now, the problem appears just as in the estimation: when calculating the predictive density of a unknown event time, there may be a complicated integral to evaluate.

1.3.3 Summary

For a single event, we have seen the problems of time-dependent covariates in the popular Cox model and parametric hazards models. If the responses are sequential multi-state events, the problems become more complicated. We aim to develop a better model to handle time-dependent covariates for phenological data. Also, we aim to extend that model to account for sequential multi-state responses.

Chapter 2

Stochastic Process Based Regression Model

We have seen that the Cox model has difficulty in prediction, and when time-dependent covariates are present, parametric proportional hazards models are computationally expensive and in some cases intractable. For the cancer data and the phenological data described in Chapter 1, when we are interested in predicting the future events, these models may not be good choices. In this chapter, we introduce a new approach for this type of data based on successive conditioning of the status of an event (occurred or not occurred) at a time point on the status of the event at all the previous time points. It turns out that for a specific type of time-to-event data where an event occurs only once for an individual, if we take the status of an event as a binary response variable at each time point, then this response over time is a Markov process. This Markovian structure will simplify the probability model for the data, and also make the prediction easy to carry out. We will first discuss this approach for a single event with time-dependent covariates in detail in Section 2.1 to 2.4. Then we will extended it to incorporate sequential events as responses in Section 2.5.

2.1 Data and notation

Suppose that there are N individuals under consideration, labelled as $i = 1, \cdots, N$. An individual could be a patient with lung cancer or an apple tree for example. An event will occur for each individual, e.g. death of a patient or blooming of an apple tree, and for each individual, we classify the status of the event as "not occurred" and "occurred". We consider the following type of single event:

2.1 Assumptions. Only one event can occur for each individual, and once that event occurs for an individual, it will stay in the "occurred" status forever.

Starting from a well-defined time origin (assuming $t_0 = 0$), for the i^{th} individual, we write the time to the occurrence of the event as T_i. In practice, we usually measure T_i on a discrete

time scale, so it is a positive integer with some time unit, e.g. day or week, etc. For each individual i, at each discrete time point $t = 0, 1, \cdots$, we create a dummy variable $Y_{i,t}$ to indicate the status of the event, and we let $Y_{i,t}$ be 0 if the status of the event is "not occurred" at time t and 1 otherwise. With this notation, for the i^{th} individual, time to event T_i and dummy variable $Y_{i,t}$ have the following relation:

$$T_i = t, \text{ where } t \text{ satisfies } Y_{i,0} = 0, \cdots, Y_{i,(t-1)} = 0, Y_{i,t} = 1, Y_{i,(t+1)} = 1, \cdots, \tag{2.1}$$

or equivalently,

$$Y_{i,0} = 0, Y_{i,1} = 0, \cdots, Y_{i,(T_i-1)} = 0, Y_{i,T_i} = 1, Y_{i,(T_i+1)} = 1, \cdots. \tag{2.2}$$

Since we have assumed that once an event occurs it will stay in the "occurred" state forever, $Y_{i,t}$ is 1 for all $t \geq T_i$. Furthermore, for every individual $i = 1, \cdots, N$, suppose there is an associated time-dependent covariate vector which is assumed to affect T_i. It is observed on a discrete time scale, and we write its value at time point t ($t = 0, 1, \cdots$) as $X_{i,t}$. Also, there might be a covariate vector which is not time-dependent, e.g. gender of a patient. However, a fixed covariate can be viewed as a time-dependent covariate with constant values over time, so to simplify the notation, we write $X_{i,t}$ as a unified notation for both types of covariates.

In practice, we can only monitor the occurrence of the event for a limited period of time. If the event has not occurred for an individual by the end of the period or the individual left the study during that period before the event occurs, we won't be able to know the exact time to the event for that individual. However, we know that the time to the event is larger than the time of completion of the study period or the time the individual leaves the study. In a survival analysis context, these are called *right censoring*. Right censoring introduces complications, and we need some extra notation to describe it. We will get into this in Section 1.1.4. As a first step, in the following section, we describe our model by assuming no censoring.

2.2 A probability model for single event

For an individual, there are known and unknown reasons for the occurrence of the event. We may partially explain the occurrence of the event by known reasons, but that explanation won't be exact due to the error caused by the presence of unknown reasons. For this reason, in statistical modeling, we usually treat the occurrence of an event as random, and use a probability model to characterize it. Here, we treat the event time T_i for individual i and the dummy indicator variable $Y_{i,t}$ for individual i at each time point t as random variables. Following convention,

we denote the value of T_i as t_i, where $t_i \in [0, \infty)$, and the value of $Y_{i,t}$ as $y_{i,t}$, where $y_{i,t}$ takes value 0 or 1. For the i^{th} individual, the covariate evaluated at time point t, $X_{i,t}$, may be considered as fixed or random depending on how it is measured. In the lung cancer example of Chapter 1, if the covariate is the prescribed amount of dose a patient needs to take every day, then we should treat it as fixed, since we know the exact value of it each day from the beginning of the study. However, in the blooming date example, if the covariate is the temperature, it might be more reasonable to treat it as a random variable, since the temperature is usually measured with error. Our main interest is to answer the question "what is the conditional probability of an event occurring at certain time given covariate values", but not to study the joint probability of response and covariate, so we only need to consider conditional probability of an event occurring at a time point given these covariate values. In this case, treating them as random or fixed is the same for our estimation problem. We will treat covariates as random in this thesis, but whenever it is appropriate to treat them as fixed, we can use the same estimation procedure described below. As before, we use lower cases $x_{i,t}$ to represent the observed value of $X_{i,t}$.

Consider the data and notation described in Section 2.1. To simplify notation, for individual i, we denote the covariate evaluated at all time points $t' = \cdots, -1, 0, 1, \cdots$, i.e. $\{\cdots X_{i,-1} = x_{i,-1}, X_{i,0} = x_{i,0}, X_{i,1} = x_{i,1}, \cdots\}$, by $X_{i,t'\in\mathbb{Z}}$. Similarly, for individual i, we write $Y_{i,0:t}$ as the dummy indicator $Y_{i,t}$ evaluated from time origin 0 to some time point t ($t = 0, 1, \cdots$), i.e. $\{Y_{i,0} = y_{i,0}, \cdots, Y_{i,t} = y_{i,t}\}$. Then the conditional probability of $Y_{i,0:t}$ given $X_{i,t'\in\mathbb{Z}}$ is

$$\mathrm{P}\left(Y_{i,0:t} \left| X_{i,t'\in\mathbb{Z}}\right.\right) = \mathrm{P}\left(Y_{i,0} = y_{i,0} \left| X_{i,t'\in\mathbb{Z}}\right.\right) \prod_{s=1}^{t} \mathrm{P}\left(Y_{i,s} = y_{i,s} \left| Y_{i,0:(s-1)}, X_{i,t'\in\mathbb{Z}}\right.\right) , \tag{2.3}$$

where $\mathrm{P}(\cdot)$ is the probability set function. Since every conditioning expression is based on all the $Y_{i,t}$ values starting from time origin 0, this equation is not simple enough to use in practice. However, Assumptions 2.3.1 means that we are only interested in the single event that will occur once and once it has occurred, will stay in "occurred" status forever. Under this condition, we have the following result, which will help us to simplify the expression.

2.2 Proposition. *For each individual i, for the single event considered in Assumptions 2.1, the stochastic process* $\{Y_{i,t} : t = 0, 1, \cdots\}$ *is a first order Markov chain, i.e.*

$$P\left(Y_{i,t} = y_{i,t} \left| Y_{i,0:(t-1)}\right.\right) = P\left(Y_{i,t} = y_{i,t} \left| Y_{i,(t-1)} = y_{i,(t-1)}\right.\right) , \tag{2.4}$$

for all $t = 1, 2, \cdots$ *and* $y_{i,t} \in \{0, 1\}$.

Proof. Since for each individual i and all $t = 0, 1, \cdots, Y_{i,t}$ only takes value 0 or 1, it suffices

to consider two cases: $Y_{i,(t-1)} = 0$ and $Y_{i,(t-1)} = 1$, separately. First, $Y_{i,(t-1)} = 0$ implies that $Y_{i,0} = 0, \cdots$, and $Y_{i,(t-2)} = 0$, so when $Y_{i,(t-1)} = 0$, $\{Y_{i,0} = 0, \cdots, Y_{i,(t-2)} = 0, Y_{i,(t-1)} = 0\}$ is the only possible probability event for $Y_{i,0:(t-1)}$, and $\{Y_{i,(t-1)} = 0\}$ is equivalent to $\{Y_{i,0} = 0, \cdots, Y_{i,(t-2)} = 0, Y_{i,(t-1)} = 0\}$. Secondly, for the type of single event under consideration, if for some $t' > 0$, $Y_{i,(t'-1)} = 1$, then $Y_{i,(t-1)} = 1$ for all $t \geq t'$. Thus, when $Y_{i,(t-1)} = 1$ ($t > 0$), we have

$$\begin{aligned} \mathrm{P}\left(Y_{i,t} = y_{i,t} \left| Y_{i,0:(t-1)} \right.\right) &= \mathrm{P}\left(Y_{i,t} = y_{i,t} \left| Y_{i,0:(t-2)}, Y_{i,(t-1)} = 1 \right.\right) \\ &= \mathrm{P}\left(Y_{i,t} = 1 \left| Y_{i,(t-1)} = 1 \right.\right) \\ &= 1 \end{aligned} \tag{2.5}$$

□

When the covariates' values at all time points $X_{i,t' \in \mathbb{Z}}$ are given, we have a similar result:

2.3 Corollary. *For each individual i, for the single event considered in Assumptions 2.1, conditioned on $X_{i,t' \in \mathbb{Z}}$, the stochastic process $\{Y_{i,t} : t = 0, 1, \cdots\}$ is a first order Markov chain, i.e.*

$$P\left(Y_{i,t} = y_{i,t} \left| Y_{i,0:(t-1)}, X_{i,t' \in \mathbb{Z}} \right.\right) = P\left(Y_{i,t} = y_{i,t} \left| Y_{i,(t-1)} = y_{i,(t-1)}, X_{i,t' \in \mathbb{Z}} \right.\right) , \tag{2.6}$$

for all $t = 1, 2, \cdots$ and $y_{i,t} \in \{0, 1\}$.

Proof. The proof is just a re-work of the proof of Proposition 2.2, with everything conditioned on $X_{i,t' \in \mathbb{Z}}$. The details are omitted. □

Using Corollary 2.3, we can simplify Equation (2.3) to

$$\mathrm{P}\left(Y_{i,0:t} \left| X_{i,t' \in \mathbb{Z}} \right.\right) = \mathrm{P}\left(Y_{i,0} = y_{i,0} \left| X_{i,t' \in \mathbb{Z}} \right.\right) \prod_{s=1}^{t} \mathrm{P}\left(Y_{i,s} = y_{i,s} \left| Y_{i,(s-1)} = y_{i,(s-1)}, X_{i,t' \in \mathbb{Z}} \right.\right) . \tag{2.7}$$

By this expression and the relationship of T_i and $Y_{i,t}$, Equation (2.2), for individual i, the conditional probability of the event occurring at time t_i given all the covariate values $X_{i,t' \in \mathbb{Z}}$ then

is

$$\begin{aligned}
\mathrm{P}\left(T_i = t_i \left| X_{i,t'\in\mathbb{Z}} \right.\right) &= \mathrm{P}\left(Y_{i,0}=0,\, Y_{i,1}=0,\, \cdots,\, Y_{i,(t_i-1)}=0,\, Y_{i,t_i}=1,\, Y_{i,(t_i+1)}=1,\, \cdots \left| X_{i,t'\in\mathbb{Z}} \right.\right) \\
&= \mathrm{P}\left(Y_{i,0}=0,\, Y_{i,1}=0,\, \cdots,\, Y_{i,(t_i-1)}=0,\, Y_{i,t_i}=1 \left| X_{i,t'\in\mathbb{Z}} \right.\right) \\
&= \mathrm{P}\left(Y_{i,0}=0 \left| X_{i,t'\in\mathbb{Z}} \right.\right) \cdot \left[\prod_{s=1}^{t_i-1} \mathrm{P}\left(Y_{i,s}=0 \left| Y_{i,(s-1)}=0,\, X_{i,t'\in\mathbb{Z}} \right.\right)\right] \cdot \\
&\quad \mathrm{P}\left(Y_{i,t_i}=1 \left| Y_{i,(t_i-1)}=0,\, X_{i,t'\in\mathbb{Z}} \right.\right) \,. \qquad (2.8)
\end{aligned}$$

Now we are ready to build a regression model based upon this probability model.

2.3 Regression model for a single event

2.3.1 Assumptions

We now assume that the occurrences of the events of different individuals are independent realizations from the same population. We require additional assumptions about the relationship of the occurrence of the event and covariate to limit the number of parameters. In Equation (2.8), the probability of an event occurring at time t_i is conditioned on covariate values evaluated at all integer time points $\cdots,\, -1,\, 0,\, 1,\, \cdots$. In real applications, the occurrence of the event usually only depends on the covariate values at and prior to the occurrence time. Furthermore, in some situations, we may assume that at a time point t ($t \geq 0$), the status of the event mainly depends on the covariate values at current time and several previous time points, or a weighted average of covariate values at previous time points. In practice, we want to make some reasonable assumptions so that the total number of covariates (and consequently the total number of parameters in regression model) is limited, and the number of covariates does not change with time. How to achieve this will become clear when we analyze the phenological data.

For the purpose of illustration, simply assume that for individual i at time point t, the status of the event, $Y_{i,t}$ only depends on the covariate values evaluated from time $t-K$ to t, i.e. $\left\{X_{i,(t-K)},\, \cdots,\, X_{i,t}\right\}$, where K is constant. Now, for individual i at time point t, no matter if $X_{i,t}$ is a vector or not, the total number of covariate values that are related to the status of the event $Y_{i,t}$ is finite and fixed, and we will put them together as a vector denoted by $\mathscr{X}_{i,t}$.

2.3.2 Regression model

Under the above assumptions about the relationship between event and covariates, in Equation (2.8), term $X_{i,t'\in\mathbb{Z}}$ on the right hand side of the equation can be replaced by $\mathscr{X}_{i,t}$. We assume

that time origin 0 is the earliest time an event can occur, otherwise the data is not useful for studying the probability of the occurrence of the event. Then we have

$$\mathrm{P}\left(Y_{i,0} = y_{i,0} \,|\mathscr{X}_{i,t}\right) = \mathrm{P}\left(Y_{i,0} = y_{i,0} \,|Y_{i,-1} = 0,\ \mathscr{X}_{i,t}\right) \ . \tag{2.9}$$

By virtue of the Markov property of $\{Y_{i,t} : t = 0,\ 1,\ \cdots\}$, to model $\mathrm{P}\left(T_i = t_i \,\middle|X_{i,t'\in\mathbb{Z}}\right)$, it is sufficient to model $\mathrm{P}\left(Y_{i,t} = y_{i,t} \,\middle|Y_{i,(t-1)} = 0,\ \mathscr{X}_{i,t}\right)$ for $t = 0,\ \cdots,\ t_i$ and $y_{i,t} \in \{0,\ 1\}$. For each fixed individual i, $\mathrm{P}\left(Y_{i,t} = y_{i,t} \,\middle|Y_{i,(t-1)} = 0,\ \mathscr{X}_{i,t}\right)$ is a function of t and $\mathscr{X}_{i,t}$. Now, our purpose is to choose a useful explicit form for this function with unknown parameters, and perform statistical inference on these unknown parameters.

For individual i at time point t, we denote

$$\mathrm{P}_{i,t} \equiv \mathrm{P}\left(Y_{i,t} = 1 \,\middle|Y_{i,(t-1)} = 0,\ \mathscr{X}_{i,t}\right) \ , \tag{2.10}$$

then we have

$$\mathrm{P}\left(Y_{i,t} = y_{i,t} \,\middle|Y_{i,(t-1)} = 0,\ \mathscr{X}_{i,t}\right) = \mathrm{P}_{i,t}^{y_{i,t}} \left(1 - \mathrm{P}_{i,t}\right)^{1-y_{i,t}} \ . \tag{2.11}$$

The right hand side of the above equation is the expression of the Bernoulli distribution. If we want to confine the functional form of $\mathrm{P}\left(Y_{i,t} = y_{i,t} \,\middle|Y_{i,(t-1)} = 0,\ \mathscr{X}_{i,t}\right)$ to be related to a linear function of parameters (as we usually do in linear/generalized linear model), then for individual i, at each time point $t = 0,\ \cdots,\ t_i$, we may consider a linear regression model for binary events, such as logistic or probit regression model. The basic idea can be described as follows.

Introduce a *link function* that is monotone $g : (0,\ 1) \to (-\infty, \infty)$. We assume that $g\left(\mathrm{P}_{i,t}\right)$ equals a linear function of the covariate vector $\mathscr{X}_{i,t}$, i.e.

$$g\left(\mathrm{P}_{i,t}\right) = \beta_t^T \mathscr{X}_{i,t} \ , \tag{2.12}$$

where β_t is a parameter vector which remains the same across different individual i, but may vary with time t. The superscript T means transpose of a vector or matrix, and the length of the vector β_t is the same with $\mathscr{X}_{i,t}$. Note that $\mathscr{X}_{i,t}$ is the design matrix (as in linear regression). Thus if an intercept term is going to be included, then $\mathscr{X}_{i,t}$ should be a vector with 1 as the first component.

From Equation (2.8) – (2.12), for individual i, we have

$$\mathrm{P}\left(T_i = t_i \,\middle|X_{i,t'\in\mathbb{Z}}\right) = g^{-1}\left(\beta_t^T \mathscr{X}_{i,t_i}\right) \prod_{s=0}^{t_i-1} \left(1 - g^{-1}\left(\beta_t^T \mathscr{X}_{i,s}\right)\right) \ , \tag{2.13}$$

where g^{-1} is the inverse function of g. For simplicity, we will take β_t to be a constant vector over time, so the subscript t of β_t in the above equation can be omitted. Under the independence assumption, the likelihood function of the data is

$$L(\beta) = \prod_{i=1}^{N} \left[g^{-1}\left(\beta^T \mathscr{X}_{i,t_i}\right) \prod_{s=0}^{t_i-1} \left(1 - g^{-1}\left(\beta^T \mathscr{X}_{i,s}\right)\right) \right] , \tag{2.14}$$

and the log-likelihood function is

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{N} \left[\log\left(g^{-1}\left(\beta^T \mathscr{X}_{i,t_i}\right)\right) + \sum_{s=0}^{t_i-1} \log\left(1 - g^{-1}\left(\beta^T \mathscr{X}_{i,s}\right)\right) \right] . \tag{2.15}$$

2.3.3 Parameter estimation

Once the basic formulation of the regression model (Equation 2.13) is given, we need to estimate the unknown parameter vector β. For example, given the likelihood function of the data (2.14), one can apply the maximum likelihood (ML) or Bayesian methods for parameter estimation. Here, we focus on the ML method.

The maximum likelihood estimator (MLE) of a parameter vector β is the estimator that maximizes the likelihood function (or equivalently minimizes the negative log-likelihood):

$$\begin{aligned} \hat{\beta}_{MLE} &\equiv \text{Argmax} L(\beta) \\ &= \text{Argmin} - l(\beta) \ . \end{aligned} \tag{2.16}$$

In our case, the likelihood function $L(\beta)$ and the log-likelihood function $l(\beta)$ are given by Equation (2.14) and (2.15).

The MLE is popular since under some mild regularity conditions (Cox and Hinkley, 1979), it has very nice large sample properties, namely consistency, efficiency, and asymptotic normality. The theory about the ML method can be found in many text books, e.g. Cox and Hinkley (1979) and Casella and Berger (2001). Under regularity conditions, the variance of the MLE can be obtained by calculating the inverse of the so-called *Fisher information matrix*. Suppose β is a p-dimensional vector, and we denote its i^{th} component by β_i. Then the Fisher information matrix, $I(\beta)$, is a $p \times p$ matrix, whose j^{th} row and k^{th} column entry ($j, k = 1, \cdots, p$), I_{jk}, is

$$I_{jk} = -E_\beta \left[\frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\beta) \right] , \tag{2.17}$$

where E_β is the expectation with β as the true parameter vector, and $l(\beta)$ is the log-likelihood

function given by (2.15). The covariance matrix of the MLE given that β is the true parameter vector is the inverse of the Fisher information matrix:

$$Cov_\beta\left(\hat{\beta}_{MLE}\right) = I^{-1}(\beta) \ . \tag{2.18}$$

The variance of the i^{th} component of the MLE, $\hat{\beta}^i_{MLE}$, is the i^{th} diagonal entry of $I^{-1}(\beta)$:

$$Var_\beta\left(\hat{\beta}^i_{MLE}\right) = \left[I^{-1}(\beta)\right]_{ii} \ . \tag{2.19}$$

In practice, Equation (2.17) cannot be used to calculate the Fisher information matrix since the true β value is unknown. Even if it were specified as a hypothetical value, the integrals involved may be hard to evaluate. In practice, we may use the *observed information matrix*, $\hat{I}(\beta)$, where

$$\hat{I}_{jk} = -\frac{\partial^2}{\partial\beta_j\partial\beta_k} l(\beta)\Big|_{\beta=\hat{\beta}} \ , \tag{2.20}$$

to approximate the Fisher information matrix $I(\beta)$. Also, Efron and Hinkley (1978) have shown that, in practice, the use of the observed information matrix is actually superior to the Fisher information matrix. After getting the observed information matrix, when calculating the variance of the MLE we can just replace $I^{-1}(\beta)$ with $\hat{I}^{-1}(\beta)$ in Equation (2.18) and (2.19).

An alternative to using the Fisher information matrix to calculate the estimation error is by using resampling methods such as the *bootstrap* (Efron and Tibshirani, 1994). The benifit of the bootstrap is that it does not rely on the strong asymptotic optimality of the MLE, thus we are not subject to the regularity conditions. We will apply the bootstrap in our data analysis.

2.3.4 Prediction

Having estimated the model parameters, we are ready to predict the time to the future event. When time-dependent covariates are present, we must know the future values of them in order to predict the time to an event. We will assume that the future values of covariates are predicted by some other statistical model, and the predictive distributions of these covariates are also given by that model.

Consider a time origin 0, and suppose the current time is t_c ($t_c \geq 0$). For a new individual (an individual that is not used for parameter estimation and the event time is unknown), suppose the event has not occurred up to current time t_c. Now we want to predict the event time for this individual.

Denote the event time for this individual as T^*, and the corresponding dummy status in-

dicator variable at time point t ($t \geq 0$) as Y_t^*. Also, we denote the covariate vector for this individual evaluated at time t as X_t^*. Similarly we write the "new individual version" of $\mathscr{X}_{i,t}$ as $\mathscr{X}_t^*$. Recall that we have assumed that the status of the event Y_t^* is only related to $\mathscr{X}_t^*$, and $\mathscr{X}_t^*$ may involve the covariates values evaluated from time 0 to time t. Since we know the covariate values up to time t_c, $\mathscr{X}_t^*$ is a function of the following two vectors: one vector $\mathscr{X}_{t,obs}^*$ consists of covariates values evaluated from time 0 to time t_c, which we observed exactly, and the other vector $\mathscr{X}_{t,pred}^*$ consists of predicted covariates values from t_c+1 to t, whose predictive distributions are given by another model. Furthermore, denote the MLE of the model parameter vector by $\hat{\beta}$, and the covariates and event status indicator variables used to estimate $\hat{\beta}$ by X^{train} and Y^{train} respectively.

If we knew the true value of β, by Equation (2.13), the predictive distribution of bloom time T^*, i.e. the probability of the event occurring at time point t_c+K for any $K \geq 1$ given $\mathscr{X}_{t,obs}^*$, the observed covariates values for the new individual, would be

$$\begin{aligned}
&\mathrm{P}_\beta\left(T^* = t_c+K \,\middle|\, \mathscr{X}_{t_c+K,obs}^*\right) \\
&= \int \mathrm{P}_\beta\left(T^* = t_c+K \,\middle|\, \mathscr{X}_{t_c+K,obs}^*,\ \mathscr{X}_{t_c+K,pred}^*\right) d\mathrm{P}\left(\mathscr{X}_{t_c+K,pred}^*\right) \\
&= \int g^{-1}\left(\beta^T \mathscr{X}_{t_c+K}^*\right) \prod_{s=1}^{K-1} \left(1-g^{-1}\left(\beta^T \mathscr{X}_{t_c+s}^*\right)\right) d\mathrm{P}\left(\mathscr{X}_{t_c+K,pred}^*\right) \ . \qquad (2.21)
\end{aligned}$$

We attach a subscript β on probability function to emphasize that we are using the true parameter values. The problem is that we do not know the true β. A crude way to estimate the predictive distribution of bloom time T^*, then, is to replace β by $\hat{\beta}$ in Equation (2.21):

$$\begin{aligned}
&\mathrm{P}_{\hat{\beta}}\left(T^* = t_c+K \,\middle|\, \mathscr{X}_{t_c+K,obs}^*\right) \\
&= \int g^{-1}\left(\hat{\beta}^T \mathscr{X}_{t_c+K}^*\right) \prod_{s=1}^{K-1} \left(1-g^{-1}\left(\hat{\beta}^T \mathscr{X}_{t_c+s}^*\right)\right) d\mathrm{P}\left(\mathscr{X}_{t_c+K,pred}^*\right) \ . \qquad (2.22)
\end{aligned}$$

This "plug-in" approach for predictive distribution is generally criticized as failing to take into account the uncertainty of the unknown parameter. But, if one takes the Bayesian approach, the uncertainty of the unknown parameter is incorporated in a natural way. Suppose in an estimation procedure, one takes the Bayesian approach and gets $\mathrm{P}\left(\beta | X^{train},\ Y^{train}\right)$, the posterior

distribution of β. Then the predictive distribution of T^* is:

$$\begin{aligned}
&\mathrm{P}\left(T^* = t_c + K \,\middle|\, \mathscr{X}^*_{t_c+K,obs},\, X^{train},\, Y^{train}\right) \\
&= \int\int \mathrm{P}\left(T^* = t_c + K \,\middle|\, \mathscr{X}^*_{t_c+K,obs},\, \mathscr{X}^*_{t_c+K,pred},\, \beta,\, X^{train},\, Y^{train}\right) \\
&\qquad d\mathrm{P}\left(\beta | X^{train},\, Y^{train}\right) d\mathrm{P}\left(\mathscr{X}^*_{t_c+K,pred}\right) \ . \qquad (2.23)
\end{aligned}$$

One may expect the Bayesian approach will in general be superior to the "plug-in" approach in terms of prediction. However, Smith (1998) showed that for many models, when assessed from the point of view of mean squared error of predictive probabilities, the "plug-in" approach is better than the Bayesian approach in the extreme tail of the distribution. It is not directly clear if this argument fits our model, but the point here is that we think both approaches make sense.

In this thesis, we will take the "plug-in" approach to estimate the predictive distribution of future events, and estimate the uncertainty of the predictive probabilities using the bootstrap method. In practice, we will use *Monte Carlo* (MC) algorithm (Liu, 2001) to approximate the integration in Equation (2.22). We generate a sample of large size L (e.g. thousands or more) from the predictive distribution of $\mathscr{X}^*_{t_c+K,\,pred}$, and denote the sample points as $\mathscr{X}^*_{t_c+K,\,pred}(l)$ ($l = 1, \cdots, L$). Then by MC algorithm, we approximate the predictive probabilities by

$$\mathrm{P}_{\hat{\beta}}\left(T^* = t_c + K | \mathscr{X}^*_{t_c+K,obs}\right) \approx \frac{1}{L}\sum_{l=1}^{L} \mathrm{P}_{\hat{\beta}}\left(T^* = t_c + K \,\middle|\, \mathscr{X}^*_{t_c+K,obs},\, \mathscr{X}^*_{t_c+K,\,pred}(l)\right) \ . \qquad (2.24)$$

In many situations, we may only want to predict the future probability of the occurrence of the event for a limited period of time up to a time point t_E, since after t_E, the probability of the occurrence of the event is small enough to be ignored. For example, very few patient will live longer than 100 years, and if an apple tree has not bloomed in October of a year, it is very likely it's not going to bloom in that year. In this case, we calculate $\mathrm{P}(T^* = t_c + 1), \cdots, \mathrm{P}(T^* = t_E)$, and $\mathrm{P}(T^* > t_E)$. By the choice of t_E, we will have $\mathrm{P}(T^* > t_E) \approx 0$, so it might be reasonable to consider only $\mathrm{P}(T^* = t_c + 1), \cdots, \mathrm{P}(T^* = t_E)$ as the predictive distribution of T^*. Now the time to the event can be predicted according to this predictive distribution. For example, we may want to take the mean or median of the predictive distribution as the predicted time to event.

2.4 Non-informative right censoring

Right censored observations introduced in Section 2.1 complicate parameter estimation. *Non-informative censoring* is when the time to the event is independent of the censoring mechanism. We will deal with non-informative right censoring using a typical method used in survival analysis (see Collett, 2003).

When writing the conditional probability of an event occurring at some time point given covariate values $X_{i,t'\in\mathbb{Z}}$, we will omit the conditioning variable. All the following probability expressions are then conditioned on $X_{i,t'\in\mathbb{Z}}$.

To describe right censored data, we will need some extra notation. For each individual i ($i = 1, \cdots, N$), we have an observed time t_i, which is either an event time, or a right censoring time. We denote this observation as a random variable τ_i, then the value of τ_i is t_i. Now, for individual i, let δ_i be a indicator variable which takes value 1 if we observe the event, and takes value 0 if we observe a right censoring. By the non-informative censoring assumption, we can assume that each individual i is associated with two independent random variables: event time T_i and censoring time C_i. If the observation for individual i is censored, we have

$$C_i < T_i \text{ and } \tau_i = C_i, \text{ when } \delta_i = 0 \,, \tag{2.25}$$

otherwise, we have

$$C_i > T_i \text{ and } \tau_i = T_i, \text{ when } \delta_i = 1 \,. \tag{2.26}$$

Now, it is easy to see that $\tau_i = \min(T_i, C_i)$, and

$$\begin{aligned} \mathrm{P}(\tau_i = t, \delta_i = 0) &= \mathrm{P}(C_i = t, T_i > t) \\ &= \mathrm{P}(C_i = t)\,\mathrm{P}(T_i > t) \,, \end{aligned} \tag{2.27}$$

where the second equality holds because of the non-informative censoring assumption. Similarly, we have

$$\begin{aligned} \mathrm{P}(\tau_i = t, \delta_i = 1) &= \mathrm{P}(T_i = t, C_i > t) \\ &= \mathrm{P}(T_i = t)\,\mathrm{P}(C_i > t) \,. \end{aligned} \tag{2.28}$$

The likelihood function for the observations $t_1, \cdots, t_N$ then is

$$\begin{aligned} L &= \prod_{i=1}^{N} \mathrm{P}\left(\tau_i = t_i,\, \delta_i\right) \\ &= \prod_{i=1}^{N} \left(\mathrm{P}\left(C_i = t_i\right)\mathrm{P}\left(T_i > t_i\right)\right)^{1-\delta_i} \left(\mathrm{P}\left(T_i = t_i\right)\mathrm{P}\left(C_i > t_i\right)\right)^{\delta_i} \\ &= \left[\prod_{i=1}^{N} \mathrm{P}\left(C_i = t_i\right)^{1-\delta_i} \mathrm{P}\left(C_i > t_i\right)^{\delta_i}\right] \left[\prod_{i=1}^{N} \mathrm{P}\left(T_i = t_i\right)^{\delta_i} \mathrm{P}\left(T_i > t_i\right)^{1-\delta_i}\right] . \end{aligned} \tag{2.29}$$

By the non-informative censoring assumption, the term $\left[\prod_{i=1}^{N} \mathrm{P}\left(C_i = t_i\right)^{1-\delta_i} \mathrm{P}\left(C_i > t_i\right)^{\delta_i}\right]$ does not involve parameters that are related to the distribution of event time T_i. Therefore, to find the MLE of the parameters of the distribution of T_i, it suffices to maximize the following function

$$L'\left(\beta\right) = \prod_{i=1}^{N} \mathrm{P}\left(T_i = t_i\right)^{\delta_i} \mathrm{P}\left(T_i > t_i\right)^{1-\delta_i} \ , \tag{2.30}$$

and

$$\hat{\beta}_{MLE} = \mathrm{Argmax} L'\left(\beta\right) \ . \tag{2.31}$$

The term $\mathrm{P}\left(T_i = t_i\right)$ in Equation (2.30) is given by Equation (2.13) (note that the conditioning variable $X_{i,t' \in \mathbb{Z}}$ has been omitted in the current expressions), while term $\mathrm{P}\left(T_i > t_i\right)$ can be calculated as follows:

$$\begin{aligned} \mathrm{P}\left(T_i > t_i\right) &= \mathrm{P}\left(Y_{i,0} = 0,\, Y_{i,1} = 0,\, \cdots,\, Y_{i,t_i} = 0\right) \\ &= \prod_{s=0}^{t_i} \left(1 - g^{-1}\left(\beta_t^T \mathscr{X}_{i,s}\right)\right) \ . \end{aligned} \tag{2.32}$$

Then we can re-write Equation (2.30) as

$$L'\left(\beta_t\right) = \prod_{i=1}^{N} \left[g^{-1}\left(\beta_t^T \mathscr{X}_{i,t_i}\right) \prod_{s=0}^{t_i-1} \left(1 - g^{-1}\left(\beta_t^T \mathscr{X}_{i,s}\right)\right)\right]^{\delta_i} \left[\prod_{s=0}^{t_i} \left(1 - g^{-1}\left(\beta_t^T \mathscr{X}_{i,s}\right)\right)\right]^{1-\delta_i} . \tag{2.33}$$

Now, we can easily estimate β_t using Equation (2.31) if it is assumed constant over time.

2.5 Regression model for sequential events

The above sections are devoted to situations with just one event. However, we frequently encounter several different events which occur in a row with a fixed order for each individual. For example, in Chapter 1, we have mentioned that these could be the different stages of lung cancer of a participant in a cancer study, or different development stages of an apple tree during one development cycle. In this section, we will discuss the modeling of this type of sequential events. The basic methodology will be an extension of the methodology for the single event case we described above.

2.5.1 Data and notation

Suppose we have N individuals under consideration, and S ($S \geq 1$) different events may occur to each individual. We make the following assumptions:

2.4 Assumptions. For each individual, the S events have the following properties:

(1) They occur in a fixed time order;

(2) For an event to occur, all the events prior to it must have occurred.

(3) For a fixed individual, no two different events occur at the same time point.

By these assumptions, we can label each event by the time order in which it occurs, using the symbol s ($s = 1, , \cdots, S$). When we talk about the occurrence of the s^{th} event, all the previous events: $1^{st} - (s-1)^{th}$ event, must have occurred.

Now, for an individual i, there are $S+1$ event statuses: no events have occurred, the first event has occurred but the second hasn't, $\cdots$, the last event have occurred (i.e. all the S events have occurred). We will denote these statuses by $0, 1, \cdots, S$, respectively. For the i^{th} individual, we will denote the random variable, the time to the s^{th} ($s = 1, \cdots, S$) event, as $T_{i,s}$, and denote its value as $t_{i,s}$. We also create a status indicator variable $Y_{i,t}$, which takes value $\{0, 1, \cdots, S\}$, with $Y_{i,t} = l$ ($l = 0, 1, \cdots, S$) indicating that the individual i is in the l^{th} status. For the i^{th} individual, starting from a time origin 0, we consider discrete time points $0, 1, \cdots, t_{i,1}, \cdots, t_{i,2}, \cdots, t_{i,S}$. The time origin 0 satisfies $0 \leq t_{i,1}$. On the other hand, the value of $Y_{i,t}$ can only be l or $l+1$ when $Y_{i,t-1} = l$ ($l = 0, 1, \cdots, S-1$), and $Y_{i,t} = S$ for all $t \geq t_{i,S}$. Then, the event times $\{T_{i,s}\}$ and status indicators $\{Y_{i,t}\}$ have the following relationship:

$$Y_{i,0} = 0, \cdots, Y_{i,t_{i,1}} = 1, Y_{i,(t_{i,1}+1)} = 1, \cdots, Y_{i,\left(t_{i,S}-1\right)} = S-1, Y_{i,t_{i,S}} = S, Y_{i,\left(t_{i,S}+1\right)} = S, \cdots \quad (2.34)$$

Furthermore, we assume that at each discrete time point, there is a covariate vector $X_{i,t}$ and its values at time points $0, \cdots, t_{i,1}, \cdots, t_{i,S}$ are observed.

Under the above notations, we will continue to use $Y_{i,0:t}$ to denote $\{Y_{i,0} = y_{i,0},\ \cdots,\ Y_{i,t} = y_{i,t}\}$, and $X_{i,t'\in\mathbb{Z}}$ to denote $\{\cdots\ X_{i,-1} = x_{i,-1},\ X_{i,0} = x_{i,0},\ X_{i,1} = x_{i,1},\ \cdots\}$, as we did for single event.

2.5.2 Regression model

For the sequential events mentioned above, for each individual i, the conditional probability of $Y_{i,0:t}$ given $X_{i,t'\in\mathbb{Z}}$ still satisfies Equation (2.3). However, the stochastic process $\{Y_{i,t} : t = 0,\ 1,\ \cdots\}$ is no longer necessarily a first-order Markov chain. Instead, the following result holds:

$$\begin{aligned}
&\mathrm{P}\left(Y_{i,t} = y_{i,t} \left| Y_{i,0:(t-1)},\ X_{i,t'\in\mathbb{Z}} \right.\right) \\
&= \begin{cases}
\mathrm{P}\left(Y_{i,t} = y_{i,t} \left| Y_{i,(t-1)} = 0,\ X_{i,t'\in\mathbb{Z}} \right.\right), & \text{if } 0 \le t \le t_{i,1} \\
\mathrm{P}\left(Y_{i,t} = y_{i,t} \left| Y_{i,(t-1)} = l,\ T_{i,1} = t_{i,1},\ \cdots,\ T_{i,l} = t_{i,l},\ X_{i,t'\in\mathbb{Z}} \right.\right), & \text{if } t_{i,l} < t \le t_{i,(l+1)},\ l = 1,\ \cdots,\ S-1 \\
1, & \text{if } t_{i,S} < t\ .
\end{cases}
\end{aligned} \tag{2.35}$$

By this result and Equation (2.3), for each individual i, we have

$$\begin{aligned}
&\mathrm{P}\left(T_{i,1} = t_{i,1},\ T_{i,2} = t_{i,2},\ \cdots,\ T_{i,S} = t_{i,S} \left| X_{i,t'\in\mathbb{Z}} \right.\right) \\
&= \left[\mathrm{P}\left(Y_{i,t_{i,1}} = 1 \left| Y_{i,(t_{i,1}-1)} = 0,\ X_{i,t'\in\mathbb{Z}} \right.\right) \prod_{t=0}^{t_{i,1}-1} \mathrm{P}\left(Y_{i,t} = 0 \left| Y_{i,(t-1)} = 0,\ X_{i,t'\in\mathbb{Z}} \right.\right) \right] \cdot \\
&\quad \left\{ \prod_{l=1}^{S-1} \left[\mathrm{P}\left(Y_{i,t_{i,(l+1)}} = l+1 \left| Y_{i,\left(t_{i,(l+1)}-1\right)} = l,\ T_{i,1} = t_{i,1},\ \cdots,\ T_{i,l} = t_{i,l},\ X_{i,t'\in\mathbb{Z}} \right.\right) \cdot \right.\right. \\
&\quad \left.\left. \prod_{t=t_{i,l}+1}^{t_{i,(l+1)}-1} \mathrm{P}\left(Y_{i,t} = l \left| Y_{i,(t-1)} = l,\ T_{i,1} = t_{i,1},\ \cdots,\ T_{i,l} = t_{i,l},\ X_{i,t'\in\mathbb{Z}} \right.\right) \right] \right\}.
\end{aligned} \tag{2.36}$$

We write

$$\mathrm{P}_{i,t}(l) \equiv \begin{cases}
\mathrm{P}\left(Y_{i,t} = 1 \left| Y_{i,(t-1)} = 0,\ X_{i,t'\in\mathbb{Z}} \right.\right), & \text{if } l = 0 \\
\mathrm{P}\left(Y_{i,t} = l+1 \left| Y_{i,(t-1)} = l,\ T_{i,1} = t_{i,1},\ \cdots,\ T_{i,l} = t_{i,l},\ X_{i,t'\in\mathbb{Z}} \right.\right), & \text{if } l = 1,\ \cdots,\ S-1\ .
\end{cases} \tag{2.37}$$

Since conditioned on $Y_{i,(t-1)} = l$, $Y_{i,t}$ can only take value l or $l+1$, once we get a model for $\mathrm{P}_{i,t}(l)$ for $l = 0, \cdots, S-1$, we can model every term in Equation (2.36).

Compared with the expression of $\mathrm{P}\left(T_i = t_i \left| X_{i,t'\in\mathbb{Z}} \right.\right)$ for a single event (Equation (2.8)),

Equation (2.36) is much more complicated. In the single event case, the Markov property states that, for each individual i, in order to model $\mathrm{P}\left(T_i = t_i \left| X_{i,t'\in\mathbb{Z}} \right.\right)$, it suffices to model conditional probability $\mathrm{P}\left(Y_{i,t} = y_{i,t} \left| Y_{i,(t-1)} = 0,\ \mathscr{X}_{i,t} \right.\right)$, which is a function of only t and $\mathscr{X}_{i,t}$. However, now we need to model $\mathrm{P}_{i,t}(l)$ for $l = 0, \cdots, S-1$, which is a function of not only t and $X_{i,t'\in\mathbb{Z}}$, but also $t_{i,1}, \cdots, t_{i,l}$, and event status l. In order to model $\mathrm{P}_{i,t}(l)$, we need to make extra assumptions on the dependences among different events. A simple way is to assume that $\{Y_{i,t} : t = 0, 1, \cdots\}$ is a Markov chain, and then we can proceed just like the case of single event. However, this assumption may be too restrictive in many cases. Below, we will provide an alternative approach based on other assumptions.

Before getting into detail, we assume as above that $Y_{i,t}$ only depends on covariate values evaluated at finite number of time points, $\mathscr{X}_{i,t}$. Thus, all the $X_{i,t'\in\mathbb{Z}}$ terms in Equation (2.37) and in right hand side of Equation (2.36) can be substitute by $\mathscr{X}_{i,t}$. Also, we write $\{Y_{i,0} = y_{i,0}, \cdots, Y_{i,t} = y_{i,t}\}$, $t = 0, 1, \cdots$, as $Y_{i,0:t}$.

Explicit successive modeling of conditional distributions

In the expression of $\mathrm{P}_{i,t}(l)$ (2.37), $T_{i,1}, \cdots, T_{i,l}$ ($l = 1, \cdots, S-1$) and covariate vector $\mathscr{X}_{i,t}$ are all conditioning variables. A simple idea would be to treat $t_{i,1}, \cdots, t_{i,l}$ as covariates, and assume an explicit form (with unknown parameters) for $\mathrm{P}_{i,t}(l)$ as a function of $t_{i,1}, \cdots, t_{i,l}$ and $\mathscr{X}_{i,t}$.

For example, let $g : [0, 1] \to (-\infty, \infty)$ be a monotonic link function, and assume $g(\mathrm{P}_{i,t}(l))$ is a linear function or a polynomial of $t_{i,1}, \cdots, t_{i,l}$ and $\mathscr{X}_{i,t}$. For different l ($l = 0, \cdots, S-1$), the numbers of conditioning event times in the expression of $\mathrm{P}_{i,t}(l)$ are different. We will take a little trick to make the number of covariates constant over time so that our mathematical expressions can be nicely formulated. For each individual i, we define the following time dependent covariates:

$$T'_{i,l}(t) = \begin{cases} 0, & \text{if } t < t_{i,l} \\ t_{i,l}, & \text{if } t \geq t_{i,l} \end{cases}, \text{ for } l = 1, \cdots, S-1 \ . \tag{2.38}$$

Now for every l ($l = 0, \cdots, S-1$), $\mathrm{P}_{i,t}(l)$ is a function of $T'_{i,1}, \cdots, T'_{i,(S-1)}$ and $\mathscr{X}_{i,t}$. If we assume $g(\mathrm{P}_{i,t}(l))$ is a linear function of $T'_{i,1}, \cdots, T'_{i,(S-1)}$ and $\mathscr{X}_{i,t}$, we define a covariate vector

$$Z_{i,t} \equiv \left(\mathscr{X}_{i,t}^T,\ T'_{i,1}(t), \cdots, T'_{i,S-1}(t)\right)^T \ . \tag{2.39}$$

Similarly, if we assume $g(\mathrm{P}_{i,t}(l))$ to be a polynomial function of them, we can define $Z_{i,t}$ as a

vector which consists of the terms of the polynomial. Under both assumptions, we can write

$$g\left(\mathrm{P}_{i,t}\left(l\right)\right)=\beta_{t,l}^{T}Z_{i,t},\ \text{ for } l=0,\ \cdots,\ S-1\ , \tag{2.40}$$

where $\beta_{t,l}$ is a parameter vector that varies with time t and event status l but remains the same across different individuals. In many situation, we may reasonably assume $\beta_{t,l}$ is constant over time. Then it is a function of only l, and we will write it as β_l.

Now, if all N individuals are independent, and for each individual, we observe all the S events (i.e. no censoring), then the likelihood function is

$$\begin{aligned}L\left(\beta_l\right)&=\prod_{i=1}^{N}\mathrm{P}\left(T_{i,1}=t_{i,1},\ T_{i,2}=t_{i,2},\ \cdots,\ T_{i,S}=t_{i,S}\left|X_{i,t'\in\mathbb{Z}}\right.\right)\\&=\prod_{i=1}^{N}\left\{g^{-1}\left(\beta_0^{T}Z_{i,t}\right)\prod_{t=0}^{t_{i,1}-1}\left(1-g^{-1}\left(\beta_0^{T}Z_{i,t}\right)\right)\prod_{l=1}^{S-1}\left[g^{-1}\left(\beta_l^{T}Z_{i,t}\right)\prod_{t=t_{i,l}+1}^{t_{i,(l+1)}-1}\left(1-g^{-1}\left(\beta_l^{T}Z_{i,t}\right)\right)\right]\right\}\end{aligned} \tag{2.41}$$

The above is just an illustration. We need not assume $g\left(\mathrm{P}_{i,t}\left(l\right)\right)$ is a linear function or polynomial of $Z_{i,t}$. However, whatever explicit form we choose for $\mathrm{P}_{i,t}\left(l\right)$ as a function of $t_{i,1},\ \cdots,\ t_{i,l}$ and $\mathscr{X}_{i,t}$, we are actually making an explicit and very strong assumption on the conditional distribution of $Y_{i,t}$ given all the previous events times. At the same time, by successive conditioning, we are actually making a strong assumption on the joint distribution of all the S event times $T_{i,1},\ \cdots,\ T_{i,S}$ (see Equation (2.36)). Usually, this assumption is hard to satisfy, and also hard to check for appropriateness. On the other hand, even if β_l is constant over l, there are $S-1$ more covariates than the single event case. When S is large compared to N, the estimates of parameters will have large standard errors.

The above impediments can be partially avoided if we assume conditional independence between $T_{i,(l+1)}$ and $\left(T_{i,(l-1)},\ \cdots,\ T_{i,1}\right)^{T}$ given $T_{i,l}=t_{i,l}$, for $l=2,\ \cdots,\ S$, i.e.

$$\mathrm{P}\left(T_{i,(l+1)}=t\left|T_{i,l}=t_{i,l},\ T_{i,(l-1)}=t_{i,(l-1)},\ \cdots,\ T_{i,1}=t_{i,1}\right.\right)=\mathrm{P}\left(T_{i,(l+1)}=t\left|T_{i,l}=t_{i,l}\right.\right)\ , \tag{2.42}$$

for $l=2,\ \cdots,\ S$, and all $t=0,\ 1,\ \cdots$. Under this assumption, Equation (2.37) can be simplified to

$$\mathrm{P}_{i,t}\left(l\right)=\begin{cases}\mathrm{P}\left(Y_{i,t}=1\left|Y_{i,(t-1)}=0,\ X_{i,t'\in\mathbb{Z}}\right.\right), & \text{if } l=0\\ \mathrm{P}\left(Y_{i,t}=l+1\left|Y_{i,(t-1)}=l,\ T_{i,1}=t_{i,1},\ X_{i,t'\in\mathbb{Z}}\right.\right), & \text{if } l=1,\ \cdots,\ S-1\ .\end{cases} \tag{2.43}$$

Now $\mathrm{P}_{i,t}\left(l\right)$ is a function of only $t_{i,l}$, l, t and $\mathscr{X}_{i,t}$. If we treat $t_{i,l}$ as covariates, and assume an

explicit functional form for $\mathrm{P}_{i,t}(l)$, we then can get a regression model just as before. However, this time, we only have one extra covariate regardless of the total number of event status S. We are still making assumption about the joint distribution of $T_{i,l+1}$ and $T_{i,l}$ for $l = 1, \cdots, S-1$. Nevertheless, this will be a little easier than making assumptions about the joint distribution of S random variables (when S is large), although great care is still needed. The key point, however, is whether the conditional independence assumption is reasonable.

2.5.3 Estimation and prediction

We consider the model defined by Equation (2.38) – (2.40). When there is no censoring, the likelihood function is given by Equation (2.41), and the subsequent estimation procedure is essentially the same with the case of single event. Now we consider non-informative right censoring. First, we may assume $T_{i,0} = 0$. Note that we have assumed $T_{i,l} \neq T_{i,l'}$ for $l \neq l'$ and $l, l' = 1, 2, \cdots, S$ in Section 2.5.1, but it is possible that $T_{i,0} = 0 = T_{i,1}$. For each individual i, we observe several times, with the last one as the event time for the last event or censoring time, and the previous times as the event times prior to the last observation. If the last observed time is the time to the last event, there is no censoring; otherwise the observation is right censored. Suppose the last observed time is associated with a random variable τ_i which has observed value t_i, and censoring time is associated with a random variable C_i. Suppose for each individual i, prior to the last observed time t_i, we observe K_i ($0 \leq K_i < S$) events. Then, for individual i, if there is no censoring, $K_i = S-1$, $\tau_i = t_i = T_{i,S}$ and $C_i > T_{i,S}$, otherwise, $\tau_i = t_i = C_i$ and $t_{i,K_i} \leq C_i < T_{i,K_i+1}$. Just as before, we define a censoring indicator δ_i, which takes value 1 if the observation is not censored, and 0 if censored. We then can easily show that, under the non-informative right censoring assumption, the MLE equals the parameter value that maximizes the following function

$$L'(\beta_l) = \prod_{i=1}^{N} \Big\{ \mathrm{P}\left(T_{i,1} = t_{i,1}, \cdots, T_{i,K_i} = t_{i,K_i}, T_{i,K_i+1} = t_i \,\middle|\, X_{i,t' \in \mathbb{Z}}\right)^{\delta_i} \cdot \mathrm{P}\left(T_{i,1} = t_{i,1}, \cdots, T_{i,K_i} = t_{i,K_i}, T_{i,K_i+1} > t_i \,\middle|\, X_{i,t' \in \mathbb{Z}}\right)^{1-\delta_i} \Big\}. \tag{2.44}$$

The first factor on the right hand side (RHS) of the above equation is 1 when $\delta_i = 0$, and when $\delta_i = 1$, it is given by Equation (2.36). The second term on the RHS is 1 when $\delta_i = 1$, and when

$\delta_i = 0$, it is

$$
\begin{aligned}
&\mathrm{P}\left(T_{i,1} = t_{i,1},\, \cdots,\, T_{i,K_i} = t_{i,K_i},\, T_{i,K_i+1} > t_i \,\middle|\, X_{i,t'\in\mathbb{Z}}\right) \\
=&\mathrm{P}\left(Y_{i,0} = 0 \middle| X_{i,t'\in\mathbb{Z}}\right) \prod_{l=0}^{K_i-1} \left[\mathrm{P}\left(Y_{i,t_{i,(l+1)}} = l+1 \,\middle|\, Y_{i,\left(t_{i,(l+1)}-1\right)} = l,\, T_{i,1} = t_{i,1},\, \cdots,\, T_{i,l} = t_{i,l},\, X_{i,t'\in\mathbb{Z}}\right) \right. \\
&\left. \prod_{t=t_{i,l}+1}^{t_{i,(l+1)}-1} \mathrm{P}\left(Y_{i,t} = l \,\middle|\, Y_{i,(t-1)} = l,\, T_{i,1} = t_{i,1},\, \cdots,\, T_{i,l} = t_{i,l},\, X_{i,t'\in\mathbb{Z}}\right) \right] \\
&\prod_{t=t_{i,K_i}+1}^{t_i} \mathrm{P}\left(Y_{i,t} = K_i \,\middle|\, Y_{i,(t-1)} = K_i,\, T_{i,1} = t_{i,1},\, \cdots,\, T_{i,l} = t_{i,K_i},\, X_{i,t'\in\mathbb{Z}}\right) . \qquad (2.45)
\end{aligned}
$$

Once model parameters are estimated, the prediction procedure is not very different from the case of single event. We only note here that it might be cumbersome to predict all the future events for a new individual at the same time. Instead, we focus on the time for the next event conditioned on konwn previous event times.

2.6 Discussion

Our main goal is to develop models capable of providing real-time prediction, i.e. whenever a new piece of time-dependent covariate information is available, we can update our prediction of future event time to make it more precise. The method developed in this Chapter achieve this goal by modeling the stochastic process of the event status indicator variable $Y_{i,t}$ on a discrete time scale. The pros and cons of our methods compared to the Cox model and parametric proportional hazards model are summarized as follows.

Compared to the Cox model, our method has the following advantages:

- Prediction is easily formulated in our model, while the Cox model is generally not suitable for prediction;
- For parameter estimation, where time-dependent covariates are present, the Cox model uses the partial likelihood, which does not use the covariate information during the gap time between event times of any two individuals. The consequence is a loss of efficiency. Cox and Oakes (1984) argue that the loss of efficiency of the partial likelihood is usually not much relative to the full likelihood unless either: (1) the model parameter is far from zero; (2) censoring is strongly dependent on covariates; or (3) there are strong time trends in the covariates. While the first two issues may not be of concerns, the third one is crucial for phenological data since the climate variable covariates have strong

seasonal variations (and will exhibit as a dominant local trend between event times within a season). Our model makes use of all information about time-dependent covariates, thus is fully efficient.

However, the Cox model allows the baseline hazard to be totally unspecified, which is a very flexible distributional assumption. In our model, although we do not explicitly assume that the event time belongs to any specific distribution family, we are still making a relatively strong distributional assumption (compared to the Cox model) by assuming the time homogeneity of the Markov chain.

Compared to parametric proportional hazards model,

- Fitting our model is computationally straightforward and less demanding, while parametric proportional hazards model may involve complicated integration when time-dependent covariates are present.

- In the parametric proportional hazards model, we usually assume that the hazard function is only related to the covariate value at the current time. However, in some cases, this is not enough and we need to assume that the hazard function is related to the covariate values at the current time and also at some previous time points. This will result in a more complicated formulation and introduce further computational issues. However in our model, we can allow at each time point without severe computational penalty, the event status indicator to be related to covariate values evaluated at several time points, both at and prior to the current time. Although the number of time points included needs to be fixed and not too large (to limit the number of parameters), we can partially incorporate the historical covariate values without much effort.

- The distributional assumption of our model is relatively less restrictive than parametric proportional hazards model.

However, our model does not directly model the distribution of event time. When that is required, our model might be less useful than the parametric proportional hazards models.

For sequential events, we simply extended our model for single event and added some extra assumptions. These assumptions, however, may not always be appropriate, so care is needed. Many techniques have been developed for different kinds of multivariate survival data (see Hougaard 2000 and Cook and Lawless 2007, for example). Our model for sequential events is especially tailored for the modeling and prediction of phenologcial data and similar survival data. Also, it is unique in the way it incorporates the time-dependent covariate information and in the way it models the event status indicators instead of event times.

In the models for both single event and sequential events, in order to simplify the problems, we have assumed that model parameters are not functions of time. In the single event case, this is equivalent to assuming a time-homogeneous Markov chain. This probably is the most restrictive distributional assumption in our models. Our future work will relax this assumption by allowing for dynamic parameters in the model. That will make our models fit into a much wider context. Another mathematical issue is that in the model for multiple events, we have assumed that no events occur at the same time points. However, in practice, this may occur, especially when the discrete time scale is coarse. We will need some further work to remove this restriction.

Chapter 3

Application to Phenological Data I – Model Building and Parameter Estimation

3.1 A brief introduction to phenology

Phenology is the study of periodic plant developmental stages and their responses to climate, especially to seasonal and interannual variations in climate. Better understanding perennial crop phenology helps to understand current and future yield distributions, and likewise to improve the assessment of agricultural risk in association with observed climate variability and extremes.

In Chapter 1, we mentioned about the development stages of apple trees as a typical example of phenogolical data. In each development cycle, an apple tree will go through many development stages from bud-bursting, blooming to fruiting. In order to study the development of the plant in further detail, this partition of developing stages may be refined. For example, between bud bursting and blooming, scientists sometimes further consider several other stages, namely side green, green tips, etc., to represent the different stages of leave growing. Also, before bud-bursting, scientists consider "invisible" stages of bud development, such as endodormancy and ecodormancy, whose starting and ending times may not be able to be observed. In each development cycle, these development stages occur sequentially, and if an apple tree fails to reach one stage, it won't reach the following stages. For example, if an apple tree fails to bloom in one year, it won't bear fruit in that year. Our primary interest is the timings of these development stages. While typically a development stage occupies a period of time, we usually choose a meaningful time point to serve as a "representative" or "landmark" for this stage. That could be the starting time of the stage or a well-defined time point that is meaningful to phenological study. In this sense, we are using one time point to represent the development stage. We then will call this development stage a *phenological event*, and call the

"representative" time point the time to the phenological event or the timing of the phenological event.

Numerous studies (e.g. Sparks et al. 2000, Schwartz et al. 2006 and Chuine et al. 2004) show that timings of phenological events are closely related to climate variables, especially temperature. However, there is no unified theory on how climate variables influence timings of phenological events, and how different development stages interact with each other. Many phenological models have been built based on different theories, but with the same ultimate goal: to predict the future phenological event using climate data that are available. A review of various phenological models can be found in Chuine (2000). In these models, values of parameters are either determined experimentally or obtained as point estimates given by least squares to yield best fits to observed data. The uncertainties associated with these estimates of parameters are often not assessed.

Here, we apply our statistical models to phenological data. The primary goal is to provide prediction for future phenological events, and alongside assessing the uncertainty associated with the prediction. Also, we hope to provide insights on the relationship between phenological events and climate variables, and provide meaningful estimates for some important parameters in phenological study. As an example, we will focus on the timing of only one event, namely blooming, and we study its response to a single covariate – air temperature. We take one classic theory about the relationship between phenological events and temperature, which states that temperature influences phenological events through the GDD as defined in Chapter 1:

$$GDD(t) = \begin{cases} \frac{T_{min}(t)+T_{max}(t)}{2} - T_{base} & \text{if } \frac{T_{min}(t)+T_{max}(t)}{2} > T_{base} \\ 0 & \text{Otherwise} \end{cases}, \tag{3.1}$$

where t stands for discrete time with the unit of day, $T_{min}(t)$ and $T_{max}(t)$ are daily minimum and maximum temperatures, and T_{base} is a thresholding constant temperature. Usually T_{base} is unknown. In fact, mathematically, the statistical model built upon this theory can be directly used to incorporate many other theories. Most of theories about the relationship between phenological events and temperature reviewed in Chuine (2000) are based on the same idea that the timing of a phenological event is related to the accumulation of some sort of "energy". Once a plant gets enough "energy", the phenological event will be triggered. In mathematical form, the variable S that influences the timing of a phenological event is written as:

$$S(t_e) = \sum_{t=t_0}^{t_e} R(x_t) \ , \tag{3.2}$$

where t is the discrete time with the unit of day, t_0 is a starting time point, t_e is the time to the phenological event, $x_t = \frac{T_{min}(t)+T_{max}(t)}{2}$ is the daily average temperature as defined in Chapter 1, and $R(x_t)$ is a function similar to GDD which stands for available daily "energy" that contributes to the occurrence of the phenological event. Different theories define different $R(x_t)$'s. We will take $GDD(x_t)$ as $R(x_t)$, but whenever one wants to use another function for $R(x_t)$, one can just replace $GDD(x_t)$ with that function and use the same procedure.

In this chapter we will model the process and discuss parameter estimation. In the next chapter we discuss prediction.

3.2 Data description and exploratory analysis

The data are the bloom dates of six high-value perennial agricultural crops (apricot, cherry, peach, prune, pear, and apple) in the Okanagan region of British Columbia. The data used in bloom date example (Table 1.1) in Chapter 1 are the bloom dates of apple trees in this dataset. The bloom dates of other corps have similar structure. In each year, blooming occurs at most once for each crop, and the bloom date is recorded as the number of days from the first day of a year to the date of the "representative" time point of blooming. For cherry, prune, pear, and apple, bloom dates in year 1937 – 1964 are recorded, and for apricot and peach the bloom dates in these years except for 1950 are available. This dataset is old, but it is a nicely collected, and has been used in many phenological studies. Also, this dataset reflects problems likely to encountered in other applications. Daily average temperatures in the Okanagan region in the corresponding years are also collected.

At a fist glance, several things seem odd to a statistician. First, in the whole Okanagan region, there is only one bloom date for each crop in each year. For each crop, this is a summary of bloom dates for all the trees in the region defined by phenologists. Here, we will just treat these bloom dates as "aggregated" data, and forget about individual trees at the current stage. Second, for each crop, the collection of bloom dates is a time series. The bloom dates in different years may be well auto-correlated. Nevertheless, the models introduced in Chapter 2 requires the observations to be independent. It's possible to extend those models to allow for correlated data, but the formulation is more complicated. Here, we will do some exploratory analysis to check if these bloom dates in different years are approximately independent. Third, since the daily maximum and minimum temperature are usually not observed exactly, the daily average temperatures defined as $\frac{T_{min}(t)+T_{max}(t)}{2}$ may have measurement errors. To keep the problems simple and focused, we will ignore this issue, and treat the observed daily average temperature as exact without any error.

Now we do some exploratory analysis to study the correlations among bloom dates of different years for each crop. Figure 3.1 shows for each crop, the plot of the sample autocorrelation coefficients (Chatfield, 2004) r_k of bloom dates against time lags k. Sample autocorrelation

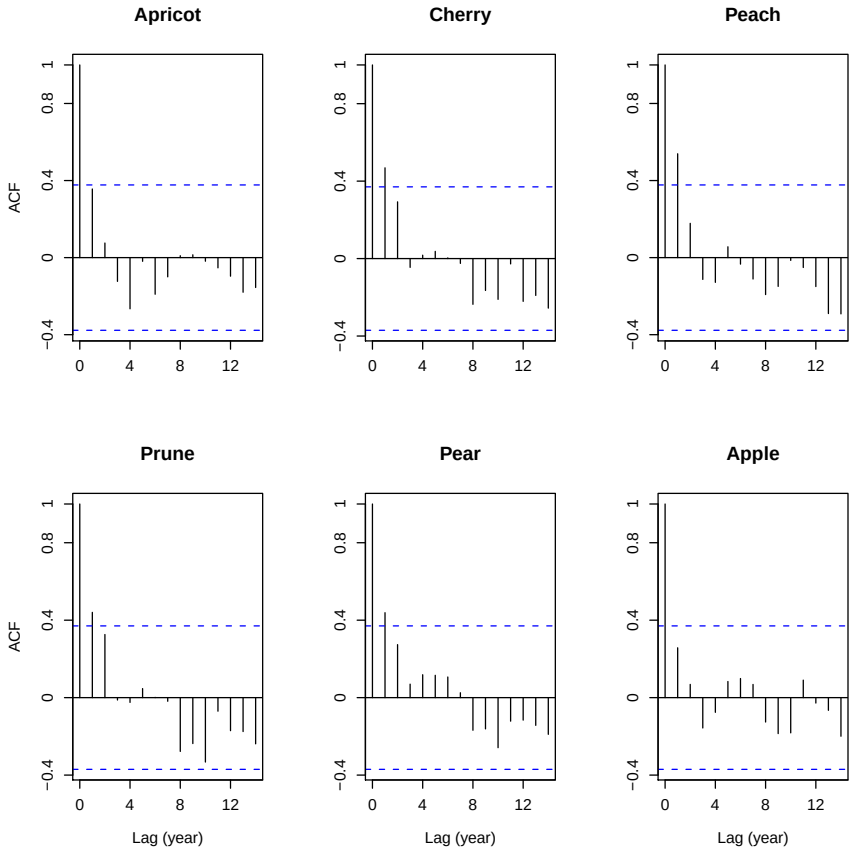

Figure 3.1: Correlograms of the bloom dates of the six crops considered in this report. No serious autocorrelation can be seen.

coefficient at lag k is an estimator of the autocorrelation function (ACF) evaluated at lag k, and this kind of plot is called a *correlogram*. It is used to discover the correlation structure of a time series. For each crop, if the observed bloom dates are realizations of independently and identically distributed (iid) random variables, then when n, the number of observations, is large, the sample autocorrelation coefficients r_k at each lag k is approximately distributed as $N(0, 1/n)$, a normal distribution with 0 mean and $1/n$ variance. This implies that when n is large, r_k is approximately 0 at all non-zero lags. The two dash lines in each correlogram are $\pm 2/\sqrt{n}$, which is roughly the 95% confidence interval (CI) of $N(0, 1/n)$. Thus, if for a

crop, the observed bloom dates are realizations of iid random variables, we expect 19 out of 20 of the values of r_k to fall between the two dash lines on the correlogram. In Figure 3.1, we see that for apple and apricot, r_k's at all lags fall between the two dash lines, so it might be reasonable to assume the observed bloom dates are realizations of iid random variables. For all other crops, the autocorrelation coefficients at lag 1 are slightly above the dashed lines, which indicates that there might be small autocorrelations, but which seem unlikely to be serious. It may be not worth adding too much complexity to our model to account for small correlations between observations. Therefore, we will assume that for every crop, the observed bloom dates are realizations from iid random variables.

We want to study the relationship between bloom dates and GDD. As we mentioned, experimental results show that the sum of the GDD from a time origin to the bloom date is the main factor that triggers bloom. For convenience, we define *accumulated GDD* (AGDD) as follows:

$$AGDD(t) = \sum_{k=t_0}^{t} GDD(k) \ , \tag{3.3}$$

where t_0 is a well-defined time origin and t is any time point that satisfies $t \geq t_0$. Scatter plots of the bloom dates against AGDD evaluated at the corresponding bloom dates for all the crops are shown in Figure 3.2. There are no obvious patterns in these plots. This fact seems counter-intuitive, but actually it doesn't conflict with the theory. If the theory is true, once the AGDD reaches some value, bloom is likely to occur. The value of the AGDD evaluated at the bloom date itself then does not reflect the time to bloom, but the whole process of how the GDD values add up to that certain value is important.

3.3 Applying the stochastic process based regression model to the data

We consider the six crops separately, so in the following discussion about methodology, we consider only one crop, but it could be any one of them.

3.3.1 Notation

For a crop, suppose we observed N bloom dates in N years (one bloom date each year). Here, years correspond to "individuals" we talked about in Chapter 2. For each year i ($i = 1, \cdots, N$), we treat the bloom date as a random variable T_i. Let t be discrete time with the unit of day. Since the bloom date of a crop in each year is counted from the first day of a year, we take

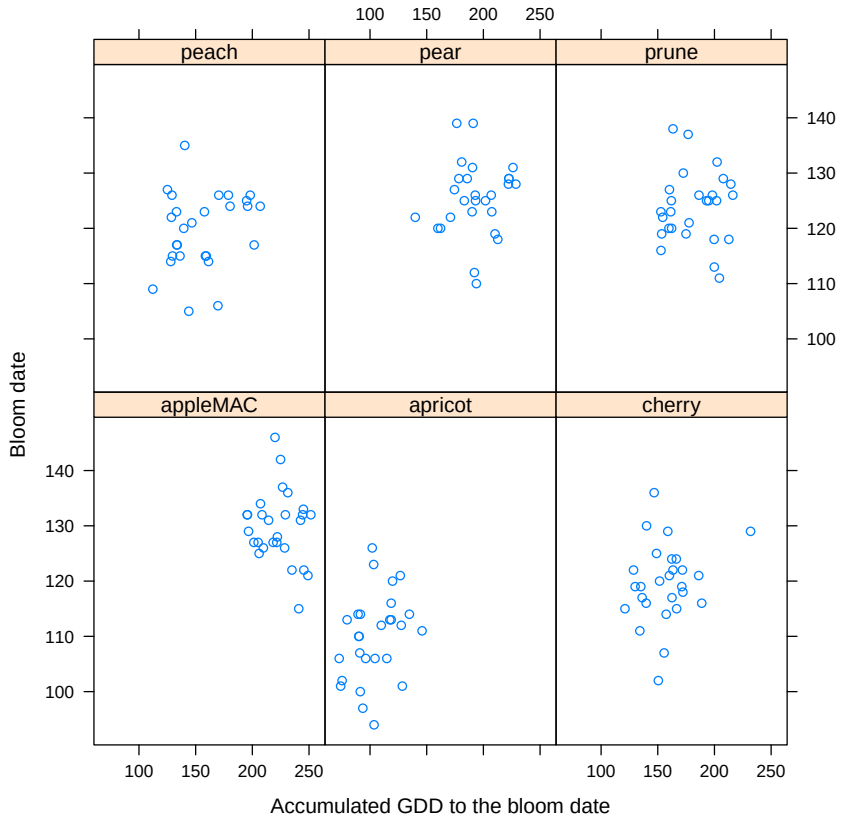

Figure 3.2: Scatter plot of the bloom dates against AGDDs evaluated at the corresponding bloom dates

the first day of a year $t = 1$ as a natural time origin. For blooming event, we denote the status indicator variable at time t as defined in Section 2.1 as $Y_{i,t}$. We also write GDD at time t in year i as $X_{i,t}$ for convenience.

3.3.2 Applying the stochastic process based regression model

At the end of section 3.2, we emphasized that before the bloom date, the evolving history of GDD is closely related to the time to bloom. To reflect this, it might be reasonable to assume the following: if blooming did not occur on the previous day $t-1$, the probability that blooming will occur today t is related to a function $f(\cdot)$ of GDD evaluated at today and possibly all or some of the previous days. This function of GDD values serves to somehow increase or

decrease the probability of blooming today. Note that this function does not need to be AGDD. AGDD is only a particular form of it.

We consider an arbitrary year i ($i = 1, \cdots, N$). On each day, whether blooming occurs or not can be expressed by the status indicator variable $Y_{i,t}$ with $Y_{i,t} = 1$ for occurrence and $Y_{i,t} = 0$ for no occurrence. We now consider $Y_{i,t}$ ($t = 1, 2, \cdots$) as the response variable and GDD $X_{i,t}$ as a time-dependent covariate. In each year, the blooming event clearly satisfies Assumptions 2.1. Hence, by Corollary 2.3, given all the GDD values, $\{Y_{i,t} : t = 1, 2, \cdots\}$ is a Markov chain. The probability that blooming occurs at time t_i given all the GDD values, then is given by Equation (2.8). Now, for a time t, if blooming didn't occur at $t-1$, i.e. $Y_{i,(t-1)} = 0$, we assume $\mathrm{P}_{i,t} \equiv \mathrm{P}(Y_{i,t} = 1 | Y_{i,t} = 0, X_{i,1}, X_{i,2}, \cdots)$, the probability of blooming occurring today given GDD evaluated at all time points, is related to $\mathscr{X}_{i,t}$, a vector consisting of GDD values at all or some of time points from time origin to current time t. We express this assumption as

$$g(\mathrm{P}_{i,t}) = f(\mathscr{X}_{i,t}; \beta) \ , \tag{3.4}$$

where $g : (0, 1) \to (-\infty, \infty)$ is a monotonic link function, and $f : (-\infty, \infty) \to (-\infty, \infty)$ is a function of $\mathscr{X}_{i,t}$ with parameter vector β, which encodes the relationship between $\mathrm{P}_{i,t}$ and $\mathscr{X}_{i,t}$. The only usage of the link function $g(\cdot)$ here is to convert the range of $\mathrm{P}_{i,t}$ to match the range of $f(\mathscr{X}_{i,t})$, a pure mathematical requirement. The logit and probit functions (Faraway, 2006) are popular choices of link functions. Although different link functions lead to different estimated parameter values, in most cases, the logit function and the probit function will produce nearly the same inference results (e.g. when testing hypotheses for the regression parameters) and have similar interpretations. Therefore, unless one has some special reasons, the choice of the link function is mostly a matter of personal preference or simple convenience. We will choose the logit function,

$$logit(p) = \log\left(\frac{p}{1-p}\right), \ 0 < p < 1 \ , \tag{3.5}$$

as the link function.

If we now settle on a particular functional form for $f(\mathscr{X}_{i,t}; \beta)$, we then can write down the likelihood function of the data, which has the same expression as Equation (2.14), only with every $\beta^T \mathscr{X}_{i,t}$ term replaced by $f(\mathscr{X}_{i,t}; \beta)$. Then, we can calculate the MLE of parameter vector β. Next, we will discuss the last but important issue: choosing $f(\mathscr{X}_{i,t}; \beta)$.

3.3.3 Incorporating GDD in the model

Now we will specify several particular functional forms for $f(\mathscr{X}_{i,t};\, \beta)$. We plug each particular function into the regression model (3.4), then calculate the MLE for the parameter β. Then we do a standard model selection by comparing Akaike information criterion (AIC) or Bayesian information criterion (BIC) (Konishi and Kitagawa, 2008) values of the models induced by these different functions.

The choices of functions under consideration are based on the idea that the accumulation of GDD values is the main factor that influences the occurrence of the blooming event. Scientists believe that this accumulation should be in the form of AGDD (Equation (3.3). This is an empirical result. Here, we will try several functions (including AGDD) which represent different ways of accumulations of GDD values, such as weighted average, and statistically explore which one yields the best model. These functions are listed as follows, where we give a name to each model that corresponds to each function.

Model GDD We take a linear function of $X_{i,t}$, the GDD evaluated at current time t:

$$f(\mathscr{X}_{i,t};\, \beta) = a + bX_{i,t}; \quad \beta = (a,\, b,\, T_{base})^T \ . \tag{3.6}$$

Model AGDD We take a linear function of AGDD evaluated at current time t:

$$f(\mathscr{X}_{i,t};\, \beta) = a + b\sum_{k=1}^{t} X_{i,k}; \quad \beta = (a,\, b,\, T_{base})^T \ . \tag{3.7}$$

Model ExpSmooth We take a linear function of a weighted sum of GDD from the time origin 1 to the current time t:

$$f(\mathscr{X}_{i,t};\, \beta) = a + b\sum_{k=0}^{t-1} (1-\gamma)^k X_{i,(t-k)}; \quad \beta = (a,\, b,\, \gamma,\, T_{base})^T \ , \tag{3.8}$$

where $0 \leq \gamma \leq 1$. We call this model "ExpSmooth" because the weighted average term is similar to the exponential smoothing used in time series (Chatfield, 2004).

Model 5Days We take a linear function of the GDD evaluated at the 5 most recent days:

$$f(\mathscr{X}_{i,t};\, \beta) = a + \sum_{k=1}^{5} b_k X_{i,(t-k+1)}; \quad \beta = (a,\, b_1,\, b_2,\, b_3,\, b_4,\, b_5,\, T_{base})^T \ . \tag{3.9}$$

Model MA5, MA10 and MA20 We take 5-day, 10-day and 20-day moving averages of GDD

series of each year i, and denote the averaged GDD series as $\bar{X}_5(i,t)$, $\bar{X}_{10}(i,t)$ and $\bar{X}_{20}(i,t)$, $t = 1, 2, \cdots$, respectively. Then we take

$$f(\mathscr{X}_{i,t};\, \beta) = a + b\bar{X}_k(i,t); \quad \beta = (a,\, b,\, T_{base})^T \ , \tag{3.10}$$

for $k = 5$, 10, and 20, respectively.

Model Spline We fit a cubic smoothing spline (Simonoff, 1998) to the GDD series of each year i. Denote the smoothed GDD series as $\tilde{X}_{i,t}$, $t = 1, 2, \cdots$. Then we take

$$f(\mathscr{X}_{i,t};\, \beta) = a + b\tilde{X}_{i,t}; \quad \beta = (a,\, b,\, T_{base})^T \ , \tag{3.11}$$

Note that in each of the above models, T_{base} is a parameter that included in the expression of GDD $X_{i,t}$.

Model GDD and AGDD serve as a basis of comparison. Model 5Days express the idea that GDD evaluated at many time points prior to current time might be important factors and each of them may have a different effect on $P_{i,t}$. In this model, we give each GDD evaluated at several days prior to and at the current day a different parameter. Given a sample size of 27 or 28, we won't be able to get sensible estimates if the number of parameters are too many. Therefore we consider only GDD evaluated at the 5 most recent days. In Model MA5, MA10, M20 and Spline, we just want to see if the "bumpy" signal in GDD series is smoothed out, what model fit we will get.

The most interesting model is Model ExpSmooth. In this model, $f(\mathscr{X}_{i,t};\, \beta)$ is a linear function of weighted sum of GDD evaluated at all the time points at and prior to the current time t. For a fixed γ ($0 \le \gamma \le 1$) value, $(1-\gamma)^k$, the weight on the GDD at lag (number of days prior to the current date) k, decays when k increases. This reflects the idea that the GDD evaluated at recent time points may contribute more to change $P_{i,t}$ than the GDD evaluated at time points long before the current time t will do. The question though is how much more, i.e. how fast the weight decays when lag increases. This is controlled by the value of γ. Figure 3.3 shows how the weight decays when the lag increases for different values of γ. When γ becomes larger, the weight decays faster. In the extreme case of $\gamma = 1$, the weighted sum is just GDD evaluated at the current time and so Model ExpSmooth becomes Model GDD. If γ becomes smaller, the weight decays slower. In the extreme case of $\gamma = 0$, i.e. no decay, Model ExpSmooth becomes Model AGDD. In all, Model GDD and AGDD are only special cases of Model ExpSmooth. Since we do not know the value of γ in advance, we treat it as a model parameter, and let the data speak, i.e. estimate its value by fitting the model.

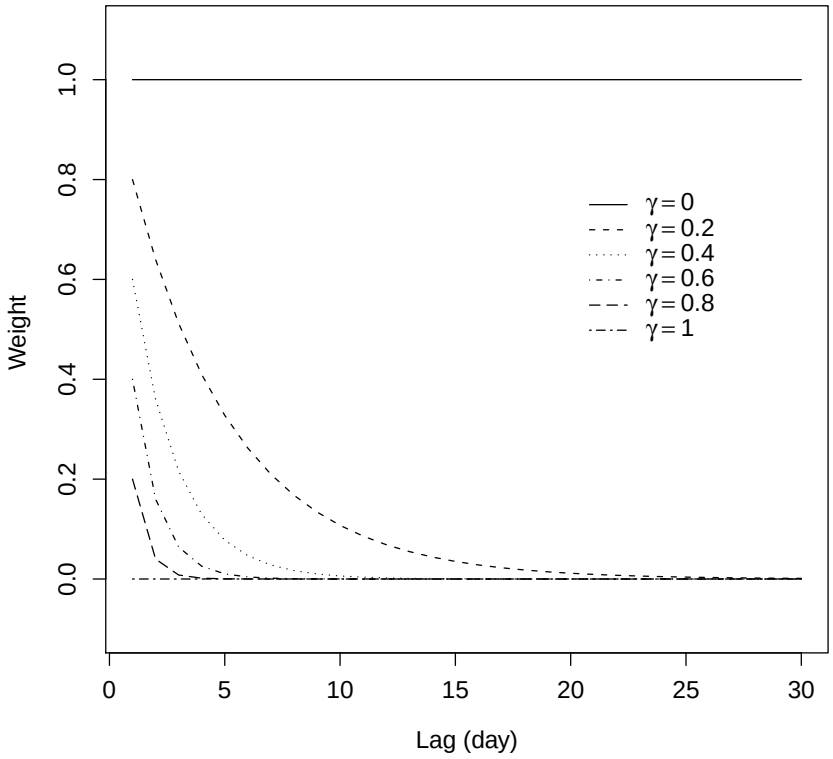

Figure 3.3: The actual weights in the weighted sum in Model ExpSmooth for different γ parameter values. The weight decays when the lag (number of days prior to the current date) increases. A larger γ value corresponds to a faster speed of decaying.

For every crop, we fit all the above models to our data. The results are reported in Table 3.1 – Table 3.6. In each table, estimated parameters a, b, T_{base} and γ for all models are reported. The smoothing parameter γ only appears in Model ExpSmooth. Model 5Days doesn't have parameter b. Instead it has five other parameters b_1 – b_5, which we don't report since the primary goal now is to compare the performance of the models but not to investigate the estimated parameter values. The last three columns in each table are the negative log-likelihood, AIC and BIC under the estimated parameter values of each model, respectively. They are useful measures of the goodness of fit of a estimated statistical model. For all these quantities, smaller values means "better" models. We say that a model performs "good" if it not only fits the observed data well, but also predicts the unknown data reasonably well. When a model has too

many parameters, it tends to fit the observed data very well or even exactly, but it may fail to give any sensible prediction for unknown data, and in this case, we consider it a "bad" model. In statistical modeling, we usually apply the "principle of parsimony", which means that we should make a model as simple as possible (i.e., make the number of parameters as small as possible) given that it can fit the observed data reasonably well, i.e. a trade-off. While negative log-likelihood is a measure of model fit to the observed data alone, AIC and BIC are two quantities that take into account both model fit and model complexity. In this regard, we usually take AIC or BIC as the criterion for model selection. BIC penalizes the model complexity more strongly than AIC does, so it favors more parsimonious models.

Table 3.1 – Table 3.6 show that, for every crop, if we look at any one of the negative log-likelihood, AIC and BIC, Model AGDD and Model ExpSmooth always perform far better than all the other models. For every crop, Model ExpSmooth has a little better performance measures than Model AGDD, but the differences are quite small. The estimated smoothing parameter γ's for Model ExpSmooth of the six crops range from 0.0036 to 0.0225. The decaying of the weights $(1-\gamma)^k$ with the increases of lags k in Model ExpSmooth of each crop is shown in Figure 3.4. We see that for every crop, the weights decay slowly, which tells us, for every crop, Model ExpSmooth gives us a very similar model to Model AGDD. This statistical result supports scientists experimental result: the accumulation of GDD is roughly in the form of AGDD.

Table 3.1: Estimated parameters, negative log-likelihood values (-logL), AIC's and BIC's of the fitted models for Apricot

Model	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$	$\hat{\gamma}$	-logL	AIC	BIC
GDD	-25.68	0.36	-51.32	NA	119.88	245.76	249.65
AGDD	-13.49	0.061	2.65	NA	69.41	144.82	148.70
ExpSmooth	-19.25	0.076	0.40	0.014	67.17	142.35	147.53
5Days	-6.87	NA	3.50	NA	107.97	229.94	239.01
MA5	-17.08	0.54	-16.60	NA	108.26	222.53	226.41
MA10	-21.19	0.68	-17.67	NA	100.07	206.15	210.03
MA20	-25.81	0.97	-15.70	NA	90.22	186.44	190.33
Spline	-11.77	0.64	-4.16	NA	102.60	211.20	215.09

Now our model candidates have been narrowed down to Model AGDD and Model ExpSmooth. We will solely study Model AGDD from now on for the following reasons: (1) AGDD is a quantity that has interested scientists for a long time. Besides, Model AGDD performs roughly the same with Model ExpSmooth; (2) Model AGDD has one less parameter;

Table 3.2: Estimated parameters, negative log-likelihood values (-logL), AIC's and BIC's of the fitted models for Cherry

Model	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$	$\hat{\gamma}$	-logL	AIC	BIC
GDD	-13.17	0.36	-15.55	NA	122.86	251.72	255.72
AGDD	-11.72	0.043	3.35	NA	76.85	159.69	163.69
ExpSmooth	-20.38	0.065	-0.30	0.020	72.04	152.08	157.41
5Days	-6.73	NA	5.30	NA	110.20	234.40	243.73
MA5	-16.34	0.55	-13.01	NA	108.96	223.92	227.92
MA10	-21.08	0.69	-15.81	NA	106.15	218.30	222.30
MA20	-22.83	1.08	-9.15	NA	93.37	192.75	196.74
Spline	-12.51	0.63	-4.13	NA	105.30	216.59	220.59

Table 3.3: Estimated parameters, negative log-likelihood values (-logL), AIC's and BIC's of the fitted models for Peach

Model	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$	$\hat{\gamma}$	-logL	AIC	BIC
GDD	-12.34	0.34	-15.01	NA	122.59	251.18	255.06
AGDD	-19.67	0.043	0.38	NA	68.40	142.79	146.68
ExpSmooth	-26.20	0.054	-1.65	0.0083	66.65	141.30	146.49
5Days	-6.53	NA	4.59	NA	116.09	246.18	255.25
MA5	-39.66	0.47	-66.25	NA	113.73	233.45	237.34
MA10	-20.90	0.59	-20.29	NA	109.73	225.45	229.34
MA20	-18.05	0.80	-9.63	NA	101.43	208.86	212.74
Spline	-26.46	0.60	-28.41	NA	108.25	222.49	226.38

with a sample size of 27 or 28, it is sensible too keep the number of total parameters under 3.

3.4 Consistency of the MLE when the likelihood function is not a continuous function of parameters

In the above discussion, we estimated the parameters of Model AGDD by the ML method. Do these estimated parameters reflect the true values of the parameters? Wald (1949) gave the famous conditions for the consistency of the MLE. Based on Wald's conditions, one can derive the regularity conditions that ensure the asymptotic normality and efficiency of the MLE. Consistency of the MLE, very loosely speaking, means that when number of independent observations N is large, the MLE will be very close to the true parameter value. The asymptotic efficiency implies that when N is large, one cannot get a better estimator than the MLE in

Table 3.4: Estimated parameters, negative log-likelihood values (-logL), AIC's and BIC's of the fitted models for Prune

Model	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$	$\hat{\gamma}$	-logL	AIC	BIC
GDD	-13.96	0.47	-10.47	NA	107.07	220.14	224.14
AGDD	-18.23	0.057	2.80	NA	66.40	138.79	142.79
ExpSmooth	-30.93	0.079	-1.14	0.016	64.29	136.58	141.91
MA5	-14.14	0.54	-9.03	NA	111.13	228.26	232.26
MA10	-625.93	0.84	-734.31	NA	107.39	220.79	224.79
MA20	-11.33	1.02	1.65	NA	99.94	205.89	209.88
Spline	-12.83	0.64	-3.76	NA	103.82	213.63	217.63
5Days	-10.48	NA	-1.04	NA	105.07	224.14	233.47

Table 3.5: Estimated parameters, negative log-likelihood values (-logL), AIC's and BIC's of the fitted models for Pear

Model	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$	$\hat{\gamma}$	-logL	AIC	BIC
GDD	-15.54	0.44	-15.15	NA	110.60	227.19	231.19
AGDD	-22.27	0.07	2.97	NA	61.29	128.58	132.57
ExpSmooth	-48.13	0.10	-4.05	0.023	59.91	127.81	133.14
5Days	-7.81	NA	5.14	NA	100.65	215.31	224.63
MA5	-19.45	0.63	-14.32	NA	101.64	209.27	213.27
MA10	-20.25	0.79	-11.03	NA	101.84	209.69	213.68
MA20	-11.76	1.16	2.41	NA	96.78	199.56	203.56
Spline	-12.83	0.71	-2.06	NA	96.80	199.61	203.60

terms of variance. The asymptotic normality states that when N is large, the MLE approximately has a normal distribution. The mean of this normal distribution is the true value of the parameter, and the covariance matrix (Equation (2.18)) is the inverse of the Fisher information matrix (Equation (2.17)). One of Wald's conditions requires that the likelihood function is a continuous function of the parameters. Regularity conditions for the asymptotic normality and efficiency also require this. Unfortunately, by the definition of GDD (Equation 3.1), the likelihood function in Model AGDD is not a continuous function of the parameter T_{base}. We then cannot directly use Wald's conditions. However, this does not imply the MLE of T_{base} is not consistent, since Wald's conditions are only sufficient conditions for the consistency of the MLE. Here, instead of trying to prove or disprove the consistency of T_{base}, we will do simulation to perform a numerical check.

In the simulation study, we generate data as follows. For each day of a year, we take the

Table 3.6: Estimated parameters, negative log-likelihood values (-logL), AIC's and BIC's of the fitted models for Apple

Model	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$	$\hat{\gamma}$	-logL	AIC	BIC
GDD	-14.04	0.34	-18.95	NA	127.56	261.11	265.11
AGDD	-26.77	0.07	2.82	NA	58.16	122.31	126.31
ExpSmooth	-29.32	0.080	2.25	0.0036	57.73	123.46	128.79
5Days	-6.92	NA	5.79	NA	115.56	245.12	254.44
MA5	-16.25	0.50	-13.93	NA	115.11	236.23	240.23
MA10	-19.64	0.68	-13.02	NA	107.19	220.37	224.37
MA20	-13.53	1.20	1.73	NA	95.86	197.72	201.71
Spline	-13.27	0.64	-3.78	NA	106.27	218.54	222.54

average of the daily average temperature from year 1916 to 2005 in the Okanagan region of British Columbia, and get one year of long term averaged daily average temperature series. We then add a noise process upon this long term averaged series. This noise process is generated from a $ARMA(3,1)$ model:

$$X_t = 1.83X_{t-1} - 0.96X_{t-2} + 0.12X_{t-3} + Z_t - 0.96Z_{t-1} \tag{3.12}$$

where the white noise Z has a normal distribution with mean 0 and variance 5.253 for any t. This ARMA model is fitted from the same daily average temperature used for extracting long term averaged series above. The details are discussed in Section 4.1 of the next chapter. Now we get a year of simulated daily average temperature data. We calculate the GDD of this generated temperature data using Equation (3.1) with parameter $T_{base} = 3.5$. Denote the GDD as X_t. Now starting from day 1, we generate a random number Y_1 from a Bernoulli distribution $Ber(p)$ with parameter $logit^{-1}\left(-13 + 0.04\sum_{k=1}^{1} X_k\right)$. If $Y_1 = 0$, we will generate

$$Y_2 \sim Ber\Big(logit^{-1}\big(-13 + 0.04\sum_{k=1}^{2} X_k\big)\Big)\,. \tag{3.13}$$

Again, as long as $Y_2 = 0$, we will proceed to generate Y_3 similarly, and so on and so forth. Once we get 1 (i.e., the first time we get 1), say at t^{th} draw, then we get simulated bloom date t. Using this procedure, we generate one year long of GDDs and a bloom date for that year as one year of data.

Now we generate 30 years of data as one sample (i.e. a sample of size 30). We generate 1000 such samples. For each sample i ($i = 1, \cdots, 1000$), we apply Model AGDD, and calcu-

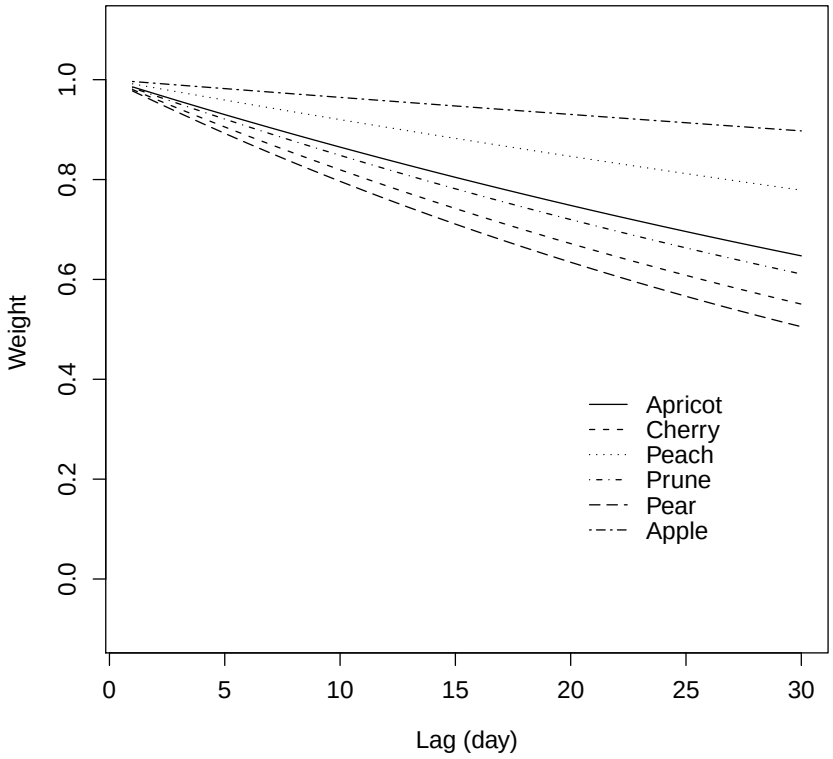

Figure 3.4: Decaying of the weights in the weighted sum in the fitted Model ExpSmooth for different crops

late the MLEs of the model parameters: $\hat{a}_i$, $\hat{b}_i$, and $\hat{T}_{base_i}$. For each parameter, for example parameter a, we calculate the estimated mean of the MLE,

$$\bar{\hat{a}} = \frac{1}{1000} \sum_{i=1}^{1000} \hat{a}_i \ , \tag{3.14}$$

the estimated variance of the MLE,

$$\frac{1}{1000-1} \sum_{i=1}^{1000} \left(\hat{a}_i - \bar{\hat{a}}\right) \ , \tag{3.15}$$

and the standard error of the mean of the MLE,

$$\sqrt{\frac{\frac{1}{1000-1}\sum_{i=1}^{1000}\left(\hat{a}_i-\bar{\hat{a}}\right)}{1000}}. \tag{3.16}$$

The estimated mean of the MLE is an estimate of the mean of the MLE of a parameter, and the estimated variance of the MLE is an estimate of the variance of the MLE. The standard error of the mean of the MLE is the standard error of the estimated mean of the MLE, which characterize how well the estimated mean of the MLE approximate the true mean of the MLE.

We repeat the above procedure for sample sizes S of 80, 150 and 400. If the MLEs are consistent, we will be able to see that for each parameter, when the sample size becomes larger, the estimated mean of the MLE becomes closer to the true value of the parameter, and the estimated variance of the MLE becomes smaller. Table 3.7 shows the estimated means of the MLEs. We see that when the sample size increases, the estimated means of the MLEs of a and b become closer to the true parameters values $a=-13$ and $b=0.04$. When the sample size reaches 400, the estimated means of the MLEs are basically the true values. For parameter T_{base}, the estimated means using different sample sizes are all fairly close to the true value of $T_{base}=3.5$. Simulation errors (Table 3.8) show that these estimated means of the MLEs as estimated means of the MLEs are realiable. The estimated variances of the MLEs are reported in Table 3.9. When the sample size increases, the estimated variances of the MLEs for all parameters become smaller. These facts show that in Model AGDD, the MLEs of all parameters might be consistent.

Table 3.7: Simulation means of the MLEs. When the sample sizes increases, the estimated means of the MLEs become closer to the true parameter values of $a=-13$, $b=0.04$ and $T_{base}=3.5$

	$S=30$	$S=80$	$S=150$	$S=400$
$\hat{a}$	-13.82	-13.23	-13.20	-13.07
$\hat{b}$	0.043	0.041	0.041	0.040
$\hat{T}_{base}$	3.50	3.50	3.48	3.51

3.5 Assessing the uncertainty of the MLEs

The MLEs are not the true parameters. Without knowing how large the errors of the MLEs are, we won't be able to draw any inference about the model parameters. Under mild regularity

Table 3.8: Simulation errors of the MLEs. Small standard errors of the means of the MLEs imply that the estimated means of the MLEs are reliable estimates of the means of the MLEs.

	$S = 30$	$S = 80$	$S = 150$	$S = 400$
$\hat{a}$	0.066	0.035	0.026	0.015
$\hat{b}$	0.0002	0.0001	0.0001	0.0001
$\hat{T}_{base}$	0.027	0.015	0.010	0.0065

Table 3.9: Simulation variances of the MLEs. When the sample size increases, the estimated variances of the MLEs become smaller.

	$S = 30$	$S = 80$	$S = 150$	$S = 400$
$\hat{a}$	4.31	1.20	0.66	0.22
$\hat{b}$	0.0001	0.0000	0.0000	0.0000
$\hat{T}_{base}$	0.70	0.21	0.11	0.04

conditions, the variances of the MLEs are given by Equation (2.19). However, we cannot use that result, because the likelihood function in Model AGDD is not a smooth function of T_{base}. Here, we will use bootstrap to assess the standard deviations of the MLEs and quantile based confidence intervals for the model parameters.

3.5.1 Consistency of the bootstrap estimators – simulation study

The validity of the bootstrap requires the convergence of the bootstrap estimate to the true parameter value. While for various estimators in many settings, the bootstrap estimator has been proved to converge to the true parameter value, for our model, when using MLE, we don't know if bootstrap is still valid. Instead of studying this issue mathematically, we do a simulation study to check the validity of the bootstrap.

Using the same simulated data that have been used in the last section, for each different sample size S, we can estimate the true variances of the MLEs by the sample variances of the MLEs obtained using the 1000 samples. The standard deviation of the MLEs then is estimated by the square root of these sample variances. The results are shown in the "Sim." fields in Table 3.10. Now we will study how much the bootstrap estimates of the standard deviations of the MLEs differ from these estimates obtained from the simulated data. For each different sample size, we randomly take one sample from the 1000 simulated samples and then take 1000 bootstrap samples from this one sample of response and predictor pairs. For each bootstrap sample, we calculate the MLEs of the parameters. For each parameter, we then take the square

root of the sample variance of the MLEs obtained from the 1000 bootstrap samples as the bootstrap estimate of the standard deviation of the MLE for that parameter. The results are shown in "Boot." fields in Table 3.10. We can see that for each parameter, when the sample size becomes larger, the bootstrap estimates and the estimates obtained using the simulated data both becomes smaller. The bootstrap estimates are always larger than the estimates obtained from the simulated data, but when the sample size gets larger, the difference between the two becomes smaller. For a sample size of 400, the two estimates are fairly close. This might be an evidence of convergence of the bootstrap estimates to the true standard deviations of the MLEs, although they don't seem to converge fast.

Table 3.10: Comparison of bootstrap estimates of the standard deviations of the MLEs and the estimated standard deviations using simulated data. "Boot." stands for the bootstrap estimates; "Sim." stands for the estimates obtained using simulated data. As the sample size increases, the estimated standard deviations calculated using the two different approaches become smaller and also closer.

	$S = 30$		$S = 80$		$S = 150$		$S = 400$	
	Boot.	Sim.	Boot.	Sim.	Boot.	Sim.	Boot.	Sim.
$\hat{a}$	2.23	2.08	1.55	1.10	0.71	0.81	0.54	0.46
$\hat{b}$	0.0102	0.0076	0.0050	0.0041	0.0034	0.0031	0.0017	0.0018
$\hat{T}_{base}$	1.36	0.84	0.49	0.46	0.27	0.33	0.25	0.21

We also want to obtain 95% confidence intervals for the model parameters. Using the MLEs obtained from the simulated data, we can get quantile-based confidence intervals for the model parameters. We also can calculate quantile-based bootstrap confidence intervals using MLEs obtained from the bootstrap samples of one simulated sample. The results are shown in Table 3.11. We see that for each parameter, the lengths of confidence intervals obtained by the two approaches are roughly the same, and as sample size gets larger, they both become smaller. However, the confidence intervals obtained using the two different approaches do not always agree – the bootstrap intervals seem to always have a bias. Fortunately, when the sample size is large, the difference between the two kinds of intervals is pretty small – it is alway smaller than 1/20 of the length of the confidence interval obtained from the simulated data when sample size is 400. We tried a bias corrected version of quantile based bootstrap confidence interval ("BC" method in Efron and Tibshirani 1986), but the results are even slightly worse than this raw version. Overall, although a small bias may exist, the quantile-based bootstrap confidence interval makes sense for our model.

Table 3.12 shows the observed range (minimum value to maximum value) of the bootstrap

Table 3.11: Comparison of quantile-based 95% confidence intervals based on bootstrap and simulated data. "Boot." stands for the bootstrap estimates; "Sim." stands for the estimates obtained using the simulated data. As the sample size increases, the confidence intervals calculated using the two different approaches both become smaller, but they do not always agree very well.

	$S = 30$	$S = 80$	$S = 150$	$S = 400$
a (Boot.)	(-17.44, -10.25)	(-17.94, -12.54)	(-13.96, -11.47)	(-14.31, -12.51)
a (Sim.)	(-17.60, -10.48)	(-15.16, -11.38)	(-14.63, -11.75)	(-13.82, -12.20)
b (Boot.)	(0.034, 0.066)	(0.036, 0.054)	(0.038, 0.050)	(0.036, 0.042)
b (Sim.)	(0.033, 0.061)	(0.035, 0.050)	(0.036, 0.047)	(0.038, 0.044)
T_{base} (Boot.)	(2.45, 5.43)	(2.02, 3.79)	(3.64, 4.60)	(2.82, 3.58)
T_{base} (Sim.)	(2.14, 5.13)	(2.72, 4.39)	(2.95, 4.12)	(3.16, 3.90)

MLEs. We see that for each parameter and all the four choices of the sample sizes S, this range covers and is much larger than the 95% confidence interval obtained using the simulated data. Without knowing the actual coverage probability, this range cannot be directly used as a confidence interval. However, the usefulness of it is that if this range does not contain a value, say θ_0, then we get stronger evidence of saying that the parameter value is not θ_0 than the possibly biased 95% bootstrap confidence interval not containing θ_0.

Table 3.12: Observed ranges of the bootstrap MLEs. These ranges always contain the quantile-based 95% confidence intervals based on the simulated data.

	$S = 30$	$S = 80$	$S = 150$	$S = 400$
$\hat{a}$	(-36.65, -6.43)	(-21.67, -8.55)	(-15.50, -8.98)	(-15.17, -7.68)
$\hat{b}$	(0.0055, 0.1527)	(0.0290, 0.0652)	(0.0324, 0.0575)	(0.0342, 0.0453)
$\hat{T}_{base}$	(-19.00, 11.44)	(0.90, 6.71)	(3.11, 6.52)	(2.51, 7.09)

3.5.2 Bootstrap estimates of the standard deviations and the 95% bootstrap confidence intervals

For each crop, we have assumed that the bloom dates of different years are independent. For each year, we treat the bloom date and the all the daily average temperatures of that year as a data pair. All these data pairs then forms a sample. We then draw 1000 bootstrap samples from this sample, and with each bootstrap sample, we calculate the MLEs of the parameters in Model AGDD. Then, we calculate the Bootstrap estimates of the standard deviations of the MLEs and the 95% confidence intervals for the model parameters.

The results for the standard deviations are shown in Table 3.13 and the results for the 95% confidence intervals are shown in Table 3.14. From the values of the estimated standard deviations and the lengths of the 95% confidence intervals, we see that the uncertainty of the MLE of parameter a is much higher than those of b and T_{base}, and the MLE of b has the smallest uncertainty. This is not a surprise, since the estimated a has the largest absolute value, the estimated b has the smallest absolute value, and usually an estimator for a parameter with a bigger absolute value has a bigger variance.

Given the data, we are interested in knowing whether the regression coefficients a and b are significantly different from 0. Since none of the 95% bootstrap confidence intervals of a and b contains 0, we have a strong evidence that both a and b are not 0 for all crops. The observed ranges of the bootstrap MLEs (Table 3.15) also support this conclusion. That is the intercept term a and AGDD both are important factors that influence the probability of the blooming event occurring on the current day given that it has not occurred on the previous day.

Table 3.13: Bootstrap estimates of the standard deviations of the MLEs

	$\hat{a}$	$\hat{b}$	$\hat{T}_{base}$
Apricot	2.66	0.012	0.85
Cherry	2.86	0.019	1.18
Peach	3.83	0.010	2.18
Prune	4.32	0.013	1.22
Pear	5.77	0.017	0.76
Apple	5.01	0.013	0.95

Table 3.14: Quantile-based 95% bootstrap confidence intervals for the model parameters

	a	b	T_{base}
Apricot	(-22.43, -12.07)	(0.051, 0.096)	(0.95, 4.00)
Cherry	(-21.18, -10.72)	(0.030, 0.095)	(1.01, 5.15)
Peach	(-31.69, -16.37)	(0.030, 0.065)	(-2.51, 1.53)
Prune	(-31.04, -14.39)	(0.046, 0.093)	(0.18, 4.70)
Pear	(-37.95, -16.62)	(0.055, 0.122)	(1.93, 3.81)
Apple	(-39.54, -14.66)	(0.060, 0.111)	(1.84, 6.95)

Table 3.15: Observed ranges of the bootstrap MLEs

	$\hat{a}$	$\hat{b}$
Apricot	(-31.46, -9.49)	(0.029, 0.142)
Cherry	(-36.13, -7.56)	(0.015, 0.150)
Peach	(-40.64, -6.86)	(0.0090, 0.0871)
Prune	(-40.48, -7.48)	(0.018, 0.178)
Pear	(-62.80, -6.85)	(0.023, 0.170)
Apple	(-50.27, -8.90)	(0.034, 0.169)

3.6 Summary

In this chapter, we applied the stochastic based regression model for single event that was developed in Chapter 2 to the bloom date data of six crops in Okanagan region. Our analysis supports scientists' theory: the temperature influences the blooming by means of AGDD. Also, our method can give a sensible estimate of T_{base}, a parameter that is important to understanding the mechanism of blooming.

In next chapter, by using the estimated parameters obtained in this chapter, we will discuss the performance of prediction of our models.

Chapter 4

Application to Phenological Data II – Prediction

Our ultimate goal is to predict future bloom dates of the six crops. We are interested in the following scenario. Suppose we are on the first day of a new year, the blooming event does not occur today and we want to know what the bloom date of a crop is going to be this year in the Okanagan region. We have built Model AGDD and estimated the model parameters using bloom dates and daily average temperatures in past years. Now, suppose we have obtained the daily average temperate of the current day. We will predict the daily average temperatures at day 2, 3, $\cdots$, until the last day of the year. With these predicted daily average temperatures, we then can use Model AGDD to predict the probabilities of blooming on each future day of this year, i.e. a predictive distribution of the blooming event. A prediction of the bloom date then can be made according to this predictive distribution. Now after day one, on the second day of this year, we again do not observe the blooming event. Using the observed daily average temperatures of day 1 and day 2, together with the predicted daily average temperatures on the remaining days of the year, we can use Model AGDD to predict the probabilities of blooming on day 3 to the last day of the year. We repeat this procedure day-by-day until one day we observe the blooming event, starting from which the prediction is no longer needed.

The above scenario has practical value. Whenever we get a piece of new information about temperature, we can update our prediction anticipating that it becomes more accurate. The above procedure is exactly how we are going to update our prediction. If Model AGDD is useful for prediction, then day-after-day, with more new information about temperature, the predictive distributions obtained each day will get more-and-more peaked and the predictions will be more-and-more accurate.

In this chapter, we will test Model AGDD with the above scenario using a leave-one-out cross validation procedure for each crop. First we will build a simple climate model to generate daily average temperatures, then we will assess the predictions of bloom dates and the uncertainties associated with those predictions.

4.1 An ARIMA time series model for predicting daily average temperature

In order to predict the bloom date of a crop using Model AGDD, we first need to predict the future daily average temperatures. There are many ways to get predicted future daily average temperatures. For example, we can use the outputs of a Global Climate Model (GCM). Here, since our main purpose is to evaluate the performance of Model AGDD for the predictions of bloom dates, we want a simple simulation model for temperature that is easy to use and one where predictive uncertainty of temperature can be assessed. In fact, we will consider an ARIMA model (Chatfield, 2004).

Climate is a very complex system. Usually in an observed temperature series, periodic and quasi-periodic signals on different time scales mingle together. By no means can we treat a temperature series as stationary: no clean way exists to remove all the periodic signals and trends. Here as a crude approximation, we will remove the most prominent periodic signal, seasonal variation, and assume the periodic signals in the residue series are so weak that we can ignore them. For each day of a year, we take the average of the daily average temperatures from year 1916 to 2005 in Okanagan, and get one year of long term averaged daily average temperatures. This averaged series then approximates the seasonal signal. We subtract this signal from the original series, and get a residue series, which is assumed to consist of a stationary part and maybe a trend. We then apply an ARIMA model to the residue series. An ARIMA model has three order parameters: the order of the autoregressive part p, the order of the moving average part q, and the order of the differencing d. We write such a model as ARIMA(p, d, q). In an ARIMA model, differencing removes trend and periodic signals of a time series so that after differencing, it appears stationary. In our case, since we assume that the periodic signals in the residue series are ignorable, the differencing is only used to remove any trend that may be present.

The question is now what p, d and q values we should take. Using the residue daily average temperature series from 1916 to 2005 in the Okanagan region, we fit ARIMA models with different combinations of p, d and q values, where p varies from 0 to 6, d varies from 0 to 4 and q varies from 0 to 6. Then we look at the AIC and BIC values for each model. For each d ($d = 0, \cdots, 4$), the ARIMA models that yield smallest AIC or BIC are shown in Table 4.1. We see that, ARIMA(3, 0, 1) without an intercept term gives us both smallest AIC and smallest BIC. We then will take this fitted model to generate daily average temperatures in the future.

It is not clear how well the daily average temperatures generated from the fitted ARIMA(3,0,1) approximate the truth. Here, we will do some simple checks using plots. Figure 4.1 shows the

Table 4.1: Comparison of AICs and BICs for different ARIMA models. A smaller AIC/BIC value corresponds to a better model.

Model	AIC	BIC
ARIMA(3, 0, 1) with an intercept	147830.05	117661.09
ARIMA(3, 0, 1) without intercept	*147828.21*	*117651.14*
ARIMA(1, 1, 2)	147885.08	117691.51
ARIMA(1, 1, 6)	147860.08	117685.31
ARIMA(6, 1, 4)	147848.66	117686.56
ARIMA(2, 2, 2)	147983.16	117790.76
ARIMA(4, 2, 6)	147982.64	117830.09
ARIMA(6, 3, 2)	150424.03	120269.10
ARIMA(0, 4, 6)	157954.62	127779.54

plots of the sample ACF (i.e. correlogram) and the partial ACF (Chatfield, 2004) of the residue series and simulated residue series. We see that the residue series and simulated residue series have similar correlation structures. However, the ACF and partial ACF do not uniquely determine a time series. The time series plots of the residue series and simulated time series are shown in Figure 4.2. We see that, the magnitudes of variations in the residue series are not symmetric about 0. At some time points, the residue series have exceptionally low values, which we do not observe in the simulated residue series. The cause of this difference might be that we didn't account for the periodic signals other than seasonal variation in the ARMIA model, or may be that the noise in the residue series are not inherently normal, issues to be addressed in future work.

4.2 Evaluation of the performance of Model AGDD on prediction

We will evaluate the performance of Model AGDD on prediction using a leave-one-out procedure.

For each crop we put aside one year of bloom data and daily average temperatures as *test data*, and we call the data in the other years *training data*. We then use a procedure similar to the one described in the beginning of this chapter, where we fit our model using the training data, and predict the bloom date in the test data. Based on the observed data and some previous knowledge about bloom dates, for each crop, the probability that a blooming event will occur on the first 60 days of a year or after the 240^{th} day of a year is assumed negligible. Thus, the probability of blooming on those days may not have much practical value. To simplify

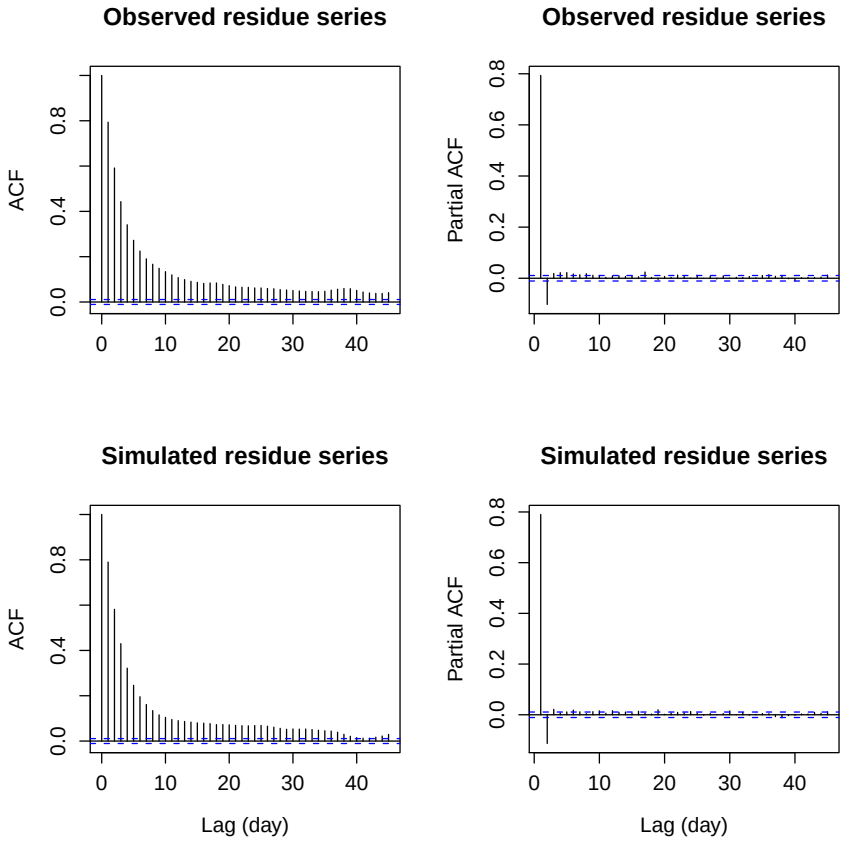

Figure 4.1: The sample ACF and PACF plots of the observed residue daily average temperature series and simulated residue daily average temperature series. The simulated residue series have similar sample ACF and PACF as the observed residue series.

computation, we will only compute the probabilities that blooming will occur for each day between the 61^{th} day and 240^{th} day of a year.

Now, imagine that we are on the first day of the left out year, and suppose we have observed the average temperature of this day. With a fitted Model AGDD and predictions of daily average temperatures from day 2 to day 240, we then can obtain a "plug-in" version of predictive distribution for the bloom date as described by Equation (2.22) by replacing the true parameters by their corresponding MLEs and then integrating over the predicted temperatures. Note that when doing this, we are assuming that the observed daily average temperatures are exact measurements without errors. The details are as follows. First, we generate 1000 replicate time series of one year of daily average temperatures using the ARIMA model we fitted in Section

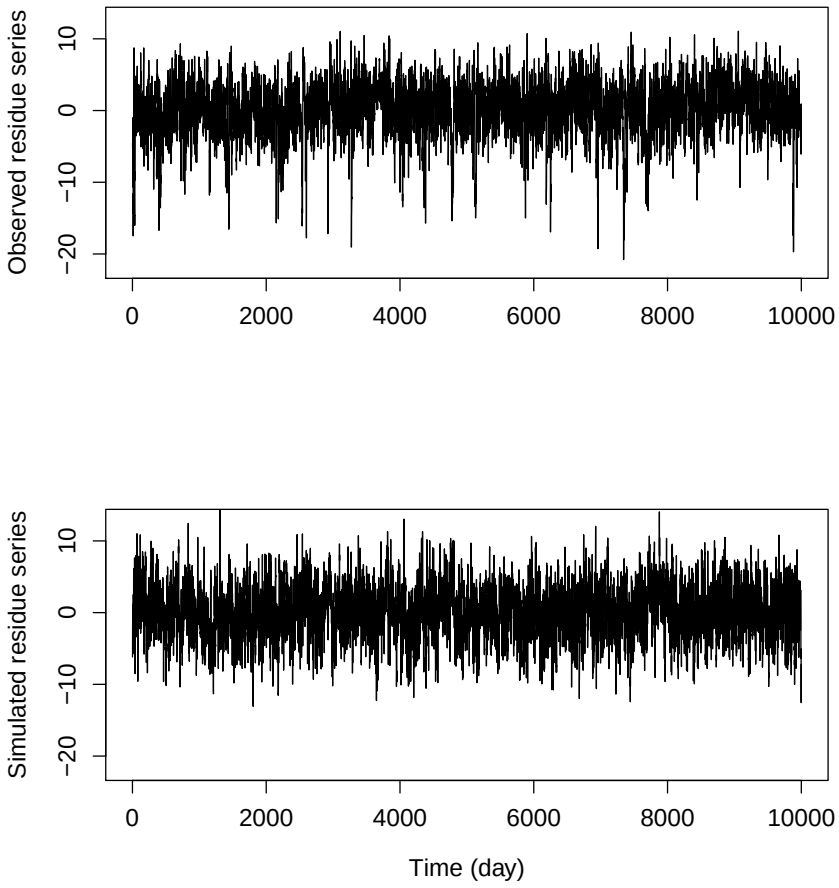

Figure 4.2: Time series plots of the observed residue daily average temperature series and simulated residue daily average temperature series. The magnitudes of variations in the observed residue series do not match those in the simulated residue series very well.

4.1. For each time series s ($s = 1, \cdots, 1000$), we denote the simulated daily temperatures as $z_1^{(s)}$, $z_2^{(s)}$, $\cdots$, $z_{365}^{(s)}$ (the length of this series will be 366 if the year to be tested is a leap year). For convenience, for any $m, n \in \{1, 2, \cdots, 365\}$ and $m \leq n$, we let

$$z_{m:n}^{(s)} = \left\{ z_m^{(s)}, \cdots, z_n^{(s)} \right\} . \tag{4.1}$$

We also denote the true daily average temperatures in the year we are going to predict by z_i^{true}

($i = 1, \cdots, 365$), and let

$$z_{m:n}^{true} = \left\{ z_m^{true}, \cdots, z_n^{true} \right\} , \tag{4.2}$$

for any $m, n \in \{1, 2, \cdots, 365\}$ and $m \leq n$. Then Equation (2.24), in a "plug-in" mode, can approximate the probability of blooming on day t ($t = 61, \cdots, 240$) given z_1^{true} as follows:

$$\mathrm{P}_{\hat{\beta}} \left(T^* = t \left| z_1^{true} \right. \right) \approx \frac{1}{1000} \sum_{s=1}^{1000} \mathrm{P}_{\hat{\beta}} \left(T^* = t \left| z_1^{true}, z_{2:t}^{(s)} \right. \right) , \tag{4.3}$$

where $\hat{\beta}$ is the MLE of the parameter vector obtained from the fitted model AGDD, and T^* is the bloom date of the test year. These probabilities of blooming on day 61, or 62 and so on to day 240 is a discrete predictive distribution of the bloom date of that year, given that we know only one true daily average temperature z_1^{true}. Now suppose we are on day 2 of the year, and we observe z_2^{true}. As on the first day, we can get predictive distribution $\mathrm{P}\left(T^* = t \left| z_{1:2}^{true} \right. \right)$ for $t = 61, \cdots, 240$. We repeat this procedure day-by-day to get a predictive distribution, up until the bloom date, t_b, after which prediction is unnecessary. On an arbitrary day n where $1 \leq n < t_b$, the general formula for the predictive distribution is

$$\mathrm{P}_{\hat{\beta}} \left(T^* = t \left| z_{1:n}^{true} \right. \right) \approx \frac{1}{1000} \sum_{s=1}^{1000} \mathrm{P}_{\hat{\beta}} \left(T^* = t \left| z_{1:n}^{true}, z_{(n+1):t}^{(s)} \right. \right) , \tag{4.4}$$

where $t \in \{\max(61, n), \cdots, 240\}$.

For one year of test data, we get a lot of predictive distributions. For each predictive distribution, we calculate a quantile based 95% (2.5% quantile – 97.5% quantile) prediction interval (PI), and see if the true bloom date of that year falls in the 95% PI. Also we calculate the mean, median and mode of this predictive distribution as possible choices of a point prediction. Then we leave out another year's as the test data, and use the remaining years' as training data. We repeat the exact same procedure as above on the new test dataset and training dataset to get PIs, means, medians, and modes of the predictive distributions. This way, every time we leave out one year of data as the test data until we finish testing our model on the data of each year. We then put all the predictive distributions for all the test datasets together, and for each method of point prediction (mean, median, or mode of a predictive distribution), we calculate the overall root mean square error (RMSE) and mean absolute error (MAE). We also calculate the ratio of the number of the 95% PIs covering the true bloom dates over the total number of PIs as an estimate of the coverage probability of the 95% PI.

The cross validation results for all the crops are shown in Table 4.2. We also provide the observed range of the bloom dates (the difference time between the maximal and minimal

Table 4.2: Cross validation results: The RMSEs and MAEs for point predictions using mode, median and mean are shown in column 2–7. The estimated coverages and average lengths of the 95% PIs are shown in the last two column respectively. The units for RMSE, MAE and average length of the 95% PI are day. The estimated coverage probabilities of these 95% PIs are generally too high.

Crop	Mode		Median		Mean		95% PI	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	Coverage	Ave. Len.
Apricot	6.74	5.12	6.58	5.06	6.62	5.14	0.99	33.24
Cherry	6.76	4.97	6.59	4.88	6.59	4.92	0.99	34.29
Peach	5.43	4.09	5.33	4.04	5.34	4.05	0.99	28.41
Prune	5.82	4.31	5.45	4.11	5.46	4.16	0.99	30.55
Pear	5.60	4.19	5.65	4.36	5.69	4.40	0.99	29.99
Apple	5.39	4.07	5.44	4.19	5.45	4.23	0.99	28.86

Table 4.3: Maximum, minimum and range of the observed bloom dates for each crop in 1937–1964 in the Okanagan region

	Apricot	Cherry	Peach	Prune	Pear	Apple
Maximum (day)	126	136	135	138	139	146
Minimum (day)	94	102	105	111	110	115
Range (day)	32	34	30	27	29	31

observed bloom dates) of each crop (Table 4.3) as a measure of natural variation of the bloom dates for each crop. We see that mean, median and mode as point predictions perform roughly the same in terms RMSE and MAE. The RMSEs for all crops fall between 5.3 and 6.8 days, and the MAEs fall between 4.0 to 5.2 days. Considering the observed ranges of the bloom dates, which vary from 27 to 34, our point predictions provide useful information about the future bloom dates. The estimated coverage probabilities of 95% PIs are higher than 98% for all crops, in conflict with our expectation of 95%. For each crop, the average length of the 95% PI is roughly the same as the observed range of the bloom dates, in accord with the high estimated coverage probability. These imply that our 95% PIs are generally too wide, possibly because the prediction of the daily average temperature is too crude. We did include a lot of variability in the series which is caused by periodic signals other than seasonal variation as the variability of the random noise in the ARIMA(3,0,1) model. The consequence of this is that we may have put too much uncertainty into the simulated daily average temperatures, which in turn induces excessively wide 95% PIs in the final predictions of the bloom dates.

We now try to reduce the variance of the white noise in the ARIMA(3,0,1) to half the estimated variance, while keeping all the other estimated parameter values unchanged. We use this new ARIMA(3,0,1) model to generate daily average temperatures, and then we perform the above cross validation procedure again. The results (Table 4.4) show that while the accuracy of the point predictions is roughly the same as before, the estimated coverage probabilities and average lengths of the 95% PIs are reduced to reasonable values. This result does not confirm that the high estimated coverage probabilities are actually caused by the high uncertainty in the predicted daily temperatures, but it at least adds weight to this explanation.

Table 4.4: Cross validation results when using variance reduced simulated daily average temperatures: The RMSEs and MAEs for point predictions using mode, median and mean are shown in column 2–7. The estimated coverages and average lengths of the 95% PIs are shown in the last two column respectively. The units for RMSE, MAE and average length of the 95% PI are day. The estimated coverage probabilities of these 95% PIs are reasonable.

Crop	Mode		Median		Mean		95% PI	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	Coverage	Ave. Len.
Apricot	6.91	5.20	6.78	5.12	6.72	5.08	0.94	24.87
Cherry	6.62	5.06	6.64	5.05	6.58	4.99	0.93	26.46
Peach	5.56	4.03	5.51	3.95	5.49	3.98	0.95	21.17
Prune	5.48	4.16	5.46	4.04	5.46	4.07	0.98	22.38
Pear	5.98	4.46	5.79	4.33	5.75	4.31	0.94	21.45
Apple	5.76	4.36	5.53	4.17	5.48	4.15	0.95	20.36

4.3 More on predictive uncertainty

As described at the beginning of this chapter, if our predictions are reasonable, when time becomes closer to the bloom date, the point predictions should become closer to the true bloom date, and the predictive distributions become more-and-more peaked. To check this, for each crop, we calculate the average lengths of 95% PIs over years at each day from 90 days prior to the bloom date (we call it lag -90) to 1 day prior to the bloom date (lag -1). The results are shown in Figure 4.3. It is clear that for all crops, the 95% PIs becomes narrower when time approaches the bloom date. We do the same thing on the MAE of the point prediction using the median of the predictive distribution. The results are shown in Figure 4.4. Although the curves are much "bumpier" than those in Figure 4.3, the decay trend with the time is obvious.

Until now, although we have calculated many quantities to characterize the predictive distribution, we haven't see the shape of it. We "randomly" pick a calculated predictive distribution:

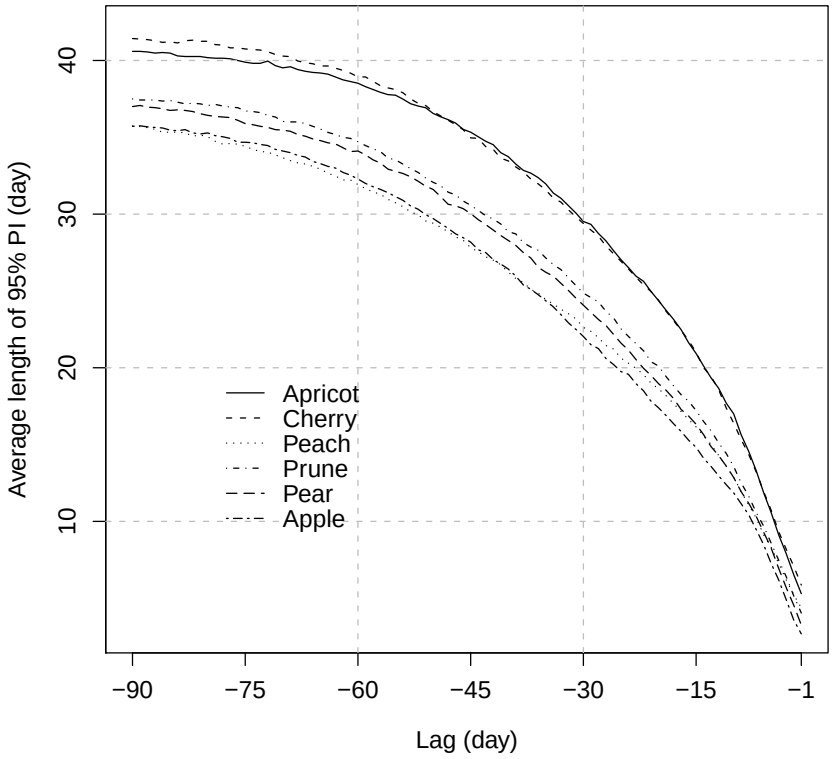

Figure 4.3: Change of the average length of 95% PIs with the change of lag. The predictive uncertainty decreases when time approaches the bloom date.

the predictive distribution of peach in year 1944 with daily average temperatures of the first 60 days of that year known. Note that the true bloom date of peach in that year is day 125. This predictive distribution (smoothed) is plotted as the solid curve in Figure 4.5. We see that it is a bell-shaped distribution which roughly looks like a normal distribution. Since the predictive distribution is calculated by plugging in the MLEs as if they were the true parameters, there is an uncertainty associated with this predictive distribution. Just as before, we again use the bootstrap to assess this uncertainty. We calculate a quantile based 95% bootstrap confidence band (shown as the shaded area in Figure 4.5) for this predictive distribution. We see that this confidence band is not too wide, so we basically can "trust" this predictive distribution.

Here, when we apply the bootstrap, we have the same problem as before: we don't know whether the bootstrap estimate of the uncertainty reflects the true uncertainty of the predictive

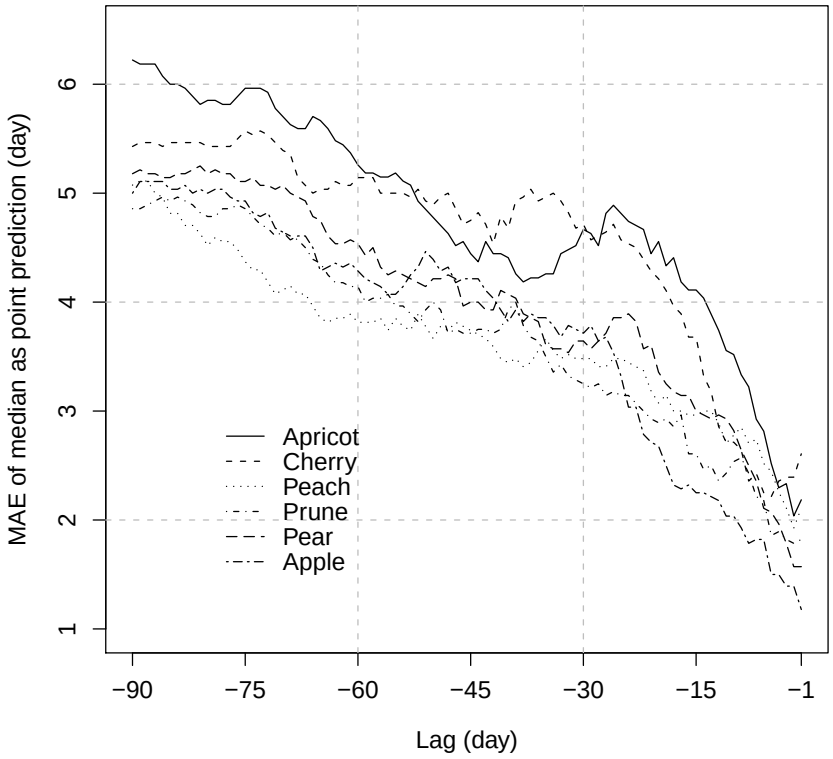

Figure 4.4: Change of the MAE of median with the change of lag. The point prediction becomes more accurate when time approaches the bloom date.

probability. We will do a simulation to investigate this issue. Take the settings for the simulation described in Section 3.4. For each sample size S, where $S \in \{30, 80, 150, 400\}$, we now generate one more year of data as test data. For a fixed sample size S, for each sample, we estimate the model parameters, and we then use this set of of parameters to predict the bloom date of the test year by assuming the first 60 days of temperatures are known. We then get 1000 predictive distributions for each sample size. For each future day, we take the sample standard deviation of these 1000 predicted probabilities as an estimate of the standard deviation of the predicted probability. we take the 2.5% and 97.5% sample quantiles of the 1000 predictive probabilities to approximate a quantile based 95% confidence interval for the predictive probability. Now, randomly pick one sample, and then take 1000 bootstrap samples of this sample, and estimate model parameters using each bootstrap sample. With each set of estimated param-

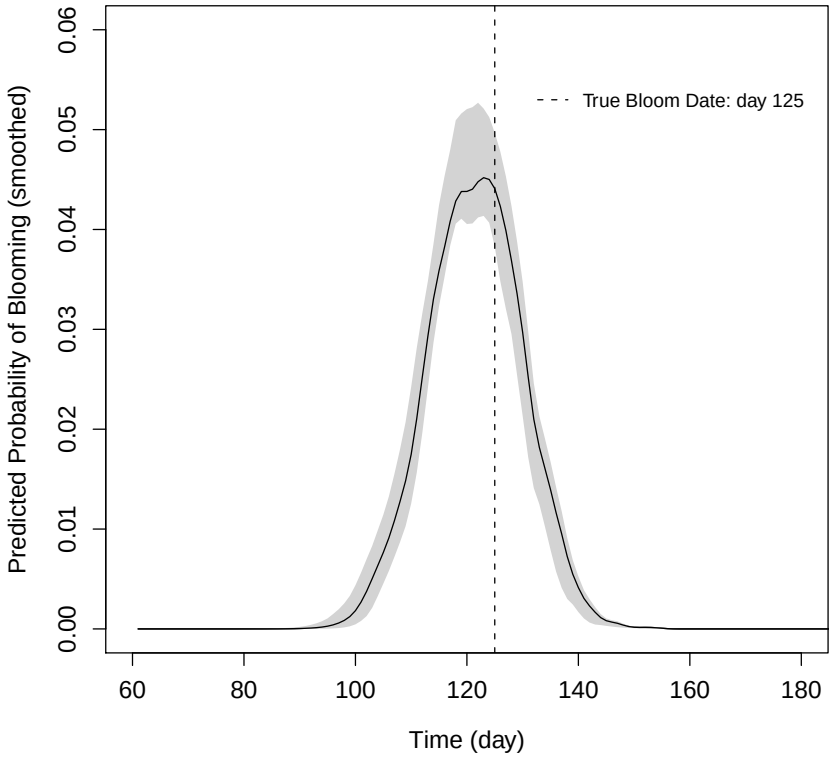

Figure 4.5: The predictive distribution (solid curve) of peach in year 1944 with daily average temperatures of the first 60 days of that year known. The shaded area is a 95% confidence band for this predictive distribution. The true bloom date of peach in that year is day 125.

eters obtained from the bootstrap, we can then make a prediction on the test year. With 1000 bootstrap samples, we get 1000 predictive distributions. For each future day, as with the simulated data, we can obtain a quantile based 95% bootstrap confidence interval for the predictive probability. We now compare the confidence intervals obtained in these two ways. Randomly picking one future day, the 95% confidence intervals for the predictive probability of blooming that day calculated using the simulated data and bootstrap are shown in Table 4.5. We can see that both types of confidence intervals become narrower when sample size becomes larger. For each sample size, the bootstrap interval is close to the interval obtained using the simulated data. Moreover, when the sample size reaches 400, the two types of intervals are basically identical. This result shows that applying bootstrap might be a sensible way to estimate the

uncertainty of the predictive probabilities.

Table 4.5: Comparison of the 95% confidence intervals for predictive probabilities obtained using bootstrap and the simulated data.

	$S = 30$	$S = 80$	$S = 150$	$S = 400$
Bootstrap	(0.0300, 0.0350)	(0.0321, 0.0358)	(0.0300, 0.0320)	(0.0305, 0.0322)
Simulation	(0.0288, 0.0348)	(0.0298, 0.0332)	(0.0302, 0.0326)	(0.0307, 0.0320)

Finally, we evaluate the accuracies of the predictions given that we know all the daily average temperatures in advance. In this situation, there is no uncertainty associated with the daily average temperatures. The cross validation results are shown in Table 4.6. When daily average temperatures are known, there is only one prediction for one test year for each crop, so the number of predictions for each crop is not big enough to give a sensible estimate for the coverage probability of the 95% PI. Hence, we don't report the estimated coverage probability here. We see that the accuracies of these predictions are much higher than those of our previous predictions, and the average lengths of the 95% PIs are much smaller. These, however, are not real predictions, since in real situations we have no way of knowing the true daily average temperatures. Nevertheless, these results tell us that Model AGDD does make sense for the prediction of the bloom dates of these crops. Also, if we can accurately predict daily average temperature, then we can significantly improve the prediction of bloom date.

Table 4.6: Cross validation results if future daily average temperatures were known: The RMSEs and MAEs for point predictions using mode, median and mean are shown in column 2–7. The average lengths of the 95% PI are shown in the last column. The units for RMSE, MAE and average length of the 95% PI are day. The point predictions are very accurate, and the average lengths of the 95% PIs are short.

Crop	Mode		Median		Mean		95% PI
	RMSE	MAE	RMSE	MAE	RMSE	MAE	Average Length
Apricot	4.18	3.30	3.65	2.93	3.62	2.90	13.48
Cherry	3.74	2.93	3.39	2.43	3.46	2.42	18.43
Peach	3.43	2.85	3.35	2.78	3.27	2.76	12.67
Prune	3.06	2.54	2.98	2.50	2.93	2.36	11.46
Pear	2.93	2.11	2.64	1.89	2.67	2.00	9.21
Apple	2.12	1.86	2.04	1.71	1.97	1.61	8.29

4.4 Summary

In this chapter, we provide a method for the predictions of bloom dates. We also provide a crude ARIMA model for predicting daily average temperatures. Leave-one-out cross validation results shows that the 95% PIs of our predictive distributions are generally wide and have bigger coverage probabilities than 95%. This may be because our ARIMA model over-estimates the uncertainty in daily average temperatures. If we reduce the uncertainty in the ARIMA model, we then get reasonable 95% PIs. Cross validation results also show that if we can get accurate predictions of the daily average temperatures, we then can significantly improve our predictions of bloom dates. In all, Model AGDD is useful for the predictions of bloom dates.

Chapter 5

Conclusions and Future Work

This thesis aims to build regression models capable of incorporating all information about time-dependent covariates and making sensible predictions for phenological data, a type of time-to-event data. Traditional models that are frequently used for dealing with time-dependent covariates are the Cox model and parametric proportional hazards models. However, these models encounter difficulties in our context. The Cox model does not use all the information in the time-dependent covariates and it is not generally suitable for prediction. At the same time the proportional hazards models involve complicated integration without a closed-form solution when time-dependent covariates are present. Also, they usually require strong distributional assumptions.

To achieve our goals, we have developed a regression model based on stochastic processes. Instead of directly modeling the hazard function, we considered dummy indicator variables which indicate the status of the event on a discrete time scale. We showed that for a single event, these indicator variables at discrete time points have a first-order Markovian structure, which helps us greatly simplify the formulation of the regression model. The parameters in this model can be estimated using many standard estimation procedures, such as the ML method. Prediction is also straightforward given that we can predict or obtain the time-dependent covariates associated with the event to be predicted. With are additional assumptions, We extended this model to account for multiple sequential events. Both models for single event and multiple sequential events are not only suitable for phenological data, but also apply to a broad class of survival data.

We applied our regression model for single event to bloom dates of six high-value perennial agricultural crops in the Okanagan region of British Columbia for two purposes: (1) studying the response of bloom dates to daily average temperatures; and (2) predicting future bloom dates. Results from a model selection procedure support the experimental result that bloom dates relate to the accumulation of GDDs. Also, we provided a sensible way of estimating T_{base}, the important thresholding parameter in the GDD function. To test our model for prediction, we performed a leave-one-out cross validation procedure. Results show that our Model AGDD makes sense. The predictive uncertainty is too high, though, probably because the ARIMA

model we employed to generate daily average temperatures is too crude. We found out that if we can manage to get accurate predictions of daily average temperatures, the accuracy of prediction of bloom dates using Model AGDD is high.

A restrictive distributional assumption in our models is that the stochastic process of the event status indicator variable must be time-homogeneous. In the future, one can try to allow model parameters to change with time to relax this assumption. Also, one can build larger models to account for the correlation among different crops at one location or even among crops at different spatial locations. Such models not only help to improve the estimation of model parameters and the predictions of future events, but also can provide a better understand the current and future distribution of perennial crops. More broadly, such results can be used to help improve the assessment of agricultural risk associated with climate variability and extremes.

Bibliography

Norman Breslow. Discussion on professor cox's paper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):216–217, 1972.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2001.

Chris Chatfield. *The Analysis of Time Series*. Chapman & Hall/CRC, 6ed edition, 2004.

I. Chuine, P. Yiou, N. Viovy, B. Seguin, V. Daux, and E. Le Roy Ladurie. Grape ripening as a past climate indicator. *Nature*, 432:289–290, 2004.

Isabelle Chuine. A unified model for budburst of trees. *Journal of Theoretical Biology*, 207(3): 337–347, December 2000.

David Collett. *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, 2 edition edition, 2003.

Richard J. Cook and Jerald F. Lawless. *The Statistical Analysis of Recurrent Events*. Springer, 2007.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall/CRC, 1st edition, 1979.

D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall, 1984.

Bradley F. Efron and David V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65:457–487, 1978.

Bradley F. Efron and Robert J. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–77, 1986.

Bradley F. Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1st edition, 1994.

Julian J. Faraway. *Extending the Linear Model with R*. Chapman & Hall/CRC, 2006.

Philip Hougaard. *Analysis of Multivariate Survival Data*. Springer, 2000.

John D. Kalbfleisch and Ross L. Prentice. Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, 2nd edition, 2002.

Sadanori Konishi and Genshiro Kitagawa. *Information Criteria and Statistical Modeling*. Springer, 2008.

Jun S. Liu. *Monte Carlo Stragegies in Scientific Computing*. Springer, 1st edition, 2001.

M. B. Murray, M. G. R. Cannell, and R. I. Smith. Date of budburst of fifteen tree species in britain following climatic warming. *J. Appl. Ecol.*, 26:693–700, 1989.

M. D. Schwartz, R. Ahas, and A. Aasa. Onset of spring starting earlier across the northern hemisphere. *Global Change Biology*, 12(2):343–351, 2006.

Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1998.

Richard L. Smith. Bayesian and frequentist approaches to parametric predictive inference. In J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 589–612. Oxford University Press, 1998.

T. H. Sparks, E. P. Jeffree, and C. E. Jeffree. An examination of the relationship between flowering times and temperature at the national scale using long-term phenological records from the uk. *International Journal of Biometeorology*, 44:82–87, 2000.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Annal of Mathematical Statistics*, 20:595–601, 1949.