The Dynamic Behavior in Work Absence*

Yuhao Li
yuli@eco.uc3m.es
Latest Version Here

Departamento de Economia, Universidad Carlos III de Madrid

July, 2019

Abstract

We use the self-exciting processes to study individuals' absence behaviors. Such behaviors are dynamic because of the firm's absence regulation, where a worker's absence records determine her absence benefit. The self-exciting process is state-dependent and enables us to include the individual's absence records into the model. We decompose an absence into an incidence event ('asking for absence') and a recovery event ('returning to work'). For each absence, we also distinguish short-term from long-term. Using firm-level data, we find that workers do consider absence records when they make short-term incidence and recovery decisions, but this is not the case for long-term events. Inspired by the empirical results, we build a simple economic model.

JEL. C51, C41, J22, J32, C13

Keywords. Work Absence, Self-Exciting Process, Count Data Regression, Duration Analysis

*I am extremely grateful to my advisor Miguel A. Delgado for his encouragement and weekly guidance throughout this project and to Winfried Stute for his inspiration. I am thankful to Tim Barmby and Michael Nolan for their generous permission to use the dataset. I would also like to thank: Carlos Velasco, Juan Carlos Escanciano, Jesus Gonzalo, Jan Stuhler, Nazarri Salish, Rui Cui, Minghai Mao, Junyi Peng Zhou, Pedro Sant'Anna, Julius Vainora, Xi Lin and seminar participants at the Universidad Carlos III de Madrid, the UC3M-Cambridge-LSE Econometrics Workshop ,the Mannheim ENTER Seminar, the ESWM Naples, the SAEe Madrid and IAAE Nicosia. All remaining errors are mine.

1 Introduction

The primary purpose of this chapter is to investigate the dynamic behaviors in work absences, where individuals may take past experiences into absence decision-making considerations. Dynamic here means preferences or constraints to future choices are altered by past experiences. A typical source of such behavior is the absence score that most firms employed in their work absence regulations. Absence scores are accumulated over time and are based on individuals' absence records. In principle, higher absence scores lead to more severe penalties (e.g., fall of income or even possible layout) and vice verse. Individuals then have to dynamically make their absence decisions. In this paper, we use the self-exciting process, a special counting process, to model the dynamic and state-dependent absence decision-making processes.

Work absence is not uncommon among both developed and developing countries. U.S Bureau of Labour Statistics (2005) data reveals that, on any given day, approximately 3.3% of the U.S. workforce does not report to work. Duflo et al. (2012) reports the absence rate in an Indian NGO teacher program could be as high as 35%. Moreover, work absences are costly for both workers and firms. For workers, although the social security covers illness-related absences in some countries in the form of sick pay, the replacement rates are in general less than 100%. For firms, arguably, labour costs are the single most considerable budgetary expense. Fister-Gale (2003) cites research showing that absenteeism costs in one survey population accounted as high as 14.3% of total payroll.

Despite the sizable number of work-hours involved and impacts on productivity, economists have paid little attention to the issue of absenteeism. Early works by Allen (1981) and Barmby et al. (1991) demonstrate the importance of financial aspects in explaining absence behavior. A group of Norwegian economists contribute significantly to this filed. Markussen et al. (2011) show that employee heterogeneity drives most of the cross-section variation in absenteeism. Fevang et al. (2014) show that Norway's social security system of short-term pay liability creates a sick pay trap: firms are discouraged from letting long-term sick workers back into work.

Applied psychologists and management specialists contribute most in the work absenteeism literature. In general, psychological literature argues, according to Steers and Rhodes (1978), that the job dissatisfaction represents the primary cause of absenteeism. In management literature, however, this view has been challenged. Increased understanding of the importance of so-called trigger absence behavior has emerged from the management literature (Steel et al., 2007). These literature argue that absence scores is a significant work absence decision-making factor. However, no satisfied empirical work has been done to support this claim.

In this paper, we aim to provide empirical evidence on the existence of this trigger absence behavior. The inclusion of absence score creates a state-dependent structure in the econometric models. We use self-exciting processes to incorporate such structure. The self-exciting process is a counting process whose filtration is generated by the counting process itself. Thus, a self-exciting process is state dependent: past experience has effects on the future events. Throughout the analysis, we try to avoid making further assumptions other than the independent, identical distributed individuals. However, within a single individual, absences are not independent.

To fully explore the features of the self-exciting process, we decompose a work absence into two decision making processes: an incidence process and a recovery process. An event in the incidence process is defined as a worker asking for an absence. An event in the recovery process happens when a worker decides to return to work from the absence. Incidence processes and recovery processes are different. Individuals may encounter shocks (e.g. illness) and seek absences, but at first, they do not have full information about the shock sizes. However, such information is available when individuals make the returning decisions. The stochastic interpretation is also different. For example, the likelihood for a hard-working individual to ask for a leave is low, while the likelihood for the same individual to return to work quickly is high. To this end, we will model these two processes separately.

We also distinguish between the short-term and the long-term absences. Short-term absences are more voluntary than the long ones. The motivation for such absences can be interpreted as maximising one's leisure time under a reasonable budget constraint. The long-term absences, on the other hand, are often related to 'involuntary' causes. A typical example is the sick-leave. Thus, in total, we will construct four types of models: short-term incidence, short-term recovery, long-term incidence and long-term recovery.

One challenge we faced is the unobserved heterogeneity. To consistently estimate the state dependent effect, one needs to separate the unobserved heterogeneity and the state dependent. We achieve so by restricting the structure of our model and perform a first ratio difference.

Comparing to the conventional econometric tools used in work absenteeism: the count data regressions (Delgado and Kniesner, 1997) and the duration models (Barmby et al., 1991; Fevang et al., 2014; Markussen et al., 2011), the self-exciting process has the advantage of modeling the dynamic decision-making process. In the count data regression literature, the study subject is the counts of events during a period. Thus, count data models lose the dynamic information by aggregating the absence records over the defined period. Duration models often assume that absence durations are i.i.d, which is incompatible with the state-dependent setting. Lagged duration models (Honoré, 1993) do exist, but they are difficult to apply.

Structural econometric models can be used to include the state-dependence. However, it would be quite complicated when one tries to model four decision making processes (short (long)-term ask for leaves, short (long)-term return to work) with only one economic model.

The modeling strategy we used (i.e., the separation of incidence and recovery decision making processes and the distinction between short-term and long-term absences) requires a raw absence records dataset, in which the researchers should have access to the details of each absence, including the beginning and ending dates as well as necessary individual demographic information. In our empirical study, a firm-level administration dataset is used. We will formally introduce the data in the later section.

This paper contributes to three strands of literature. First, we provide substantial evidence on the existence of dynamic behavior in the work absences. Specifically, we observe workers dynamically make absence decisions in short-term absences in both incidence and recovery processes. While in the long-term absences, such dynamic behavior plays an insignificant role in the decision making processes.

Second, we provide a modeling method that complements to the conventional methods (structural models, count data regression and duration analysis models.) Instead of focusing on the events per se, we try to model the whole behavior process. Thus we require i.i.d of individuals, but not the events.

Last, we propose a new way to swipe out the unobserved heterogeneity. Unlike the existing literature where the Independence assumption on the unobserved heterogeneity and other covariates is essential for identification and estimation, our method relax such restriction and can be categorized as the fixed effect way.

The paper is structured as follow. Section 2 introduces the data and provides some preliminary results based on conventional count data regression and duration models. The aims of these preliminary results are mainly to show the existence of dynamic behavior in work absences, to highlight the incompatibility of the conventional methods and to illustrate the nature of the problem. In section 3, we first introduce some notations and basics about the self-exciting process, followed by the presentation of our model. We also discuss the difficulties to include the unobserved heterogeneity in the model and our solution. In section 4 the estimating results are presented and discussed. Based on the empirical findings, we develop a simple economic model in section 5. Section 6 compares the self-exciting process to count data regression and duration analysis models. Finally, section 7 concludes the whole paper.

2 Data and Preliminary Results

In this section, we briefly introduce the data and present some preliminary results based on conventional count data regression and duration models. Following the procedure proposed by Heckman (1981), we also provide some evidences that support the existence of state dependence in the data. At the end of this section, we will illustrate the nature of the work absenteeism problem.

2.1 The Data

The data we used come from a UK based manufacturing firm, which produces a homogeneous product using production lines. Other publications that use the same data (or a subset of the data) are Barmby et al. (1991),Barmby et al. (1995), etc. In 1983, the firm introduced an experience rated sick-pay scheme where workers with less cumulative absence scores receive a better sick-pay benefit. More specifically, the scheme provides the sick-pay benefit at three levels: Grade A workers are paid with their full normal wage including bonuses less the statutory sick-pay (SSP) of the UK social security; Grade B workers are paid with their basic wages less SSP; Grade C workers receive no benefits from the firm. All the workers are eligible to the SSP.

To be eligible to the SSP, workers should be absent from work for more than three consecutive days. Because of this requirement, we define the short-term absences as the ones whose duration are less or equal to 3 days, and the others are categorised as long-term.

Workers are categorised into these three grades based on the absence records over the previous two years: at any given time, individuals need to consider both last year's and current year's absence scores, since these scores will decide next year's benefit. Each day of absence attracts a certain number of 'points', mostly 1 point, depending on the cause of this absence. To simplify our analysis, we assume that one day off is 1 point of absence score. Grade A workers have less than 21 points, Grade B workers have 21 to 41 points, and Grade C workers are those above 41 points.

We believe there is no abnormal behavior occurs around the cut-off points 21 and 41. To show it, we non-parametrically estimate the absence score density function at the end of the year 1987 and 1988. Figure 2 plots the result. The P.D.Fs are smooth around these cut-off absence scores. Some possible explanations to this smoothness could be 1) It is difficult to foresee the occurrence of a future absence, 2) the absence regulation renews every two years, the last year's absence records (1988) to determine 1989's sick pay benefit is also the middle year to determine the benefit for 1990, hence the absence score is updated in a 'smooth' way, and 3) the absence score will only affect the sick-pay benefit (which is stochastic: only receive the benefit when ill) not the salary (which is deterministic), hence the incentive to 'control' the absence scores around the cut-off points are not strong.

[Insert Figure 2 Here]

A worker's decision to be absent will not only lead to a loss of earnings[1] but also affect the eligibility for the sick-pay at some point in future, usually in a stochastic fashion. The incentives to take a leave and to return to work from an absence created by this scheme are complex and raise challenges for econometric analysis.

[1]That is not the case for class A workers whose benefit will not be affected during an absence. However, for the other two classes, some loss of income is a certain.

The data consists of detailed absence records: the beginning and ending dates of absences, type of absences (sick-leave, maternity release, jury service, work accident etc.) as well as individual characteristics such as age, gender, contract type, etc. In this paper, we will deal with the 'working age', which is the real age subtracts the legal working age (16 in the UK). Some common covariates such as education, wage and job hierarchy are not included in this dataset. However, we do not think these covariates could play significant roles: most workers are blue-colour, who have similar education backgrounds, receive similar wages and their job levels are more or less the same. We use the data from calendar year 1987 to 1988. In total, we have 749 workers with 5718 absence records.

Figure 3 shows the histogram distribution of the length of absences. Among all the absences, 1-day off leaves account for more than half. Around 78.1% are short-term absences. Long-term absences, especially those longer than ten days are rare.

[Insert Figure 3 Here]

2.2 Preliminary Results

Conventionally, count data regression (Delgado and Kniesner, 1997) and duration models (Markussen et al., 2011) are commonly used in the analysis of sickness absences. In this subsection, we provide some preliminary results using these methods.

The subject under study in the count data regressions is the counts of occurred events over a period. In our application, this subject would be the number of absence records in the year 1988. We use four count data regressions: the Poisson, the negative binomial, the zero-inflation and the hurdle models. The Poisson regression is the basic model for count data analysis. One restriction to this model is the equidispersion: the mean of the counts must be equal to the variance. To overcome this restriction, researchers have proposed more general over-dispersion model. Negative binomial model is particularly popular. One source of over-dispersion is the excess of zeros. Two models are often used to deal with this property: zero-inflation and hurdle models. The general idea is first to use binomial distribution to describe the zeros and then to use another probability distribution to describe positives. In the zero-inflation model, the second probability distribution can generate both zeros and positives. While in the hurdle model, this probability distribution is truncated at zero. We left the technical description of these count data regression models in Appendix A. One important trait that we include in the models are the absence counts in the previous year (1987). The goal of this trait is to obtain some insights on how past experiences could alter future decisions.

Table 1 summarises our count data regression results. One crucial explanatory variable is the number of times of absences in 1987, which are used as an approximation of heterogeneities of individuals. The results are quite similar across different models. This conclusion is consistent with previous literature (Delgado and Kniesner, 1997).

[Insert Table 1 Here]

Another commonly used tool is the duration analysis. Here, we study the duration of attendance until the first absence in a year. The workhorse in the duration analysis is the hazard rate, which is the ratio of the probability density function to the survival function. It can be interpreted as the failure rate or the force of mortality. We study a baseline duration model, where the hazard rate is constant over time and no presence of the unobserved heterogeneity. Appendix A documents the details of this model. The first column in Table 2 reports the estimation results of this standard duration model.

[Insert Table 2 Here]

We also study a more commonly used duration model where the unobserved heterogeneity is introduced. This hazard rate has a multiplicative form of the unobserved heterogeneity term, a random variable, and the remaining part. As proposed by Heckman and Singer (1984), we use discrete distribution to approximate the true random variable distribution, and obtain the non-parametric maximum likelihood estimator (NPMLE). Detailed description about this extension model as well as the NPMLE can also be found in Appendix A. Note that one requirement to use the NPMLE is the independent of the unobserved heterogeneity with all other covariates. In our application, this is clearly violated, as the absence counts in the previous year is correlated with the heterogeneity term. Nevertheless, we still present the results.

Column 2 of Table 2 presents the estimation results for this model. The log-likelihood value for two mass points and three mass points are almost the same and the probability associated with the third mass point is close to zero. Based on these information, we believe, two mass points would be good enough.

Count data regressions and duration models are incapable of studying the strategic behavior. For count data regressions, the information is aggregated at the end of one year. Hence the dependent structures among events are lost. For duration models, one needs to maintain the events independence assumption. Thus, by design, the duration models assume that past events are uncorrelated with future ones. Notice some multiple-spell models break the independence assumption and allow lagged duration dependence (e.g.,Honoré (1993)). However, this lagged duration model is in a panel setting, and its hazard rate can be very difficult to study, since one needs to separate the state dependence from the unobserved heterogeneity. This difficulty is even more intimidating when one distinguishes short-term and long-term absences, as these two panels are shocks to each other.

From the preliminary results of both count data regression and duration analysis, we have seen that the counts of previous year's absences are positively correlated with the dependent variable (count data regression) and the hazard rate. It would be wired to interpret these results as causal since it implies the more absences one took last year, the more absences one would ask for in this year; or the more absences one took last year, the higher hazard (hence, the shorter the attendance duration) one would have. This interpretation contradicts the intention of the firm's absence benefit program. The proper interpretation of this trait should be an approximation of heterogeneities of individuals: frequent-absence workers tend to have more absences all the time, while less frequent workers should have fewer absence records in the future.

2.3 The Nature of the Problem

In this subsection, we try to illustrate the econometric challenges when modeling the strategic behavior. To have a better understanding, considering Figure 4, which demonstrates a possible realization of work absences. The dash lines here are absence periods, and the solid lines are the attendance periods. Lower case 's' and 'r' are the starting and recovery dates of a short-term absence respectively, and upper case 'S' and 'R' are the starting and recovery dates for a long-term absence. In this example, we have two short-term absences before a long-term absence.

[Insert Figure 4 Here]

Suppose now we are at $t \in [r_1, s_2)$ and the goal is to investigate how likely the next absence is going to occur at time $t + dt$. To account for the strategic behavior, one needs to include previous absence records. In our application, it is the cumulative absence time that matters most. Hence $d_1 = r_1 - s_1$ should be in the model. We would expect the coefficient of the cumulative absence time is negative if the firm's absence benefit program is working. That is, larger cumulative absence time will discourage any further absence behavior.

Besides, we would also like to investigate whether asking for leaves is duration dependent. Therefore, we need to include a time dependence term $t - r_1$.

The distinction between short term and long term absences also creates a challenge. Since the economic motivations behind these two different absences are disparate, one should model them separately. However, the cumulative absence time is the summation of these two. Thus these two models depend on each other, creating a quasi-simultaneous equation system.

At this moment, it is evident that conventional micro-econometric tools such as count data regression and duration analysis offer no satisfying solution to this problem. The nature of this dynamic behavior is the state dependence. In the next section, we are going to introduce the self-exciting process, that by definition is state dependent and can model the dynamic behavior among work absences.

3 Econometric Models for Work Absenteeism

In this section, we use self-exciting processes to construct our work absenteeism models. We first present some introductions to the self-exciting process, followed by a detailed discussion about the models. We will end this section by illustrating how to estimate the models.

3.1 Introducing the Self-Exciting Process

As mentioned before, a self-exciting process is a particular counting process. A counting process (non-tied) can be regarded as a step function:

$$N(t) = \sum_{i=1}^{\infty} \mathbb{I}\{t_i \leq t\} \tag{1}$$

Its value adds one at time t if and only if an event occurs at this time. The counting process not only tells how many events have occurred before time t but also indicates exact occurrence times for each event.

The Doob-Meyer decomposition theorem states that any counting process can be decomposed into a predictable cumulative intensity part and a martingale part, i.e.

$$\begin{aligned} N(t) &= \Lambda(t) + M(t) \\ N(dt) &= \lambda(t)dt + M(dt) \end{aligned} \tag{2}$$

$\Lambda(t)$ is the predictable cumulative intensity, also known as the compensator. In brief, predictability means conditional on the information just before present (say $t-$), we should know the value of $\Lambda(t)$. Rigorously, $\Lambda(t)$ is predictable w.r.t a filtration $\mathcal{F}_{t-}$ if and only if $\Lambda(t)$ is $\mathcal{F}_{t-}$-measurable.$\lambda(t) = \Lambda(dt)/dt$ is the associated intensity and $M(t)$ is the martingale satisfying $M(0) = 0$.

The decomposition is unique: if there exists another decomposition:

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

we then have $\Lambda(t) = \tilde{\Lambda}(t)$ and $M(t) = \tilde{M}(t)$. In addition, we have

$$\mathbb{E}N(t) = \mathbb{E}\Lambda(t)$$

The proofs of the uniqueness and this equation can be found in Appendix B.

The (cumulative) intensity is usually conditioned on a filtration $\mathcal{F}_{t-}$:

$$\lambda(t|\mathcal{F}_{t-}) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N([t, t+\Delta t])|\mathcal{F}_{t-})}{\Delta t} \tag{3}$$

If the filtration is generated by the underlying counting process itself:

$$\mathcal{F}(t-) = \sigma(N(s) : s < t) \tag{4}$$

we call the counting process $N(t)$ a self-exciting process. One may study the process $N(t)$ through modeling and estimating $\lambda(t|\mathcal{F}_{t-})$ $(\Lambda(t|\mathcal{F}_{t-}))$. We postpone the estimation method in later subsection.

We may generalize the filtration by including other relevant information. Let $\mathcal{H}_{t-} = \mathcal{H}_0 \vee \mathcal{F}_{t-}$ be the conditional filtration, where $\mathcal{H}_0$ is the $\sigma - algebra$ generated by some external covariates such as age, sex, etc. We interpret this filtration as the 'whole history'. Notice that $\mathcal{H}_0$ can also be time-dependent, i.e., $\mathcal{H}_{t-} = \mathcal{H}_0(t-)$ and contains external measurements, shocks or impulses. Thus $\mathcal{H}_{t-}$ strictly includes $\sigma(N(s) : s < t)$ in most cases.

3.2 Modeling the State-Dependent Effect

Recall that by the eligible condition for the SSP, we category any absences that is less or equal to 3 days as short-term and other absences as long term. Define three alternative states, $k = 1, 2, 3$, that an individual can occupy in our model: attendance ($k = 1$), short-term absence ($k = 2$) and long-term absence ($k = 3$). Thus $\lambda_{12}(t)$ is the short-term incidence intensity function (from attendance state to short-term absence), and $\lambda_{21}(t)$ is the short-term recovery intensity function, other two long-term intensity functions follow the same index rule.

Furthermore, we define attendance periods as any time intervals between the last recovery dates and the next starting dates of absences. Define absence periods as any time intervals between the starting dates and the recovery dates of absences. Figure **??** in section 2 describes the situation. We assume that a new absence cannot occur without the end of current absence. That is the incidence intensity is zero in the absence period. Similarly, recovery events cannot occur before any absences ever started: the recovery intensity is zero in attendance period.

Recall that the cumulative intensity can be expressed as the summation of cumulative hazard rate:

$$\Lambda(t) = \Lambda(T_k) + \int_0^{t-T_{k+1}} \frac{F_k(dx)}{1 - F_k(x)}, t \in (T_k, T_{k+1}]$$

where $F_k(x)$ is the distribution of k^{th} duration. Thus, we are going to study individual's work absence via different duration and build our econometric through cumulative hazard.

There are two reasons behind this decision. First, duration and their corresponding cumulative hazard rates are well studied and documented in the literature, and second, more importantly, we are going to convert the cumulative hazard rate into an accelerated failure time structure to swipe out the unobserved heterogeneity.

Our general settings (that is, incidence and recovery processes as well as the distinguishing of short and long term absences) creates 4 types of duration: short (long) term incidence duration and short (long) term recovery duration. Before going to the details of our model, it is worthwhile to give these duration a clear definition and the following figure would help to achieve it.

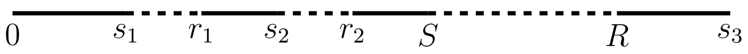


Figure 1: A possible realization of absences

We define the short-term incidence duration as the duration from previous short (long)-term recovery (and the starting point) to the next short (long)-term absence. In the figure above, the short-term incidence duration are $(0, s_1), (r_1, s_2), (r_2, s_3)$, the long-term incidence duration is $(0, S)$; The short (long)-term recovery duration are the ones from the starting dates of current short (long)-term absences to the recovery dates. The short-term recovery duration are $(s_1, r_1), (s_2, r_2)$, the long-term recovery duration is (S, R). Notice here we allow both short-term and long-term incidence intensities to be positive during the same attendance period. Furthermore, we assume that conditional on current common filtration, the occurrences of short-term and long-term incidences are independent. That is when studying the short-term (long-term) incidence process, we consider absences that are due to short-term (long-term) causes and long-term (short-term) causes are independent censoring. This assumption resembles the cause-specific hazard model rather than the competing risk model in the duration analysis.

To begin with, we first present our hazard model for the short-term incidence duration, other types of duration share a similar structure. For individual i and the j^{th} short-term incidence duration, the hazard rate is specified as

$$h^i_{12}(t_j) = t_j^{\alpha} \exp(-\beta H_{i,j} + \gamma' X_i + \nu_i), t_j \in \text{ attendance period} \tag{5}$$

where $H_{i,j}$ is the absence score for this individual before j^{th} short-term absence, X_i is a vector of individual time-invariant covariates, and ν_i is the unobserved heterogeneity for individual i. We do not impose any restriction on ν_i except it is time persistent, this means the unobserved heterogeneity might be correlated with other explanatory variables, in fact, ν_i is correlated with $H_{i,j}, \forall j$. t_j^{α} here measures duration dependence effect, we restrict $\alpha > -1$ to make the model meaningful [2].

If t_j is in absence period, then $\lambda^i_{12}(t_j) = 0$. For example, in figure 1, for short-term incidence duration (r_2, s_3), if t_j is within periods (r_2, S) or (R, s_3), $\lambda^i_{12}(t_j) > 0$, but during (S, R), the hazard rate is zero. Incidence duration like (r_2, s_3) are quite special, because there are external shocks within the period: the hazard rate is zero for part of the duration and the absence score is different for t_j in (r_2, S) and in (R, s_3).

[2] The cumulative hazard rate should be non-decreasing, if $\alpha < -1$, the cumulative hazard is a decreasing function.

For any normal short-term incidence duration (the ones without external shocks), we can convert the cumulative hazard rate into an accelerated failure time (AFT) structure:

$$t_j^{\alpha+1} = (\alpha+1)\exp(\beta H_{i,j} - \gamma^{'} X_i - \nu_i) u_{i,j} \tag{6}$$

where $u_{i,j} \overset{i.i.d}{\to} \exp(1)$

Similar to the first-difference GMM estimator in the dynamic panel models, we perform a first-ratio transformation to swipe out the unobserved heterogeneity term ν_i.

$$\tilde{t}_j^{\alpha+1} = (\frac{t_j}{t_{j-1}})^{\alpha+1} = \exp(\beta\Delta_{i,j})\tilde{u}_{i,j} \tag{7}$$

where $\Delta_{i,j} = H_{i,j} - H_{i,j-1}$ and $\tilde{u}_{i,j} = u_{i,j}/u_{i,j-1}$.

It is easy to show that the distribution of $\tilde{u}$ is:

$$F_{\tilde{u}}(x) = \frac{x}{1+x}$$

hence, the distribution of our new first-ratio duration $\tilde{t}_j$ is

$$F_{\tilde{t}_j}(x) = \frac{\exp(-\beta\Delta_{i,j})x^{\alpha+1}}{1+\exp(-\beta\Delta_{i,j})x^{\alpha+1}} \tag{8}$$

Notice that it is impossible to convert the hazard rate of a special short-term incidence duration (the ones suffer from external shocks) into an AFT structure and perform the followed first ratio transformation, thus we have to drop those duration. In the example of Figure 1, the workable short-term incidence duration are $(0, s_1)$ and (r_1, s_2).

After the first-ratio transformation, we end up with a series of transformed duration $\{\tilde{t}_j\}_j$, which can be used to construct a new counting process $\tilde{N}(\tau)$. The cumulative intensity of this counting process is

$$\Lambda_{12}(\tau) = \Lambda_{12}(\tilde{T}_{k-1}) + \int_0^{\tau-\tilde{T}_k} \frac{F_{\tilde{t}_k}(dx)}{1-F_{\tilde{t}_k}(x)}, t \in (\tilde{T}_{k-1}, \tilde{T}_{k+1}] \tag{9}$$

where $\tilde{T}_k$ is the new 'calender time':

$$\tilde{T}_k = \sum_{j=1}^{k} \tilde{t}_j$$

The new cumulative hazard rate $H(t_k) = \int_0^{\tau-\tilde{T}_k} \frac{F_{\tilde{t}_k}(dx)}{1-F_{\tilde{t}_k}(x)}$ for k^{th} first-ratio duration is

$$H(t_k) = \log(1+\exp(-\beta\Delta_{i,j})t_k^{\alpha+1}) \tag{10}$$

We can verify that the above stated cumulative intensity is indeed the cumulative intensity for the counting process $\tilde{N}(\tau)$ by using the random time change theorem.

The random time change theorem states that for any counting process whose cumulative intensity $\Lambda(t)$ is almost surely continuous, the random variable $\Lambda(T_j) - \Lambda(T_{j-1}) \overset{i.i.d}{\to} \exp(1)$, where T_j are random variables for occurrence times of events. In our application, it is trivial to demostrate that

$$P(H(t_k) \leq x) = 1 - \exp(-x)$$

Using the same model, we can build first-ratio long-term incidence duration and first-ratio short (long)-term recovery duration. Notice that for long-term incidence duration, almost all of them are suffer from external shocks and are not able to do the first-ratio transformation. However, as shown in the data section, long-term absences are rare and we are more interested in the long-term recovery duration rather than their incidence duration. To this end, we are not going to model the long-term incidence duration. On the other hand, the recovery duration, both short and long term, are free from external shocks. We are able to perform the first-ratio transformation without any loss of information.

3.3 Modelling the Time-Invariant Effect

The first-ratio transformation not only swipe out the unobserved heterogeneity, but also cancels the time-invariant term. To estimate the individual time-invariant effect for short-term incidence duration, we take advantage that the initial duration is not under the influence of absence score effect, since $H_{i,1} = 0 \forall i$. We may then use the standard Heckman and Singer's NPMLE to analysis these effects.

We use a subset of our data that only consists of newly-hired workers, the hazard rate for initial short-term incidence duration is

$$h(\mathbf{X}_i, \nu_i) = \exp(\nu_i) exp(\gamma^{'} \mathbf{X}_i) \tag{11}$$

To perform the NPMLE, we need to assume that the random variable ν_i is independent with other observed covariates $\mathbf{X}_i$.

For short-term recovery duration, we are more interested in investigating the state-dependent effect than the time-invariant effect, since the recovery duration are quite short, hence we have no intention to build such model. The situation is more complicated for the long-term recovery duration, it is worth to study the time-invariant effect, however almost all long-term duration begin with a non-zero absence score, which is correlated with the unobserved heterogeneity, violating the independent assumption in NPMLE. However, if our state-dependent model could provide evidence that long-term recovery duration do not have state-dependent effect, then it is straightforward to pool all the long-term duration and perform the NPMLE. In fact, most sick-leave literature indeed assume that recovery duration are free from state-dependent effect and are i.i.d (Markussen et al., 2011).

3.4 How to Estimate the Models

Recall the Doob-Meyer decomposition, we have $\mathbb{E}N(t) = \mathbb{E}\Lambda(t)$. One may obtain the estimator by minimising the distance between the counting process and its cumulative intensity. Inspired by this idea, Kopperschmidt and Stute (2013) developed a minimum distance estimation method. This method only requires the observations (individuals) to be i.i.d. It does not assume the differentiability of the cumulative intensity and allow unexpected jumps in the intensity function. Here, we provide a summary. Technical details can be found in their paper.

Formally, let $N_1, ..., N_n$ be i.i.d copies of n observed counting process that are conditional on the increasing filtrations $\mathcal{H}_i(t), 1 \leq i \leq n$, which are comprised by the counting process N_i as well as some other external information. Let $\Lambda_{v,i}(t|\mathcal{H}_i(t-))$ with $v \in \Theta \subset \mathbb{R}^d$ be a given class of parametric cumulative intensities.

We set,

$$< f, g >_\mu = \int_0^T fg d\mu \tag{12}$$

where T is the terminating time. If f and g are square integrable functions w.r.t. μ. The corresponding semi-norm is,

$$||f||_\mu = [< f, f >_\mu]^{1/2} \tag{13}$$

Let,

$$\bar{N}_n = \frac{1}{n}\sum_{i=1}^{n} N_i; \bar{\Lambda}_{v,n} = \frac{1}{n}\sum_{i=1}^{n} \Lambda_{v,i} \tag{14}$$

We call the former the averaged counting process and the later the averaged cumulative intensity. Naturally the associated averaged innovation martingale is,

$$d\bar{M}_n = d\bar{N}_n - d\bar{\Lambda}_{v_0,n} \tag{15}$$

If, for μ, we take $\mu = \bar{N}_n$, the quantity $||\bar{N}_n - \bar{\Lambda}_{v,n}||_{\bar{N}_n}$ is then an overall measurement of fitness of $\bar{\Lambda}_{v,n}$ to $\bar{N}_n$. The estimator v_n is computed as,

$$v_n = arg \inf_{v \in \Theta} ||\bar{N}_n - \bar{\Lambda}_{v,n}||_{\bar{N}_n} \tag{16}$$

Kopperschmidt and Stute (2013) have shown the consistency and asymptotic normality of this estimator. For the statement of the asymptotic theories, we refer readers to the Appendix B.

Thus for incidence processes, we collect starting dates of absences that have previous absent records to construct individual counting processes $N_{1k}(t), k \in \{2, 3\}$. The distance function is then:

$$||\bar{N}_{1k,n} - \bar{\Lambda}_{1k,n}||_{\bar{N}_{1k,n}}$$

where

$$\bar{\Lambda}_{1k,n}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \lambda_{i,1k}(s)ds$$

Similarly, the recovery intensities have the following distance function:

$$||\bar{N}_{k1,n} - \bar{\Lambda}_{k1,n}||_{\bar{N}_{k1,n}}$$

where $N_{k1}(t), k \in \{2,3\}$ are made of ending dates of absences, and

$$\bar{\Lambda}_{k1,n}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \lambda_{i,k1}(s)ds$$

Conventionally, a likelihood-based method is used to obtain estimates. One may do so by exploiting the fact that $f(t) = \lambda(t)\exp(-\Lambda(t))$. For example, the Hawkes process, a special case of the self-exciting process that is widely used in the high-frequency financial analysis, uses MLE to obtain the consistent estimators (e.g., Aït-Sahalia et al. (2015),Bacry and Muzy (2014) and **?**). However, it requires the data to be stationary, in complicated situations, we have little intention to maintain this assumption. In our application, the absence score is calculated as the summation of all the past absence duration for a defined duration and is mostly likely not stationary.

4 Main Results

Before going to the main results, some explanation of data is in order. We only use absence records that have previous records, that is, $\{j : H_{i,j} > 0, \forall i\}$ when estimating the state-dependent effects. This is because the initial incidence duration only occurs to newly hired workers, and it is reasonable to believe that the duration dependence of an initial absence is different from the others. For example, we would expect that a worker is more likely to ask another absence if the attendance time between these two absences is short (or an absence is more likely to trigger another one if the these two absences are close to each other). But before the initial absence, there is no previous absence record, hence there is no source of triggering. Compared to a usual incidence duration, a short attendance duration in an initial incidence duration does not have the same implication.

On the other hand, to estimate the individual time-invariant effect for short-term incidences, we only use the initial short-term incidence duration of newly hired workers (we have in total 223 new workers in the data) to avoid separating the unobserved heterogeneity from the state dependence.

4.1 State-Dependent Effect Results

Table 3 presents the main results. Column (1) reports the state dependent effects for short-term incidence duration. Both duration dependent and absence score effects are significant. We have a negative duration dependence, meaning previous absence is more likely to trigger the next one if they are close enough, the link effect between two consecutive absences is decreasing as the attendance duration increases. For firm managers, this means a absence regulation that takes attendance duration into consideration is key to avoid frequent short-term absences.

[Insert Table 3 Here]

The absence score effect is positive, a higher absence score leads to a longer attendance duration (or a lower intensity for asking a short-term leave). This result is hardly a surprise given the firm has established the experience-rated absence regulation.

As mentioned before, most of the long-term incidence duration suffer from the external shocks and are unable to perform the first-ratio transformation to swipe out the unobserved heterogeneity term. To investigate the state dependent effects of long-term absences, we pooled all absence together. The goal is to investigate whether the state dependent coefficients are still significant when we do not distinguish short and long-term incidence duration. Column (2) displays the results.

The duration dependent effect is still negative and significant, but the absence score coefficient is no longer significant. These results imply that workers do not respond to the absence score when she/he asks for a long-term leave. Possible explanation is that unlike the short-term leaves, which are more likely to be 'voluntary', the long-term absences are mostly sick-leaves, and when dealing with heath-related issues, future financial benefit is less important. Previous long-term absence has stronger effect on the next long-term absence when they are close to each other, as illness episodes are prone to occur in a cluster way.

To sum up, for the decisions to ask for the leaves, individuals will respond to the absence score for short-term absences, but this is not the case for the long-term ones. Duration dependence exist in both short and long-term incidences, both types of incidences tend to have a cluster structure.

Column (3) and (4) report the results of short and long-term recovery duration. Both of them have negative duration dependence: the longer one stays in absence, the less likely she/he is going to return to work. However, individuals only respond to the absence score when they make short-term recovery decisions. Notice this means it is safe to assume that long-term recovery duration are i.i.d and we can employ the standard duration model. We postpone the description and the estimating results for such model to the next subsection, where individual time-invariant effects are discussed.

The results in Table 3 are based on the setting that the criteria of a short and a long-term absence is three-days absence duration. Although this criteria comes from a solid reasoning, one may still want to verify the validity of this setting. To this end, we re-define a short-term absence as the ones individual will respond to the absence score when making incidence and recovery decisions, other absences fail to satisfy these conditions are long-term ones. We increase the cut-off duration (c) from three to four and Table 4 displays the results.

[Insert Table 4 Here]

As we increase the cut-off duration, β is still significant in short-term incidence but fail to be significant in short-term recovery, further confirming that $c = 3$ is the proper cut-off criteria.

4.2 Individual Time-Invariant Effect Results

As mentioned before, we are focusing on modelling and estimating the individual time-invariant effect for short-term incidence duration. In addition, since individuals do not respond to the absence score in the long-term recovery duration, we could also treat these duration as i.i.d and employ the conventional duration model.

As mentioned before, the duration dependence for the initial short-term incidence duration does not have the same interpretation as the usual ones, hence we assume a constant hazard rate described in previous section. The individual likelihood contribution is:

$$L_i(\nu_i) = exp(-H_i(t))h_i^{y_i}$$

where

$$h_i = h_i(X_i, \nu_i) = \exp(\beta' X_i + \nu_i)$$

$$H_i(t) = th_i$$

and y_i is a censoring indicator, which equals to zero if censored. Heckman & Singer's NPMLE likelihood is

$$L = \prod_{i=1}^{N} \mathbb{E}[L_i(\nu_i)] = \prod_{i=1}^{N}\sum_{l=1}^{Q} p_l L_i(\nu_l), \text{ WITH } \sum_{l=1}^{Q} p_l = 1$$

where Q is the number of mass points used to approximate the true distribution of ν_i.

For long-term recovery duration, we specify the following hazard rate,

$$h_i(t_j) = t_j^{\alpha} \exp(\beta' X_i + \nu_i)$$

$$H_i(t_j) = \frac{1}{1+\alpha} t_j^{1+\alpha} \exp(\beta' X_i + \nu_i)$$

here the subscript i denotes individual i and j stands for the j^{th} long-term recovery duration of this individual.

The individual likelihood contribution is:

$$L_i(\nu_i) = \prod_{j \in S_i} exp(-H_i(t_j))h_i(t_j)$$

where S_i the set of observed long term duration for individual i. The fact that an absence has already occurred implies that we do not have the censoring problem here. The overall NPMLE likelihood function can be then easily written.

Table 5 reports the results. In this application, we use the 'working age', which is the actual age net the legal working age (16). Column (1) reports the individual time-invariant effect for short-term incidence duration. Young and most senior workers are more likely to ask for short-term leave, mid-age workers have the lowest hazard rate to be absent. Other covariates seem play insignificant roles.

Column (2) displays the results for long-term recovery duration. Again, we have observed the negative duration dependence, mid-age workers have the highest returning intensity, while in comparison, young and most senior workers tend to stay in the absence longer. Male workers are more likely to return to work than their female counterparts.

[Insert Table 5 Here]

5 An Economic Model for Work Absenteeism

In this section, inspired from the empirical results, we present a simple economic model. We first provide a narrative approach to describe the incentives to the strategic behavior in work absences. Next, we modify a standard labour-leisure model to characterise the decision-making process of asking for leave and returning to work. We also construct a structural model to describe how individuals optimise the long-term absence durations.

5.1 The Incentive to the Strategic Absence behavior

It is known that a work search is costly. A worker may accept a job offer even though the contracted wage is not equal to the marginal rate of substitution between leisure and income. If a worker accepts such a job offer, she remains an incentive to consume more leisure, one common way to do so is, of course, to be absent from work.

Even if the marginal rate of substitution between income and leisure is equal to the contracted wage, a worker may occasionally prefer to be absent due to external accidents. A worker will choose to be absent when the (expected) size of a shock is large, and the alternative activities are more attractive.

The last element is the worker's personal absence history. Working discipline regulations in most firms specified particular reward/punishment schedules for work absences. These rules usually reward 'good reputation' workers (those who have less cumulative absence time) and punish 'bad reputation' workers (those who have more cumulative absence time). The shadow costs for workers in different positions of the cumulative absence time spectral are different. This creates the incentive to consume more (or less) absences depending on one's absence score.

Here we provide empirical evidence that support our labor-leisure hypothesis. We are going to investigate, conditional on individual's decision to take a short-term absence and other covariates, including absence score, whether individuals are more likely to ask leaves on Friday or Monday. Monday/Friday leaves could maximize an individual's leisure time while keeping the cost at minimum. If an individual's short-term absence is made not based on leisure, then the conditional probability of taking Monday/Friday leaves should be 40%.

We use the standard logistic regression to calculate the predicted probability of Monday/Friday leaves. We find that the mean conditional predicted probability of Monday/Friday leaves is 46.3% with standard deviation of 0.029.

5.2 Decision to Ask for Absence

Suppose the worker's utility is a linear function of $\omega, C, R(A)$ and some other unobserved factors. ω is the general well-being, and C is consumption. $R(A)$ is the reputation. It is a function of cumulative absence time A with $R^{'}(\cdot) < 0$ and $R^{''}(\cdot) < 0$.

Note in the incidence intensity model, we express the reputation as $\exp(-\beta A)$, whose first and second derivative are $-\beta \exp(-\beta A) < 0$ and $\beta^2 \exp(-\beta A) > 0$ respectively. We do not think this setting contradicts our economic assumptions on the reputation function. Since individual may not necessarily map utility to absence actions linearly. If current utility is quite low, one more absence may make little difference to individuals.

At the first stage, workers accept the job offers and have the same reputation. Random shocks $e \in [0, \infty)$ hit all individuals. Notice that $e = 0$ means no accident shock, and a higher value of e indicates a more severe accident. Workers can observe the existence of the shocks but cannot observe the sizes of them without further information. We assume at this stage, after observing the shocks, workers will always ask for absence. After that, further information is given, the size of the shock is known, and workers choose the duration of the absence (The decision process for how to choose the length of an absence spell will be described later). Cumulative absence time is updated from $0 \to A$ (different values of A for different workers). The well-being $\omega^{'}$ evolves as follow:

$$\omega^{'} = \omega - e + g(A)$$

where $g(\cdot)$ is the well-being generating function with $g(0) = 0, g^{'}(\cdot) > 0$ and $g^{''}(\cdot) < 0$.

In the second stage, individuals again observe the existence of shocks. But in this stage, a worker has to decide whether to ask for the absence ($D = 1$ for absence, $D = 0$ otherwise) based on her history and the expectation on the size of the accident by:

$$\begin{aligned} D &= \mathbb{I}\{U(R(A + a(\mathbb{E}(e))), \omega^{'} - \mathbb{E}(e) + g(a(\mathbb{E}(e))), C_1) + \epsilon_1 > U(R(A), \omega^{'} - \mathbb{E}(e), C_2) + \epsilon_2\} \\ &= \mathbb{I}\{U^1 + \epsilon_1 > U^0 + \epsilon_2\} \end{aligned}$$

where $U(\cdot)$ is the utility function. Let $U_R, U_\omega, U_C > 0$ be the partial derivatives of the utility function. $a(\cdot)$ is the duration of the absence and is determined by the size of an accident with $a^{'}(\cdot) > 0$, ϵ_1, ϵ_2 represent unobserved factors that might effect the utility function.

In the case of long-term absences, individuals do not respond to the reputation, the absence decision is then governed by

$$D = \mathbb{I}\{U(\omega^{'} - \mathbb{E}(e) + g(a(\mathbb{E}(e))), C_1) + \epsilon_1 > U(\omega^{'} - \mathbb{E}(e), C_2) + \epsilon_2\}$$

This decision rule specifies that an individual will ask for a leave if and only if the expected utility for being absent is higher than the utility of attendance. Individuals' absence decisions are then depended on (a) their cumulative absence time and (b) their beliefs about the size of the accidents.

Taking the expectation, we have

$$\begin{aligned} Pr(D = 1) &= Pr(\epsilon_2 - \epsilon_1 < U^1 - U^0) \\ &= F_\epsilon(U^1 - U^0) \end{aligned}$$

where $\epsilon = \epsilon_2 - \epsilon_1$

Since $F^{'}(\cdot) > 0, R^{'}(\cdot) < 0$ and $R^{''}(\cdot) < 0$. We have:

$$\frac{\partial Pr(D = 1)}{\partial A} < 0$$

5.3 Decision to Recovery

Conditional on the fact that individuals have decided to take absences, they will receive information about the size of shocks. This further information is given by, for example, doctors if workers went to hospitals. The workers then have to decide the duration of their absences. Since the empirical results suggest that only in the short-term recovery processes, workers tend to have strategic behavior, we assume that reputations will only be a part of the equation if the size of an accident is within some level. That is if $e \leq e^*$, $a(e, R) \in [0, a(e^*)]$. If $e > e^*$, $a(e)$ is then a deterministic function of accident e that can not be altered by the reputation R.

Within such size range, a worker's problem is:

$$\max_a U(R(A+a), \omega - e + g(a), C)$$
$$s.t$$
$$I + w(t^c - a) + R(A+a) - C = 0 \tag{17}$$

where I is non-labour income, w is wage, t^c is the contracted working time.

First order condition with respect to a yields:

$$\begin{aligned} &U_R R' + U_\omega g' - U_C(w - R') = 0 \\ &U_C(w - R') = U_R R' + U_\omega g' > 0 \end{aligned} \tag{18}$$

By differentiating the first order condition (17) through (18), one can show that

$$\frac{\partial a}{\partial A} < 0$$

That is, as long as the accident is small ($e < e^*$), the shorter the cumulative absence time, the longer absence duration one may choose.

Notice that in the case of scheduled absence, there is no stochastic in a 'shock'. The size of this 'shock' is observed all the time. And the decisions to ask for leave and to return to work should be made simultaneously: workers do not need the decision process for asking for absence, she only need to decide the duration of such absence (0,1,2 or 3 days,0 days absence means no absence).

5.4 A Structural Model for Long-Term Recovery

If the sizes of accidents are greater than the threshold e^*, a worker may recognize this event as a 'major' and will leave the reputation out of the equation. Statistically, this means that the duration of a long-term absence is memoryless, and there is no harm to treat each of them as independent and identical distributed.

The task of a worker under this circumstance consists of choosing an optimal duration to maximize her utility without the consideration of reputation. This task is mostly a discrete choice problem under continuous time. And the independence assumption inspires us to build a simple structural model for the long-term absence duration decision making process. This structural model is a simplified version of Honor and De Paula (2010) and de Paula and Honore (2017), in which the authors study the couple's interdependent retirement duration.

For individual i who is now in j^{th} long-term recovery period, she has a positive utility flow $K_{ij}Z_1(t)\phi_1(X_i)$, where K_{ij} is a positive random variable that could represent initial health. At any point, she may choose to 'switch' to the alternative state: returning to work, with a utility flow $Z_2(t)\phi_2(X_i)$. Assuming individuals are myopic and an exponential discount rate ρ, individual i's utility for taking part in the j^{th} long-term recovery period until time t_{ij} is:

$$\int_0^{t_{ij}} K_{ij} Z_1(s)\phi_1(X_i)e^{-\rho s}ds + \int_{t_{ij}}^{\mathbb{E}(T)} Z_2(s)\phi_2(X_i)e^{-\rho s}ds \tag{19}$$

where $\mathbb{E}(T)$ is the expecting beginning time of a next long-term absence.

The first order condition for maximizing this with respect to t_{ij} is:

$$\Big[K_{ij}Z_1(t_{ij})\phi_1(X_i) - Z_2(t_{ij})\phi_2(X_i)\Big]e^{-\rho t_{ij}}$$

Thus the optimal T_{ij} is given by:

$$\begin{aligned} T_{ij} &= \inf\{t_{ij} : [K_{ij}Z_1(t_{ij})\phi_1(X_i) - Z_2(t_{ij})\phi_2(X_i)]e^{-\rho t_{ij}} < 0\} \\ &= \inf\{t_{ij} : K_{ij} - Z(t_{ij})\phi(X_i) < 0\} \end{aligned} \tag{20}$$

where $Z(\cdot)\phi(X_i) = Z_2(\cdot)\phi_2(X_i)/\big(Z_1(\cdot)\phi_1(X_i)\big)$.

Notice the above equation is in the spirit of discrete choice structure model under a latent variable framework in the sense that individual compares the instant utility between two states: $\nu^\star = K_{ij} - Z(t_{ij})\phi(X_i)$. If $\nu^\star \leq 0$, individuals will return to work, $\nu^\star > 0$ otherwise. The multiplicative structure of $Z(t)$ and $\phi(X_i)$ is explicitly designed to have the accelerated failure time model as a special case. There is no difficulty in estimation to lose this structure. To sum up, the individual will switch at

$$T_{ij} = Z^{-1}(K_{ij}/\phi(X_i)) \tag{21}$$

Notice that in this structure model, the source of randomness is K_{ij}. We can re-write equation 21 as the following:

$$\ln Z(T_{ij}) = -\ln\phi(X_i) + \epsilon \tag{22}$$

where $\epsilon = \ln K_{ij}$. Equation 22 is a typical accelerated failure time (AFT) model. Assume $Z(t) = t$, $\phi(X_i) = e^{-X_i^T\beta}$ and $K_{ij} \sim \exp(1)$, we may end up with the exponential AFT model. The cumulative distribution function of T_{ij} is given by

$$\begin{aligned} F_{T_{ij}}(t) &= Pr[K_{ij}e^{X_i^T\beta} \leq t] \\ &= Pr[K_{ij} \leq te^{-X_i^T\beta}] \\ &= 1 - \exp(-t\exp(-X_i^T\beta)) \end{aligned} \tag{23}$$

The corresponding hazard rate is

$$\begin{aligned} h_{T_{ij}}(t) &= \frac{f_{T_{ij}}(t)}{1 - F_{T_{ij}}(t)} \\ &= \exp(-X_i^T\beta) \end{aligned} \tag{24}$$

These assumptions are mainly made to compare with our reduced form model from the previous section. Note that the signs of coefficients are opposite. Intuitively, a higher hazard leads to a shorter duration.

Table 6 presents the estimates for this exponential AFT model. Not surprisingly, the results are consistent with the reduced form hazard model.

[Insert Table 6 Here]

6 Discussion: Self-Exciting Process as a Complementary Tool to Conventional Methods

In this section, we compare the differences between the self-exciting process and two widely used conventional econometric tools in microdata analysis: count data regression and duration analysis. We argue that many major issues in these two conventional tools can be easily overcome by using a self-exciting process. We also highlight the fact that despite the numerous advantages of using self-exciting process, it can not replace conventional methods completely. Researchers should adopt proper econometric tools to their specific needs.

6.1 Compare to the Count Data Regression

Many count data display over-dispersion property: the variance of data exceeds the mean of data. One source of such over-dispersion is excess zeros: the dataset may have more zero observations than is consistent with the basic Poisson model.

Unlike the count data regression, where the discrete counts y is treated as a random variable, in a self-exciting process, the outcome is a time depended counting process $N(t)$. The additional time dimension enables us to generate excess zeros. The intuition of our argument is quite simple: if the terminated time is small (relative to the intensity), we can easily generate a high proportion of zeros. More precisely, we treat the zero event as an end-of-study censoring problem: events will happen in the future, but they are censored due to an end of the study. Also in the generalised count data models (e.g. Zero inflation and Hurdle), zeros and non-zeros (positives) are assumed to come from two different data generating processes (DGPs). Whereas in the self-exciting process, zeros and positives are generated from the same stochastic process.

We use two DGPs to illustrate our argument. The first is the standard Poisson process and the second is a self-exciting process. The Poisson process serves as our baseline model (same as the Poisson regression in count data). Simulations will show that although by setting a small time interval, we can generate a high proportion of zeros, the Poisson process is still equidispersion. The self-exciting process, on the other hand, can mimic the over-dispersion property of data through excess zeros.

Poisson DGP The intensity for a (homogeneous) Poisson process is a constant $\lambda = \mu$. We set $\mu = 5.5$, and let the time interval to be $[0, T^*] = [0, 0.2]$. We run 100 trials of simulation. The simulation procedure is detailed in Appendix. For each Poisson process, we record its corresponding counts: $Y_i = N_i(T^*), i = 1, 2, \cdots, 100$. The following histogram displays our simulation results.

[Insert Figure 5 Here]

Of all 100 runs, we are able to generate 33 zeros. However, the data is still equidispersion: its sample mean is $\bar{Y} = 1.05$ and sample variance is $\hat{V}(Y) = 1.067$.

Self-Exciting DGP The DGP for the self-exciting process we picked is the ETAS (epidemic-type aftershock sequence) model, it is first introduced by Ogata and Katsura (1988) and ever since it has been widely studied in seismology (e.g. Zhuang et al. (2002)). It characterises the earthquakes occurrence times and magnitudes and belongs to a marked Hawkes process family.

The intensity of a ETAS model, for its simplest form, could be:

$$\lambda(t|\mathcal{F}_{t-}) = \mu + \sum_{i:t_i<t} e^{\alpha x_i} \left(1 + \frac{t - t_i}{c}\right)^{-p} \tag{25}$$

where x_i is the magnitude of an earthquake occurring at time t_i, and the mark density for simplicity is assumed to be independent and follow a exponential distribution.

$$f(x|t, \mathcal{F}_{t-}) = \delta e^{-\delta x}$$

Notice that without the exciting part, the intensity degenerates to a standard homogeneous Poisson intensity. We set the parameters as $\mu = 0.01$, $\alpha = 1.98$, $c = 0.018$, $p = 0.94$ and $\delta = log(10)$. The time interval is $[0, T^*] = [0, 100]$.

The simulation method we used is called the *thinning method*, introduced by Ogata (1981),Lewis and Shedler (1979). Briefly, this method first calculates an upper bound for the intensity function in a small time interval, simulating a value for the time to the next possible event using this upper bound, and then calculating the intensity at this simulated point. However, these 'events' are known to be simulated too frequently (Lewis and Shedler, 1979). To fix this problem, the method will compare the ratio of the calculated rate with the upper bound to a uniform random number to randomly determine whether the simulated time is treated as an event or not (i.e. thinning). A full description of the algorithm is detailed in Appendix C.

Like before, we run 100 simulations and record their corresponding counts at the terminal time T^*. With these parameters, we can generate 44 zero observations out of 100. The largest count is at 92. We plot its histogram as below.

[Insert Figure 6 Here]

The self-exciting data exhibits the over-dispersion property: $\bar{Y} = 3.27$, $\hat{V}(Y) = 108.5425$.

6.2 Compare to the Duration Analysis

Unlike the counting process where the interested subject is the time stamps of events (by modeling the intensity function), in duration analysis, the subject under investigation is the duration of a default state (by modeling the hazard rate). The intensity function and hazard rate are, in some sense, quite similar but conceptually different.

Consider a self-exciting process, let τ be the time of the last event before time t and $\mathcal{F}$ be the filtration. Denote the conditional distribution of the time of the next event as:

$$G(t|\mathcal{F}(\tau)) = Pr(T \geq t|\mathcal{F}(\tau))$$

and $g(t|\mathcal{F}(\tau))$ as the corresponding conditional density function. Then from the definition of intensity (equation (3)),

$$\lambda(t|\mathcal{F}(\tau)) = \frac{g(t|\mathcal{F}(\tau))}{1 - G(t|\mathcal{F}(\tau))} \tag{26}$$

Now, consider a system begins in time 0 and fails at some random time $T > 0$. The hazard rate (or hazard function) $h(t)$ is defined as:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{Pr\{T \in (t, t + \Delta t)\}}{Pr\{T > t\}\Delta t} \\ &= \frac{f_T(t)}{1 - F_T(t)} \end{aligned} \tag{27}$$

Where t here is the duration of a state. The hazard rate tells us the conditional probability of the system failing in the interval $(t, t + \Delta t]$ conditioned on the system being in working at time t.

Despite the similarity between (26) and (27), the intensity and the hazard rate are conceptually different. Intensity deals with reoccurring arrivals with a focus on the timing per se, while the hazard rate deals with the duration or the length of only one spell. Most duration analysis can only study the recurrent events with the i.i.d of events assumption holds. It is difficult to employ this method when recurrent events are state dependent. A self-exciting process, on the other hand, is free from these problems since the state dependence is included in the filtration.

Another feature of the self-exciting process is its ability to generate quite different individual behaviors even without the unobserved heterogeneity. The behaviors of a self-exciting process are largely shaped by its history.

Using the self-exciting DGP (the ETAS model, with its intensity as in equation 25) mentioned before as our example: The exciting part of the intensity $\sum_{i:t_i<t} e^{\alpha x_i}(1 + \frac{t-t_i}{c})^{-p}$ governs the individual heterogeneity. Figure 7 presents three quite different individual events histories simulated by our ETAS DGP using the same parameter settings stated before as an example. Individual 1 has the most frequent events

experience, the total number of events is 92. Individual 2 is somewhat moderate, with 37 events. Individual 3 has the least frequent events with only two during the time interval $[0, 100]$. Despite the hugely different behavior, they are governed by the same intensity function.

[Insert Figure 7 Here]

7 Conclusion

In this paper, a series of self-exciting process models are constructed to study the work absenteeism. A minimum distance estimation method is employed. This estimation method, unlike the conventional likelihood-based method, allows including external shocks into the intensity.

In the empirical study, firm-level data is used. The firm introduced an experience rate sick pay scheme that links sick pay benefit with worker's absence history. We find the worker's decision makings are entirely different in short-term, long-term incidence and recovery processes. Specifically, we found substantial evidence supporting the existence of dynamic behavior in both short-term incidence and recovery process. The dynamic behavior is generated by the cumulative absence time. However, in the long-term recovery process, we have to reject the existence of dynamic behavior and state-dependent structure. Instead, we adopt a conventional duration analysis and employ Heckman and Singer's NPMLE to complete the analysis.

References

Aït-Sahalia, Y., J. Cacho-Diaz, and R. J. Laeven (2015): "Modeling financial contagion using mutually exciting jump processes," *Journal of Financial Economics*, 117, 585–606.

Allen, S. G. (1981): "An empirical model of work attendance," *The Review of Economics and Statistics*, 77–87.

Bacry, E. and J.-F. Muzy (2014): "Hawkes model for price and trades high-frequency dynamics," *Quantitative Finance*, 14, 1147–1166.

Barmby, T., C. Orme, and J. Treble (1995): "Worker absence histories: a panel data study," *Labour Economics*, 2, 53–65.

Barmby, T. A., C. D. Orme, and J. G. Treble (1991): "Worker absenteeism: An analysis using microdata," *The Economic Journal*, 101, 214–229.

de Paula, A. and B. Honore (2017): "A New Model for Interdependent Durations," *Quantitative Economics*.

Delgado, M. A. and T. J. Kniesner (1997): "Count data models with variance of unknown form: An application to a hedonic model of worker absenteeism," *Review of Economics and Statistics*, 79, 41–49.

Duflo, E., R. Hanna, and S. P. Rya (2012): "Incentives work: Getting teachers to come to school," *The American Economic Review*, 102, 1241–1278.

Fevang, E., S. Markussen, and K. Røed (2014): "The sick pay trap," *Journal of Labor Economics*, 32, 305–336.

Fister-Gale, S. (2003): "Sickened by the cost of absenteeism, companies look for solutions," *Workforce*, 82, 72–75.

Gaure, S., K. Røed, and T. Zhang (2007): "Time and causality: A Monte Carlo assessment of the timing-of-events approach," *Journal of Econometrics*, 141, 1159–1195.

Heckman, J. and B. Singer (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica: Journal of the Econometric Society*, 271–320.

Heckman, J. J. (1981): "Heterogeneity and state dependence," in *Studies in labor markets*, University of Chicago Press, 91–140.

Honor, B. E. and Á. De Paula (2010): "Interdependent durations," *The Review of Economic Studies*, 77, 1138–1163.

Honoré, B. E. (1993): "Identification results for duration models with multiple spells," *The Review of Economic Studies*, 60, 241–246.

Kopperschmidt, K. and W. Stute (2013): "The statistical analysis of self-exciting point processes," *Statistica Sinica*, 1273–1298.

Lewis, P. A. and G. S. Shedler (1979): "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics (NRL)*, 26, 403–413.

Markussen, S., K. Røed, O. J. Røgeberg, and S. Gaure (2011): "The anatomy of absenteeism," *Journal of health economics*, 30, 277–292.

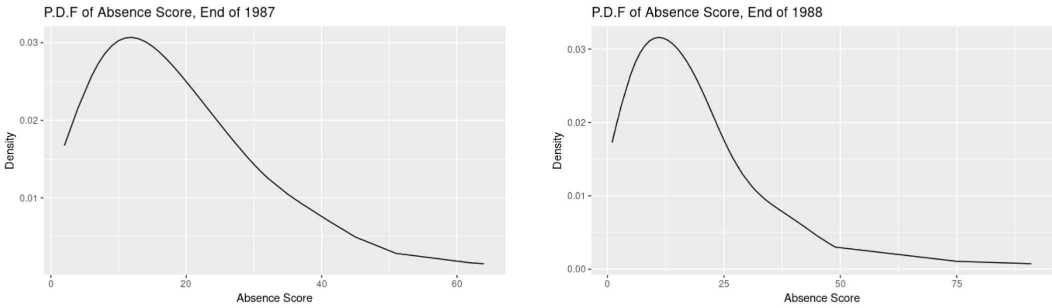
Ogata, Y. (1981): "On Lewis' simulation method for point processes," *IEEE Transactions on Information Theory*, 27, 23–31.

Ogata, Y. and K. Katsura (1988): "Likelihood analysis of spatial inhomogeneity for marked point patterns," *Annals of the Institute of Statistical Mathematics*, 40, 29–39.

Steel, R. P., J. R. Rentsch, and J. R. Van Scotter (2007): "Timeframes and absence frameworks: A test of Steers and Rhodes'(1978) model of attendance," *Journal of Management*, 33, 180–195.

Steers, R. M. and S. R. Rhodes (1978): "Major influences on employee attendance: A process model." *Journal of applied Psychology*, 63, 391.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002): "Stochastic declustering of space-time earthquake occurrences," *Journal of the American Statistical Association*, 97, 369–380.



(a) At the end of 1987
(b) At the end of 1988

Figure 2: Non-parametric P.D.F of absence scores

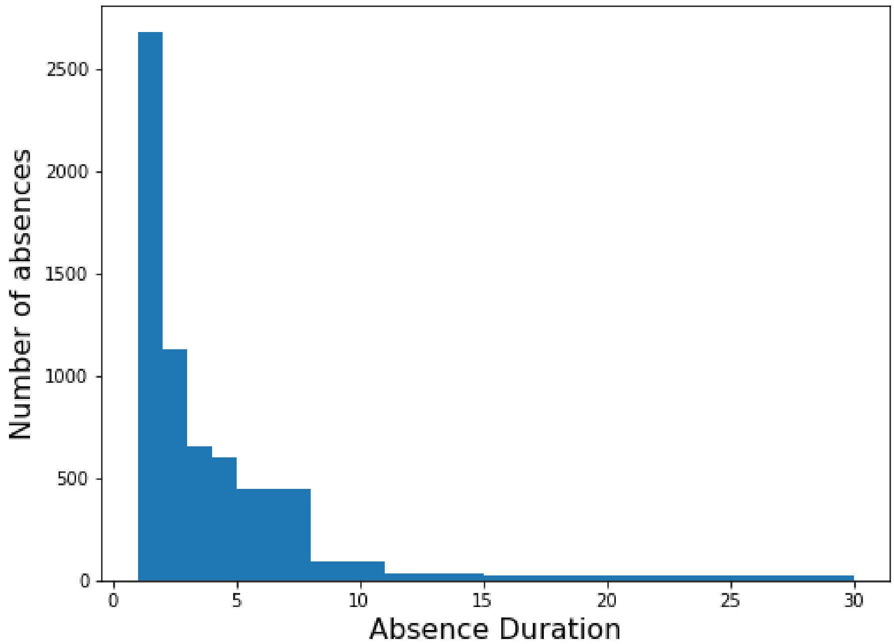


Figure 3: Most frequent absence durations

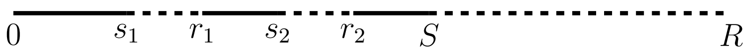


Figure 4: A possible realization of absences

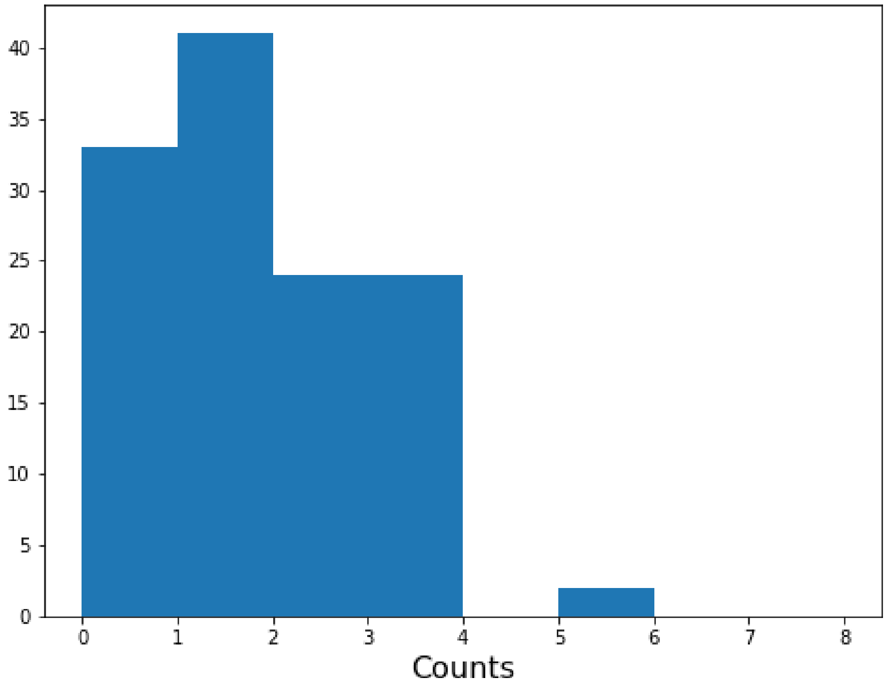


Figure 5: Result of Poisson Process Simulation

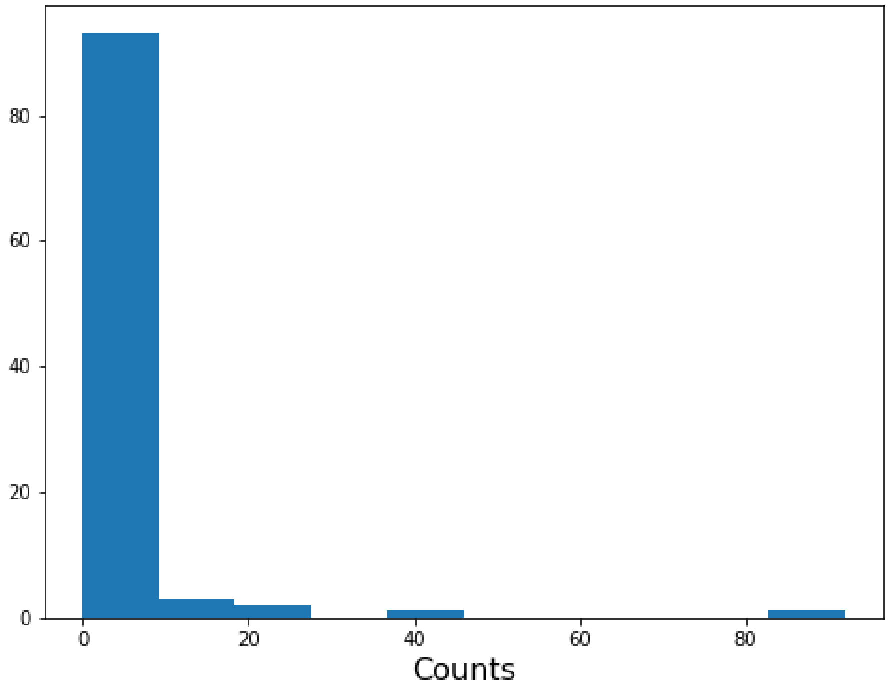
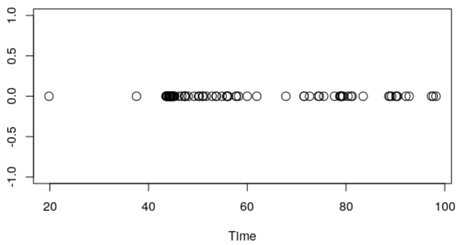
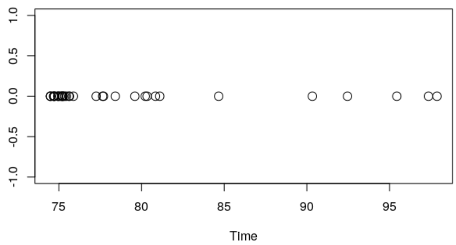


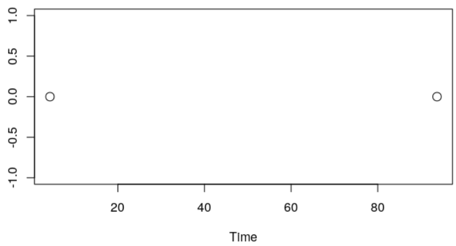
Figure 6: Result of Self-Exciting Process Simulation



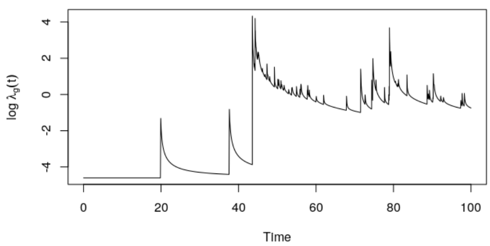
(a) Individual 1, Event Time



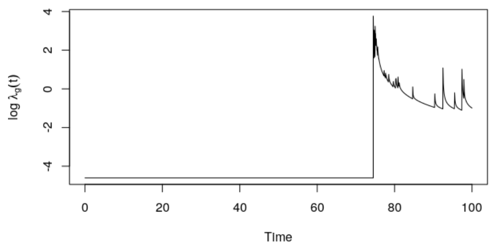
(b) Individual 2, Event Time



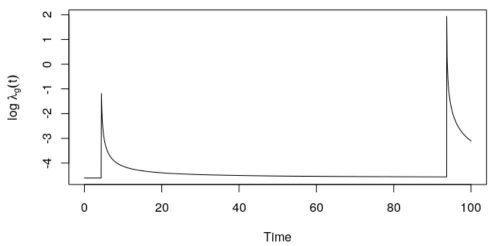
(c) Individual 3, Event Time



(d) Individual 1, log of intenstiy



(e) Individual 2, log of intenstiy



(f) Individual 3, log of intenstiy

Figure 7: Three individual events history

Table 1: Count Data Regression Results

	Dependent variable:			
	count88			
	Poisson	*negative binomial*	*hurdle count part*	*zero-inflated count part*
	(1)	(2)	(3)	(4)
age	−0.005	−0.006	−0.017	−0.005
	(0.011)	(0.016)	(0.012)	(0.012)
age2	0.007	0.008	0.015	0.0002
	(0.014)	(0.019)	(0.014)	(0.016)
male	−0.249***	−0.224***	−0.230***	−0.236***
	(0.045)	(0.065)	(0.046)	(0.048)
full	0.104**	0.115	0.094*	0.115**
	(0.049)	(0.074)	(0.050)	(0.052)
marriage	−0.066	−0.076	0.002	−0.011
	(0.052)	(0.075)	(0.056)	(0.059)
count87	0.131***	0.156***	0.086***	0.101***
	(0.005)	(0.008)	(0.006)	(0.006)
Constant	0.944***	0.866***	1.565***	1.234***
	(0.193)	(0.284)	(0.205)	(0.220)
Observations	874	874	874	874
Log Likelihood	-1,991.314	-1,878.365	-1,965.877	-1,940.922
θ		3.445*** (0.383)		
Akaike Inf. Crit.	3,996.627	3,770.731		

Note: $age2 = age^2/100$. *p<0.1; **p<0.05; ***p<0.01. This table presents the four counting data regression results. For the zero-inflation and hurdle models, we only present the count parts. The dependent variables are the counts of absences in the year 1988. One important trait is the counts of absences in the previous year, which is positively correlated with the dependent variable. It would be wrong to interpret the results as causal, since otherwise it implies the work discipline regulation play exactly the opposite role: encourage more absences. Instead, this trait should be interpreted as the approximation of the unobserved heterogeneity.

Table 2: Duration Analysis Results

	Dependent variable:	
	duration	
	Standard	*Heckman & Singer*
age	-0.028***	-0.068***
	(0.006)	(0.026)
age2	0.047***	0.094***
	(0.010)	(0.032)
male	-0.115	-0.133
	(0.093)	(0.113)
full	0.163	0.147
	(0.105)	(0.133)
marriage	0.076	0.119
	(0.099)	(0.127)
count87	0.254***	0.264***
	(0.010)	(0.013)
Observations	878	878
Log Likelihood	−248.668	-224.8397
χ^2	576.961*** (df = 5)	
Number of Mass Points		2

Note: This table presents the duration analysis results. The subject under study is the attendance duration before 1988's first absence. No short and long-term absence distinguishing in this table. Heckman and Singer's NPMLE is employed to approximate the distribution of unobserved heterogeneity. We found 2 mass points are good enough. The counts in 1987 is positive, indicating that the more absences in the previous year, the higher the likelihood to ask leaves. This result can not be interpreted as casual, instead, it approximate the unobserved heterogeneity. *p<0.1; **p<0.05; ***p<0.01

Table 3: State-Dependent Effect

	short-term incidence	*pooled incidence*	*short-term recovery*	*long-term recovery*
	(1)	(2)	(3)	(4)
α	-0.637768***	-0.761345***	-0.652089***	-0.245150***
	(0.017122)	(0.016289)	(0.026568)	(0.062854)
β	0.446411**	0.000038	-0.189703*	-0.005980
	(0.190365)	(0.167986)	(0.103588)	(0.080234)
Distance	0.277426	0.240651	0.326466	0.212823

Note: Column (1) reports the results for short-term incidence duration. Column (2) reports the result for pooled incidence duration, where we do not distinguish short and long-term incidence duration. Column (3) and (4) report the state-dependent results for short and long-term recovery duration respectively. α is the duration dependent coefficient, while β is the absence score coefficient. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 4: β under different cut-off duration

cut off	*short-term incidence*	*short-term recovery*	*long-term recovery*
	(1)	(2)	(3)
$c = 3$	0.446411**	-0.189703^{*}	-0.005980
	(0.190365)	(0.167986)	(0.080234)
$c = 4$	0.430791**	-0.077864	0.128624
	(0.205761)	(0.132090)	(0.120405)

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 5: NPMLE Results

	short-term incidence	*long-term recovery*
	(1)	(2)
age	-0.056225^{***}	0.220771^{***}
	(0.014975)	(0.017542)
age2	0.107390^{***}	-0.391739^{***}
	(0.032095)	(0.032903)
male	0.001947	0.233080^{***}
	(0.137773)	(0.079700)
full time	0.042278	-0.00121
	(0.150764)	(0.086639)
marriage	0.051824	-0.01931
	(0.208319)	(0.091401)
α	$--$	-0.129389^{***}
	$--$	(0.030884)
Log Likelihood	-249.0849	3288.113
Number of Mass Points	2	2

Note: $age2 = age^2/100$, *p<0.1; **p<0.05; ***p<0.01

Table 6: AFT Model Results

	Dependent variable:
	long-term duration
age	-0.26440^{***}
	(0.00685)
age2	0.44302^{***}
	(0.01569)
male	-0.58349^{***}
	(0.07593)
full time	-0.09395
	(0.08192)
married	0.14154^{*}
	(0.08066)
Observations	1,204
Log Likelihood	2,971.49500
χ^2	-362.47690 (df = 4)

Note: The structure econometric model in our setting is in fact an accelerated failure time model. This table reports the results. Comparing to the long-term duration analysis (table **??**), 1) the coefficient sign are opposite, but the economic meaning are the same; and 2) lack of the heterogeneity term. $age2 = age^2/100$. *p<0.1; **p<0.05; ***p<0.01

A Count Data Regressions and Duration Models

A.1 Four Count Data Regressions

The dependent variable in these models is the counts of events in an interval of time. The most basic count data regression model is the Poisson, where $Pr(C_i = c \mid X_i) = \exp(-\mu(X_i))\mu(X_i)^c/c!$ $\mathbb{E}(C_i|X_i) = \mu(X_i) = Var(C_i|X_i)$, C_i and X_i are counting numbers and covariates for individual i respectively. Normally, $\mu(X_i) = \exp(X_i'\boldsymbol{\beta})$.

The equality between the mean and the variance in the Poisson model is restrictive. A popular generilisation of over-dispersion model is the negative binomial, whose density is given by

$$f_{nb}(c_i \mid X_i) = \frac{\Gamma(c_i + \psi_i)}{\Gamma(\psi_i)\Gamma(c_i + 1)} \left(\frac{\psi_i}{\lambda_i + \psi_i}\right)^{\psi_i} \left(\frac{\lambda_i}{\lambda_i + \psi_i}\right)^{c_i}$$

where $\lambda_i = \exp(X_i'\boldsymbol{\beta})$ and the precision parameter ψ_i^{-1} is specified with $\psi_i = \lambda_i/\alpha$ and a positive over-dispersion parameter α. This specification yields the mean function $\mathbb{E}[C_i \mid X_i] = \lambda_i$ and the variance function $Var[C_i \mid X_i] = (1 + \alpha)\lambda_i$.

Zero-inflation and hurdle models are good at explaining the excess of zeros. The zero-inflation model considers a mixture distribution of a degenerated distribution concentrated on zero and a negative binomial distribution. In particular,

$$\begin{aligned} Pr(C_i = 0 \mid X_i, Z_i) &= \phi(Z_i) + (1 - \phi(Z_i))f_{nb}(0 \mid X_i), \\ Pr(C_i = c_i \mid X_i, Z_i) &= (1 - \phi(Z_i))f_{nb}(c_i), \end{aligned}$$

where Z_i is a vector of zero-inflated covariates,$\phi(\cdot))$ is the binomial probability. The zero-inflation model can be treated as a special case of the latent class model.

The Hurdle model, on the other hand, can be interpreted as the first part concerns the decisions to ask for leave as a binary outcome process, while the second part models the positive number of work absences conditional on the individual seeking a leave. In particular,the first part of the two-part hurdle structure is specified as

$$\begin{aligned} Pr(C_i = 0 \mid X_i) &= \left(\frac{\psi_{h,i}}{\lambda_{h,i} + \psi_{h,i}}\right), \\ Pr(C_i > 0 \mid X_i) &= 1 - \left(\frac{\psi_{h,i}}{\lambda_{h,i} + \psi_{h,i}}\right) \end{aligned}$$

where the subscript h denotes parameters associated with the “hurdle distribution”. The likelihood function associated with this stage of the hurdle process can be maximized independently of the specification of the second stage. The second part of the model is given by the truncated negative binomial distribution:

$$f(c_i \mid X_i, C_i > 0) = \frac{\Gamma(c_i + \psi_i)}{\Gamma(\psi_i)\Gamma(c_i + 1)} \left[\left(\frac{\lambda_i + \psi_i}{\psi_i} \right)^{\psi_i} - 1 \right]^{-1} \left(\frac{\lambda_i}{\lambda_i + \psi_i} \right)^{c_i}.$$

A.2 Duration Models

As mentioned in the paper, the workhorse in the duration analysis is the hazard rate $h(t) = f(t)/S(t)$, where $f(t), S(t)$ are probability density function and its survival function respectively. In a basic duration model, for every individual, define the constant hazard rate and its cumulative hazard rate as:

$$h_i(X_i) = \exp(X_i'\boldsymbol{\beta})$$
$$H_i(T) = h_i(X_i)T$$

Notice from the definition of the hazard rate, we have:

$$-h(t) = \frac{dlog(S(t))}{dt}$$
$$-\int_0^T h(t)dt = log(S(T))$$
$$S(T) = exp(-\int_0^T h(t)dt)$$

Hence the likelihood function is

$$L = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \exp(-H_i(t))[h_i]^{y_i}$$

where y_i is the censoring indicator: if censored, $y_i = 0$, otherwise $y_i = 1$.

One concern regarding this model is the unobserved heterogeneity among individuals. The usual way to account for this is to include a random variable $\nu \sim G$ in the hazard rate.

$$h_i(\nu_i, X_i) = \exp(X_i'\boldsymbol{\beta} + \nu_i)$$

Integrate out the random variable ν, we end up with the marginal hazard rate,

$$\begin{aligned} h(t|X) &= \frac{\int_0^\infty h(X,\nu)S(t|X,\nu)dG(\nu)}{S(t|X)} \\ &= \exp(X\boldsymbol{\beta})\mathbb{E}(\exp(\nu)|T > t, X) \end{aligned} \tag{28}$$

where $S(t|X)$ is the associated survival function.

The second equation comes from the fact that

$$
\begin{aligned}
g(\nu|T > t, Z) &= \frac{Pr\{T \geq t|Z, \nu\}g(\nu)}{Pr\{T > t|Z\}} \\
&= \frac{S(t|Z, \nu)g(\nu)}{S(t|Z)}
\end{aligned}
$$

and

$$
\mathbb{E}(\exp(\nu)|T > t, Z) = \frac{\int_0^\infty \exp(\nu)S(t|Z, \nu)g(\nu)d\nu}{S(t|Z)}
$$

Assume ν is independent from X_i, one may use Heckman and Singer (1984)'s non-parametric maximum likelihood estimator (NPMLE) to avoid unjustified assumptions about the distribution G. Instead, one may approximate G in terms of a discrete distribution.

Let Q be the (prior unknown) number of support points in this discrete distribution and let $\nu_l, p_l, l = 1, 2, \cdots, Q$ be the associated location scalars and probabilities. The likelihood contribution is:

$$
\mathbb{E}[L_i(\nu_i)] = \sum_{l=1}^{Q} p_l L_i(\nu_l), \sum_{l=1}^{Q} p_l = 1
$$

where $L_i(\nu_l) = \exp(-H_i(t|\nu_l, X_i))[h_i(t|\nu_l, X_i)]^{y_i}$.

The likelihood function is

$$
L = \prod_{i=1}^{N} \mathbb{E}[L_i(\nu)] = \prod_{i=1}^{N} \sum_{l=1}^{Q} p_l L_i(\nu_l), \sum_{l=1}^{Q} p_l = 1
$$

The estimation procedure consists of maximising the likelihood function with respect to β as well as the heterogeneity parameters ν_l and their probabilities p_l for different values of Q. Starting with $Q = 1$, and then expanding the model with new support points until there is no gain in likelihood function value.

Heckman and Singer (1984) has proven that such an estimator is consistent, but its asymptotic distribution has not been discussed yet. Gaure et al. (2007) provide Monte Carlo evidence indicating the parameter estimates obtained by NPMLE are consistent and approximately normally distributed and hence can be used for standard inference purpose.

B Properties of the Doob-Meyer Decomposition and the Inference Theorems

B.1 Properties of the Decomposition

Uniqueness of the Doob-Meyer Decomposition Suppose we have an alternative decomposition:

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

Conclude that

$$\Lambda(t) - \tilde{\Lambda}(t) = M(t) - \tilde{M}(t)$$

whence

$$\mathbb{E}[\Lambda(t) - \tilde{\Lambda}(t)|\mathcal{F}(t-)] = \mathbb{E}[M(t) - \tilde{M}(t)|\mathcal{F}(t-)]$$

Since $\Lambda(t)$ and $\tilde{\Lambda}(t)$ are measurable w.r.t $\mathcal{F}(t-)$ and $M(t)$ and $\tilde{M}(t)$ are martingales, the last equation yields

$$\Lambda(t) - \tilde{\Lambda}(t) = M(t-) - \tilde{M}(t-) = \Lambda(t-) - \tilde{\Lambda}(t-)$$

By induction, it follows that

$$\Lambda(t-) - \tilde{\Lambda}(t-) = \Lambda(0) - \tilde{\Lambda}(0) = 0 - 0 = 0$$

and, finally, $M(t) = \tilde{M}(t)$.

Proof of $\mathbb{E}N(t) = \mathbb{E}\Lambda(t)$ Before going into the proof, we first introduce the following lemma:

Lemma Let the random variable X_i records the occurrence time of i^{th} event and let $F_i(\cdot)$ be the distribution function of X_i. Denote $\mathbb{I}(\cdot)$ the indicator function. We have

$$\mathbb{E}(N(t)|\mathcal{F}(t-)) = N(t-) + \mathbb{I}(X_{N(t-)+1} > t-)\frac{F_{N(t-)+1}(dt)}{1 - F_{N(t-)+1}(t-)}$$

Proof. Note

$$N(t) = N(t-) + \mathbb{I}(t- < X_{N(t-)+1} \leq t)$$

The first component is measurable w.r.t $\mathcal{F}(t-)$. For the second component we get

$$\begin{aligned}
&\mathbb{E}[\mathbb{I}(t- < X_{N(t-)+1} \leq t)|\mathcal{F}(t-)] \\
&= \frac{\mathbb{I}(t- < X_{N(t-)+1}) \int_{\{t- < X_{N(t-)+1}\}} \mathbb{I}(t- < X_{N(t-)+1} \leq t) dF_{N(t-)+1}(x)}{1 - F_{N(t-)+1}(t-)} \\
&= \mathbb{I}(X_{N(t-)+1} > t-)\frac{F_{N(t-)+1}(dt)}{1 - F_{N(t-)+1}(t-)}
\end{aligned}$$

□

Now we prove the main claim.

Proof. Let $N(0) = \Lambda(0) = M(0) = 0$. By the Doob-Meyer decomposition, the proof is completed if $\mathbb{E}M(t) = 0, \forall t > 0$. For $\forall t > 0$, set, by recursion,

$$\Lambda(t) = \Lambda(t-) - N(t-) + \mathbb{E}(N(t)|\mathcal{F}(t-))$$

$$M(t) = M(t-) + N(t) - \mathbb{E}(N(t)|\mathcal{F}(t-))$$

From the above lemma, the martingale part of the counting process then satisfies the recursion

$$M(t) = M(t-) + \mathbb{I}(t- < X_{N(t-)+1} \leq t) - \mathbb{I}(X_{N(t-)+1} > t-)\frac{F_{N(t-)+1}(dt)}{1 - F_{N(t-)+1}(t-)}$$

$$M(dt) = \mathbb{I}(t- < X_{N(t-)+1} \leq t) - \mathbb{I}(X_{N(t-)+1} > t-)\frac{F_{N(t-)+1}(dt)}{1 - F_{N(t-)+1}(t-)}$$

Taking integral on both side w.r.t a Lebesgue measure leads to

$$M(t) = \mathbb{I}(X_{N(t-)+1} \leq t) - \int_0^t \frac{\mathbb{I}(x \leq X_{N(t-)+1})}{1 - F_{N(t-)+1}(x-)} F_{N(t-)+1}(dx)$$

The proof is completed by taking expectation on both sides on the last equation. □

B.2 Inference Theorems

The following theorems come from Kopperschmidt and Stute (2013). let $v_0 \in \Theta \subset \mathbb{R}^d$ be the true parameters, and let $\Lambda_{v,i}$ with $v \in \Theta \subset \mathbb{R}^d$ be a given class of parametric cumulative intensities.

Theorem (Consistency) Let $\Theta \in \mathbb{R}^d$ be a bounded open set and for each $\epsilon > 0$, we assume,

$$\inf_{||v-v_0|| \geq \epsilon} ||\mathbb{E}\Lambda_{v_0} - \mathbb{E}\Lambda_v||_{\mathbb{E}\Lambda_{v_0}} > 0 \tag{29}$$

$$\textit{The process}(t, v) \to \Lambda_v(t) \textit{ is continuous with probability one} \tag{30}$$

Then

$$\lim_{n\to\infty} v_n = v_0 \textit{ with probability one} \tag{31}$$

Condition 29 is a weak identifiability condition, while condition 30 guarantees continuity (but not differentiability) of Λ_v in t and allows for unexpected jumps in the intensity function λ_v as well.

Theorem (Asymptotic behavior) Let

$$\Phi_0(v) = \frac{\partial}{\partial v} \int_E (\mathbb{E}\Lambda_{g,v}(t) - \mathbb{E}\Lambda_{g,v_0}(t))\mathbb{E}\frac{\partial}{\partial v}\Lambda_{g,v}(t)^T \mathbb{E}\Lambda_{g,v_0}(dt) \tag{32}$$

a matrix-valued function, where T denotes transposition, $E = [\underline{t}, \bar{t}]$. And suppose (29) and (30) hold, furthermore, assume that

$$\| \frac{\partial}{\partial v}(\mathbb{E}\Lambda_{g,v}(t) - \mathbb{E}\Lambda_{g,v_0}(t)\mathbb{E}\frac{\partial}{\partial v}\Lambda_{g,v}(t)^T) \| \leq C(t) \tag{33}$$

for all v in a neighborhood of v_0, function C is integrable w.r.t $\mathbb{E}\Lambda_{v_0}$, and

$$\phi(x) = \int_{[x,\bar{t}]} \mathbb{E}\frac{\partial}{\partial v}\Lambda_{g,v}(t)\mathbb{E}\Lambda_{g,v_0}(dt) \mid_{v=v_0}, \underline{t} \leq x \leq \bar{t} \tag{34}$$

is square integrable w.r.t. $\mathbb{E}\Lambda_{v_0}$. Then as $n \to \infty$

$$\sqrt{n}\Phi_0(v_0)(v_n - v_0) \to \mathscr{N}_d(0, C(v_0)) \tag{35}$$

where $C(v_0)$ is a $d \times d$ matrix with entries

$$C_{ij}(v_0) = \int_E \phi_i(x)\phi_j(x)\mathbb{E}\Lambda_{g,v_0}(dx) \tag{36}$$

Remark Let Φ_n be the empirical analogue of Φ_0,

$$\Phi_n(v) = \frac{\partial}{\partial v} \int_E (\bar{\Lambda}_{v,n}(t) - \bar{\Lambda}_{v_0,n}(t))\frac{\partial}{\partial v}\bar{\Lambda}_{v,n}(t)^T \bar{\Lambda}_{v_0,n}(dt) \tag{37}$$

Since all $\bar{\Lambda}_{v,n}$ are sample means of i.i.d non-decreasing processes, a Glivenko-Cantelli argument yields, with probability one, uniform convergence of $\bar{\Lambda}_{v,n} \to \mathbb{E}\Lambda_v(t)$ in each t on compact subsets of Θ, we have the expansion,

$$\Phi_n(v) = \Phi_0(v) + op(1) \tag{38}$$

Such expansion guarantees that in a finite sample situation, we can replace the unknown matrix $\Phi_0(v_0)$ by $\Phi_n(v_n)$ and $C(v_0)$ by $C^n(v_n)$ without destroying the distributional approximation through $\mathscr{N}_d(0, C(v_0))$, where C^n is the sample analog of C. In practice, one need to plug and replace the true ones with estimators and replace $\mathbb{E}\Lambda_{v_0}(dt)$ with its empirical counterpart $\bar{N}(dt)$.

C Simulation Procedure

Here we describe the simulation procedures of the Poisson and the ETAS processes.

C.1 Simulation of Poisson Process

We use the fact that the inter-event durations $d_i = t_i - t_{i-1}, i = 2, 3, \cdots$ of a Poisson process are exponentially distributed:

$$F(d) = 1 - \exp(-\lambda d)$$

Procedure

1. Draw a uniformly distributed random variable $U \in [0, 1]$, the duration until next event is then generated by:

$$nextTime = \frac{-\log(1-U)}{\lambda}$$

2. Update the time list by:

$$\text{TimeList} = \text{append(TimeList,TimeList+nextTime)}$$

3. if the latest time stamps is below the terminated time $TimeList \leq tMax$, repeat steps 1 and 2, otherwise, terminate the program.

C.2 Simulation of ETAS Process

The detailed thinning method steps can be summarised as:

1. Let τ be the start point of a small simulation interval
2. Take a small interval $(\tau, \tau + \delta)$
3. Calculate the maximum of $\lambda_g(t|\mathcal{F}_{t-})$ in the interval as

$$\lambda_{max} = \max_{t \in (\tau, \tau+\delta)} \lambda_g(t|\mathcal{F}_{t-})$$

4. Simulate an exponential random number ξ with rate λ_{max}
5. if

$$\tfrac{\lambda_g(\tau+\xi|\mathcal{F}_{t-})}{\lambda_{max}} < 1$$

go to step 6.

Else no events occurred in interval $(\tau, \tau + \delta)$, and set the start point at $\tau \leftarrow \tau + \delta$ and return to step 2

6. Simulate a uniform random number U on the interval $(0, 1)$
7. If

$$U \leq \frac{\lambda_g(\tau+\xi|\mathcal{F}_{t-})}{\lambda_{max}}$$

then a new 'event' occurs at time $t_i = \tau + \xi$. Simulate the associated marks for this new event.

8. Increase $\tau \leftarrow \tau + \xi$ for the next event simulation
9. Return to step 2

D Optimization Procedure

Due to the highly non-linear nature of our intensity function (hence the distance function), there is no obvious closed-form solution. A set of numerical optimization routines are employed.

It is well known that for non-linear optimization, the starting points (or guess points) are important. Different starting points may lead to different optimization results. To minimize such impacts, we first use heuristic algorithms such as simulated annealing to find a set of 'proper' starting points. These algorithms, however, are usually not speedy enough, we thus break the process after 24 hours.

Using these 'proper' starting points, we next perform the usual optimization routine such as BFGS ,L-BFGS-B, Nelder-Mead and other Newton-based methods. We fail to find a single optimization routine that can 'fit' all the distance functions. For example, in the short-term incidence, although the BFGS routine can find the smallest distance value, the estimator for the absence score is positive (and significant), i.e., the higher the absence score is, the more likely to ask for leaves. This result is obviously wrong. Instead, we perform a bounded optimization using the L-BFGS-B routine. The new distance value is larger, but difference is quite small.

We write the code in Python with several packages installed, among them, the most important ones are numpy and numba.