Sick but at Work.
An Econometric Approach to Presenteeism

Benjamin Huver * Sébastien Richard * Nicolas Vaneecloo *
Thomas Delclite * Ingrid Bierla†

15^{th} *IZA European Summer School in Labor Economics*
Buch am Ammersee – April 23-29, 2012

Introduction

Companies facing the problem of absenteeism often undertake a complete overhaul of their absence management and sickness compensation policies. In order to minimize the economic losses and the organizational failures related to high absence rates, an increase in the psychological and/or financial cost of absence appears as a rational decision.
Thus, some companies even pay "attendance bonuses" to employees who were never absent over a given time. But, by fostering (directly or indirectly) their employees to attend to work regardless of their health status, companies forget a crucial point. Absence could namely be more legitimate than presence: for an employee, taking two days off from work to fully recover is sometimes the best way to carry out his functions efficiently again. This behavior of an employee going to work while sick is commonly called "presenteeism". For himself, the consequences are obvious: worsening health conditions, increased risks, lack of well-being, etc. But some studies show that presenteeism is also harmful to firms. Goetzel et al. (2004) highlighted that for companies, presenteeism costs could represent up to 61% of the aggregated costs related to health. This would mainly be due to a decreased on-the-job productivity.
But the identification of presenteeism remains uneasy. Regarding absenteeism, some simple descriptive statistics may be sufficient to provide a first impression: the number of days absent for each employee is usually a free and easily available information for firms. By contrast, measuring presenteeism requires very costly (both in time and money) qualitative surveys and sometimes results in potentially unreliable declarative data.

*University of Lille – CLERSÉ. Corresponding author: huver.benjamin@gmail.com.
†IÉSEG School of Management

Hence, our paper's starting point is a very simple observation. It would be relevant to work out a quantitative tool designed to measure presenteeism based on employees absence data. So far, such instrument is not defined in the academic literature (section 1), but a "zero-inflated negative binomial model" is well suited to this task. After a detailed presentation of its theoretical underpinnings (section 2), we propose an empirical application to the case of a French hospital with about 3600 employees (section 3). We then analyze our results and provide evidence on the reliability of our measures, thanks to longitudinal absence data (Section 4).

1. Sick workers

Before presenting the model, it is worth to describe presenteeism. To our knowledge, up to now only qualitative studies explored it, especially through large-scale survey data (mostly from Nordic countries). Some of the questions in these surveys directly focus on presenteeism: they therefore constitute the best sources of knowledge on the subject.

1.1 What is presenteeism?

A definition of presenteeism was widely accepted in the last years (Aronsson & Gustafsson 2005, Johns 2009, Bergström et al. 2009); it characterizes the behavior of an employee who goes to work even though his health status would justify a sick leave. From this fundamental definition, two alternative statistical approach can be deduced. First, in line with Hansen & Andersen (2008), presenteeism and absenteeism should be understood as two results of the same process. Let us consider that there is a standard absence level, or in other words an average absence level, corresponding to a normal behavior: "if I am sick, I stay at home, otherwise I show up at work". In reference to this standard behavior, a group of individuals has an abnormally low absence level (presenteeism) and another group experiences an abnormally high absence level (absenteeism). The benefit of this simplified approach is to clarify the links between the studied phenomena and to anticipate our econometric procedure.
A second standpoint emerges when analyzing presenteeism over time. When an employee is used to attend work despite being sick, his past absence records should be characterized by very low or zero values. Thus, if longitudinal data is available, a diachronic analysis is another way to measure presenteeism. Some studies indeed recognize the relevance of such longitudinal approach, especially to explore the links between presenteeism and further sickness absence (Kivimäki et al. 2005, Bergström et al. 2009, Demerouti et al. 2009).

1.2 What are the causes of presenteeism?

Two types of factors are identified to explain presenteeism: personally-related and work-related factors (Aronsson & Gustafsson 2005). Demographic variables fall into the first category. Initially, it may be observed that the effects of gender and age are not subject to a real consensus. Older employees would be more likely to attend work while sick: according to Hansen & Andersen (2008), presenteeism and age are positively correlated. Concerning gender, and focusing on a hospital context, men seem to be more at risk than women (Demerouti et al. 2009). A possible explanation is that women put more effort into family life and childcare. But Aronsson et al. (2000) get contradictory results: employees with dependent children are more likely to experience presenteeism. The link between presenteeism and family life remains to be demonstrated.
However, studies don't develop these demographic factors and often use them as mere control variables. In fact, job-related causes have a more decisive effect on presenteeism (Hansen & Andersen 2008). Presenteeism is prevalent in jobs where attendance has a great influence on other people and on their primary needs (Aronsson et al. 2000). This is the case in the hospital sector. According to Demerouti et al. (2009) working every day closely with patients leads nurses to believe that their tasks cannot be postponed or delegated: this paves the way for presenteeism. Team membership produces the same effect. Caverley et al. (2007) suggest that the decision to be at work despite being sick is related to two key issues. First, if replacement is impossible, employees want to spare their colleagues an additional workload. Second, if they manage a team, the feel responsible for their subordinates and want to ensure that the activities run smoothly. It is, in fact, widely accepted that presenteeism behavior more often occurs for managers.
But, beside the working environment, an individual's job situation can also produce presenteeism. Perfect work attendance, whatever the circumstances, is sometimes perceived by employees as organizational commitment and as an evidence of loyalty to the company (Hansen & Andersen 2008). Presenteeism is for instance higher for employees hired on fixed-term contracts (Caverley et al. 2007), who expect to achieve a permanent status. Such behavior is a kind of "voluntary" presenteeism (resulting from an individual's decision) that some authors (among them Gosselin & Lauzier (2011)) distinguish from "involuntary" presenteeism, imposed upon employees by organizational demands.

1.3 What are the consequences of presenteeism?

Presenteeism may first lead to longer sick leaves in the following months or years. Bergström et al. (2009) show that employees reporting that they attend frequently while sick at the beginning of the study, are significantly

more absent 18 months and three years later. These respondent's perceived health status clearly deteriorates during the study. The harmful impact of presenteeism on health is also highlighted by Kivimäki et al. (2005): a three year follow-up of British civil servants reveals that individuals who were never sick during the period have an "increased risk of serious coronary events" (Kivimäki et al. 2005, p.102) and a poorer health compared to others. "Costs of presenteeism" are also identified. The background to these costs is a decreased on-the-job productivity, which accounts for a large part in the total health-related expenses of employers. Goetzel et al. (2004) even emphasize that these presenteeism costs are significantly higher than medical or absence costs. Moreover, according to Demerouti et al. (2009), in the hospital sector, presenteeism could affect the quality of care.

2. Model overview

These outcomes will constitute a reference point when it comes to evaluate the relevance of our own results. But in this section, we first give theoretical considerations on zero-inflated models. The starting point is a simple statistical distribution: the Poisson distribution. After several improvements, we will show that a particular treatment of zero values is well suited to the measure of presenteeism.

2.1 Absence data and probability distributions

Poisson distribution

Absence variables are "count data": they count, for instance, the number of days absent for each employee in a year. The Poisson distribution is used to model the number of events occurring over a given time, and therefore seems to be well suited for absence data. More formally, if a random variable called Y_i follows a Poisson distribution with mean $\lambda > 0$, then

$$P(Y_i = y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{1}$$

and

$$E(y_i|x_i) = V(y_i|x_i) = \lambda_i = \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k) = e^{X'\alpha} \tag{2}$$

where y_i is the observed (discrete) variable, λ_i the mean (expected value), $X = [x_1\ x_2\ ...\ x_k]$ a vector of k explanatory variables and $\alpha = [\alpha_0\ \alpha_1\ ...\ \alpha_k]$ a vector of k coefficients.
In a Poisson distribution mean and variance are assumed to be equal. However, for absence data, this assumption is never satisfied: it is, on the contrary, characterized by an important overdispersion (the variance is larger

than the mean). A single Poisson process with given mean cannot explain the population's absence behavior: there is too much heterogeneity among individuals. That could be explained by differences in the health status, for example.

Negative binomial distribution

The traditional way to deal with unobserved heterogeneity is to use a negative binomial distribution. It is an improvement or a generalization of the simple Poisson model. The main change lies in the input of an heterogeneity parameter devoted to the capture of overdispersion (Hilbe 2011). In the Poisson distribution (see equation 1), $\lambda_i = \exp(X'\alpha)$, that's to say $\log(\lambda_i) = X'\alpha$. We introduce u_i as a multiplicative parameter:

$$f(y_i|x_i, u_i) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!}$$

To model this parameter, we apply a Gamma distribution[1], which fits to the particular nature of u_i. This dispersion parameter is namely multiplicative: since λ is always positive, u_i has to be positive. Moreover, because it is designed to capture dispersion, u_i should allow a higher variance and leave the Poisson mean unchanged. The Gamma is well suited to do so: it models real positive variables of various shapes (Hilbe 2011), and if we consider that u_i follows a Gamma distribution accepting two equal parameters, so that $u_i \sim \Gamma(\theta, \theta)$, we find $E(u_i) = 1$ (and $V(u_i) = 1/\theta$). Thus, the mean of the negative binomial distribution will finally be equal to the Poisson mean (λ), and the variance will be $V(y_i) = \lambda + (\lambda^2/\theta)$. Identity between the mean and variance is no more needed and the link between mean and variance depends on parameter θ. It may also be noticed that the Gamma distribution provides an analytical solution for the maximum likelihood estimation.
If we integrate the Gamma distribution in the Poisson model, we get the negative binomial probability mass function:

$$P(Y_i = y_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \left(\frac{\theta}{\theta + \lambda_i}\right)^{\theta} \left(\frac{\lambda_i}{\theta + \lambda_i}\right)^{y_i} \tag{3}$$

2.2 Modeling excess zeros

Another feature of absence data is that nil values are overrepresented. A majority of employees are actually never absent during a given year. The real difference between basic models and zero-inflated models lies precisely in a special treatment of zero values. This property allows also an estimation of presenteeism behavior. Zero-inflated models assume that two separate processes may produce nil values (Winkelmann 2008). In the first situation,

[1]The negative binomial model is a Poisson-Gamma mixture.

zeros result from a "normal" count data process, which produces nil values, but also positive values. These are "incidental" zeros (Staub & Winkelmann 2011): the employee was not absent because he was not sick. If sickness had occurred, he would obtain a positive value. In the second case, the zeros come from a specific process (called "perfect state" by some authors) that produces only nil values. The employee was sick, but, despite his poor health status, he voluntary decided to attend work. These zero values are "strategic" We call such behavior "presenteeism".
To our knowledge, two papers concerning absenteeism and using zero-inflated models also noticed this particular behavior, but don't clearly identified it as presenteesim. A French article (Missègue 2007, p.11) distinguishes these two sources of nil values. On the one hand, "individuals with health problems, who show up even if a sick leave would be justified" and, on the other hand, individuals who were not sick. As well, Frick & Malo (2008, pp.517-518) assume that "some individuals will have zero absence days because they follow an absolute rule of no voluntary absenteeism".
The probability distribution of a zero-inflated model also shows the separate treatment of nil values:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)f(0) & \text{if } y_i = 0 \\ (1-p_i)f(y_i) & \text{if } y_i > 0 \end{cases} \quad (4)$$

Some general remarks can be made here about the model's components. First of all, the function noted $f(.)$ can be a Poisson or a negative binomial probability function[2]. The only difference with basic models is that $f(0)$ (incidental zeros) and $f(y_i) \forall y_i > 0$ (other positive values) are distinguished. Furthermore, if $y_i = 0$, we see that two situations are taken into account: p_i is the probability to be in a "perfect state" where only zero values are generated and $(1-p_i)$ covers the situations where y_i can be either nil or positive. This raises the question of the interpretation of these three components: p_i, $(1-p_i)f(0)$ and $(1-p_i)f(y_i)$.
p_i and $(1-p_i)f(0)$ differentiate between two sources of nil values. p_i is the probability that zeros are "structural" (Staub & Winkelmann 2011) and $(1-p_i)f(0)$ captures "incidental" zeros. p_i depends on a specific vector of independent variables (denoted by Z): we consider that this set of variables includes the explanatory factors for presenteeism and that p_i is the presenteeism probability for individual i. It is important to notice that this vector Z can be identical or different from the vector of explanatory variables X used to estimate the count data model (Winkelmann 2008). The probability p_i is estimated through a logistic regression model. We can write:

$$p_i = \frac{exp(Z'\beta)}{1 + exp(Z'\beta)} \quad (5)$$

[2]Or even more generally, any other statistical probability function.

where $Z = [z_1\ z_2\ ...\ z_k]$ is a vector of k explanatory variables and $\beta = [\beta_0\ \beta_1\ ...\ \beta_k]$ a vector of k coefficients.
Thus, the "Zero-Inflated Negative Binomial" (ZINB) can be written:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)\left(\frac{\theta}{\theta+\lambda_i}\right)^{\theta} & \text{if } y_i = 0 \\ (1-p_i)\frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\lambda_i}\right)^{\theta}\left(\frac{\lambda_i}{\theta+\lambda_i}\right)^{y_i} & \text{if } y_i > 0 \end{cases} \quad (6)$$

And the log-likelihood function is:

$$\mathcal{L} = \begin{cases} \sum_{i=1}^{n} \log\left[p_i + (1-p_i)\left(\frac{\theta}{\theta+\lambda_i}\right)^{\theta}\right] & \text{if } y_i = 0 \\ \sum_{i=1}^{n} \log(1-p_i) + \log\Gamma(y_i+\theta) + \log\Gamma(y_i+1) - \log\Gamma(\theta) & \\ \quad +\theta\log(\theta) + y_i\log(\lambda_i) - (\theta+y_i)\log(\theta+\lambda_i) & \text{if } y_i > 0 \end{cases} \quad (7)$$

2.3 ZINB models: further illustrations

Two articles using the ZINB model to deal with absence data have already been mentioned above (Missègue 2007, Frick & Malo 2008). Other studies make use of zero-inflated models, either to show their robustness (Garay et al. 2011) or to apply them empirically in various fields. A few examples will allow a better understanding of the way presenteeism can be inferred from a zero-inflated model.
In the seminal paper of Lambert (1992) a zero-inflated is applied to the number of defects in a production process. If, over a given time, there is no defect, two explanations are possible: the process can either be fitted properly ("perfect state") or the process remains maladjusted but no defect occurred. In this latter case, the number of defects could have been greater than zero: it follows a Poisson distribution.
Another example, in the field of road safety. Allain & Brennac (2001) study the number of accidents on various road sectors. The two groups of zero values are described as follows: "the studied road sector can be divided in two states, two groups. In the 'zero-accident state', the road sectors are basically safe, at low risk, that is to say that no accident has been observed. In the 'accident state', the number of accidents follows a given distribution (Poisson or negative binomial) and, among them, one can find road sectors where no accident was observed."

3. Empirical specification

The approach developed in these few examples will now be transferred to presenteeism behavior. To show the practical interest of the ZINB model, it will be applied to micro-data coming from a French hospital.

3.1 Explanatory variables

The database consists of demographic variables, job-related variables and absence data for each employee. It covers a 30-year time period (1981-2010): such time depth is quite rare for micro-data and will be useful to give evidence on our model. These explanatory variables are presented in Table 1.

Table 1: Explanatory variables in the model

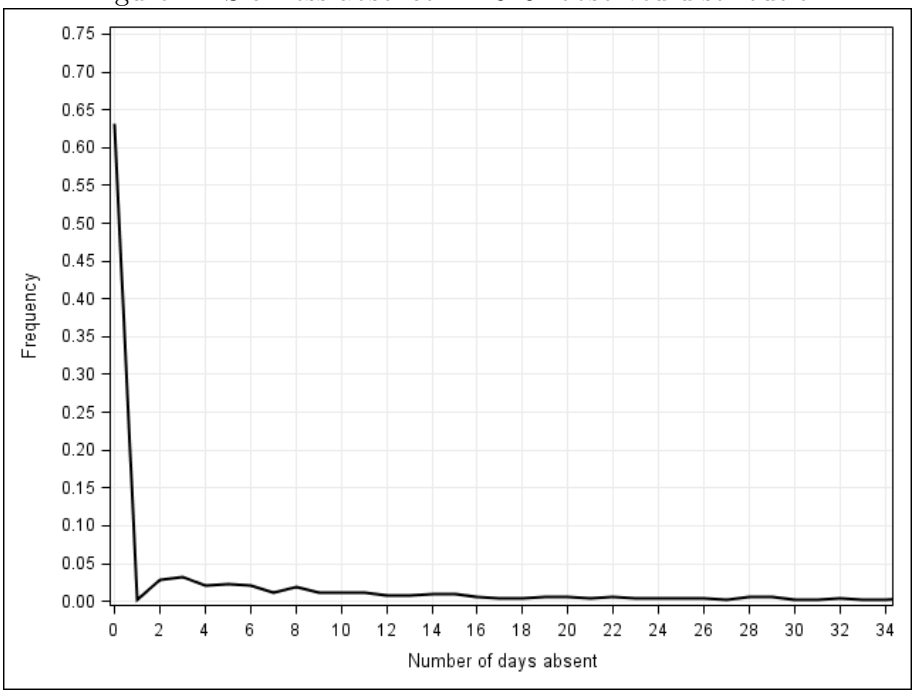
Category	Variables	Type	Statistics
Demographic variables	Man	B	(1) 21.2%
	Age	C	($\bar{x}$) 43 years
	Age squared	C	
	Single	B	(1) 17.0%
	Married in 2009 or 2010	B	(1) 2.3%
	No children	B	(1) 18.7%
	One child	B	(1) 24.1%
Job-related variables	Job seniority	C	($\bar{x}$) 5.6 years
	Job mobility in 2008 or 2009	B	(1) 24.1%
	Manager	B	(1) 4.4%
	Non-permanent staff	B	(1) 8.8%
	Clerical employee	B	(1) 10.4%
	Service staff	B	(1) 8.6%
	Caregiver	B	(1) 23.7%
	Nurse (higher class)	B	(1) 10.6%

Notes: B=binary, C=continuous, ($\bar{x}$)=mean, (1)=share in total population

3.2 Dependant variable

The dependent variable in the ZINB model is the number of days absent for sickness in 2010. Individuals with more than 70 days absent were excluded for two reasons. First, the purpose of this paper is to study presenteeism: it is obvious that beyond a given severity level (in the case of an accident or severe disease, for instance), sick workers cannot attend work, even if they are willing to do so. From a statistical point of view, in our dataset, this level can be set at 70 days absent in one year: beyond this threshold only a few individuals are concerned, and they all have very long absence periods. On the other hand, presenteeism is considered as a decision. Including long-term sickness absences would therefore result in a biased estimation. Because these events occur independently from an individual's choice and because they necessarily imply a withdrawal, it seems relevant to exclude

Figure 1: Sickness absence in 2010: observed distribution



them from the database.
In addition, the dependant variable has the classic features of absence data. It is discrete, overdispersed (the empirical mean is 6.63 versus a variance of 183.22) and nil values are overrepresented (63.17% of the total population was never absent in 2010). Figure 1 presents the observed distribution.

3.3 Model estimation

The estimation output is presented in Table 2. One can see that the coefficients are divided into two blocks: the first block consists in the negative binomial explanatory variables (vector X as noted above), the second contains the variables of the logistic model (vector Z). Only the most significant variables are retained. The coefficients will be analyzed in the next section (section 4.1 page 11).
Several other indicators appear in the model output. First, the dispersion parameter α. It is simply the inverse of θ in the ZINB model log-likelihood function (see equation 7). As a link between mean and variance, the α parameter directly indicates the degree of overdispersion in the model (Hilbe 2011). When it tends to 0 (which means that the variance tends to λ), the model returns to a Poisson distribution (no dispersion). This value in the ZINB model (0.978) has to be compared to the value in a basic negative

binomial model without zero inflation, where $\alpha = 7.499$. The overdispersion of the data is obvious, but it is significantly reduced by a specific treatment of zero values. Parameter α is fundamental. It namely allows an efficient model estimation, even if missing variables lead to unobserved heterogeneity. In our case, dispersion among employees could be explained by heterogeneous health levels, for example, which remains an unobserved information.
Two popular indicators were helpful to select this model among other attempts: the log-likelihood and the AIC (Akaike Information Criterion). The log-likelihood ($\mathcal{L}$) is provided through the estimation. The AIC additionally takes into account the number of variables in the model[3].

Table 2: Zero-inflated negative binomial regression

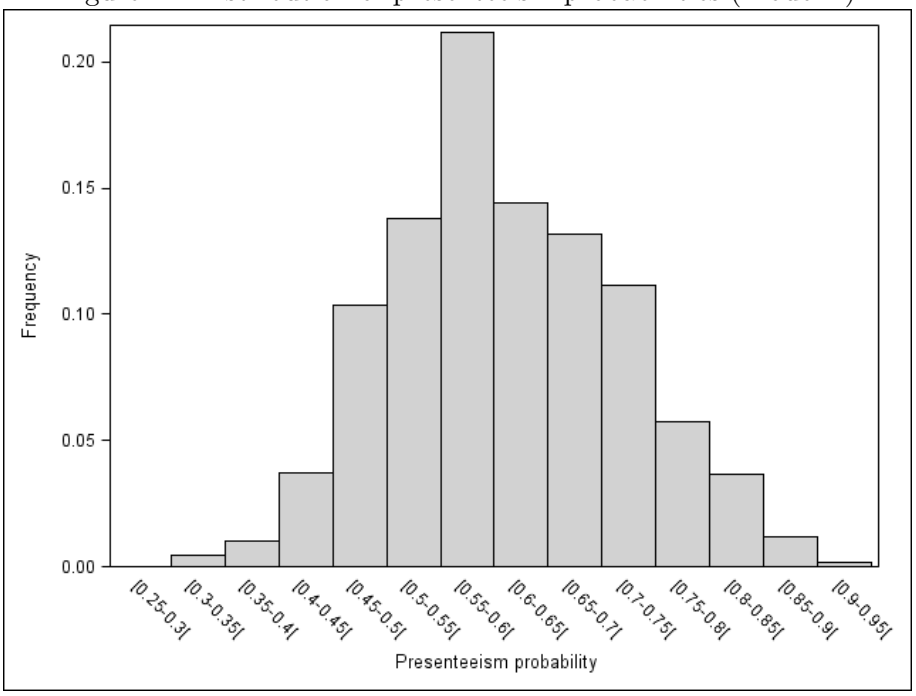
Parameter	Type	Estimate	
Intercept		5.160	***
Single	Binaire	-0.264	**
Clerical employee	Binary	-0.316	**
Service staff	Binary	0.179	.
Caregiver	Binary	0.153	*
Age	Continuous	-0.109	**
Age squared	Continuous	0.001	**
Intercept		-0.328	
Man	Binary	0.321	**
Age	Continuous	0.010	*
Married in 2009 or 2010	Binary	-0.466	.
No children	Binary	0.313	**
One child	Binary	-0.224	*
Job seniority	Continuous	0.046	***
Job mobility in 2008 or 2009	Binary	-0.260	**
Service staff	Binary	-0.503	***
Nurse (higher class)	Binary	0.500	***
Manager	Binary	0.983	***
Non-permanent staff	Binary	0.706	***
α		0.978	***
Log-likelihood		-5971	
AIC		11983	

P-value: <0.1%=*** <1%=** <5%=* <10%=.

One may also compute a presenteeism probability for each employee (denoted by p_i in the previous section, cf. equation 4). The presenteeism probability distribution is presented in Figure 2.

[3] AIC $= -2\mathcal{L} + 2k$ (where k is the number of variables): the lower it is, the better the model.

Figure 2: Distribution of presenteeism probabilities (Model 1)



4. Results and relevance analysis

In this final section, we first show the relevance of using a ZINB model to measure presenteeism. In general, our results are consistent with other existing outcomes. As well, we will see that the variables included in the model confirm by themselves that the captured phenomenon is undoubtedly presenteeism. We then approach this question from another angle, by comparing presenteeism probabilities to past absence records. The model estimation could also be significantly improved by introducing a new set of explanatory variables: previous absences of employees.

4.1 Empirical results

Variables and estimates

The ZINB estimates should be read in the light of a common analytical framework. Let us first raise a preliminary matter: when, in a given year, an employee has no absence, what are the reasonable grounds? There are two compelling reasons: either the individual's health status is very good and he is never sick, or he attends work while sick, adopting a behavioral rule of voluntary presence.
For the independant variables in Table 2 (second block), the first explana-

tion does not hold up. Age is a key example: the coefficient is positive. This means that the occurrence of the observed phenomenon increases with age. Therefore, it cannot merely be explained by a good health or physical strength because health generally declines with age and because the incidence of many diseases increases in the same time. It is more rational to consider that older employees adopt more frequently a behavioral rule, i.e. working when sick. Each other variable should also be analyzed in the light of this reasoning. All things being equal, why would non-permanent employees (coefficient equal to 0.706) have a better health status than others? Why would childless employees (0.313) have less health problems than parents? Even when it comes to job-related variables, a better health status is not able to explain such differences. From this, one can conclude that the measured phenomenon here is presenteeism.

Results are in line with other studies

In section 1.1, some outcomes of other studies were presented. The results of the ZINB model are all consistent with them. Further key information is provided by the "European Working Conditions Survey" (EWCS) conducted in 2010 on about 40,000 respondents in 34 countries. Presenteeism is captured through the following question: "Over the past 12 months did you work when you were sick?".
To that question, 48% of French respondents and 40% of all respondents answered positively. In the French part of the EWCS, men and women seem more or less on the same level playing field towards presenteeism. The ZINB model finds that men are more often subject to presenteeism, as well as older employees. This confirms the specificity of the hospital sector for the gender variable (Demerouti et al. 2009) and, regarding age, the finding is congruent with several other studies (Hansen & Andersen 2008, among them). ZINB estimates show also that family context is an important element in an individual's decision. Employees whose marital status has changed in 2009 or 2010 (wedding) or those with one (young) child are less exposed to presenteeism. Those who are childless, in contrast, are more likely to come to work being sick.
Finally, the effect of job-related variables is clearly in line with three recurring findings. First, a high responsibility level implies a high risk of presenteeism (Caverley et al. 2007). In the ZINB model, a managerial position is related to increased presenteeism. Second, in the hospital sector, employees who are daily involved in direct care for patients are more likely to work being sick (Demerouti et al. 2009). In the hospital we study, nurses (especially the more experienced ones[4]) are more subject to presenteeism. The reverse is true for the service staff, whose relationship with patients is

[4]In the EWCS, presenteeism increases with qualification level: 57.1% of the qualified employees reported for presenteeism, versus 46.2% with lower qualification levels.

Table 3: Presenteeism probability and absence mean over 10 years

$N = 1954$	$[0]$	$]0;7[$	$[7;17[$	$[17;+[$	Total
$[0.25;0.5[$	3.2%	25.5%	25.9%	**45.4%**	100%
$[0.5;0.7[$	6.0%	39.4%	29.3%	25.3%	100%
$[0.7;0.95[$	**21.4%**	**49.2%**	17.2%	12.2%	100%
Total	9.6%	40.4%	25.8%	24.2%	100%

Notes: 45.4% of employees having a presenteeism probabily lower than 0.5 in 2010, also have an absence mean higher than 17 days between 2001 et 2010.

more distant. Third, job insecurity leads to more presenteeism (Hansen & Andersen 2008). In our model, non-permanent staff is in fact more exposed to presenteeism.

4.2 Comparison with panel data

The relevance of our results can also be highlighted by crossing the presenteeism probabilities obtained with the ZINB model in 2010 and the past absence records.

As discussed above, a way to detect presenteeism behavior is to explore an individual's past absence records. If the model is relevant, employees with high presenteeism probabilities, should not have very much absence days in the years before. To verify that, we put together employees who were with the hospital at least for 10 years (2001-2010): that is the case for 1954 employees among 2911 included in the ZINB estimation. In Table 3 the average number of days absent over 10 years is crossed with the presenteeism probabilities observed in 2010.

As we can see, 70.6% of employees who have a presenteeism probability higher than 0.7 were also absent less than 7 days in average over the past ten years[5]. In the same way, 45.4% of employees whose probabilities are lower than 0.5 have in average more than 17 days off[6]. A χ^2 test of independence shows that significant differences exist in the distribution[7]. As expected, the test confirms that the most important contributions to the total χ^2 value concern two subpopulations. Employees with a high presenteeism probability are clearly overrepresented under 7 days absence mean and underrepresented beyond. Conversely, employees with a low presenteeism probability have also very high absence means.

In addition, column percentages show that 96.28% of the 188 employees who have never been absent during 10 years were found with a presenteeism

[5]The median is 6.95.

[6]The third quartile value is 16.58.

[7]$\chi^2 = 210.9$. It is significant at a 0.1% level with 4 degrees of freedom.

probability higher than 0.5, and 56.91% with a probability higher than 0.7. Thus, without any information about previous absences, and by using only cross-sectional data for 2010, our model achieved a result which is consistent with the absence records from past years. This is true in a ten-year analysis, as we have seen, but remains coherent over 20 years (In Appendix A page 20, the same table for the period 1991-2010 is provided).
Based on the presenteeism probability distribution (Figure 2 page 11), it is also possible to study successively subpopulations, that is to say, groups of employees divided into tranches according to their probabilities. Appendix B (page 21) presents several absence indicators for these subpopulations over the past three years, and Appendix C (same page) provides a similar analysis for employees who were with the firm between 2001 and 2010. In these two tables, special consideration is given to the number of years without any absence.
Again, it appears that the estimated presenteeism probabilities are very consistent when comparing with previous absences. Thus, subpopulations with a presenteeism probability lower than 70% are characterized by low absence averages and by a high number of "zero absence" years.
However, these highest categories apart, sub-populations are quite heterogeneous. This makes especially an issue if we aim to divide the total population into two homogeneous groups, those who are "at risk" of presenteeism and those who are not. It is *a priori* difficult to set a level beyond which individuals can be considered "at risk" and below which they are not. Central values don't give a clear indication because the boundary cannot be placed similarly according to the different indicators. Ideally, a new explanatory variable should be introduced in order to further reduce heterogeneity between individuals and to create a clear separation. This is the subject of the last part.

4.3 Previous absences

Finally, we propose an improvement of our specification to allow a better identification of employees who are particularly subject to presenteeism. The model contains the same independent variables as the previous one ("model 1"). The only difference of this new model ("model 2") is that, for each individual, the absences recorded during the two past years are used as explanatory variables. Table 4 describes the variables used in addition.
Past absences are specific variables. First, they reduce heterogeneity related to health. Employees namely differ in their physical condition, resistance to disease, etc. Past absences give further indications on this individual specificities. "As worker absence reflects, to some extent the worker's health state, the probability of an event of absence occurring at time t will potentially depend on whether or not the worker was absent at time $t-1$" (Barmby 2002, p.471). In this line, previous absences also have another property: when

Table 4: Additional variables

Category	Variables	Type	Statistics
Past absences	Maternity leave (2010)	B	(1) 4.3%
	Absence mean (2008 and 2009)	C	($\bar{x}$) 12.5 days
	No absence in 2008 and 2009	B	(1) 39.0%

Notes: B=binary, C=continuous, ($\bar{x}$)=mean, (1)=share in total population

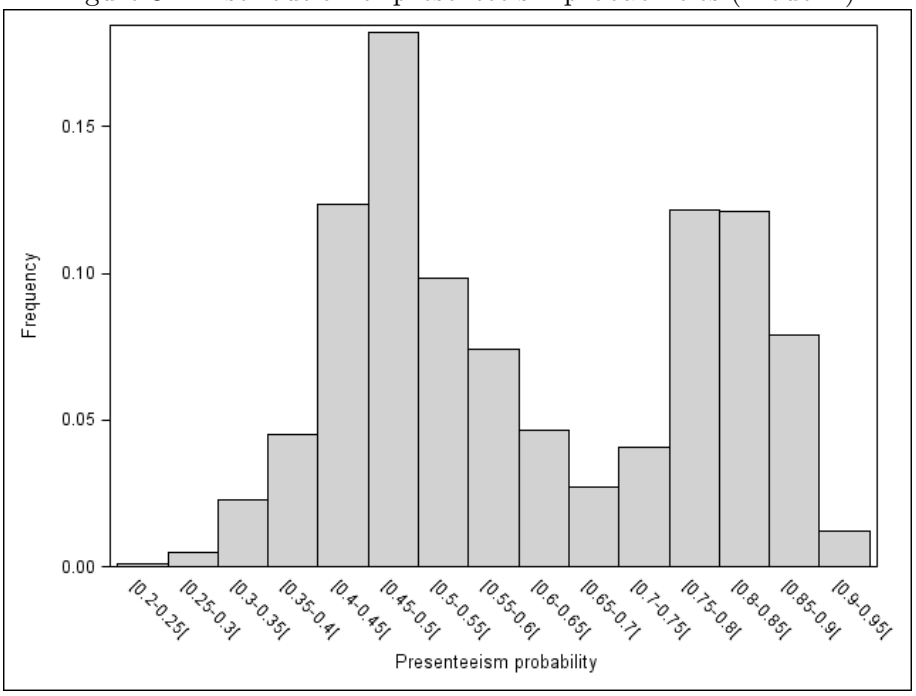
used as explanatory variables, they allow a better discrimination between employees. The probability distribution of model 1 (without previous absences) had a nearly Gaussian shape (Figure 2 page 11): it was hard to set a boundary between two sub-populations. When the past absences are included (model 2), the probability distribution becomes bimodal. Two sub-populations are markedly distinguished and a threshold emerges around 70% (Figure 3).

The reduction of heterogeneity is also illustrated by the decrease in parameter α: it changes from 0.978 in Model 1 to 0.913 in Model 2.

A comparison between model 1 and model 2 (see Appendix D page 22) shows that when past absences are included, other variable's estimates decrease and/or lose some of their significance. That is the case for age: as well as past absences, age is a proxy for health. Therefore, it seems normal that the estimate and the p-value of this variable decrease.

As noted earlier, 40% of respondents to the EWCS reported that they had worked being sick during the 12 past months. In model 2, individuals particularly exposed to presenteeism represent a share of 37.5% in the total population, which seems relevant. When past absences are used, coefficients signs remain unchanged, values are stable, even if slightly pulled down. This may be due to the important weight of the variable "No absence in 2008 and 2009": it is the highest of all estimates. But the findings of Model 1 (Section 4.1) remain entirely valid.

Figure 3: Distribution of presenteeism probabilities (Model 2)



Conclusion

The central purpose of our zero-inflated negative binomial model (ZINB) is to allow a more systematic and less costly analysis (both in time and money) of presenteeism behavior. An empirical application to corporate data, to various industries and even to large-scale surveys could possibly help scholars and practitioners to identify more accurately the factors underlying the decision to attend work while sick.

To do so, a goal of this paper was to look beyond a mere econometric approach. Even if a clarification of the ZINB model's theoretical underpinnings was absolutely necessary, especially to put forward its two essential properties (the capture of heterogeneity and a special treatment for nil values), we emphasized the empirical outcomes and evidences. That was the core of our analysis. Thanks to a long-term longitudinal dataset (1981-2010) we could compare presenteeism probabilities computed by the model (only based on cross-sectional data for year 2010) with the past absence records of the studied population. Such bifocal point of view proved to be fruitful, since the two approaches identified nearly the same at-risk individuals.

Of course, lines of improvement emerged as well. The last section of this paper outlines one of them. In order to separate more strictly two homogeneous sub-populations, the at-risk individuals and others, we introduced

past absences (for the two previous years) in the model as explanatory variables. This attempt is promising, but still requires some adjustments. In further studies, we will also try to apply the ZINB model each year successively to cross-sectional data. Preliminary tests result in stable estimates. This is another reason to believe that the model is robust. This could also provide an opportunity to examine individual cases and to verify with more accuracy if presenteeism is a factor of future absenteeism.

References

Allain E. & Brennac T. (2001), "Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers: le modèle de poisson et ses extensions", *Recherche Transports Sécurité* **72**, pp. 3–18.

Aronsson G. & Gustafsson K. (2005), "Sickness presenteeism: prevalence, attendance pressure factors, and an outline of a model for research", *Journal of Occupational and Environmental Medicine* **47**(9), pp. 958–956.

Aronsson G., Gustafsson K. & Dallner M. (2000), "Sick but yet at work. An empirical study of sickness presenteeism", *Journal of Epidemiology and Community Health* **54**, pp. 502–509.

Barmby T. (2002), "Worker absenteeism: a discrete hazard model with bivariate heterogeneity", *Labour Economics* **9**, pp. 469–476.

Bergström G., Bodin L., Hagberg J., Lindh T., Aronsson G. & Josephson M. (2009), "Does sickness presenteeism have an impact on future general health?", *International Archives of Occupational and Environmental Health* **82**(10), pp. 1179–1190.

Caverley N., Cunningham J. B. & MacGregor J. N. (2007), "Sickness presenteeism, sickness absenteeism, and health following restructuring in a public service organization", *Journal of Management Studies* **44**(2), pp. 304–319.

Demerouti E., Le Blanc P., Bakker A. B., Schaufeli W. B. & Hox J. (2009), "Present but sick: a three-wave study on job demands, presenteeism and burnout", *Career Development International* **14**(1), pp. 50–68.

Frick B. & Malo M. A. (2008), "Labor market institutions and individual absenteeism in the European Union: the relative importance of sickness benefit systems and employment protection legislation", *Industrial Relations* **47**(4), pp. 505–529.

Garay A. M., Hashimoto E. M., Ortega E. M. & Lachos V. H. (2011), "On estimation and influence diagnostics for zero-inflated negative binomial regression models", *Computational Statistics and Data Analysis* **55**, pp. 1304–1318.

Goetzel R. Z., Long S. R., Ozminkowski R. J., Hawkins K., Wang S. & Lynch W. (2004), "Health, absence, disability, and presenteeism cost estimates of certain physical and mental health conditions affecting U.S. employers", *Journal of Occupational and Environmental Medicine* **46**(4), pp. 398–411.

Gosselin E. & Lauzier M. (2011), "Le présentéisme : lorsque la présence n'est pas garante de la performance", *Revue Française de Gestion* **37**(211), pp. 15–27.

Hansen C. D. & Andersen J. D. (2008), "Going ill to work - What personal circumstances, attitudes and work-related factors are associated with sickness presenteeism?", *Social Science and Medicine* **67**, pp. 956–964.

Hilbe J. M. (2011), *Negative Binomial Regression - Second Edition*, Cambridge University Press, New York.

Johns G. (2009), "Presenteeism in the workplace: a review and research agenda", *Journal of Organizational Behavior* **31**(4), pp. 519–542.

Kivimäki M., Head J., Ferrie J. E., Hemingway H., Shipley M. J., Vahtera J. & Marmot M. G. (2005), "Working while ill as a risk factor for serious coronary events: the Whitehall II study", *American Journal of Public Health* **95**, pp. 98–102.

Lambert D. (1992), "Zero-inflated poisson regression, with an application to defects in manufacturing", *Technometrics* **34**(1), pp. 1–14.

Missègue N. (2007), "Les arrêts de travail des séniors en emploi", *Dossiers Solidarité et Santé (Direction de la recherche, des études, de l'évaluation et des statistiques)* **2**, pp. 1–23.

Staub K. E. & Winkelmann R. (2011), "Consistent estimation of zero-inflated count models", Working Paper SOI0908, Socioeconomic Institute - University of Zurich.

Winkelmann R. (2008), *Econometric Analysis of Count Data - Fifth Edition*, Springer-Verlag, Berlin Heidelberg.

Appendix A Presenteeism probability and absence mean over 20 years

$N = 1059$	[0]	]0; 7[	[7; 17[	[17; +[	Total
[0.25; 0.5[	1.4%	28.8%	23.3%	**46.6%**	100%
[0.5; 0.7[	1.6%	41.5%	34.4%	22,5%	100%
[0.7; 0.95[	**12.6%**	**54.0%**	22.6%	10.9%	100%
Total	5.2%	44.8%	29.7%	20.3%	100%

Notes: 46.6% of employees having a presenteeism probabily lower than 0.5 in 2010, also have an absence mean higher than 17 days between 1991 et 2010.

Appendix B Presenteeism probability and previous absences ($N = 2911$)

	0.2-0.55	0.55-0.6	0.6-0.65	0.65-0.7	0.7-0.75	0.75-0.8	0.8-0.95	Total
Absence mean in 2010	9.37	6.88	7.10	5.00	3.88	3.04	2.60	6.63
No absence in 2010	51.5%	60.2%	63.8%	68.7%	74.7%	79.6%	83.4%	63.2%
No absence in 2008	47.3%	53.9%	61.0%	64.5%	62.3%	71.3%	80.7%	57.6%
No absence in 2008 and 2009	25.1%	35.2%	40.0%	46.7%	47.8%	60.5%	69.0%	39.0%
In total population	856	616	420	383	324	167	145	2911
In total population (%)	29.4%	21.2%	14.4%	13.2%	11.1%	5.7%	5.0%	100.0%

Appendix C Presenteeism probability and previous absences ($N = 1954$)

	0.2-0.55	0.55-0.6	0.6-0.65	0.65-0.7	0.7-0.75	0.75-0.8	0.8-0.95	Total
Absence mean over 5 years	19.18	13.81	11.55	10.19	8.01	7.34	5.23	12.44
Absence mean over 10 years	18.60	14.69	11.07	11.05	8.83	7.05	4.70	12.62
Years without absence (5 years)	2.24	2.57	2.86	3.05	3.27	3.61	3.88	2.85
Years without absence (10 years)	4.38	4.94	5.52	5.72	6.28	7.04	7.78	5.50
In total population	440	435	285	294	249	130	121	1954
In total population (%)	22.5%	22.3%	14.6%	15.0%	12.7%	6.7%	6.2%	100.0%

Appendix D ZINB models with and without previous absences

Parameter	Type	Model 1		Model 2	
Intercept		5.160	***	3.822	***
Single	Binary	-0.264	**	-0.153	.
Clerical employee	Binary	-0.316	**	-0.258	*
Service staff	Binary	0.179	.	0.144	
Caregiver	Binary	0.153	*	0.163	*
Age	Continuous	-0.109	**	-0.062	.
Age squared	Continuous	0.001	**	0.001	.
Maternity leave in 2010	Binary			0.627	***
Absence mean (2008 and 2009)	Continuous			0.007	***
Intercept		-0.328		-0.508	*
Man	Binary	0.321	**	0.225	*
Age	Continuous	0.010	*	0.006	
Married in 2009 or 2010	Binary	-0.466	.	-0.583	*
No children	Binary	0.313	**	0.128	
One child	Binary	-0.224	*	-0.168	
Job seniority	Continuous	0.046	***	0.037	***
Job mobility in 2008 or 2009	Binary	-0.260	**	-0.189	.
Service staff	Binary	-0.503	***	-0.507	**
Nurse (higher class)	Binary	0.500	***	0.367	**
Manager	Binary	0.983	***	0.682	**
Non-permanent staff	Binary	0.706	***	0.486	**
No absence in 2008 and 2009	Binary			1.359	***
α		0.978	***	0.913	***
Log-likelihood		-5971		-5826	
AIC		11983		11699	