CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

# Poisson Models with Employer-Employee Unobserved Heterogeneity: An Application to Absence Data

Jean-François Angers
Denise Desjardins
Georges Dionne
Benoit Dostie
François Guertin

June 2007

CIRRELT-2007-15

**Bureaux de Montréal:**
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone: 514 343-7575
Télécopie: 514 343-7121

**Bureaux de Québec:**
Université Laval
Pavillon Palasis-Prince, local 2642
Québec (Québec)
Canada G1K 7P4
Téléphone: 418 656-2073
Télécopie: 418 656-2624

**www.cirrelt.ca**

UNIVERSITÉ LAVAL
UQÀM Université du Québec à Montréal
HEC MONTRÉAL
ÉCOLE POLYTECHNIQUE MONTRÉAL
Université de Montréal

# Poisson Models with Employer-Employee Unobserved Heterogeneity: An Application to Absence Data

**Jean-François Angers[1,2], Denise Desjardins[1], Georges Dionne[1,3], Benoit Dostie[4,*]**
**François Guertin[1,5]**

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada H3C 3J7

[2] Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada H3C 3J7

[3] Canada Research Chair in Risk Management, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine Montréal, Canada H3T 2A7

[4] Institute of Applied Economics, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine Montréal, Canada H3T 2A7, and Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi (CIRPÉE)

[5] Réseau québécois de calcul de haute performance (RQCHP), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada, H3C 3J7

**Abstract.** We propose a parametric model based on the Poisson distribution that permits to take into account both unobserved worker and workplace heterogeneity as long as both effects are nested. By assuming that workplace and worker unobserved heterogeneity components follow a gamma and a Dirichlet distribution respectively, we obtain a closed form for the unconditional density function. We estimate the model to obtain the determinants of absenteeism using linked employer-employee Canadian data from the Workplace and Employee Survey (2003). Coefficient estimates are interpreted in the framework of the typical labor-leisure model. We show that omitting unobserved heterogeneity on either side of the employment relationship leads to notable biases in the estimated coefficients. In particular, the impact of wages on absences is underestimated in simpler models.

**Keywords**. Absenteeism, linked employer-employee data, employer-employee unobserved heterogeneity, count data models, Dirichlet distribution.

**Acknowledgements.** We would like to thank Lene Kromann for excellent research assistance and Robert Clark for comments. G. Dionne and B. Dostie acknowledge funding from the "Fonds québécois de la recherche sur la société et la culture" (FQRSC).

Results and views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect those of CIRRELT.

Les résultats et opinions contenus dans cette publication ne reflètent pas nécessairement la position du CIRRELT et n'engagent pas sa responsabilité.

[*] Corresponding author: benoit.dostie@hec.ca

Dépôt légal – Bibliothèque nationale du Québec,
Bibliothèque nationale du Canada, 2007

# 1 Introduction

In this paper, we test a serie of parametric models for count data that allow for overdispersion along more than one dimension. We test our models on absenteeism linked employer-employee data from the most recent version of Statistics Canada Workplace and Employee Survey 2003. While linked employer-employee data sets are becoming more and more available (see Abowd and Kramarz (1999)) and many dependent variables that are of interest to economists (e.g. days of vacations, days of absenteeism, days of training, number of job offers) take the form of a count, we know of no count data parametric application that takes into account overdispersion at both the employee and employer levels.

We propose a new parametric model which can account for observable and unobservable characteristics of both firms and employees. The extension adds a random employee effect to the firm random effect of the negative binomial model. The potential gain of using parametric models is an increase in efficiency when the specific parametric assumptions are not rejected. One caveat is that, in order to estimate the model providing the best fit, the employer and employee effects need to be nested. Linked employer-employee data fitting this description would be the Workplace and Employee Survey from Statistics Canada or the Workplace Employee Relations Survey from Britain.

We show this model peforms better than popular alternatives to the Poisson model like the negative binomial model or the negative binomial model with random effects as proposed by Hausman, Hall, and Griliches (1984). Our other main finding is that the impact of the wage rate on absenteeism is greatly underestimated when using models that do not correctly take into account the characteristics of the data. We also show that the model performs very well by ruling out any significant correlation between the residuals and $\hat{\beta}X_{ij}$ or $\exp\left(\hat{\beta}X_{ij}\right)$.

In the next section, we jump directly to the empirical specification and discuss the data in Section 3. In Section 4, we provide a detailed assessment of the goodness of fit of each model and thoroughly discuss the coefficient estimates in the context of the absenteeism literature. A brief conclusion follows.

## 2 Empirical specification

We use an econometric model to evaluate the absenteeism probabilities for employees within firms. Our dependent variable is a count of days of absenteeism by a particular individual in a given year.[1] Absenteeism can be explained by both individual and firm characteristics that are observable and unobservable.

We eschew the simple Poisson model that does not perform well in our setting. We thus use, as our base model for the firm effect, the traditional negative binomial model (Gouriéroux, Monfort, and Trognon (1984)) and test its performance against two other more sophisticated alternatives. The first alternative is the random effect negative binomial model as developed by Hausman, Hall, and Griliches (1984). The second alternative (that we call 'Full model') extends Angers, Desjardins, Dionne, and Guertin (2006) to capture both worker and firm unobserved heterogeneity. We describe each model below.

### 2.1 The negative binomial (NB) model

Let $y_{ij}$ be the number of days of absenteeism for employee $i$ in firm $j$. The basic Poisson model is

$$P(y_{ij} \mid \lambda_{ij}) = \frac{e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}}{\Gamma(y_{ij}+1)}, \tag{1}$$

where

$$E\left[y_{ij}\right] = Var\left(y_{ij}\right) = \lambda_{ij}, \tag{2}$$

[1] For different applications of count data models in different settings, see (Cameron and Trivedi (1998) and Winkelmann (1997)).

with $\lambda_{ij} > 0$. $\lambda_{ij}$ controls for the observed employee-employer heterogeneity. We first consider unobserved firm heterogeneity.

One can introduce unobserved firm heterogeneity in the Poisson model in a multiplicative form through $\lambda_{ij}$. If we write

$$\lambda_{ij} = \psi_j \gamma_{ij} \tag{3}$$

with $\gamma_{ij} = \exp(\beta X_{ij})$ and assume that $\psi_j$ follows a gamma distribution with parameters $\left(\delta^{-1}, \delta^{-1}\right)$, we obtain the standard negative binomial model with parameters $\left(\delta^{-1}, \delta^{-1}/\gamma_{ij}\right)$. $\psi_j$ can be interpreted as a workplace effect and could represent safety-risk specific to the workplace that affects the absence decisions of all its workers similarly. Note in this model that $E[y_{ij}|X_{ij}] = \gamma_{ij}$ and $Var\left(y_{ij}|X_{ij}\right) = \gamma_{ij} + \delta\gamma_{ij}^2$.

The $\delta$ parameter captures unobserved firm heterogeneity by introducing overdispersion. As is the case with traditional panel data analysis, this model might be inadequate because it does not capture the potential persistence of the firm effect when there are many employees in a given workplace. One way to solve this problem is by adding an additional source of randomness.

## 2.2 NB model with firm random effects

Now, assume $y_{ij}$ follows a negative binomial distribution with parameters $\left(\gamma_{ij}, \delta_j\right)$ where

$$\begin{aligned} E[y_{ij}|\delta_j, X_{ij}] &= \gamma_{ij}\delta_j^{-1} && (4) \\ Var\left(y_{ij}|\delta_j, X_{ij}\right) &= \gamma_{ij}\frac{1+\delta_j}{\delta_j^2}. && (5) \end{aligned}$$

The typical random effects (RE) negative binomial model follows from the assumption that $\frac{\delta_j}{1+\delta_j}$ is distributed as a beta distribution with parameters $(a, b)$.

The unconditional expressions for the expectation and variance become

$$E[y_{ij}|X_{ij}] = \frac{b}{a-1}\gamma_{ij}, \tag{6}$$

$$Var(y_{ij}|X_{ij}) = \gamma_{ij}\frac{b(a+b-1)}{(a-1)(a-2)}\left(1+\frac{\gamma_{ij}}{a-1}\right). \tag{7}$$

This setup adapts the model of Hausman, Hall, and Griliches (1984), proposed for panel data, to workplace heterogeneity. Here, instead of having repeated observations on individuals over time, we have repeated observations of workplace characteristics over all sampled employees. Many of these characteristics are observable, but other are not.

However, it is not clear that this reinterpretation of Hausman, Hall, and Griliches (1984) is sufficient for our purpose because it does not take into account employee unobserved heterogeneity. In applications to panel data, repetitions are for the same subject over time. These repeated observations are almost interchangeable over time, and time varies in the same way for all observations. In our setting where we observe repeated workplace characteristics over employees, this means that the NB model with random effects implicitly assumes that workplaces are interchangeable and come from the same population. But in our empirical application, the relation between the firm and the worker may be more complex in the sense that these actions or events may also be affected by unobserved worker heterogeneity. Therefore, we propose another generalization of the NB model that integrates both workplace and worker unobserved heterogeneity.

## 2.3 Model with workplace and worker effects

We now use the following parameterization for $\lambda_{ij}$

$$\lambda_{ij} = \psi_j\theta_{ij}\gamma_{ij}, \tag{8}$$

with

$$\gamma_{ij} = \exp(\beta X_{ij}). \tag{9}$$

Parameter $\psi_j$ is still the random effect associated with firm $j$; that is, unobservable heterogeneity related to firm $j$ and affecting the absenteeism decisions of its employees. $\theta_{ij}$ is the random effect of employee $i$ in firm $j$ and represents unobserved heterogeneity at the individual level. Here we assume that the random effects are nested (we do not observe employees moving from firm to firm) in accordance with the structure of the data set we are using for the estimation.[2]

Within each firm, we assume that

$$\sum_{i=1}^{E_j} \theta_{ij} = 1, \tag{10}$$

where $E_j$ is the total number of employees sampled from firm $j$. We assume (to obtain a parametric model) that $\theta_{ij}$ follows a Dirichlet parametric distribution $\left(\nu_1, \nu_2, ..., \nu_{E_j}\right)$ and that $\psi_j$ follows a gamma parametric distribution with parameters $(E_j \tau_j, \tau_j)$. This parametrization makes it possible to obtain an average workplace effect that increases with the number of workers in the firm.

With these assumptions, we have that

$$\begin{aligned} E\left[y_{ij}|X_{it}\right] &= \gamma_{ij} E_j \frac{\nu_i}{\sum_{i=1}^{E_j} \nu_i} && (11) \\ Var(y_{ij}|X_{it}) &= \gamma_{ij} E_j \frac{\nu_i}{\sum_{i=1}^{E_j} \nu_i} \left(1 + \gamma_{ij}\left(E_j + \tau^{-1}\right) \frac{(1+\nu_i)}{\left(1 + \sum_{i=1}^{E_j} \nu_i\right)}\right) && (12) \end{aligned}$$

For a firm $j$ with $E_j$ employees, the joint distribution of weeks of absenteeism

[2] Note that we subscript $\theta$ with both $i$ and $j$ to emphasize that the worker effect is nested withing the workplace effect.

is given by

$$\begin{aligned}\Pr\left(y_{1j},...,y_{E_jj}\right) &= \\ &= \int_{\sum_{i=1}^{E_j}\theta_{ij}=1}...\int \Pr\left(y_{1j},...,y_{E_jj}\mid\theta_{1j},...,\theta_{E_jj},\gamma_{1j},...,\gamma_{E_jj}\right)\times \\ &\times f\left(\theta_{1j},...,\theta_{E_jj}\right)d\theta_{1j}...d\theta_{E_j-1j}, \qquad (13)\end{aligned}$$

where, from equation (10), we have

$$\theta_{E_jj} = 1-\sum_{i=1}^{E_j-1}\theta_{ij}. \qquad (14)$$

The conditional density is written as

$$\begin{aligned}\Pr\left(y_{1j},...,y_{E_jj}\mid\theta_{1j},...,\theta_{E_jj},\gamma_{1j},...,\gamma_{E_jj}\right) &= \\ &= \int_0^\infty \Pr\left(y_{1j},...,y_{E_jj}\mid\psi_j,\theta_{1j},...,\theta_{E_jj},\gamma_{1j},...,\gamma_{E_jj}\right)\times \\ &\times f\left(\psi_j\right)d\psi_j. \qquad (15)\end{aligned}$$

If we integrate out the firm effect $\psi_j$, we obtain a Dirichlet compound multinomial (Johnson and Kotz (1969)) distribution for the number of weeks of absenteeism whose joint conditional distribution is equal to

$$\begin{aligned}\Pr\left(y_{1j},...,y_{E_jj}\mid\theta_{1j},...,\theta_{E_jj},\gamma_{1j},...,\gamma_{E_jj}\right) &= \\ &= \left[\prod_{i=1}^{E_j}\frac{\left(\gamma_{ij}\right)^{y_{ij}}\left(\theta_{ij}\right)^{y_{ij}}}{\Gamma(y_{ij}+1)}\right]\left[\frac{\tau_j^{E_j\tau_j}}{\Gamma(E_j\tau_j)}\right]\frac{\Gamma\left(E_j\tau_j+\sum_{i=1}^{E_j}y_{ij}\right)}{\left(\tau_j+\sum_{i=1}^{E_j}\theta_{ij}\gamma_{ij}\right)^{E_j\tau_j+\sum_{i=1}^{E_j}y_{ij}}}. \qquad (16)\end{aligned}$$

If we substitute for this multivariate density in (13) and if we assume that the worker effects follow a Dirichlet distribution with parameters $\left(\nu_1,...,\nu_{E_j}\right)$, we obtain

$$\begin{aligned}\Pr\left(y_{1j},...,y_{E_jj}\right) &= \\ &= \left[\prod_{i=1}^{E_j}\frac{\left(\gamma_{ij}\right)^{y_{ij}}}{\Gamma(y_{ij}+1)}\right]\left[\frac{\tau_j^{E_j\tau_j}\Gamma\left(E_j\tau_j+\sum_{i=1}^{E_j}y_{ij}\right)\Gamma\left(\sum_{i=1}^{E_j}\nu_i\right)}{\Gamma(E_j\tau_j)\prod_{i=1}^{E_j}\Gamma(\nu_i)}\right]\times \\ &\times \quad \int_{\sum_{i=1}^{E_j}\theta_{ij}=1}\cdots\int\frac{\prod_{i=1}^{E_j}(\theta_{ij})^{\nu_i+y_{ij}-1}}{\left(\tau_j+\sum_{i=1}^{E_j}\theta_{ij}\gamma_{ij}\right)^{E_j\tau_j+\sum_{i=1}^{E_j}y_{ij}}}d\theta_{1j}...d\theta_{E_j-1j}. \end{aligned} \tag{17}$$

In order to solve this equation, one must find the solution to the multidimensional integral in (17). Following Angers, Desjardins, Dionne, and Guertin (2006), we approximate it through the use a hypergeometric function. We could also use Monte-Carlo approximation.

## 2.4 Estimation

To estimate the statistical model with the hypergeometric approximation, we make the simplifying assumption that it is possible to separate the workers into two groups (for example) and define $G_1 = 1, ..., g$ as all the workers in the first group with

$$\gamma_{g_1j} = \frac{\sum_{i=1}^{g}\gamma_{ij}}{g}, \tag{18}$$

and $G_2 = g+1, ..., E_j$ as all the workers in the second group with

$$\gamma_{g_2j} = \frac{\sum_{i=g+1}^{E_j}\gamma_{ij}}{E_j - g}. \tag{19}$$

The integral of equation (17) thus becomes

$$\int_{\sum_{i=1}^{E_j}\theta_{ij}=1}...\int \frac{\left[\prod_{i=1}^{g}(\theta_{ij})^{c_i-1}\prod_{i=g+1}^{E_j}(\theta_{ij})^{c_i-1}\right]}{\left(\tau_j+\gamma_{g_1j}\sum_{i=1}^{g}\theta_{ij}+\gamma_{g_2j}\sum_{i=g+1}^{E_j}\theta_{ij}\right)^d}\ d\theta_{1j}...d\theta_{E_j-1j} \tag{20}$$

with

$$\begin{aligned} c_i &= \nu_i + y_{ij} & (21)\\ d &= E_j\tau_j + \sum_{i=1}^{E_j} y_{ij}. & (22) \end{aligned}$$

If we define

$$\begin{aligned} u_i &= \frac{\theta_{ij}}{\sum_{i=1}^{g}\theta_{ij}},\ i=1,...,g-1 & (23)\\ v &= \sum_{i=1}^{g}\theta_{ij} & (24)\\ w_i &= \frac{\theta_{ij}}{1-\sum_{i=1}^{g}\theta_{ij}},\ i=g+1,...,E_j, & (25) \end{aligned}$$

then we can rewrite equation (20) and substitute it in equation (17) to obtain an approximation for the distribution of the number of days of absence in establishment $j$ (under the restriction $\gamma_{g_2j} \geq \frac{\gamma_{g_1j}-\tau}{2}$):

$$
\begin{aligned}
\Pr\left(y_{1j},...,y_{E_j j}|\gamma_{1j},...,\gamma_{E_j j}\right) \quad &\approx \\
\approx \prod_{i=1}^{E_i}\left(\tfrac{\left(\gamma_{ij}\right)^{y_{ij}}\Gamma(y_{ij}+\nu_i)}{\Gamma(y_{fi}+1)\Gamma(\nu_i)}\right)\left(\tau_j\right)^{E_j\tau_j} \times \\
\times \frac{\Gamma\left(E_j\tau_j+\sum_{i=1}^{E_j} y_{ij}\right)}{\Gamma(E_j\tau_j)} \times \left(\tfrac{1}{\tau_j+\gamma_{g_2 j}}\right)^{E_j\tau_j+\sum_{i=1}^{E_j} y_{ij}} \times \\
\times \frac{\Gamma\left(\sum_{i=1}^{E_j}\nu_i\right)}{\Gamma\left(\sum_{i=1}^{E_j}(\nu_i+y_{ij})\right)} \times \\
\times {}_2F_1\left(\begin{array}{c} \sum_{i=1}^{g}(\nu_i+y_{ij}), E_j\tau_j+ \\ +\sum_{i=1}^{E_j} y_{ij}, \sum_{i=1}^{E_j}\left(\nu_i+y_{ij}\right), \left(\tfrac{\gamma_{g_2 j}-\gamma_{g_1 j}}{\tau_j+\gamma_{g_2 j}}\right) \end{array}\right), \qquad (26)
\end{aligned}
$$

where ${}_2F_1$ is the hypergeometric function. It should be noted that this procedure for estimating the integral can be generalized to several homogeneous groups, but it is not obvious that the precision gained would be greater. Indeed, Angers, Desjardins, Dionne, and Guertin (2006) show that the two-group approximation yields results very similar to a Monte Carlo approximation of the multivariate integral of equation (20).

# 3 Data

We use data from the Workplace and Employee Survey (WES) 2003 conducted by Statistics Canada. The survey is both longitudinal and linked in that it documents the characteristics of the workers and of the workplaces over time. The target population for the “workplace” component of the survey is defined as the collection of all Canadian establishments who had paid employees in March of the year of the survey, excluding Yukon, the Northwest territories and Nunavut.[3] For the “employee” component, the target population is the collection of all employees working, or on paid leave, in the workplace target population.

The sample for the workplaces comes from the “Business registry” of Statistics Canada which contains information on every business operating in Canada. Employees are then sampled from an employees list provided by the selected workplaces. For every workplace, a maximum of twenty-four employees are selected, and for establishments with less than four employees, all employees are sampled. In the case of total non-response, respondents are withdrawn entirely from the survey and sampling weights are recalculated in order to preserve representativeness of the sample.

WES selects new employees in odd years. For workplaces, the initial 1999 sample is followed over time and is supplemented at two-year intervals with a sample of births selected from units added to the Business Register since the last survey occasion. In order to control for the design effect in our estimations, we use the final sampling weights for employees as recommended by Statistics Canada.

In 1999, workplace data were collected in person; subsequent workplace surveys were conducted by means of computer assisted telephone interviews. For the employee component, telephone interviews were conducted with individuals who had agreed to participate in the survey by filling out and posting an employee participation form.

Individuals who did no work throughout the year are included but we control for their limited exposure to the risk of being absent in our regression framework. Finally, we drop workers who were absent for more than fifty days of work in

[3] Establishments operating in fisheries, agriculture and cattle farming are also excluded.

the past year.[4]

The rich structure of the data set allows us to control for a variety of factors determining absenteeism decisions. From the worker questionnaire, we are able to extract detailed demographic characteristics including measures of health, human capital, and income from other sources. Moreover, we use detailed explanatory variables on the employment contract including wage and contracted hours. From the workplace questionnaire, we are able to construct firm size indicators and build measures of layoff and vacancy rates. Finally, our regressions include industry (13) and occupation (6) dummies.

Summary statistics on all explanatory variables are presented in Table 1 for the dependent variable, Table 2 for the employees and Table 3 for the employers. Note that the number of weeks absent in Table 1 refers to a five day workweek. Thus zero means the worker was absent less than five days during a year.

# 4 Results

We first discuss which model should be preferred in terms of goodness of fit and then turn to coefficient estimates and their relation to absence decisions.

## 4.1 Goodness of fit

Table 6 provides several traditional measures for goodness of fit that allow comparisons of the three models. First, note that the Full model yields the highest value for the likelihood function. Likelihood ratio tests also reject simpler models in favor of the Full specification. It should be noted, however, that these models are not nested.

Since it is always possible to increase the goodness of fit of the model by increasing the number of parameters, therefore, we also computed the BIC

[4] Results are robust to other cutoff points for eliminating outliers.

($= -2\ln L + k\ln(N)$) and AIC ($= -2\ln L + 2k$) where $\ln L$ is the value of the log-likelihood function, $k$ the number of parameters, and $N$ the number of observations. Both the BIC and AIC favor the Full model and this conclusion holds even for non-nested models.

Hausman, Hall, and Griliches (1984) argue that a good econometric random effect model for count data should also yield estimates so that computed residuals are orthogonal to $\hat{\beta}X_{ij}$ and $\exp\left(\hat{\beta}X_{ij}\right)$. We test for the presence of such correlations with two types of residuals. First, we define the raw residuals as

$$e_{ij} = y_{ij} - E[y_{ij}|X_{ij}]. \tag{27}$$

Next, we define the Pearson residuals as

$$e_{ij}^{P} = \frac{y_{ij} - E[y_{ij}|X_{ij}]}{\sqrt{Var\left(y_{ij}|X_{ij}\right)}}. \tag{28}$$

Using the above definitions, we have the following formulas for the residuals of each model. For the negative binomial model, we have

$$e_{ij} = y_{ij} - \exp(X_{ij}\beta), \tag{29}$$

$$e_{ij}^{P} = \frac{y_{ij} - \exp(X_{ij}\beta)}{\sqrt{\exp(X_{ij}\beta) + \delta\left(\exp(X_{ij}\beta)\right)^{2}}}. \tag{30}$$

For the random effect negative binomial model, we can compute

$$e_{ij} = y_{ij} - \left(\frac{b}{a-1}\right)\exp(X_{ij}\beta), \tag{31}$$

$$e_{ij}^{P} = \frac{y_{ij} - \left(\frac{b}{a-1}\right)\exp(X_{ij}\beta)}{\sqrt{\exp(X_{ij}\beta)\frac{b(a+b-1)}{(a-1)(a-2)}\left(1 + \frac{\exp(X_{ij}\beta)}{a-1}\right)}}. \tag{32}$$

Finally, for the Full model, one can verify that

$$e_{ij} = y_{ij} - \left(\frac{E_j \nu_i}{\sum_{i=1}^{E_j} \nu_i}\right) \exp(X_{ij}\beta), \tag{33}$$

$$e_{ij}^P = \frac{y_{ij} - \left(\frac{E_i \nu_i}{\sum_{i=1}^{E_j} \nu_i}\right) \exp(X_{ij}\beta)}{\sqrt{\exp(X_{ij}\beta)\frac{E_j \nu_i}{\sum_{i=1}^{E_j} \nu_i}\left(1 + \exp(X_{ij}\beta)(E_j + \tau^{-1})\frac{(1+\nu_i)}{\left(1+\sum_{i=1}^{E_j} \nu_i\right)}\right)}}. \tag{34}$$

For both types of residuals, we test whether they are significantly correlated with both $\widehat{\gamma}_{ij} = \exp(\hat{\beta} X_{ij})$ and $\hat{\beta} X_{ij}$.

We note that, in this respect, both the negative binomial and the Full models perform well, and we can rule out a statistically significant correlation between either the raw and Pearson residuals and $\exp(\hat{\beta} X_{ij})$ or $\hat{\beta} X_{ij}$. However, our adaptation of the negative binomial model with random effects as proposed by Hausman, Hall, and Griliches (1984) performs especially badly, even though it provides a better fit for the data than the simple binomial negative model. One possible explanation of this negative result may be linked to the implicit assumption that the random workplace effect is not nested with the worker effect and is assumed to be interchangeable for all workers whatever the characteristics of the firm, such as size. This implicit assumption does not hold in the simple negative binomial model where the workplace effect is taken into account by a single common parameter.

## 4.2 Determinants of absence

It is very interesting to look at absenteeism decisions for several reasons. The first main reason is that despite its rising frequency and associated cost (Akyeampong (2005)), there are relatively few studies on the determinants of absenteeism. The second is that it is pretty straightforward to obtain theoretical predictions with respect to many variables of interest using the typical labor supply framework. In fact, it is relatively easy to show that, following the maximization of a utility function with respect to budget and time constraints, absences should decrease with wages and increase with both non-labor income and contracted hours. Absences should also decrease with their cost (other than wages) (Hausman (1980); Blomquist (1983); Dionne and Dostie (2007)).

These predictions have been tested with mixed success by a variety of studies. In one of the first study on the topic, Allen (1981) uses the 1972-73 Quality of Employment Survey and obtains results that validate the theoretical predictions except for a positive non-labor income effect. Although the estimated wage effect is negative, it is too low to make it realistic for workplaces to use wage increases to diminish absences.

In part because of this finding, there have been relatively few studies that focused on the core predictions of the labor-leisure model and many studies instead tried to find more important determinants of absences. For example Wilson and Peel (1991) have data on a sample of 52 firms in the engineering and metal industry in the United Kingdom and focus mainly on the impact of profit-sharing and other forms of employee participation. Drago and Wooden (1992) work on a sample of 15 firms from the U.S., Canada, and New-Zealand but focus on workgroup cohesion. Vistnes (1997) uses the 1987 National Medical Expenditure Survey and does look at the traditionnal economic determinants but finds that health is a better predictor of absence than the typical economic predictors. She also finds that, in the case of women, having kids of pre-school age also leads to more absences.[5]

More recent studies have tried to get better estimates to test the real importance of economic determinants on absences. This is the case with Barmby

[5] Barmby, Orme, and Treble (1991) use data on four factories of an unidentified firm and are missing information to test all the above predictions. Delgado and Kiesner (1997) look at London bus operators and are also missing a lot of information about the labor contract.

(2002) who has data on only one UK manufacturing firm but detailed information on sick pay entitlement. He finds evidence of a significant impact of the cost of absenteeism as measured by the difference between daily earnings and and sick pay entitlement. Johansson and Palme (2002) use data from the 1991 Swedish Level of Living Survey (SLLS) and program evaluation methods to obtain similar results i.e. that higher costs of absenteeism lead to more absences. Henrekson and Persson (2004) use aggregate data from the National Social Insurance Board of Sweden and find that absences are inversely related to the generosity of sick leave.[6]

Turning to our results, we find that all our coefficient estimates match the predictions from the theoretical model of Dionne and Dostie (2007) for example. We find a negative impact of wages and positive impacts for both contracted hours and non-labor income although the last two effects are pretty small. Since we use log wages as an explanatory variable, the coefficient can be interpreted directely as the elasticity of days of absence with respect to wages. The estimated coefficient ($-0.11$) thus implies that workplaces have considerable leverage in diminishing absences through wage increases.

Unfortunately, our measures for the cost of absence are not statistically significant. In the literature, the cost of absenteeism is usually related to an increased likelihood of being fired or being passed up for promotion. Therefore, we settle on an indicator of the layoff rate (defined as the number of workers laid off in the past year divided by average employment) and the vacancy rate (defined as the number of positions available in the firm divided by average employment). These variables are interpreted as indicating the willingness of the workplace to use layoffs as a way to discipline employees. For example, if the vacancy rate is high, the employer might be reluctant to fire employees even

[6] Another recent study is Kauermann and Ortlieb (2004). They have absenteeism data from one German firm and focus more on whether absenteeism increases before a downsizing occurs.

if they misbehave.

It should be noted that using a model that does not take into account unobserved heterogeneity at both the worker and workplace levels leads to biases in the estimates. For example, this is especially the case for wages where the estimated impact is positive in both the simple binomial negative model and binomial negative model with random effects. We also observe that some parameters for explanatory varibles at the workplace level become significant in the NB model with random effects such as layoff rate and the indicator variable for workplaces with $100-499$ employees. This may be explained by the observed correlation between the computed residuals and $\exp(\hat{\beta}X_{ij})$ or $\hat{\beta}X_{ij}$.

Among other results of interest, we find that higher absences are related with being a women, being in worse health (measured as the presence of activity limitations), and having more numerous pre-school aged kids (for both gender). Also noteworthy is that we find, in the Full model, no impact of race, workplace size (as measured by the number of employees) and no consistent relation with education.

# 5 Conclusion

We estimate two extensions to the standard negative binomial model for linked employer-employee count data. We find that a parametric model that incorporates both worker and workplace unobserved heterogeneity (the Full model) provides a better fit for the data than both the simple binomial negative model and a binomial negative model modified to incorporate random effects along the lines of Hausman, Hall, and Griliches (1984). This improvement is not obtained at the cost of a higher correlation between the residuals and predicted values for the dependent variable.

Turning to the economic results, we find that both characteristics of the labor contract and demographic characteristics of the individual are important predictors of absence. This is particularly the case for the wage rate and demographic characteristics such as gender (women), the number of pre-school aged kids and health.[7]

As an extension, it would be useful to modify the model in order to be able to estimate it on panel data. The use of panel data would allow separate identification of the worker and workplace effects that is not dependant on the parametric assumptions.

# References

Abowd, J. M. and F. Kramarz (1999). The analysis of labor markets using matched employer-employee data. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, vol 3B*, Chapter 40, pp. 2629–2710. Elsevier Science North Holland.

Akyeampong, E. B. (2005). Fact sheet on work absences. *Perspectives on Labour and Income 6*(4), 21–30.

Allen, S. G. (1981). An empirical model of work attendance. *Review of Economics and Statistics 63*(1), 77–87.

Angers, J.-F., D. Desjardins, G. Dionne, and F. Guertin (2006). Vehicle and fleet random effects in a model of insurance rating for fleets of vehicles. *Astin Bulletin 36*(1), 25–77.

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

[7] In the case of the wage rate, we should also mention that its impact is not statically significant when parameter estimation is done using a model with both worker and workplace unobserved heterogeneity normally distributed and thus with no closed from solution. This confirms the superiority of our parametric closed-form model. This result is not presented here but is available

Barmby, T., C. Orme, and J. G. Treble (1991). Worker absenteeism: An analysis using microdata. *Economic Journal 101*(405), 214–229.

Blomquist, N. S. (1983). The effect of income taxation on the labor supply of married men in sweden. *Journal of Public Economics 22*(2), 169–197.

Cameron, A. and P. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

Delgado, M. A. and T. J. Kiesner (1997). Count data models with variance of unknown form: An application to a hedonic model of worker absenteeism. *Review of Economics and Statistics 79*(1), 41–49.

Dionne, G. and B. Dostie (2007). New evidence on the determinants of absenteeism using linked employer-employee data. *Industrial and Labor Relations Review, forthcoming*.

Drago, R. and M. Wooden (1992). The determinants of labor absence: Economic factors and workgroup norms across countries. *Industrial and Labor Relations Review 45*(4), 764–778.

Gouriéroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood functions: Applications to Poisson models. *Econometrica 52*(3), 701–720.

Hausman, J., B. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-r&d relationship. *Econometrica 52*, 909–938.

Hausman, J. A. (1980). The effect of wages, taxes, and fixed costs on women's labor force participation. *Journal of Public Economics 14*(2), 161–194.

Henrekson, M. and M. Persson (2004). The effects on sick leave of changes in the sickness insurance system. *Journal of Labor Economics 22*, 87–113.

Johansson, P. and M. Palme (2002). Assessing the effect of public policy on worker absenteeism. *Journal of Human Resources 37*, 381–409.

Johnson, N. and K. Kotz (1969). *Discrete Distribution*. Houghton Mifflin, Boston.

Kauermann, G. and R. Ortlieb (2004). Temporal pattern in number of staff on sick leave: the effect of downsizing. *Journal of the Royal Statistical Society Series C - Applied Statistics 53*, 355–367.

Vistnes, J. P. (1997). Gender differences in days lost from work due to illness. *Industrial and Labor Relations Review 50*(2), 304–323.

Wilson, N. and M. J. Peel (1991). The impact on absenteeism and quits of profit-sharing and other forms of employee participation. *Industrial and Labor Relations Review 44*(3), 454–468.

Winkelmann, R. (1997). *Econometric Analysis of Count Data*. Springer, Berlin.

Table 1: Weighted summary statistics on absenteeism in Canada (2003)

| | Mean | Std.Dev. |
|---|---:|---:|
| Days absent | 3.691 | 6.665 |
| | | |
| Weeks absent | Freq. | % |
| 0 | 9,717,342 | 90.16 |
| 1 | 669,090 | 6.21 |
| 2 | 185,702 | 1.72 |
| 3 | 25,927 | 0.24 |
| 4 | 31,343 | 0.29 |
| 5 | 7,437 | 0.07 |
| 6 | 22,676 | 0.21 |
| 7 | 14,159 | 0.13 |
| 8 | 15,538 | 0.14 |
| 9 | 10,675 | 0.10 |
| 10 | 3,986 | 0.04 |
| | (...) | (...) |
| Total | 10,777,543 | 100.00 |

Table 2: Summary statistics - Employees

| | 2003 | |
|---|---|---|
| | Mean | Std.Dev. |
| **Demographic characteristics** | | |
| Women | 0.506 | 0.500 |
| Black | 0.012 | 0.109 |
| Other race | 0.276 | 0.447 |
| Married | 0.568 | 0.495 |
| Number of pre-school aged kids | 0.228 | 0.554 |
| **Health** | | |
| No activity limitation | 0.958 | 0.200 |
| **Human Capital** | | |
| High school degree | 0.178 | 0.382 |
| Less than bachelor degree | 0.574 | 0.450 |
| Bachelor degree | 0.130 | 0.337 |
| Some higher education | 0.054 | 0.092 |
| Seniority | 9.106 | 8.403 |
| Experience | 16.470 | 10.660 |
| **Income** | | |
| Income from other sources | 2120.118 | 11226.875 |
| **Wage Contract** | | |
| Natural logarithm of hourly wage | 2.819 | 0.504 |
| Contracted hours | 37.077 | 9.120 |
| **Work arrangement** | | |
| Covered by a collective bargaining agreement | 0.321 | 0.469 |
| **Technology** | | |
| Use computer | 0.627 | 0.484 |
| Use computer assisted design | 0.130 | 0.337 |
| Use other technology | 0.249 | 0.432 |
| Number of employees | 18671 | |

Table 3: Summary statistics - Workplace

| | 2003 | |
|---|---|---|
| | Mean | Std.Dev. |
| **Cost of absenteeism** $E(w^a)$ | | |
| Vacancy rate | 0.027 | 0.061 |
| Layoff rate | 0.099 | 0.376 |
| **Size** | | |
| 10-19 employees | 0.451 | 0.478 |
| 20-99 employees | 0.472 | 0.499 |
| 100-499 employees | 0.067 | 0.250 |
| 500 employees and more | 0.010 | 0.099 |
| Number of workplaces | 3767 | |

Table 4: Count model with unobserved heterogeneity on days of absence

| | NB | RE NB | Full |
|---|---|---|---|
| Log-wage | 0.015 | 0.005 | -0.111 |
| | (0.035) | (0.028) | (0.019) |
| Contracted hours | 0.002 | 0.007 | 0.002 |
| | (0.001) | (0.001) | (0.000) |
| Income from other sources | 0.000 | 0.000 | 0.000 |
| | (0.001) | (0.001) | (0.000) |
| Vacancy rate | -0.335 | 0.113 | -0.336 |
| | (0.377) | (0.307) | (0.500) |
| Layoff rate | -0.004 | -0.058 | -0.003 |
| | (0.027) | (0.025) | (0.029) |
| Constant | 0.867 | -1.371 | 1.272 |
| | (0.144) | (0.122) | (0.102) |

Standard errors in parantheses
Include occupation and industry dummies

Table 4: Cont'd

| | NB | RE NB | Full |
|---|---|---|---|
| Women | 0.227 | 0.247 | 0.198 |
| | (0.028) | (0.023) | (0.012) |
| Black | 0.086 | 0.099 | -0.006 |
| | (0.109) | (0.081) | (0.043) |
| Other race | -0.010 | -0.041 | 0.006 |
| | (0.029) | (0.023) | (0.013) |
| Married | 0.036 | 0.010 | 0.054 |
| | (0.027) | (0.021) | (0.012) |
| Number of pre-school aged kids | 0.055 | 0.034 | 0.055 |
| | (0.026) | (0.021) | (0.011) |
| Women * pre-school kids | 0.012 | 0.025 | -0.002 |
| | (0.043) | (0.032) | (0.017) |
| No activity limitation | -0.228 | -0.196 | -0.214 |
| | (0.040) | (0.031) | (0.015) |
| Seniority | 0.011 | 0.013 | 0.009 |
| | (0.005) | (0.004) | (0.002) |
| Seniority squared / 100 | -0.041 | -0.045 | -0.031 |
| | (0.014) | (0.011) | (0.006) |
| Experience | -0.014 | -0.008 | -0.012 |
| | (0.004) | (0.003) | (0.002) |
| Experience squared / 100 | 0.024 | 0.005 | 0.023 |
| | (0.009) | (0.008) | (0.004) |
| High school degree | -0.017 | 0.013 | -0.024 |
| | (0.046) | (0.039) | (0.020) |
| Less than bachelor degree | 0.074 | 0.098 | 0.086 |
| | (0.042) | (0.036) | (0.018) |
| Bachelor degree | -0.006 | 0.039 | -0.001 |
| | (0.054) | (0.045) | (0.023) |
| Some higher education | 0.003 | -0.038 | 0.008 |
| | (0.066) | (0.054) | (0.028) |
| Collective bargaining agreement | 0.279 | 0.210 | 0.276 |
| | (0.028) | (0.022) | (0.015) |
| Use a computer | -0.028 | 0.097 | -0.043 |
| | (0.031) | (0.026) | (0.014) |
| Use computer assisted design | 0.100 | 0.117 | 0.080 |
| | (0.034) | (0.027) | (0.014) |
| Use other technology | 0.112 | 0.069 | 0.103 |
| | (0.027) | (0.021) | (0.011) |
| 20-99 employees | -0.006 | 0.047 | -0.003 |
| | (0.044) | (0.037) | (0.049) |
| 100-499 employees | 0.052 | 0.138 | 0.066 |
| | (0.047) | (0.038) | (0.051) |
| 500 employees or more | 0.036 | 0.070 | 0.045 |
| | (0.053) | (0.043) | (0.056) |

Standard errors in parantheses
Include occupation and industry dummies

Table 5: Dispersion parameters

| | NB | RE NB | Full |
|---|---|---|---|
| $\delta$ | 1.765 | | |
| | (0.026) | | |
| $a$ | | 42.502 | |
| | | (16.313) | |
| $b$ | | 261.597 | |
| | | (104.537) | |
| $\nu$ | | | 0.593 |
| | | | (0.010) |
| $\tau_1$ | | | 2.261 |
| | | | (0.194) |
| $\tau_2$ | | | 2.344 |
| | | | (0.157) |
| $\tau_3$ | | | 2.106 |
| | | | (0.093) |
| $\tau_4$ | | | 2.072 |
| | | | (0.108) |
| $\tau_5$ | | | 1.630 |
| | | | (0.286) |

Standard errors in parantheses

Table 6: Goodness of fit

| | NB | RE NB | Full |
|---|---|---|---|
| **Goodness of fit** | | | |
| ln(L) | -38602.21 | -38313.73 | -37714.07 |
| BIC | 77641.03 | 77083.49 | 75922.99 |
| AIC | 77294.41 | 76721.46 | 75530.15 |
| **Correlation with Pearson residuals (p-value)** | | | |
| $\widehat{\gamma}_{ij}$ | -0.003 | -0.050 | -0.001 |
| | (0.727) | (<.0001) | (0.908) |
| $\widehat{\beta}X_{ij}$ | 0.000 | -0.052 | 0.002 |
| | (0.955) | (<.0001) | (0.775) |
| **Correlations with raw residuals (p-value)** | | | |
| $\widehat{\gamma}_{ij}$ | -0.010 | -0.052 | -0.010 |
| | (0.210) | (<.0001) | (0.208) |
| $\widehat{\beta}X_{ij}$ | -0.0061 | -0.05221 | -0.007 |
| | (0.435) | (<.0001) | (0.405) |