A model for the analysis of sick leave in Sweden
Inference using the HUS data

David Edgerton and Curt Wells
Economics Department
Lund University
Box 7082
S-200 07 Lund
Sweden

January 28, 2000

1 Introduction

The following paper addresses the question of how to model individual behavior in the face of changes in a set of rules governing the social welfare system in Sweden. To this end, the Swedish sickness insurance provides an excellent study object as the system has been changed often during the past decade. As the question employee compensation for sick leave is one the more widely discussed aspects of Swedish social welfare legislation, it is of interest to examine how individuals respond to changes in this legislation. It is therefore not surprising that a number of studies have appeared in the literature. Using data for 1991, Broström, Johansson and Palme (Broström et al., 1998) model the transition from work to work absence and *vice versa* using a proportional hazard model. They use the models to examine to changes following the reduction in the rate of compensation in March 1991. Their framework is a neoclassical utility maximizing model similar to the one presented below where working time and sick leave time vary but leisure is given.[1]

In this paper we ask the question as to what extent individual response to changes in system depend upon differences in socioeconomic and demographic characteristics. The basic theoretical model is presented in Section 2; the data is presented in third section. The fourth section presents the models considered; the fifth some results from these models and the final section draws conclusions.

2 The economic model

Sick leave will be analyzed in the context of the usual neo-Classical model. This is the approach followed by Palme–Johansson (Johansson and Palme, 1996), Brose (Brose, 1995) and others. The basic model posits a utility function which depends positively upon leisure time (L), consumption of the composite good, x, as well as individual characteristics, K, which will be represented by socio–economic variables such as marital status, number of children, education, working conditions and so on:

$$u = u(x, L, K) \tag{1}$$

Leisure time is "purchased" by abstaining from working time z costs the going wage rate, w. Total time, T, is the sum of leisure time and working time.

$$T = z + L \tag{2}$$

The budget restriction is derived from the identity that expenditure must equal income. On the debit side we find the cost of goods — p is the aggregate price level. On the credit side we find earned income after taxes (assumed proportional with a rate t), $w(1-t)z$, and unearned income, F. Thus the budget equation will be

$$x = w(1-t)z + F \tag{3}$$

[1] They divide time into t^c, contracted working time, t^l, contracted leisure time, assumed constant, and t^a, time absent. These three variables seem to be measured in hours.

Equation (3) follows from the accounting definition with the price level normalized to unity.

To solve for the utility maximizing consumption, substitute (2 and(3) into the utility function (1) and set the derivative with respect to working time to zero:

$$\frac{du}{dz} = \frac{\partial u}{\partial x} w(1-t) - \frac{\partial u}{\partial L} = 0 \tag{4}$$

When utility is maximized, assuming an interior solution, the marginal rate of substitution between leisure time and working time will be equal to the real wage rate net of tax. This of course the familial solution from a basic course in micro economics. However, the real world differs from this approach in a number of ways.

First of all, the individual may be unable to work because of illness. However, because of the existing social security system in Sweden, the individual is insured against income loss resulting from absence due to illness. This insurance system requires that the restrictions placed on the utility function must be modified. The time restriction must now include sick time, s.

However, working time and sick time cannot be chosen indepently. One has a certain amount of contracted time and sick time must be deducted from this given amount. Thus the individual chooses sick time to maximize his utility and working time becomes a residual. If we call contracted time C, then $C \equiv z + s$ then the budget equation can be written

$$T = z + s + L = C + L \tag{5}$$

The second adjustment concerns the budget restriction. The individual is compensated for income loss with a percentage of income, δ.[2] Thus sick time becomes a poorer paid substitute for working time. With a given amount of contracted time, the modified budget equation is as follows:

$$\begin{aligned} x &= w(1-t)(z+\delta s) + F \\ &= w(1-t)[C-(1-\delta)s] + F \end{aligned} \tag{6}$$

As pointed out above, the individual now has a different variable under his control: the amount of sick leave taken. Illness and its effect on the individual is personal. There are of course cases when one is laid out and cannot work; on the other hand, a common cold is not really a hinder if one feels that he cannot afford the reduction in income which occurs from staying at home.[3] Using equation (6),first order conditions are now

$$\frac{du}{ds} = -\frac{\partial u}{\partial x} w(1-t)(1-\delta) + \frac{\partial u}{\partial L} = 0 \tag{7}$$

[2] This is not quite true: one is compensated with $\delta \times 100$ percent of income up to 7.5 "base amount" which is an inflation adjusted amount set by the government each year. However, for those with income above this amount, there is an additional insurance paid by the employer. Thus as a first approximation, we assume that the individual receives the same compensation regardless of income.

[3] A more personal note here would explain how my youngest daughter was infected shortly after birth by a nurse who was working at the clinic despite a cold.

This expression is not as familiar as (4). However, in (7) working time is a residual. This first order conditions mean that the individual chooses *sick time* until, at the margin, the rate of substitution between leisure and consumption is equal to the net real wage decreased by what the individual "pays" for the sick time.

For the remainder of this paper, however, we will treat contracted time — and thus leisure time — as given as let the individual chose the time he is absent from work due to illness.

As we will be doing an econometric model using variants of linear models, we would like the utility function to be of such a form that the solution to the utility maximization problem would be linear in the relevant variables. A utility function which meets these requirements has been derived by Hausman (Haussman, 1980) and used by others in particular Johansson – Palme (Johansson and Palme, 1996), (Cassel et al., 1996). The function is derived by Hausman by beginning with a linear function and proceeding backwards through the indirect utility function to the direct one. In the present context, this function would appear as follows:

$$u(x,s) = -\left[1 + \frac{\beta(x + \frac{k}{\beta} - \frac{\alpha}{\beta^2})}{\frac{\alpha}{\beta} - T + L + s}\right] + \ln(T - L - s - \tfrac{\alpha}{\beta}) - \ln(\beta) \tag{8}$$

Substitution (6) into this utility function yields the following demand for sick leave equation:

$$s = C - \alpha w(1-t)(1-\delta) - \beta(F + w(1-t)\delta C) + k \tag{9}$$

This is essentially Johannson – Palme's equation (4) and Hausman's equation (2). There is, however, one essential difference. In the other two studies, individual characteristics are seen to enter the demand equation linearly through what is here interpreted as a constant, k, but is, with these authors, a vector of individual characteristics. Their approach would seem to exclude interaction terms involving wage and non–wage income as well as gender effects in the case of Johansson–Palme. We prefer to add the individual effects through the parameters in (9). Secondly, we allow the constant term, k, to vary across individuals as well. This will allow other variables to enter the equation.

$$s_i = C_i - a_i w_i(1-t_i)(1-\delta) - a_i(F_i + w_i(1-t_i)\delta C_i) + k_i \tag{10}$$

The parameters, in turn, will be linear combinations of other variables:

$$a_i = \alpha_0 + \boldsymbol{q}_i\boldsymbol{\alpha}_i \tag{11}$$

$$b_i = \beta_0 + \boldsymbol{q}_i\boldsymbol{\beta}_i \tag{12}$$

$$k_i = k_0 + \boldsymbol{q}_i\boldsymbol{\gamma}_i \tag{13}$$

This specification is very general. It will allow the Johansson–Palme equation by assuming that all terms other than α_0 and β_0 in equations (11) and (12) are zero. In addition to the interactions involving the wage variables, equation (13) allows both other variables as well as other interactions to enter the demand for sick leave in a linear manner.

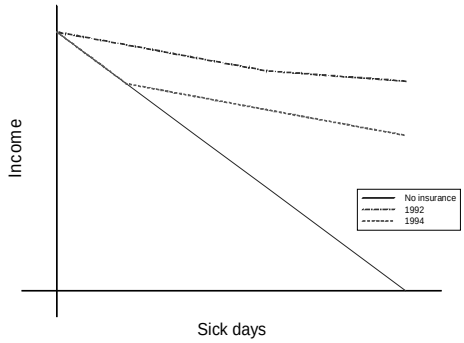

Figure 1: The budget line given a fixed number of working days per week. The individual can vary the number of sick days he takes.

There is a problem with equation (10). While absence for sickness is a positive variable, there is nothing in (10) to prevent s_i from becoming negative. We assume then that this equation is a linear approximation of the one actually estimated. Count models assume that the regression is of the form $\ln y = \boldsymbol{x\beta}$ and we take (10) to be an approximation of this equation which is the one estimated.

The budget line used in the above analysis is illustrated in Figure 1. We have assumed that the number of working days is set contractually; therefore, income for a person never ill is the intercept on the vertical axis. The solid line would be income in the absence of sick insurance for an increasing number of sick days. At some point, the number of sick days would equal the contracted time and income would fall to zero. A full insurance would compensate for all income loss due to illness. This was the case in Sweden in for the first year in the study (1986).[4] In 1992, the system was changed. First, the level of compensation was reduced to 75% of income for the first three sick days. This was increased to 90% from the fourth sick day. This change is represented by the uppermost line in the diagram, labeled 1992. Note the kink in the line after the first three days. The second change is that sick pay for the first fortnight became the responsibility of the employer. The national sick insurance paid out compensation from day 15. This is the relevant budget line for the second wave used here (with data from 1992).

From April 1993 the system changed again: now a qualifying day was added: one got no compensation for the first sick day while the rate of compensation remained unchanged for the second and third day of illness but was reduced from 90% to 75% for the remainder of the first two weeks. This budget line is labeled "1994" in the diagram and is the one relevant for the third wave used with data from 1995.

3 HUS data

The first wave of the Swedish *Household Market and Nonmarket Activities* (HUS) was completed in 1984 following a pilot study initiated by Anders Klev-

[4]This is note exactly true. Blue collar workers had a qualification day and 90% renumeration rate. White collar workers had no qualification day and 100% compensation.

marken in 1982. While the pilot survey was based on a random sample of three western counties (*sic*)— Göteborgs- och Bohuslän, Älfsborgs län and Värmlands län — the entire survey is based on a stratified random sample of individuals in the entire Kingdom. The household to which the individual belonged was then included in the sample. both the husband and wife in the household — as well as the person selected if different from these two — were interviewed. In all, 2131 households and a total of 3757 individuals were selected.[5] Net sample consisted of 1993 households with 3552 individuals.

The 1986 wave included all those in the previous study plus a few other categories. The first of these was the *nonresponse in 1984*: those who should have been in the study but for one reason or other were excluded. Further, those who moved into the household after the 1984 interview were included. Here a net of 1949 individuals were included. Finally, all members of the household born 1966 or 1967 were included. If they had entered a new household, then the head of that household and his (or her) partner was interview as well. this sample, called the *Supplementary sample*, encompassed 528 individuals.

The 1988 (2291 individuals) and 1991 (2052 individuals) waves were much more limited in scope and were really follow ups on the two previous waves recording changes in household composition, housing and labor market conditions. The main difference between these two latter waves was that the 1991 wave attempted to include new household members (those who had moved into the household after the 1986 survey) as well as those who had recently turned 18.

The 1993 wave was a repeat of the 1986 survey. The same decision rules for including new subjects were applies here (with, of course, the year of birth adjusted to 1973 or 1974). There were a net total of 1811 individuals in the panel, a net of 1643 in the supplementary survey and 733 in the nonresponse survey.

The 1995 wave as well as the 1997 wave is again a duplication of the earlier, larger waves. The last wave is at this writing note complete and is thus not included in this paper[6] The 1995 panel includes 2963 individuals and 276 in the Supplement.

This paper will consider a total of 8921 observations. The sample is composed of the individuals who have been in the work force.[7] An important subsample will be the 7081 who have answered the questions on their working environment.

[5] The sample procedure is described in Klevmarken (Klevmarken, 1984). More detail is found in the first volume of the Codebook (Klevmarken and Olovsson, 1993).

[6] This wave will be available Summer 1999 and will be included in the final report.

[7] As some of the questions asked were not answered by all those interviewed, there may be a slight discrepancy between the actual number of observations used in a regression of a table and the stated total. For example, table 5 contains information on 8817 rather than 8921 individuals.

Table 1: The dependent variable: weeks of sick leave (s) for own illness.

Survey	Size of sample	Number of Women	Prop $s > 0$ Men	Prop $s > 0$ Women	Mean $s > 0$ All	Mean $s > 0$ Men	Mean $s > 0$ Women
All	8921	4465	0.327	0.400	4.907	4.330	5.378
1986	2591	1289	0.439	0.471	4.435	3.834	5.000
1993	3766	1928	0.307	0.391	5.465	4.485	6.199
1996	2564	1248	0.245	0.341	4.668	4.938	4.464

4 The data, a description

To the uninitiated, the extraction of data from a database seems a simple matter. However, this has turned out not to be the case. There are many problems involved in finding the data in there files. One hinder is that questions referring to the entire household are asked on to the head of the household. For example, the question as to the type of housing — rental apartment, purchased apartment or own home — is only asked to the head of the household. At times this value is then also administratively given to other household members in the survey, at times not. Then, in the panel surveys, the head of household is asked if they have moved since the last interview. If the response is no, then the question on housing ownership is not asked and the researcher must go backward through all the surveys until he finds an answer to the question. Failure to do so reduces the sample size by quite a sizable amount.

Of the almost 9000 individuals in the survey, about 50% are female. In Table 1 we note that the proportion of both men and women who have taken sick leave has fallen during the three samples. However, the trend is for longer sick leave periods for men but not so for women. This latter group has increase the number of sick weeks in the 1992 and then decreased them by about 30% in 1995. The reason for this increase could be the changing age composition of the survey: the individuals get older and the input of younger individuals observed as children grow up and form families of their own is not enough to offset this trend. Table 2 shows this phenomenom especially in the transition from the 1986 to the 1993 surveys. For all three surveys, the proportion of those under 40 decreases and those over increases. And sick leave tends to increase with age.

Table 2: The age distribution of the sample.

	All surveys		1986 survey		1993 survey		1996 survey	
Variable	Men	Women	Men	Women	Men	Women	Men	Women
Age $<$ 20	0.020	0.012	0.066	0.039	0.000	0.000	0.002	0.002
$20 \leq$ Age $<$ 30	0.074	0.081	0.128	0.126	0.072	0.093	0.024	0.017
$30 \leq$ Age $<$ 40	0.188	0.181	0.197	0.184	0.198	0.205	0.165	0.140
$40 \leq$ Age $<$ 50	0.241	0.240	0.269	0.268	0.210	0.212	0.258	0.253
$50 \leq$ Age $<$ 60	0.231	0.239	0.210	0.238	0.248	0.235	0.229	0.247
$60 \leq$ Age $<$ 65	0.092	0.091	0.072	0.087	0.099	0.083	0.101	0.107
Age $\geq$ 65	0.153	0.156	0.057	0.058	0.173	0.171	0.221	0.233
Total	4456	4465	1302	1289	1831	1920	1316	1248

Table 3: The proportion of those in the population taking sick leave for own illness sorted by age group and sex.

	All surveys		1986 survey		1993 survey		1996 survey	
Variable	Men	Women	Men	Women	Men	Women	Men	Women
Age $<$ 20	0.001	0.000	0.003	0.001	0.000	0.000	0.000	0.001
20$\leq$ Age $<$ 30	0.023	0.030	0.055	0.057	0.015	0.030	0.001	0.002
30$\leq$ Age $<$ 40	0.074	0.081	0.092	0.094	0.083	0.093	0.043	0.048
40$\leq$ Age $<$ 50	0.085	0.097	0.115	0.112	0.085	0.099	0.055	0.081
50$\leq$ Age $<$ 60	0.078	0.106	0.115	0.134	0.063	0.093	0.062	0.095
60$\leq$ Age $<$ 65	0.031	0.043	0.036	0.052	0.029	0.038	0.030	0.041
Age $\geq$ 65	0.036	0.043	0.022	0.021	0.032	0.038	0.055	0.073
Total	4456	4465	1302	1289	1838	1928	1316	1248

There is a problem with the dependent variable in Table 1. The respondents were asked if they had taken sick leave during the previous year. If the answer was positive, then they were asked "how many weeks" were you absent from your job. They were also asked to round off their answer to the nearest number of weeks. Thus, if they were absent one or two days they were to answer "zero weeks". Thus some of those in the count who are registered as not having sick leave have in reality been absent up to a couple of days. There are 532 in the entire sample cases where a 'zero' answer is actually a rounded down answer. This is about 9.3% of the those in the entire sample who have answered 'zero'. For the three waves individually, the corresponding percentages are 12.4% in 1985, 7.4% in 1992 and 9.5% in 1995. This is unfortunate but that is the way the interview was conducted.[8] However, in the econometric model, we will attempt to adjust for this problem by estimating a *zero adjusted* count model.

When one considers those in the sample who took sick leave we find differences between ages and sexes. Table 3 shows a reduction of the proportion of those taking such leave in all age groups save the one for those over 65 and still working. Especially men between 40 and 60 and women between 20 and 30 show large decreases between 1985 and 1992 (and even 1995 for that matter). The changes in the social security system that occurred between 1986 and 1992 seem to have had a direct and lasting impact.

[8] The formulation used is perhaps understandable. Individuals are asked about their actions the previous year and it is doubtful if they could recall the exact number of days they were absent from work for sickness. By allowing the interviewee to round off to the nearest week probably increases the accuracy of the answers.

Table 4: Weeks of sick leave (s) for care of another family member.

Survey	Size of sample	Number of Women	Prop s $>$ 0 Men	Prop s $>$ 0 Women	Mean s $>$ 0 All	Mean s $>$ 0 Men	Mean s $>$ 0 Women
All	8926	4459	0.123	0.190	7.746	2.570	11.122
1986	2597	1286	0.115	0.187	5.689	2.099	7.938
1993	3769	1931	0.122	0.190	10.729	2.817	15.559
1996	2560	1242	0.134	0.191	5.431	2.659	7.489

Table 5: Family composition. The proportion of families with children in the indicated age groups.

Men	none	≤ 6	7-12	13-18	Total
All	0.532	0.216	0.118	0.133	4412
1986	0.496	0.221	0.136	0.147	1283
1993	0.558	0.204	0.113	0.126	1813
1996	0.532	0.228	0.109	0.131	1316
Women	none	≤ 6	7-12	13-18	Total
All	0.500	0.227	0.129	0.144	4405
1986	0.465	0.224	0.148	0.163	1252
1993	0.527	0.214	0.122	0.138	1905
1996	0.495	0.251	0.120	0.134	1248

Given the decrease in sick leave observed above, it is natural to ask whether individuals have substituted other types of absence. As the changes in social welfare legislation did not effect those who took sick leave to care for their children, we ask whether the decrease noted in Table 3 is balanced by an increase in paid absence to care for others.

In Table 4, we find a slight and trending increase in the proportion of men that have taken sick leave to care for family members — almost surely own children.[9] There is a similar increase for women but it is so small as to be negligible. However, in 1992, the number of weeks of sick leave for care of family members increased slightly for men and tremendously for women. This period then fell back to previous levels in 1995 — ending up about 25% above the 1985 level for men and 5% below for women. There is another interesting item in this table: while the proportion of those taking compensated sick leave decreased from the 1986 to the 1993 survey, the length of the sick periods increased.

The question as to why the large increase in the average number of sick weeks for women noted in Table 4 occurs. One suggestion would be that the proportion of women with children less than six years old was large that year compared to the other years. However, it turns out that there were relatively fewer women (22.0% in 1986, 21.3% in 1993 and 24.7% in 1996) in that year.

Another suggestion is that the number of children born in the early 1990's exceeded the number born in the mid–1980's. More children for the same num-

[9] I include here also ones partner's children in a previous relationship.

Table 6: Average weeks of sick leave (s) for care of another family member sorted according to family composition.

Survey	Women			Men		
	≤ 6	7–12	13–18	≤ 6	7–12	13–18
All	13.358	2.224	4.643	2.765	1.306	1.714
1986	10.179	1.815	1.000	2.303	1.308	1.000
1993	18.152	2.939	1.667	3.185	1.214	2.000
1996	8.913	1.610	8.000	2.644	1.444	2.150

ber of women mean more frequent absence for caring for sick children.

Table 5 does, however, give us a hint. The variable *sick leave for care of another family member* includes maternity leave. The increase in the proportion of individuals with children six years old or younger increases in the 1996 compared to the previous one. This could indicate an increase in the number of births in the 1993 survey compared to the 1986. However, it could also be an error in the data.

5 Modeling count data

The dependent variable in this study is discrete and thus a model which reflects this should be chosen. There are a number of possible models. One of these is the Poisson regression model. However, for this model the theoretical mean and variance are equal, an equality that is seldom observed on actual data. There are a number of alternatives open to the econometrician when confronted with a Poisson model exhibiting *overdispersion* — that is, a variance in excess of the mean. As this overdispersion results in biased estimates of the variances of the estimated parameters, one alternative would be to correct these variances for the overdispersion.

A second alternative would be to estimate a model where the variance is theoretically greater than the mean. There are a number of such models available. One is the negative binomial model that in essence adds a gamma–distributed noise term to the Poisson model. A second would be a Hurdle model which assumes that the data are generated by two independent Poisson processes: one that determines a zero occurrence and one that determines the non–zero level of the count variable. A third would be a "zero–inflated" count model where a binomial model determines whether a binary or a count process has generated the observation. In this case, a zero count may be the result of the binomial model or of the count model.

Below we discuss these different alternatives.

5.1 The Poisson model

The traditional model for count data is the Poisson model. It models the number of events that take place during a given interval. The frequency function for the process is

$$P(y_i = j) = \frac{e^{-\lambda_i}\lambda_i^j}{j!} \tag{14}$$

It is well known that the mean and the variance of this process are both equal to λ_i. When this parameter is modeled using exogenous variables, it is usual to define the log of λ_i as a linear combination of these variables:

$$\ln \lambda_i = x_i\beta \tag{15}$$

x_i is a 1 by k matrix of the independent variables at observation i and β a k by 1 parameter vector. That the equality between the mean and the variance is

not often observed is one of the problems with this model. I will return to this below.

The estimation of the Poisson model is easy as the expressions for the first and second derivatives are not at all complicated.[10] The likelihood function is

$$\ln L_{pi} = -\lambda_i + y_i x_i \beta - \ln y! \tag{16}$$

The matrixes of the first and derivatives are

$$\frac{\partial \ln L_{pi}}{\partial \beta'} = (y_i - \lambda_i) \cdot x_i \tag{17}$$

$$\frac{\partial \ln L_{pi}}{\partial \beta \partial \beta'} = -\lambda_i x_i' x_i \tag{18}$$

That the variance is larger than the mean is called *overdispersion* in the literature. The most troubling aspect of this problem is that the standard errors of the estimated coefficients are biased downwards. This is similar to the problem of heteroscedasticity in the usual regression model.

Given overdispersion, there are a couple of alternatives avaliable. One is to estimate a model which produces a variance in excess of the mean. The is the subject of the following section. A second alternative is to correct the standard errors in the Poisson model. Such a consistent estimate of the variance covariance of the parameters would be

$$\text{cov}(\hat{\beta}) = H^{-1} x' V(y) x H^{-1} \tag{19}$$

where x is the complete n by k matrix of the independent variables. V is a diagonal matrix with

$$V(y_i) = \lambda_i + \sigma^2 \lambda_i^2 \tag{20}$$

on the diagonal. This is made operational by replacing $V(y_i)$ by $\hat{V}(y_i) = (y_i - \hat{\lambda})^2$, by replacing λ_i in equation (20) with the $\hat{\lambda}_i$ estimated in the sample, and by estimating σ^2 using *OLS*.

5.2 Over–dispersed models

Below I present three count models where the variance is theoretically exceeds the mean. The first of these, the *Negative Binomial*, extends the Poisson by adding an error term to equation (15) and then assuming that this residual has a Gamma distribution. The second of the three models, the *Poisson Hurdle* model, assumes that the observed count is the outcome of two different Poisson processes, one of which produces the zero result and the other the positive counts. The final model presented, the *Zero Augmented Count* model, also

[10] All of the models presented in this paper have been estimated using the Newton-Raphson algorithm with analytical first and second derivatives. In a program such as Gauss or Matlab — or indeed C++ or Fortran — routines for the likelihood function and the two derivatives are simple to implement. I have programmed all four routines in Gauss and these programs are available on request.

assumes that the observation is generated by two processes: one a binary model whose zero outcome is the observed zero count and whose no zero output is a usual count processes with both zero and positive outcomes. The count model may be either a Poisson or a Negative Binomial.

5.2.1 The Negative Binomial model

One way to model overdispersion is to chose a function where the mean is less than the variance. Such a model is the *Negative Binomial*. It obtains by assuming, in the Poisson model, that λ_i is generated with an error term:

$$\ln \lambda_i = x_i\beta + \varepsilon_i \tag{21}$$

The *Negative Binomial* obtains by assuming that ε_i follows a gamma distribution with parameters (ψ_i, ν_i) so that (Cameron and Trivide, 1986)

$$f(\lambda_i) = \frac{1}{\Gamma(\nu_i)} \left(\frac{\nu_i\lambda_i}{\psi_i}\right)^{\nu_i} \exp\left(\frac{\nu_i\lambda_i}{\psi_i}\right) \frac{1}{\lambda_i} \tag{22}$$

This leads to the frequency function for the *negative binomial* distribution:

$$P(Y_i = j) = \frac{\Gamma(j+\nu_i)}{\Gamma(j+1)\Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i+\psi_i}\right)^{\nu_i} \left(\frac{\psi_i}{\nu_i+\psi_i}\right)^{j} \tag{23}$$

where

$$\mathrm{E}(Y_i) = \psi_i \tag{24}$$

$$\mathrm{Var}(Y_i) = \psi_i + \frac{1}{\nu_i}\psi_i^2 \tag{25}$$

Cameron and Trivedi define $\nu_i = (1/\alpha)(\exp(x_i\beta))^k$ and distinguish two basic models. Type I sets $k = 1$ and Type II has $k = 0$. Here I will work with Type II. The former implies a constant mean–variance ratio while the latter has a linear mean–variance ratio.[11] Defining $\phi_i = \exp(x_i\beta)$, the frequency function then becomes

$$\mathrm{Prob}(Y_i = h) = \frac{\Gamma\left(h + \frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right) \cdot h!} (\alpha\phi_i)^h (1+\alpha\phi_i)^{-(h+1/\alpha)} \tag{26}$$

After a bit of algebra to remove the gamma functions, the log–likelihood for a single observation is

$$\ln L_i = \sum_{m=0}^{y_i-1} \ln\left(\frac{1}{\alpha} + m\right) - \ln y_i! + y_i \ln(\alpha\phi_i) - \left(y_i + \frac{1}{\alpha}\right) \ln(1+\alpha\phi_i) \tag{27}$$

[11] (Cameron and Trivide, 1986, p. 33)

The first derivatives are rather easy to calculate:

$$\frac{\partial \ln L_i}{\partial \beta} = y_i x_i - \left(y_i + \frac{1}{\alpha}\right) \frac{\alpha x_i e^{x_i\beta}}{1 + \alpha e^{x_i\beta}} \tag{28}$$

$$\frac{\partial \ln L_i}{\partial \alpha} = -\frac{1}{\alpha^2} \sum_{m=0}^{y_i-1} \frac{1}{\frac{1}{\alpha} + m} + \frac{y_i}{\alpha} + \frac{1}{\alpha^2} \ln\left(1 + \alpha e^{x_i\beta}\right)$$

$$\times \left(y_i + \frac{1}{\alpha}\right) \frac{e^{x_i\beta}}{1 + \alpha e^{x_i\beta}} \tag{29}$$

But the second derivatives are hairier:[12]

$$\frac{\partial^2 \ln L_i}{\partial \beta \partial \beta'} = -\left(y_i + \frac{1}{\alpha}\right) \frac{\alpha e^{x_i\beta}}{\left(1 + \alpha e^{x_i\beta}\right)^2} x_i' x_i \tag{30}$$

$$\frac{\partial^2 \ln L_i}{\partial \alpha^2} = \frac{2}{\alpha^3} \left(\sum_{m=0}^{y_i-1} \frac{1}{\frac{1}{\alpha} + m} - \ln\left(1 + \alpha e^{x_i\beta}\right) \right) - \frac{y_i}{\alpha^2}$$

$$+ \frac{1}{\alpha^4} \left(\frac{2\alpha^2 e^{x_i\beta}}{1 + \alpha e^{x_i\beta}} + \sum_{m=0}^{y_i-1} \frac{1}{\left(\frac{1}{\alpha} + m\right)^2} \right)$$

$$+ \left(y_i + \frac{1}{\alpha}\right) \frac{e^{2x_i\beta}}{\left(1 + \alpha e^{x_i\beta}\right)^2} \tag{31}$$

$$\frac{\partial^2 \ln L_i}{\partial \beta \partial \alpha} = \frac{1}{\alpha^2} - \left(y_i + \frac{1}{\alpha}\right) \frac{x_i e^{x_i\beta}}{\left(1 + \alpha e^{x_i\beta}\right)^2} \tag{32}$$

5.2.2 Modified counts

It is not at all unreasonable to posit that the process returning a count value of zero is different from the one that yields positive integers. For example, the number of children in a marriage may well be none at all while the couple plans a family in the future. Again, the number of bottles of beer a person has

[12] Routines that maximize a function using only numerical derivatives are of course an alternative to the work involved in calculating and programming first and second derivatives. However, convergence is much faster using Newton's method and analytical firsts and seconds. Indeed I find that the is a such a great improvement using both of these derivatives even compared to routines using only analytical firsts such as the BHHH routine which approximates the Hessian with the product of the score matrix. The back side of my assertion is that programming these derivatives is seldom error free: on must compare results to those returned by numerical differentiation. The Gauss package for doing maximum likelihood estimation, MAXLIK, version 3, contained a routine that did such checking. I use it extensively.

consumed in the past month may well be zero or any positive number within limits. In both cases one suspects that the process where the count variable is zero and the one where the variable is positive are in fact two different processes. The standard reference here is Mullahy (Mullahy, 1986).

The Poisson hurdle model In this specification we study two separate processes. The first one will determine whether the output is zero or positive; the second will determine the positive count. Here we may hypothesize two separate count processes: one is active when the observed count is zero and the other when the count is a positive integer. A second alternative would be to consider the first process a probit and the second a Poisson.

For the zero outcome, using two Poisson processes, we have

$$P(y_i = 0) = e^{-\theta_i} \tag{33}$$

where $\theta_i = \exp(z_i \cdot \gamma)$.

If the zero process is specified as a Probit, we would have

$$P(y_i = 0) = \Phi(z_i \cdot \gamma) \tag{34}$$

where $\Phi_i \equiv \Phi(z_i \cdot \gamma)$. In either case, z_i is n by k_0 matrix of the variables working on the zero process and γ is a conformal column vector.

For the positive outcome, $j = 1, 2, \ldots$, we have

$$P(y_i = j) = \frac{1 - e^{-\theta_i}}{1 - e^{-\lambda_i}} \frac{e^{-\lambda_i} \lambda^j}{j!} \tag{35}$$

where $\lambda_i = \exp(x_i \cdot \beta)$. Here x_i is n by k_1 matrix of the variables working on the zero process and β is a conformal column vector.

This model allows for both over– and underdispersion. Winkelmann and Zimmermann show that the mean may be expressed as

$$E(y_i) = \sum_{j=1}^{\infty} j \cdot \frac{e^{-\lambda_i} \lambda^j}{j!} \cdot \Psi_i \tag{36}$$

where Ψ_i is the ratio of the probability that the first process is not zero to the probability that the second process is equal to or greater than one. For (33) the expected value is $\lambda_i \cdot (1 - \exp(\theta_i))/(1 - \exp(\lambda_i))$; for (34) it is $\lambda_i \cdot (1 - \Phi(z_i\gamma))/(1 - \exp(\lambda_i))$.

The variance of the hurdle process is

$$Var(y_i) = \Psi_i \cdot \lambda_l(\lambda_l + 1) - \Psi_i^2 \lambda_i^2 \tag{37}$$

The log likelihood for an observation is

$$\begin{aligned} \ln(L_{hi}) &= -r_i\theta_i + (1 - r_i) \times \\ &\quad \left[\ln\left(1 - e^{-\theta_i}\right) - \ln\left(1 - e^{-\lambda_i}\right) + y_i x_i \beta - \lambda_i - \ln\left(y_i!\right)\right] \end{aligned} \tag{38}$$

The first derivatives of $\ln(L_{hi})$ are

$$\frac{\partial \ln(L_{hi})}{\partial \gamma'} = \frac{\theta_i \left(e^{-\theta_i} - r_i\right)}{1 - e^{-\theta_i}} \cdot z_i \tag{39}$$

$$\frac{\partial \ln(L_{hi})}{\partial \beta'} = (1 - r_i)\left(y_i - \lambda_i - \frac{\lambda_i}{e^{\lambda_i} - 1}\right) \cdot x_i \tag{40}$$

Here r_i is an indicator variable equaling one when the observed count is zero and zero otherwise.

As the derivative of $\ln(L_{hi})$ with respect to γ' or β' does not depend upon the other variable, the Hessian will be block diagonal. Thus

$$\frac{\partial^2 \ln(L_{hi})}{\partial \gamma \partial \gamma'} = \left[\frac{\theta_i \left(e^{\theta_i}\left(1 - \theta_i\right) - 1\right)}{\left(e^{\theta_i} - 1\right)^2} - \frac{r_i \theta_i \left(1 - e^{-\theta_i} - \theta_i e^{-\theta_i}\right)}{\left(1 - e^{-\theta_i}\right)^2}\right] \cdot z_i' z_i \tag{41}$$

$$\frac{\partial^2 \ln(L_{hi})}{\partial \beta \partial \beta'} = (1 - r_i) \cdot \left[\frac{\lambda_i \left(e^{\lambda_i}\left(1 - \lambda_i\right) - 1\right)}{\left(e^{\lambda} - 1\right)^2}\right] \cdot x_i' x_i \tag{42}$$

Starting values could be the least squares estimates. Newton's method will give rapid convergence; otherwise one could use the BHHH algorithm and only the first derivatives.

The Zero–Inflated Count models The basic difference between the *Zero–inflated* models and the Hurdle model described in section 5.2.2 is the way the zero observation is modeled. As before, we consider two processes; but here, one is a binary Probit model that always returns a zero with probability Φ_i. However, with probability $1 - \Phi_i$ the count process is the ruling one; and this process may well return a zero count. Thus the probability of observing a zero is

$$P(y_i = 0) = \Phi_i + (1 - \Phi_i) f_{0i} \tag{43}$$

The probability of observing a positive y_i is simply:

$$P(y_i = j) = (1 - \Phi_i) f_{+i} \tag{44}$$

I use f_{0i} to represent the frequency function of the count process when $h = 0$. Similarly, f_{+i} represents the truncated at zero frequency function of the count.

Letting $\lambda_i = \exp(x_i \beta)$, the mean this distribution is

$$\mathrm{E}(y_i) = (1 - \Phi_i)\lambda_i \tag{45}$$

If the count is a Poisson, the variance is

$$\mathrm{Var}(y_i) = \lambda_i (1 - \Phi_i)(1 + \lambda_i \Phi_i) \tag{46}$$

If the count is negative binomial, the variance will be

$$\mathrm{Var}(y_i) = \lambda_i(1-\Phi_i)(1+\lambda_i[\Phi_i+\alpha]) \tag{47}$$

The *ZIP* model exhibits overdispersion as the variance is larger than the mean by a factor of $1+\lambda_i \cdot \Phi_i$. (Greene, 1995, p. 573). Greene[13] points out that the relationship of the variance to the mean in the *ZAP* model is very similar to that observed in the *negative binomial* model. This relationship is, for the *ZAP*

$$\frac{\mathrm{Var}(y_i)}{\mathrm{E}(y_i)} = 1 + \frac{\Phi_i}{1-\Phi_i}\mathrm{E}(y_i)$$

while for the *negative binomial* it is

$$\frac{\mathrm{Var}(y_i)}{\mathrm{E}(y_i)} = 1 + \gamma\mathrm{E}(y_i)$$

where γ is the extra parameter in the *negative binomial* model. Note that these two expressions are quite similar. Testing one against the other is dicey as the models are non–nested. However, Vuong (Voung, 1989) has proposed a test which Greene asserts as "some power"[14] is distinguishing the overdispersion due to the *ZAP* and to the *negative binomial* specifications. The test statistic is

$$Z_v = n^{0.5}\frac{\bar{m}}{s_m} \tag{48}$$

where $\bar{m}$ is the mean and s_m is the standard deviation of the log of the series formed by dividing the frequency function of, say, the *ZAP* with that of the *negative binomial*:

$$m_i = \ln\left(\frac{f_{ZAP}(y_i)}{f_{NB}(y_i)}\right)$$

The asymptotic distribution of Z_v is standard normal so that a statistic in excess of 1.96 would favor the *ZIP* model, one less than -1.96 would favor the *negative binomial* and one between these to values would not allow one to reject the null of no difference. However, one does have to estimate both models.

We assume now that $\Phi_i \equiv \Phi(z_i\gamma)$, the distribution function of the standard normal, represents the Probit process. As above, it depends on the k_0 variables z_i with the conformal parameter vector γ. Further, f_i is a Poisson process as defined in equation (14), or a Negative binomial process defined in 23. As usual, $\ln(\lambda_i) = x_i\beta$. r_i is as defined above.

The log likelihood function for the *Zero–Inflated Count* is

$$\ln(L_{zi}) = r_i \ln\left(\Phi_i + (1-\Phi_i)\, f_{0i}\right) + (1-r_i)\ln(1-\Phi_i) + (1-r_i)\ln(f_{+i}) \tag{49}$$

The first and second derivatives for both alternatives complicated but still rather straight forward.

$$\frac{\partial \ln(L_{zi})}{\partial \gamma'} = \left(r_i\frac{1-f_{0i}}{p_{0i}} + \frac{1-r_i}{1-\Phi_i}\right)\phi_i z_i \tag{50}$$

$$\frac{\partial \ln(L_{hi})}{\partial \theta'} = r_i\frac{1-\Phi_i}{p_{0i}}\frac{\partial f_{0i}}{\partial \theta'} + \frac{\partial \ln(f_{+i})}{\partial \theta'} \tag{51}$$

[13](Greene, 1995, p. 573)
[14]*ibid.*

Here I have used θ as the vector of the count parameters. If the model is Poisson, then θ and β are identical; for the negative binomial, θ also includes the over–dispersion parameter α.

Greene[15] notes that epically the second derivatives are rather messy and chooses to use the BHHH algorithm which uses on the first derivatives. However, I have found that convergence is much more rapid using Newton's method and analytical second derivatives. In fact, both the first and second derivatives have been given above. All that will be new is the cross–partials.

The Zero–Inflated Count models with selection Winkelmann (Winkelmann, 1998, pp. 347–350) argues that ZIP models estimated assuming independence of the selection and the count processes will result in incorrect inference if there is indeed correlation between the two processes. Given correlation, he suggests estimating the Poisson model as the Negative Binomial model will be difficult to estimate as the first two moments will not be sufficient to identify both the overdispersion and the heterogeneity parameters. As above, the model will have both a selection and a count part.

In the Poisson model, with $y^\star$ being the count, individual heterogeneity is introduced as in the Negative Binomial model:

$$E(y^\star|\boldsymbol{x}_i, u_i) = \exp(\boldsymbol{x}_i\boldsymbol{\beta} + u_i) \tag{52}$$

In (21) above we assumed that the heterogeneity followed a gamma distribution. Here we assume it to normally distributed and correlated with the error part of the selection equation. Here we define y_i according to

$$y_i = \begin{cases} y_i^\star & if\ c_i = 1 \\ 0 & if\ c_i = 0 \end{cases}$$

where c_i is the latent process

$$c_i = \boldsymbol{z}_i\boldsymbol{\gamma} + \epsilon_i \tag{53}$$

The point here is that u_i and ϵ_i will be joint normal with zero expectations and the covariance matrix

$$cov(u_i, \epsilon_i) = \begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix}$$

The variance of ϵ_i is normalized to unity as $\boldsymbol{\gamma}$ can only be estimated in relation to the variance of ϵ_i in any case. Defining $\tilde{\lambda}_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta} + u_i)$, the probability for an observation, conditional on u_i, becomes (Winkelmann, 1998, p. 349)

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \int_{-\infty}^{+\infty} \left\{ (1 - \Phi_i^\star)\delta + \Phi_i^\star \frac{\exp(-\tilde{\lambda}_i y_i^{\tilde{\lambda}_i}}{y_i!} \right\} \frac{1}{\sigma}\phi(\frac{u_i}{\sigma})du_i \tag{54}$$

[15] *ibid.*, p. 579.

Following Winkelmann I have used

$$\Phi_i^\star \equiv \Phi\left(\frac{\boldsymbol{z}_i\boldsymbol{\gamma} + \rho u_i/\sigma}{\sqrt{1-\rho^2}}\right)$$

This cannot be solved analytically but a Gauss-Hermite quadrature may be used to evaluate the integral numerically.[16]

5.3 A Poisson Random Effects model

Another way to model the individual heterogeneity is to assume that the in the Poisson model that $\tilde{\lambda}_{it} = \lambda_i t\alpha_i$ where there is now both a individual index (i) and a time index (t). The individual effects are in the α_i which are assumed to be independent of the independent variables $\boldsymbol{X}_i t$.

The usual specification is $\tilde{\lambda}_{it} = \exp(\boldsymbol{X}_i t\boldsymbol{\beta} + \mu_i)$ where the indepentent variables include a constant. The individual effects $\alpha_i = \exp(\mu_i)$ is assumed to follow a *gamma* distribution (δ, δ) so that the expectation of α_i is unity.[17] With the individual effects integrated out, the likelihood becomes

$$L_i = \left[\prod_i \frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right]\left[\frac{\delta}{\sum_t \lambda_{it} + \delta}\right]\left[\sum_t \lambda_{it} + \delta\right]^{\sum_t y_{it}} \frac{\Gamma(\sum_t y_{it} + \delta}{\Gamma(\delta)} \qquad (55)$$

6 A ZIP model for the HUS data

Table 7 presents a first estimate of the model. The model reported is that with correlation between the error terms in 54. No attempt has been made to find interaction terms. The probit model estimates the probability that one will be sick during the year. The additional zero in the model comes from those who are sick but do not take sick leave.[18]

During the three years studies (1986, 1992 and 1995), there have been three different sets of rules for sick leave. In the earliest year, the blue collar but not white collar workers had a qualification day. The renumeration rate was 90% of ones wage. In 1992, the renumeration rate was reduced to 75% for the first two sick days and increased to 90% for the third and following days. Finally, in 1995, the was a qualification day and the renumeration rate for the third and following sick days was reduced to 75%.

It is therefore of interest to note the signs on the coefficients for the 1992 effect – caught by the variable *Post 1991* in the Poisson part of model and the 1995 effect in the variable *Post 1993* in the Probit — the selection — part. Reducing renumeration reduces the number of sick weeks for the individuals while the introduction of a qualifying day reduced the probability that one would take sick leave.

[16] The reader will perhaps have noted that (54 and (49) have used the binary process in opposite meanings. Lambert uses the binary as the probability that a zero will occur whereas Winkelmann uses the binary for the probability that a non–zero result occurs.

[17] The derivation is detailed in Hausman *et. al*, (Hausmann et al., 1984).

[18] As explained above, the data also contains an extra zero: this will be those who are sick but have had two or less paid sick days.

Table 7: A ZIP model with correlated processes. The independent variable is the number of weeks absent from work with paid sick leave.

	The Probit part		
Variable	*Coef*	*Std.err.*	*t–stat*
Constant	1.411	0.220	6.420
Gender	0.525	0.131	4.004
Disp inc	-0.635	0.129	-4.927
Capt inc	-0.162	0.076	-2.143
Rural	-0.316	0.182	-1.741
Post 1993	-0.460	0.108	-4.259
	The Poisson part		
Variable	*Coef*	*Std.err.*	*t–stat*
Constant	-0.510	0.119	2.028
Age	0.330	0.064	5.283
Disp inc	0.134	0.081	3.223
Post 1991	-0.194	0.070	-2.759
Hectic	0.021	0.085	-2.056
Monoton	0.303	0.070	2.608
Uncomft	0.221	0.058	4.841
Cust	-0.286	0.070	-2.835
Edu_voc	0.147	0.060	2.147
Young_ch	-0.098	0.030	-2.247
	The bivariate normal part		
σ	1.287	0.028	46.479
ρ	0.142	0.088	1.620

The gender effect is as expected: being female increases the probability that one is sick; the age effect works through the count part of the model to increase the time spend away from work. Income works 2 ways: both current and capital income have a negative effect on the probability of being sick while higher current income is associated with longer periods of absence. Finally, those living in rural areas seem to be healthier — or at least have a lover probability of being sick — than those in other areas.

Finally, note that those variables on the working environment which increase the burden of work (a stress–filled job, a job requiring uncomfortable movements, monotonous tasks) seem also to increase the length of sick periods. Customer contact, which should add variety to work, decreases sick leave. [19]

References

Baltagi, B. H. (1995). *Econometric Analysis of Panel Data.* John Wiley & Sons Ltd, west Sussex, England.

[19] Here we could also hypothesize that the added responsibility implied by customer contact perhaps entices one to work even if one is a bit sick.

Brose, P. (1995). Sick absence: an empirical analysis of hus panel. Department of Economics, Uppsala University.

Broström, G., Palme, M., and Johansson, P. (1998). Assessing the effect of economic incentives on incidence and duration of work absence. Working paper series in Economics and Finance, no. 288, Stockholm School of Economics, Stockholm.

Cameron, A. C. and Trivide, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1:29–53.

Cameron, A. C. and Trivide, P. K. (1998). *Regression Analysis of Count data.* Econometric Society Monographs No. 30. Cambridge University Press, New York, New York.

Cassel, C.-M., Johansson, P., and Palme, M. (1996). A dynamic discrete choice model of blue collar worker absenteeism in sweden, 1991. Umeå Economic Studies No. 425.

Greene, W. H. (1995). *LIMDEP, version 7.0.* Econometric Software, inc., Bellport, New York.

Hausmann, J., Hall, B. H., and Grilliches, Z. (1984). Econometric models for count data with an application to the patents–r&d relationship. *Econometrica*, 52(4):909–938.

Haussman, J. (1980). The effect of wages, taxes and fixed costs on women's labor force participation. *Journal of Public Economics*, 14(1):161–194.

Hsiao, C. (1986). *Analysis of Panel Data.* Econometric Society Monographs No. 11. Cambridge University Press, New York, New York.

Johansson, P. and Palme, M. (1996). Do economic incentives affect work absence? empirical evidence using swedish micro data. *Journal of Public Economics*, 59(2):195–218.

Klevmarken, A. (1989). Panel studies: What can we learn from them? an introduction. *European Economic Review*, 33:523–529.

Klevmarken, N. A. (1984). Household market and nonmarket activities: The first year of a swedish panel study. *Vierteljahrsfefte zur Wirtschaftsforschung*, (4):452–457.

Klevmarken, N. A. and Olovsson, P. (1993). *Household market and nonmarket activities: Procedures and codes 1984–1981.* The Industrial Institute for Economic and Social Research, Box 5501, S–114 85, Stockholm, Sweden. Distributed by Almqvist & Wiksell International, Stockholm.

Mullahy, J. (1986). Specification and testing of some modefied count data models. *Journal of Econometrics*, 33:341–365.

Voung, Q. H. (1989). Likelihood ratio tests for model selection and non–nested hypothesis. *Econometrica*, 57(2):307–334.

Winkelmann, R. (1998). Count data models with selectivity. *Econometric Reviews*, 17(4):339–359.