Research Report
Statistical Research Unit
Department of Economics
University of Gothenburg
Sweden

When does Heckman's two-step procedure for censored data work and when does it not?

Robert Jonsson

Research Report 2008:2
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Statistical Research Unit	Nat: 031-786 12 74	Nat: 031-786 00 00	http://www.statistics.gu.se/
P.O. Box 640 SE 405 30 Göteborg Sweden	Int: +46 31 786 12 74	Int: +46 31 786 00 00	

When does Heckman's two-step procedure for censored data work and when does it not?

Robert Jonsson

Department of Economics, University of Gothenburg, Box 640, 405 30 Göteborg, Sweden

Abstract:

Heckman's two-step procedure (Heckit) for estimating the parameters in linear models from censored data is frequently used by econometricians, despite of the fact that earlier studies cast doubt on the procedure. In this paper it is shown that estimates of the hazard *h* for approaching the censoring limit, the latter being used as an explanatory variable in the second step of the Heckit, can induce multicollinearity. The influence of the censoring proportion and sample size upon bias and variance in three types of random linear models are studied by simulations. From these results a simple relation is established that describes how absolute bias depends on the censoring proportion and the sample size. It is also shown that the Heckit may work with non-normal (Laplace) distributions, but it collapses if *h* deviates too much from that of the normal distribution. Data from a study of work resumption after sick-listing are used to demonstrate that the Heckit can be very risky.

Keywords:

Censoring, Cross-sectional and panel data, Hazard, Multicollinearity

1. Introduction

When studying the relation between a dependent variable Y^* and a set of explanatory variables it sometimes occurs that a large proportion of the observations falls on $Y^* = a$, and no observations are found below the known constant *a*. The consequences of this are that standard conditions for efficient estimation of the parameters are violated. This may be termed the problem of *border-observations*. One way to deal with the latter is to use the fact, or just make the assumption, that it has originated from censoring of some latent variables. (According to Kruskal and Tanur (1978) data are censored if observations are measured only in some interval, while observations outside the interval are counted but not measured). The relation between Y^* and the latent variables can be expressed in several ways, the simplest being the Tobit model (Tobin, 1958)

$$Y^* = \begin{cases} Y, \text{if } Y > a \\ a, \text{if } Y \le a \end{cases} \tag{1}$$

The Tobit model was later generalized by Heckman who introduced a further latent variable to take account of selection effects (Heckman 1976, 1979). Consider e.g. the variable Y^* = 'Number of sick-listed days per person' where many observations are zeros. To deal with the problem of border observations at $a = 0$ one may introduce the latent variable Y = 'State of health' which can be measured in several ways (cf. e.g. Hansson *et al*, 2004). For those interested in the actual and private budgetary consequences of sick-listening there is no reason to include selection effects because the zeros are true zeros. However, persons with zero sick-listed days may be different from others in several respects. E.g. in a Swedish study women with extremely low household incomes returned to work after sick-listening earlier than others and after 90 days nearly all had returned (Bergendorff *et al.* 2001, p. 33). For those interested in studying the potential outcome that would follow if incomes were changed, it seems natural to take account of the selection effect that derives from household income. The problem of choosing a proper model for the censoring in the latter case may be termed the *selection-effect* problem and is separated from the border-observation problem mentioned above. A clarifying discussion on the problem of border observations and selection effects has been given by Dow and Norton (2003).

Objections may be raised against introducing a latent variable, the meaning of which may be unclear, such as 'State of health' but this gives anyhow a simple solution of a complicated problem. The introduction of a latent variable in the selection-effect situation is even more delicate, especially if it is generally stated that the two latent variables has a bivariate normal distribution (cf. e.g. Flood and Gråsjö, 2001). In the latter paper simulation studies were performed that showed that the simple Tobit model can be as good as more sophisticated selection-effects models, and sometimes even better. In this paper only the censoring in Eq. (1) is studied.

Eq. (1) contains two types of data, counting data and observations on *Y*. When *Y* depends on explanatory variables in a regression relation it is possible to find the Maximum Likelihood (ML) estimates of the parameters by using both types of data under suitable assumptions, such as linearity of the regression and normality (Rosett and Nelson, 1975, Nelson, 1984). The computational difficulties involved in solving the ML equations led Heckman (1976, 1979) to propose a simple two-step method (Heckit). Although it was originally designed for censoring due to selection effects in cross-sectional data, it can be used for data free from selection effects and for panel data. The Heckit requires in a first step an estimate of a censoring proportion *p* from counting data. This in turn gives estimates of the hazard (*h*) for approaching *a* (or inverse Mills ratio). In a second step the parameters in the linear model are obtained by regressing the observations on the explanatory variables and on estimates of *h*.

It is peculiar that the Heckit never seems to have been used by biostatisticians, although problems with censoring occur frequently in this area. Also pure statisticians seem to have ignored the procedure. It is typical that in a recent PhD thesis in statistics including four papers on the subject, the Heckit is not mentioned (Karlsson, 2005). But, among econometricians the Heckit is still popular despite of the fact that an extensive amount of Monte Carlo studies casts doubt on the procedure. (See Puhani, 2000 for an overview). But, from these studies it is hard to find guide lines which can be used in practice

Heckman's two-step procedure involves several critical moments. It is the aim of this paper to clarify the following issues: (*i*) Which are the properties of the estimated hazard that is used later in the second step? (*ii*) Which are the properties (bias and variance) of the regression estimates obtained with three different linear models? Furthermore, is it possible to adjust for the bias? In earlier studies the performance of the Heckit estimators have been compared with other alternatives such as the Tobit ML estimator and several semiparametric estimators (Kim and Lai, 2000, Lee, 1996, Newey, 2001 and Powell, 1994). This paper will focus only on the Heckit. The aim is to find simple guide lines for when the Heckit works and when it does not.

2. Notations, assumptions and some theoretical results

Let Y_{tj} denote an observation on the latent variable from the *j*:th subject at time *t*, *j*=1,…,*n* and *t*=1,…,*T*. For cross sectional data the index *t* is omitted The observations for each subject are represented by a transposed vector $\mathbf{y}'_j = \left(Y_{1j} ... Y_{Tj}\right)$ and it is assumed that the latter are independent over the *j*'s. The problem considered is to estimate a linear regression function $E\left(Y_{tj} \middle| \mathbf{x}_t\right) = \mu_{\mathbf{x}}$, where $\mathbf{x}_t$ is a vector of *p* explanatory variables possibly depending on *t*, when observations are obtained only in the interval (a, ∞) and it is known how many observations that fall below *a*. The function $\mu_{\mathbf{x}}$ is written $\alpha + \mathbf{x}'_t\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a vector of regression coefficients.

2.1 Three linear models with different random structures

Consider the following models, where random variables are denoted by capitol letters, fixed values by small letters and parameters by Greek symbols.

$$(a)\ Y_{tj} = \alpha + \mathbf{x}_t'\boldsymbol{\beta} + U_{tj},\ (b)\ Y_{tj} = A_j + \mathbf{x}_t'\boldsymbol{\beta} + U_{tj},\ (c)\ Y_{tj} = A_j + \mathbf{x}_t'\mathbf{b}_j + U_{tj} \quad (2)$$

Here the U_{tj}'*s* are independent and identically distributed (iid) disturbances with mean 0 and variance σ_U^2. A_j is a random intercept that is specific for the *j*:th subject with mean α and variance σ_A^2, while $\mathbf{b}_j$ is a vector of random regression coefficients specific for the *j*:th subject with mean $\boldsymbol{\beta}$ and variance $\sigma_{B_r}^2$ for the *r*:th component. All A_j's and $\mathbf{b}_j$'s are iid and U_{tj} is independent of A_j and $\mathbf{b}_j$. The latter two may be correlated with $Cov(A_j, B_{rj}) = \sigma_{AB_r}$. All random variables are assumed to be normally distributed.

The models in Eq. (2) have been widely used (see e.g. Swamy, 1971 and Hsiao, 2003) and have been termed (*a*) Gauss-Markov (GM), (*b*) Error Components Regression (ECR) and (*c*) Random Coefficient Regression (RCR), just to mention a few names. The GM-model is intended for cross-sectional data or panel data without within-subject correlations. ECR- and RCR models are intended for panel data. Tests for uncensored data in order to establish a proper random structure have been suggested by several authors (see e.g. Honda, 1985, Lundevaller and Laitila, 2002, Hsiao, 2003), but no such test seems to have been suggested for censored data.

The Heckit requires that the censored variable is normally distributed. This can be tested by Pearson's chi-square statistic or the likelihood-ratio statistic also called the deviance, provided that data can be sorted by the explanatory variables. For each combination of the latter, the observed proportion of censored observations are compared with the estimates of the corresponding theoretical proportion $p_{\mathbf{x}}$ defined by

$$p_{\mathbf{x}} = P(Y_{tj} \le a) = \Phi(u_{\mathbf{x}}), \text{ with } u_{\mathbf{x}} = \frac{a - \mu_{\mathbf{x}}}{v_{\mathbf{x}}} \text{ where } v_{\mathbf{x}} = \sqrt{V(Y_{tj})} \quad (3)$$

These tests are supplied by several statistical packages such as SAS (SAS Online Guide, 2006).

Below it is shown that the performance of Heckman's estimation procedure is dependent on the magnitude of the standardized variable $u_{\mathbf{x}}$ rather than on $\mu_{\mathbf{x}}$ or $\mathbf{x}$. In order to simplify the simulation studies (Sect. 3) it was therefore decided to consider just one explanatory variable, that was chosen as *t*, *t*=1,…,*T*, so the expressions in Eq. (2) simplifies to

$$(a)\ Y_{tj} = \alpha + \beta t + U_{tj},\ (b)\ Y_{tj} = A_j + \beta t + U_{tj},\ (c)\ Y_{tj} = A_j + B_j t + U_{tj} \quad (4)$$

with variances $V(Y_{tj}) = \sigma_U^2$ (a), $\sigma_U^2 + \sigma_A^2$ (b), $\sigma_U^2 + \sigma_A^2 + 2t\sigma_{AB} + t^2\sigma_B^2$ (c) and covariances $Cov(Y_{sj}, Y_{tj}) = 0$ (a), σ_A^2 (b), $\sigma_A^2 + (s+t)\sigma_{AB} + st\sigma_B^2$ (c).

2.2 Results on expectations of censored variables

2.2.1 Normally distributed censored variables

Let ϕ be the density of a standardized normal variable and consider the function

$$h_{\mathbf{x}} = \phi(u_{\mathbf{x}}) / (1 - p_{\mathbf{x}}) \tag{5}$$

This is often referred to as the inverse Mills ratio. Since $h_{\mathbf{x}}$ is the limit of $\delta^{-1} P\left(Y_{tj} \in (a, a+\delta) \middle| Y_{tj} > a\right)$ as $\delta \to 0$ it can be interpreted as the hazard for approaching the censoring limit a for a given vector $\mathbf{x}_t$. The behaviour of $h_{\mathbf{x}}$ as a function of $u_{\mathbf{x}}$ is seen in Figure 1. Notice that $h_{\mathbf{x}}$ is roughly linear when $u_{\mathbf{x}}$ is large. From the inequality $u_{\mathbf{x}} < h_{\mathbf{x}} < u_{\mathbf{x}} + 1/u_{\mathbf{x}}$ (Gordon, 1941), it follows that the asymptotic slope for large $u_{\mathbf{x}}$ is 1. In Figure 1 the range of $u_{\mathbf{x}}$ is from -2 to 2. The latter corresponds to a range of the censoring proportion from 2.3 % to 97.7 % and this will cover most situations that occur in practice.

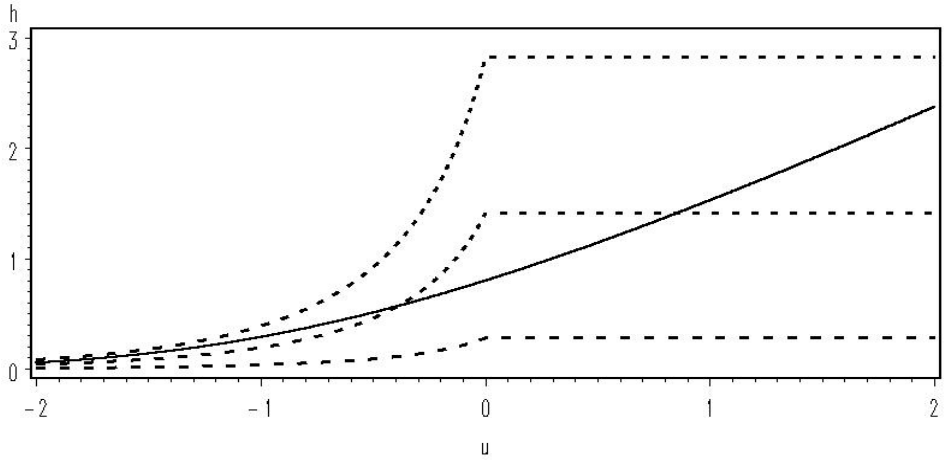

Figure 1. The solid line is the hazard in Eq. (5) (normal observations). The three dotted lines are the hazards for Laplace distributed observations (cf. Section 2.2.2) with v = 0.5 (upper curve), v =1.0 and v =5.0 (lower curve).

The expectation of the Y_{tj}'s that are found above a is related to $\mu_{\mathbf{x}}$ in the following way (Johnson *et al*, 1994)

$$E\left(Y_{tj} \middle| Y_{tj} > a\right) = \mu_{\mathbf{x}} + v_{\mathbf{x}} \cdot h_{\mathbf{x}} \tag{6}$$

As Heckman noticed, the latter relation makes it possible to obtain estimates of the parameters in $\mu_{\mathbf{x}}$ by regressing $\left(Y_{tj}\middle|Y_{tj} > a\right)$ on the explanatory variables and on the estimated hazard. The expectation of the observed variable Y_{tj}^{*} can finally be obtained by putting Eq. (6) into the obvious relation

$$E\left(Y_{tj}^{*}\right) = a \cdot p_{\mathbf{x}} + E\left(Y_{tj}\middle|Y_{tj} > a\right) \cdot (1 - p_{\mathbf{x}}) \tag{7}$$

All these results are based on the assumption of normality of the censored variables and the two-step procedure described above would therefore be termed *normal-Heckit*. Below (Sect.3) it will be found that, if the normal-Heckit is applied to data that are not normally distributed, it may collapse.

2.2.2 Non-normally distributed censored variables: *The Laplace distribution*

Under normality assumptions the hazard $h_{\mathbf{x}}$ is separated from $\mu_{\mathbf{x}}$ in Eq. (6) in an additive way. For other distributions this decomposition is seldom possible. Consider e.g. the case when the Y_{tj} 's in the GM model (4a) has the Laplace (or double exponential) distribution with the following density $f(y)$ and cdf $F(y)$:

$$f(y) = \frac{1}{2\sigma} \cdot \begin{cases} \exp(-z), \text{if } z \geq 0 \\ \exp(z), \text{if } z \leq 0 \end{cases} \quad F(y) = \begin{cases} 1 - \exp(-z)/2, z \geq 0 \\ \exp(z)/2, \quad z \leq 0 \end{cases}, \text{ with } z = \frac{y - \mu_{\mathbf{x}}}{\sigma}.$$

The expectation and variance of Y_{tj} is $\mu_{\mathbf{x}}$ and $2\sigma^2$, respectively (cf. Johnson *et al*, 1994). The normal density and the Laplace density are both symmetric around $\mu_{\mathbf{x}}$ but compared to the normal density the Laplace density has a sharper peak at $\mu_{\mathbf{x}}$ and longer tails. In terms of $u_{\mathbf{x}}$ defined in Eq. (3), the hazard for approaching the censoring limit *a* is

$$h_{\mathbf{x}} = \begin{cases} \left[\sigma\left(2\exp(-u_{\mathbf{x}}\sqrt{2}) - 1\right)\right]^{-1}, \text{for } u_{\mathbf{x}} \leq 0 \\ \sigma^{-1} \qquad\qquad\qquad\qquad\quad , \text{for } u_{\mathbf{x}} \geq 0 \end{cases} \tag{8}$$

This function is shown in Figure 1 for $v = \sigma\sqrt{2}$ = 0.5, 1.0 and 5.0. When $u_{\mathbf{x}} \leq 0$ the hazard is increasing and for some values of *v* the hazard is rather close to that of the normal distribution. For $u_{\mathbf{x}} \geq 0$ the hazard is completely different and is identical to the hazard of the exponential distribution with a constant level. It also follows that

$$E\left(Y_{tj}\middle|Y_{tj} > a\right) = \frac{\int_a^{\mu_{\mathbf{x}}} yf(y)dy + \int_{\mu_{\mathbf{x}}}^{\infty} yf(y)dy}{1 - \frac{1}{2}\exp(-(\mu_{\mathbf{x}} - a)/\sigma)} = \mu_{\mathbf{x}} + h_{\mathbf{x}}\left(\sigma(\mu_{\mathbf{x}} - a) + \sigma^2\right) \text{for } a \leq \mu_{\mathbf{x}}$$

$$E\left(Y_{tj}\middle|Y_{tj} > a\right) = \frac{\int_a^{\infty} yf(y)dy}{\frac{1}{2}\exp(-(a - \mu_{\mathbf{x}})/\sigma)} = a + \sigma, \text{for } a \geq \mu_{\mathbf{x}}$$

In the last expressions $\mu_{\mathbf{x}}$ and $h_{\mathbf{x}}$ can not in general be expressed in separate terms as in Eq. (6). Only when a equals $\mu_{\mathbf{x}}$ they have the same structure.

Thus, if the normal-Heckit is applied to data where the censored variable in fact is Laplace distributed, estimates can be expected to be very unreliable for two reasons. First, estimates of the hazard are uncertain since the form of the hazard is incorrectly specified and second, the hazard is not additively separated from $\mu_{\mathbf{x}}$, so the regression relation is incorrectly specified in Heckman's second step.

2.3 Heckman's two-step procedure

The first step in Heckman's procedure is to estimate the hazard in the definition (5), and this in turn requires the estimates of $p_{\mathbf{x}}$ or $u_{\mathbf{x}}$ in Eq. (3). The most basic way to estimate $p_{\mathbf{x}}$ is to count the number of observations that falls below a for a given $\mathbf{x}_t$ out of a total of $n_{\mathbf{x}}$. This suggests the estimator

$$\hat{p}_{\mathbf{x}} = \text{Proportion observations} < a \text{ at } \mathbf{x}_t \text{ and from this } \hat{u}_{\mathbf{x}} = \Phi^{-1}(\hat{p}_{\mathbf{x}}) \qquad (9a)$$

The estimator of the hazard that is based on Eq. (9a) will be termed semi-parametric. In practise the latter is only feasible when the model has a small number of explanatory variables, each with a limited state space. Alternatively one can perform a probit analysis that fits the relation in Eq. (3) to data. In this way one gets estimators of $(a-\alpha)/v_{\mathbf{x}}$ and $\boldsymbol{\beta}/v_{\mathbf{x}}$ (being of less value when $v_{\mathbf{x}}$ is unknown), but also of $p_{\mathbf{x}}$ and $u_{\mathbf{x}}$,

$$\hat{p}_{\mathbf{x}} \textit{ and } \hat{u}_{\mathbf{x}} \textit{ from probit analysis} \qquad (9b)$$

The latter estimator will be termed probit-based. The essential difference between the two types of estimators is that the one in (9b) makes full use of the normality assumption, while that in (9a) only uses the normality assumption for estimating the numerator in the definition (5). The estimates of α and $\boldsymbol{\beta}$ are finally obtained in the second step by regressing $\left(Y_{tj}\middle|Y_{tj} > a\right)$ on $\mathbf{x}_t$ and on the estimated hazard $\hat{h}_{\mathbf{x}}$.

In Figure 1 $h_{\mathbf{x}}$ is roughly linear for large values of $u_{\mathbf{x}}$, say $h_{\mathbf{x}} \approx \lambda + \theta \cdot u_{\mathbf{x}}$, where $\theta \in (0,1)$ and $\lambda < 0$. Putting this into Eq. (7) and using Eq. (3) gives

$$E\left(Y_{tj}\middle|Y_{tj} > a\right) \approx \left[\alpha(1-\theta) + v_{\mathbf{x}}\lambda + \theta \cdot a\right] + \mathbf{x}_t'\boldsymbol{\beta} \cdot (1-\theta) \qquad (10)$$

From this it is obvious that estimates of α and $\boldsymbol{\beta}$ can be seriously biased by performing the second step in Heckman's procedure since one is estimating the slope vector $\boldsymbol{\beta}(1-\theta)$ rather than $\boldsymbol{\beta}$. Provided that $\boldsymbol{\beta}(1-\theta)$ is estimated without

bias, it follows that $-\theta$ can be interpreted as the relative bias of the $\boldsymbol{\beta}$-components. If θ is known this can be used to adjust for the bias when estimating $\boldsymbol{\beta}$ by simply dividing the estimate by $(1-\theta)$. An example of this will be given in Section 4.3

2.4 Specific problems to be considered

The theoretical exposition above raises some questions that will be dealt with in the next section:
(i) Which are the properties of the semi-parametric and the probit-based estimates of the hazard under normal- and non-normal distributional assumptions? (ii) For which range of $u_{\mathbf{x}}$-values, or alternatively for which censoring proportions, are estimates obtained by Heckman's procedure reliable? (iii) Under which of the three random structures, GM, ECR and RCR, are estimates obtained by Heckman's procedure reliable?

3. Monte Carlo simulations

3.1 Design of the simulation study

Data were generated according to the three models in (4) with $E(Y_{tj}) = \mu_t = \alpha + \beta \cdot t$, t = 1,2,3,4 and $V\left(Y_{tj}\right) = v^2$ with $v^2 = \sigma_U^2$ for GM data and $v^2 = \sigma_U^2 + \sigma_A^2$ for ECR data. For RCR data the variance depends on t, $V(Y_{tj}) = v_t^2 = \sigma_U^2 + \sigma_A^2 + 2t\sigma_{AB} + t^2\sigma_B^2$. The censoring limit was a = 0 and the Heckit was studied within the ranges $u_t \in [-2,0] = I_-$, $u_t \in [-1,1] = I_0$ and $u_t \in [0,2] = I_+$.

For GM and ECR data the parameters were $\beta = -10, -30$ and $v = -3\beta/2$ (=15, 45). For $u_t \in I_-$ $\alpha = -4\beta$ (=40, 120) yielding $u_t = (2t-8)/3$. For $u_t \in I_0$ $\alpha = -5\beta/2$ (=25, 75) yielding $u_t = (2t-5)/3$, and for $u_t \in I_+$ $\alpha = -\beta$ (=10, 30) yielding $u_t = 2(t-1)/3$. The expected proportion censored observations was : 0.22 for $u_t \in I_-$, 0.50 for $u_t \in I_0$ and 0.78 for $u_t \in I_+$.

For ECR data two sets of variance components were used $(\sigma_U^2 = 200, \sigma_A^2 = 25)$ and $(\sigma_U^2 = 25, \sigma_A^2 = 200)$ giving $v = 15$, and furthermore $(\sigma_U^2 = 1772, \sigma_A^2 = 253)$ and $(\sigma_U^2 = 253, \sigma_A^2 = 1772)$ giving $v = 45$. Since v_t^2 depends on t in the RCR model it is not possible to find parameter values such that $V\left(Y_{tj}\right)$ is exactly the same as for the GM- and ECR data. The following parameter choices made the results for the RCR model roughly comparable with the former models: $\beta = -10, \sigma_U^2 = 25, \sigma_A^2 = 200,\ \sigma_B^2 = 10$. For $u_t \in I_-$, $\sigma_{AB} = -18.45$, so v_t varied between 14.1 and 15.4 and for $u_t \in I_+$, $\sigma_{AB} = -31.55$ with v_t varying between 11.5 and 13.1.

Simulations were also performed to study the performance of the normal-Heckit when in fact the observations with GM data were Laplace distributed. Three cases were considered: (*i*) $v = 0.5, \beta = -0.25$, (*ii*) $v = 1, \beta = -0.5$, (*iii*) $v = 5, \beta = -2.5$. For $u_t \in I_-$ $\alpha = -4\beta$ giving $u_t = (t-4)/\sqrt{2}$, *t*=1,2,3,4). For $u_t \in I_+$ $\alpha = -\beta$ giving $u_t = (1-t)/\sqrt{2}, t = 1,2,3,4$. The hazards for these three values of *v* are shown in Figure 1.

Estimates of p_t and u_t that are required in order to estimate the hazard h_t in the first step of Heckman's procedure were obtained from probit analysis. Based on the results from a preparatory study of the bias of the estimated hazard outlined below, the sample sizes were chosen as *n* = 100 and 400 when studying bias and variance of the α and β estimates. All simulations were performed with 10,000 replicates, using random number functions and procedures in SAS version 9.1. A computer program is available from the author on request.

3.2 The estimated hazard

The bias of the estimated hazard h_t was studied at *t* = 1, 2, 3, 4 when data were generated by the GM model with normally distributed disturbances. For both estimators in (9a) and (9b) the bias decreased rapidly with increasing *n*. For small *n*, the bias could be substantial, especially for $u_t \in I_+$ and *t* = 4. However, it was concluded that for practical purpose when estimating h_t, the bias could be ignored when *n* is 100 or larger. The same conclusions were drawn about the variances of the h_t estimates. Here the probit-based estimator had a slightly smaller variance and the variance decreased more rapidly than the bias with increasing *n*. A similar pattern was obtained for the ECR and RCR models. So, under normality assumptions the probit-based estimator is at least as good as the semi-parametric estimator, and for *n*=100 or larger the influence from bias can be ignored and the variance remains small.

Now, consider the case when the disturbances are Laplace-distributed. The absolute relative bias was smallest for $v = 1$ With increasing *n* the bias persisted and the variance decreased. The latter was more than five times larger for *n* = 100 than for *n* = 400. The results show that both proportional-based and probit-based estimators of the hazard can be seriously biased if the hazard is far from that of the normal and this can not be compensated for by increasing *n*.

In the sequel, when the properties of estimates of β and α are studied under normality, *n* is chosen as 100 and 400. From the results above it follows that possible biases of the estimates can not be caused by poor estimates of the hazard in the first step in the Heckit, but purely on the fact that $\mu_{\mathbf{x}}$ and $h_{\mathbf{x}}$ in Eq. (6) are both linear which in turn leads to the structure in Eq. (10).

Since the Heckit is so closely tied up with normality it was furthermore studied whether two commonly used tests of normality for censored data, Pearson's chi-square and maximum likelihood-ratio (SAS Online Guide, 2006), were able to detect deviations from normality. When the observations were

Laplace distributed ($\nu = 0.5, \beta = -0.25, \alpha = -4\beta$) it was found that the p-values of both tests were roughly the same. However, for n=100 only 20 % of the p-values were below 0.10 (the recommended significance level) and 36 % were below 0.20. For n=400, 58 % of the p-values were lower than 0.10 and 72 % were lower than 0.20. It is beyond the scope of this paper to go into details about these tests, but it is clear that the powers of the tests are unsatisfactory low when the alternative to the normal distribution is that of Laplace and $n \leq 400$.

3.3 Estimates of β and α

Tables 1a and 1b summarize the properties of the β and α estimates when the Heckit was applied to GM data. Both bias and variance of the estimates increased as the range of the u_t values moved upwards, and decreased with increasing n. Especially for $u_t \in I_+$, bias and variance were considerable, up to 15 times larger than for $u_t \in I_-$. As expected, both bias and variance was larger for $\beta = -30$ than for $\beta = -10$ since the former value makes $V(Y_{tj})$ larger. However, it is interesting that the absolute relative bias turned out to be independent of the magnitude of β for given n and a given range of u_t.

Table 1a. Relative bias (%) with the GM model.

		Relative bias of $\hat{\beta}$			Relative bias of $\hat{\alpha}$		
β	n	I_-	I_0	I_+	I_-	I_0	I_+
-10	100	5	28	71	3	6	61
	400	0.3	4	53	4	8	51
-30	100	5	29	70	3	5	60
	400	0.6	5	50	4	4	50

Table 1b. Variances with the GM-model.

		Variance of $\hat{\beta}$			Variance of $\hat{\alpha}$		
β	n	I_-	I_0	I_+	I_-	I_0	I_+
-10	100	19	128	289	19	27	84
	400	2.9	34	270	3.7	5.9	88
-30	100	163	2282	3316	166	298	1188
	400	26	392	2033	34	65	679

Similar results, when the Heckit was applied to ECR data, are seen in Tables 2a and 2b. Bias and variance were roughly the same as for the GM data. For $u_t \in I_-$ and $u_t \in I_0$ bias and variance of the β -estimator were larger when the ratio σ_A^2 / σ_U^2 is large. As for the GM model, the absolute relative bias seemed to be roughly independent of the magnitude of β .

Table 2a. Relative bias (%) with the ECR-model. The first and second figures represent the cases when σ_A^2 / σ_U^2 is small and large, respectively.

		Relative bias of $\hat{\beta}$			Relative bias of $\hat{\alpha}$		
β	n	I_-	I_0	I_+	I_-	I_0	I_+
-10	100	6, 14	33, 36	69, 61	2, 2	5, 11	71, 90
	400	0.5, 2	6, 5	53, 51	4, 3	8, 8	55, 77
-30	100	6, 12	32, 33	67, 60	2, 2	5, 10	67, 90
	400	0.4, 1	6, 16	53, 47	4, 3	8, 9	57, 73

Table 2b. Variances with the ECR-model. The upper and lower figures in the cells represent the cases when σ_A^2 / σ_U^2 is small and large, respectively.

		Variance of $\hat{\beta}$			Variance of $\hat{\alpha}$		
β	n	I_-	I_0	I_+	I_-	I_0	I_+
-10	100	29	161	298	24	23	135
		97	196	291	29	30	411
	400	3.6	43	233	4.0	4.7	93
		14	89	179	4.1	9.7	250
-30	100	228	1286	3581	194	222	1310
		708	1683	2159	235	229	3165
	400	34	504	3851	37	45	1446
		142	993	2071	39	107	3012

Tables 3a and 3b show the pattern for the RCR data. Compared with the results in the Tables 1 and 2, bias and variance are smaller.

Table 3a. Relative bias (%) with the RCR-model.

	Bias of $\hat{\beta}$		Bias of $\hat{\alpha}$	
N	I_-	I_+	I_-	I_+
100	-5	46	4	57
400	-5	30	5	53

Table 3b. Variances with the RCR-model.

	Variance of $\hat{\beta}$		Variance of $\hat{\alpha}$	
n	I_-	I_+	I_-	I_+
100	1.7	230	3.6	33
400	0.34	212	0.91	34

From Tables 1-3 it is concluded that the Heckit works quite well for $u_t \in I_-$, (22 % censored) and is less good when $u_t \in I_0$ (50 % censored), especially regarding bias of the β estimator. For $u_t \in I_+$ (78 % censored), Heckman's procedure is very poor but seems to perform slightly better with RCR data.

In Section 2.3 it was noticed that the absolute relative bias when estimating the $\boldsymbol{\beta}$-components can be expressed by θ in Eq. (10). Since θ in Tables 3-5 is roughly independent of the magnitude of β and thus also of ν and only dependent on n and on the censoring proportion p, it is challenging to search for a relation that describes how θ depends on n and p. From the results in Tables 1 and 2 (GM- and ECR data) the following relation was established,

$$\theta = p^{\Psi\sqrt{n}} \tag{11}$$

where $\Psi = 0.1966$ (GM), 0.1791 (ECR with σ_A^2 / σ_U^2 small), 0.1324 (ECR with σ_A^2 / σ_U^2 large). The constant Ψ was determined by fitting the linearized version of Eq. (11) to the estimates obtained in Tables 1-2 by ordinary least squares. The coefficient of determination (R^2) ranged from 99.3 % to 99.8 %. The relation in Eq. (11) is illustrated in Figures 2a,b. From Figure 2a it is concluded that when n = 1000 or larger the censoring proportion p has less impact on the magnitude of θ as far as p is below 50 %. E.g. n = 1000 and p = 0.5 gives $\theta = 0.01$. If the censoring proportion is small, say below 20 %, then Figure 2b tells us that the absolute relative bias can be ignored for sample sizes above 250. However, for large p and small n the absolute relative bias can be substantial.

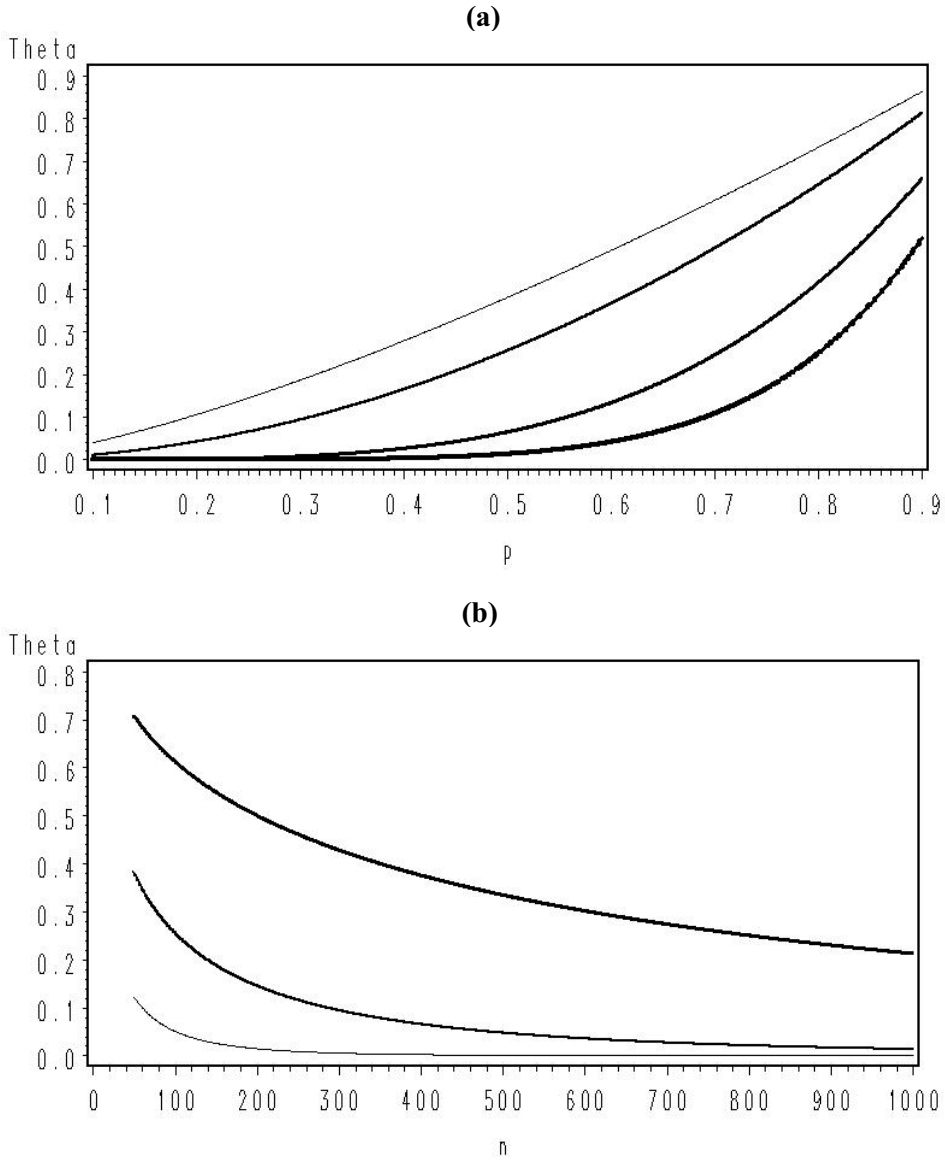

Figure 2. Illustration of the dependency of the absolute relative bias θ (Theta) on p and n in Eq. (11) when $\Psi = 0.1966$ (GM-model). (a) The upper to the lower curves show the dependency for n =50, 100, 400 and 1000. (b) The upper to the lower curves shows the dependency for the censoring proportions p =0.78, 0.50 and 0.22.

Since θ can be estimated from data by means of Eq. (11) it is possible to remove a great part of the bias by dividing the β estimate obtained from the second step in the Heckit by $(1-\theta)$ (cf. Eq. (10)). This was also confirmed in simulation experiments where the absolute relative bias was about three times smaller after the adjustment. A similar adjustment for bias when estimating α requires an estimate of v. Although v is an estimable parameter in the second

step of Heckman's procedure, the estimates of the latter seems to be extremely unreliable. In the simulation study the estimates of v had a serious negative bias and the variances of the v-estimates were 5-15 times larger than the variance of $\hat{\beta}$. For this reason no attempt was made to adjust for bias of the α parameter.

The (normal-)Heckit estimates of α and β was furthermore studied when the disturbances in fact were Laplace distributed using the parameters v = 0.5, 1.0, 5.0. The corresponding hazards are shown in Figure 1. For $u_t \in I_-$ it is concluded from Table 4a that for given n the absolute relative bias of the estimates are roughly the same for the three values of v. Whith increasing n much of the bias persists and the variances are reduced. A comparison between Table 4a and Table 1a for $u_t \in I_-$ shows that absolute relative bias is very much the same for n = 100. The difference is that in Table 1a, where the Heckit is applied to normally distributed observations, the bias is reduced much more for n = 400. The normal-Heckit seems yet to be surprisingly robust for Laplace distributed observations provided that $u_t \leq 0$. On the other hand, for $u_t \geq 0$ it is seen from Table 4b that the normal-Heckit collapses with Laplace distribute data.

Table 4a Relative bias (%) and variance of estimates obtained by the normal-Heckit when in fact the data are Laplace distributed with $u \in I_-$.

n	β	v	Relative bias of $\hat{\beta}$	Variance of $\hat{\beta}$	Relative bias of $\hat{\alpha}$	Variance of $\hat{\alpha}$
	-0.25	0.5	5	0.02	4	0.01
100	-0.5	1.0	5	0.07	5	0.04
	-2.5	5.0	6	3.20	5	1.85
	-0.25	0.5	3	0.00	5	0.00
400	-0.5	1.0	3	0.01	5	0.01
	-2.5	5.0	2	0.18	5	0.18

Table 4b Relative bias (%)and variance of estimates obtained by the normal-Heckit when in fact the data are Laplace distributed with $u \in I_+$.

n	β	v	Relative bias of $\hat{\beta}$	Variance of $\hat{\beta}$	Relative bias of $\hat{\alpha}$	Variance of $\hat{\alpha}$
	-0.25	0.5	94	1.09	36	0.50
100	-0.5	1.0	100	2.33	41	1.43
	-2.5	5.0	101	45.42	41	41.54
	-0.25	0.5	97	0.30	38	0.33
400	-0.5	1.0	99	1.14	40	0.96
	-2.5	5.0	100	13.90	42	14.74

It is interesting to compare these results with those obtained by Paarsch (1984). Here the normal-Heckit was applied to Laplace distributed observations using two sets of parameters: $\alpha = -2.94, \beta = 1, v = 10$ giving $u_t = (2.94 - t)/10$ for $t = 0,1,\ldots 20$ and $u_t \in (-1.706, 0.294)$ (25 % censoring) and $\alpha = -10$ and same β and v giving $u_t = 1 - t/10$ and $u_t \in (-1,1)$ (50 % censoring). For $n = 100$ the relative bias of the β-estimator was found to be 32 % (25 % censoring) and 68 % (50 % censoring). Although these figures were based on simulations with only 100 replicates, they agree well with the results in this paper.

3.4 Comparison between the efficiency obtained with censored and uncensored data

When data are censored it is obvious that some information is lost when estimating the parameters. Although this is inevitable it may be of some interest to compare the variances in Tables 1-3 with those that are obtained with uncensored data. Such a comparison may be considered to be of purely academic interest, but one reason for doing it is to set up a standard that allows for comparisons between the normal-Heckit and alternative methods. Let the optimal estimator of β with uncensored data be $\hat{\beta}_{OPT} = \sum_{j=1}^{n} \hat{\beta}_j$, where $\hat{\beta}_j = w_{tY} / w_{tt}$ with $w_{tY} = \sum_{t=1}^{T} (t - \bar{t})(Y_{tj} - \bar{Y}_j), w_{tt} = \sum_{t=1}^{T} (t - \bar{t})^2$ (cf. Rao, 1965, Ch. IV in Swamy, 1971 and Ch. 3 in Hsiao). Then $V\left(\hat{\beta}_{OPT}\right) = \sigma_U^2 / n w_{tt}$ for the GM and ECR models, and $V\left(\hat{\beta}_{OPT}\right) = \left(\sigma_B^2 + \sigma_U^2 / w_{tt}\right) / n$ for the RCR model. From this one obtains the relative efficiency $\mathrm{RE} = 100 \cdot V(\hat{\beta}_{OPT}) / V(\hat{\beta}_{Heck})$, where $V(\hat{\beta}_{Heck})$ is the variance of $\hat{\beta}$ obtained from the Heckit and is determined from the simulations. For $u_t \in I_0$ and $u_t \in I_+$ the relative efficiency is below 1 % for all three models. But for $u_t \in I_-$, RE is 11.0 % when *n*=400 and 8.8 % when *n*=100 for the RCR-model, compared with RE of 3.4 % (*n*=400) and 2.4 % (*n*=100) for the GM-model. Also from this point of view, Heckman's procedure seems to produce the best estimates when it is applied to the RCR-model.

4. Using the Heckit for analysing recurrence of lower back problems among sick-listed men

4.1 Background

In 1993 the International Social Security Association initiated the Work Incapacity and Reintegration project, primarily because of high levels of

expenditure on sickness in many industrialized countries (Hansson and Hansson, 2000). In the Swedish part of the project sick-listed men and women due to lower back or neck problems were followed during 2 years. One purpose of the study was to analyze the effects of commonly practiced medical interventions upon work resumption. The Swedish data base also contains information about the person's health during a further 2-year period after the 2-year follow-up. Results from this post follow-up period have not been published elsewhere. Of special interest was to study the number of sick-listed days during the post follow-up due to the same diagnosis as in the follow-up.

4.2 The post follow-up

Data from the post follow-up will be used to illustrate some undesirable consequences of the Heckit. $n = 203$ men with unspecified lower back diagnoses who had returned to work within the follow-up period were observed during the post follow-up. Men with specific back diagnoses (about 10 % of all cases, Bergendorff et al. 2001, p. 46) were excluded since these had back surgery and were thereafter free from back problems with the same diagnosis. The dependent variable of interest is DAYS = 'Number of sick-listed days during the post follow-up due to the same diagnosis as in the follow-up'. One important explanatory variable was EQT = 'Value on EuroQol Thermometer scale', obtained at the end of the 2-year follow-up. The latter is a health-related quality of life measure obtained from a visual scale on which the respondent is asked to mark his health from 0 (worst function) to 100 (best function) (Hansson et. al., 2005). The variable EQT was negatively associated with DAYS. Another explanatory variable was STATE1Y (= 1 if the person had returned to work within 1 year during the previous follow-up, and = 0 otherwise). Rather unexpectedly, there was a significant positive association between *not* returning to work within 1 year and DAYS = 0 (p-value= 0.01, Chi-square test). In fact, 89 % (31/35) of those who did not return within 1 year had zero days during the post follow-up period, while the corresponding figure for those who returned within 1 year was 68 % (115/168). No further explanatory variables, such as demographic and socio-economic factors, work environment, co-morbidity and treatment received, were found to be associated with DAYS.

The major part of the observations are found on the border DAYS = 0, and it is obvious that the standard conditions for performing a regression analysis, such as normality or at least symmetrically distributed disturbances, are violated. Therefore, a latent variable Y is introduced such that

$$DAYS = \begin{cases} 0, \text{if } Y \le 0 \\ Y, \text{if } Y > 0 \end{cases}$$

and Y is a variable that is related to a person's state of health. It is assumed that for the j : th person, $Y_j = \alpha + \beta_1 \cdot STATE1Y + \beta_2 \cdot EQT + U_j$, $j = 1,\ldots,203$.

4.3 Applying Heckman's two-step approach

Below the data is analyzed by the Heckit and in order to clarify the different steps they are numbered *(i)-(iii)*.

(i) Estimation of $h_{\mathbf{x}}$ *in* (5) *by means of probit analysis*

The probit model is

$$p_{\mathbf{x}} = P(Y_j \le 0) = \Phi(u_{\mathbf{x}}), u_{\mathbf{x}} = \theta_0 + \theta_1 \cdot STATE1Y + \theta_2 \cdot EQT$$

where $\theta_0 = -\alpha / v, \theta_1 = -\beta_1 / v, \theta_2 = -\beta_2 / v$. The fit of the model was tested by Pearson's chi-square statistic and the Maximum Likelihood Ratio (MLR) statistic, giving the p-values 0.33 and 0.20, respectively, so the probit model should not be rejected at the 10 % level. The estimates that were obtained from the probit analysis were $\hat{\theta}_0 = 0.5649, \hat{\theta}_1 = -1.1037, \hat{\theta}_2 = 0.0148$. The observed censoring proportion was 146/203 = 0.72. Much of the *u*-range is located to the part where the hazard is roughly linear, especially for STATE1Y = 0, where $u_{\mathbf{x}}$ ranges from 0.56 to 2.04. The range of the $u_{\mathbf{x}}$ -values indicates that the Heckit may give unreliable estimates (cf. Section 2.3).

(ii) Regressing $(Y_j | Y_j > 0)$ *on* $\mathbf{x}'$ = (STATE1Y, EQT) *and* $h_{\mathbf{x}}$

The estimated regression relation in Eq.(6), by using OLS, is

$$\hat{E}(Y_j | Y_j > 0) = -3563 + 2788 \cdot \text{STATE1Y} - 37.5 \cdot \text{EQT} + 3095 \cdot \hat{h}_{\mathbf{x}} \qquad (12)$$

Here all estimated coefficients are significantly different from zero at the 5 % level as judged by two sided t-tests.

(iii)Calculation of expected number of sick-listed days during the post follow-up, according to Eq.(7).

The expected number of sick-listed days is $\hat{E}(DAYS) = \hat{E}(Y_j | Y_j > 0) \cdot (1 - \hat{p}_{\mathbf{x}})$. Here the first factor is given in Eq. (12) and an estimate of $p_{\mathbf{x}}$ is obtained from the estimated probit model. The estimates have little in common with the actual data. E.g. at EQT = 20, $\hat{E}(DAYS)$ is about 800, but in the actual data no one had more than 650 days. From Eq. (12) the estimate of θ is $(0.72)^{0.1966\sqrt{203}} = 0.40$, i.e. the β -coefficients have been estimated with an absolute relative bias of 40 %. This figure can be used to correct for the bias of the β -parameters by using Eq. (10):

$$\hat{\beta}_1(1-\theta) = \hat{\beta}_1(1-0.40) = 2788 \Rightarrow \hat{\beta}_1 = 4647$$

$$\hat{\beta}_2(1-\theta) = \hat{\beta}_2(1-0.40) = -37.5 \Rightarrow \hat{\beta}_2 = -62.5$$

5. Conclusions and suggestions for further research

This paper has studied the performance of Heckman's two-step approach when it is used to solve the problem with border-observations without selection effects and when data are censored from below. From the simulations it was concluded that the Heckit performed quite well for n larger than 100 and when the censoring proportion was 0.22, provided that the censored variable was normally distributed. With increasing censoring proportion the estimates gradually became more biased and the variance increased. However, it is possible to compensate for this by increasing the sample size.

By means of Eq. (11) it is possible to estimate θ, the absolute relative bias of the β-estimates, and to adjust for the bias in the way that was done in Section 4.3. Eq. (11) can also be used in the planning of a study. By first taking a pilot sample one gets a rough estimate of the censoring proportion p. The final proper sample size n can then be determined from restrictions on θ. E.g. if it is required that θ is at most 1 % for the GM model, then n should be at least 62 if $p = 0.05$ and at least 1142 if $p = 0.50$. From considerations of space Eq. (11) had to be considered for two special cases of the ECR model. This gives some practical guide lines, but more detailed studies should be performed on the effect of the variance ratio upon the relation in Eq. (11).

Since the Heckit inevitably gives more or less biased estimates one should compare the estimated expectation of the observed variable with the observed data in a final step. A warning practical example was given in Section 4 where the censoring proportion was 0.72, leading to an estimated absolute relative bias of the regression estimates of 40 %, and this in turn led to gigantic over-estimates of the actual costs for sick-listing.

When the censored variable has a distribution that is not normal Heckman's two-step procedure may collapse for at least two reasons. One is that estimates of the hazard (or Mills ratio) used in the first step are biased. A second is that the regression function of interest and the hazard no longer are added to each other. From considerations of space the effects of misspecification was only studied for Laplace distributed disturbances, but such effects should be further investigated for a variety of distributions.

Acknowledgements

The author would like to thank two anonymous referees for their valuable comments. The research was supported by the National Social Insurance Board in Sweden (RFV), Dnr 3124/99 –UFU.

References

Dow, W.H. and Norton, E.C. (2003), Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions, *Health Services & Outcomes Research Methodology* 4, 5-18.

Flood, L. and Gråsjö, U. (2001), A Monte Carlo simulation study of a Tobit model, *Applied Economics Letters* 8, 581-584.

Gordon, R.D. (1941), Values of Mills' ratio of area to boarding ordinate and of the normal probability integral for large values of the argument, *Annals of Mathematical Statistics* 12, 364-366.

Hansson, T. and Hansson, E. (2000), The Effects of Common Medical Interventions on Pain, Back Function, and Work Resumption in Patients With Chronic Low Back Pain, *SPINE* 25, No 23, 3055-3064.

Bergendorff, S., Hansson, E., Hansson, T. and Jonsson, R. (2001), Vad kan förutsäga utfallet av en sjukskrivning? (Predictors of health status and work resumption) (in Swedish), Rygg och Nacke 8. Stockholm: RFVand Sahlgrenska Universitetssjukhuset.

Hansson, E, Hansson, T. and Jonsson, R. (2004), Predictors for work ability and disability in men and women with low-back or neck problems, accepted for publication in *European Spine Journal*.

Heckman, J. (1976), The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator of such models, *Annals of Economic and Social Measurement* 5, 475-492.

Heckman, J. (1979), Sample Selection Error as a Specification Error, *Econometrica* 47, 153-161.

Honda, Y. (1985), Testing the Error Components Model with Non-Normal Disturbances, *The Rewiev of Economic Studies* 52, 681-690.

Hsiao, C. (2003), *Analysis of panel data*, Cambridge University Press, Cambridge.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994), *Continuous univariatedistributions, vol I (2nd ed.)*, Wiley, New York.

Karlsson, M. (2005), *Estimators of Semiparametric Truncated and Censored Regression Models*, Statistical Studies 34, PhD thesis, Department of Statistics, Umeå University.

Kim, C.K. and Lai, T.L. (2000), Efficient score estimation and adaptive M-estimators in censored and truncated regression models, *Statistica Sinica* 10, 731-749.

Kruskal, W.H. and Tanur, J.M. (Ed.) (1978), *International Encyclopedia of Statistics, vol 2*, McMillan, New York.

Lee, M.J. (1996), *Method of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*, Springer, New York.

Lundevaller, E.H. and Laitila, T. (2002), Test of random subject effects in heteroscedastic linear models, *Biometrical Journal* 44, 825-834.

Nelson, F.D. (1984), Efficiency of the two-step estimator for models with endogenous sample selection, *Journal of Econometrics* 24, 181-196.

Newey, W.K. (2001), Conditional moment restrictions in censored and truncated regression models, *Econometric Theory* 17, 863-888.

Paarsch, H.J. (1984), A Monte Carlo comparison of estimators for censored regression models, *Journal of Econometrics* 24, 197-213.

Powell, J.L. (1994), Estimation of semiparametric models. In: Engel, R.F. and McFadden, D.L. (Eds.), *Handbook of Econometrics,* Vol 4, pp 2444-2521, North-Holland, Amsterdam.

Puhani, P.A. (2000), The Heckman correction for sample selection and its critique, *Journal of Economic Surveys* 14, No 1, 53-68.

Rao, C.R. (1965), The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves, *Biometrica* 52*,* 447-458.

Rosett, R.N. and Nelson, F.D. (1975), Estimation of the two-limit probit regression model, *Econometrica* 43, 141-146.

SAS Online Guide (2006), http://support.sas.com/91doc/getDoc/statug.hlp/probit_sect.5/htm.

Swamy, P.A.V.B. (1971), *Statistical inference in random coefficient regression model,* 55, Springer, Berlin.

Tobin, J. (1958); Estimation of relationships for limited dependent variables, *Econometrica* 26, 24-36.

Research Report

2007:2	Frisén, M.:	Optimal Sequential Surveillance for Finance, Public Health and other areas.
2007:3	Bock, D.:	Consequences of using the probability of a false alarm as the false alarm measure.
2007:4	Frisén, M.:	Principles for Multivariate Surveillance.
2007:5	Andersson, E., Bock, D. & Frisén, M.:	Modeling influenza incidence for the purpose of on-line monitoring.
2007:6	Bock, D., Andersson, E. & Frisén, M.:	Statistical Surveillance of Epidemics: Peak Detection of Influenza in Sweden.
2007:7	Andersson, E., Kühlmann-Berenzon, S., Linde, A., Schiöler, L., Rubinova, S. & Frisén, M.:	Predictions by early indicators of the time and height of yearly influenza outbreaks in Sweden.
2007:8	Bock, D., Andersson, E. & Frisén, M.:	Similarities and differences between statistical surveillance and certain decision rules in finance.
2007:9	Bock, D.:	Evaluations of likelihood based surveillance of volatility.
2007:10	Bock, D. & Pettersson, K.	Explorative analysis of spatial aspects on the Swedish influenza data.
2007:11	Frisén, M. & Andersson, E.	Semiparametric surveillance of outbreaks.
2007:12	Frisén, M., Andersson, E. & Schiöler, L.	Robust outbreak surveillance of epidemics in Sweden.
2007:13	Frisén, M., Andersson, E. & Pettersson, K.	Semiparametric estimation of outbreak regression.
2007:14	Pettersson, K.	Unimodal regression in the two-parameter exponential family with constant or known dispersion parameter
2007:15	Pettersson, K.	On curve estimation under order restrictions
2008:1	Frisén, M.	Introduction to financial surveillance