ELSEVIER

Agricultural Economics 11 (1994) 171–189

AGRICULTURAL ECONOMICS

Estimating random effects production function models with selectivity bias: an application to Swedish crop producers

Almas Heshmati

Department of Economics, University of Gothenburg, Viktoriagatan 30, S-411 25 Gothenburg, Sweden

Accepted 26 April 1994

Abstract

In this paper, the estimation of production functions and measurement of the rate of technical change is performed when selectivity bias is expected. A sample selection model consisting of a selection and a regression equation is estimated using Heckman's two-stage method. It is discussed in the context of a production function where the underlying technology is represented by a translog functional form. For the regression, a random effects model with heteroscedastic variances is assumed. This model and an alternative conventional model retaining heteroscedasticity without considering selectivity bias are estimated using the Generalized Least Squares method. The data used are a large rotating panel data set from Swedish crop producers over the period 1976–1988. The empirical results from the comparison between these two models show that the introduction of heteroscedasticity and the integration of sample selection in the production relationship is important. The impact of a correction for selectivity bias on the results, in terms of input elasticities and returns to scale is found to be significant.

1. Introduction

A farm utilizes a large number of inputs in order to produce a certain number of outputs in an often cyclical production process. The production of different types of outputs cannot proceed completely independently. The overall production process requires the joint utilization of some inputs. Since the process cannot be broken down into production stages and lines, we are facing a non-separable technology. In order for all inputs and outputs to be accounted for, some degree of aggregation is required. Depending on the level of aggregation, the multi-output problem is reduced to a single output or a few output problems. Another advantage of aggregation is that it reduces measurement error due to the non-separability of some inputs like capital. However, aggregation truncates the dependent variable to the production of a certain output. Sample selection bias may thus arise due to aggregation/truncation, which makes the sample non-random.

In collecting data from a population of micro units and taking into account the heterogeneous characteristics of the production units, stratified sampling is a desirable design used by data collecting agencies. Stratification is a key to the solution of the problem of how well the sample selected represents the population. It is based on a number of characteristics important to the de-

SSDI 0169-515000017-V

gree of heterogeneity such as location, size and specialization in the production of a certain number of outputs. This type of sampling is common in agricultural surveys. Depending on the purpose of the data collection, some conditions are imposed such that the units included in the population have certain characteristics, e.g. being a family farm. This will result in truncation of the population and possibly create selection bias.

In the real world, the structure of the population of farms, in terms of number (i.e. entry and exit), concentration, specialization etc is continuously changing. This implies the introduction of rotation in the sampling. In a rotating sampling design, the data collecting agency follows a procedure of dropping a fraction of the sample selected in previous years, replacing it with new farms from the population. Rotation of samples has the advantage of reducing the degree of non-response and it improves the quality of the data. New farms become part of the rotating sampling procedure by random while the exclusion of farms from the sample is non-random [1]. The non-random exclusion of farms from the sample is another possible source of sample selection bias. However, this type of non-randomness might be ignorable due the sampling design and the expected positive effect on sample representation of the real population.

Thus, a sample selection bias may emerge for three reasons: (a) inclusion of units specialized in the production of a certain output, (b) imposing conditions on units having certain characteristics to be included in the population, and (c) the non-random exclusion of individuals from the rotating sample. The importance of sample selection and the implication of selectivity bias has been discussed at the theoretical level and considered frequently in some areas, e.g. labor supply (for a survey, see Killingsworth and Heckman, 1986) but in the area of applied production studies selectivity bias seems to be neglected. A priori, farming should be an obvious case.

The main objective of this study is to investigate the production structure of Swedish crop production. This particular activity may be of little general interest. However, the contribution from the modeling side should be more relevant to a wider audience. First, I will take into account selectivity bias arising from non-randomness of the sample as presented above, by the integration of sample selection in the production relationship. A sample selection model is estimated using Heckman's two-stage method (Heckman, 1979). The more efficient Generalized Least Squares technique is used to estimate parameters of the model in the second step. Second, I will compare the sample selection and the conventional production function model where no correction for selectivity bias is undertaken. Third, I will use a rotating panel data model estimating this on data from Swedish crop producers over the period 1976–1988. I will compare the productive performance of the farms by estimating elasticities of output with respect to different inputs, returns to scale as well as the rate of technical change. Finally, I will estimate a production function with a generalized error component structure having heteroscedastic disturbances suggested by Baltagi and Griffin (1988). I will introduce some changes in the procedure used to estimate the variance components, that reduces the frequency of negative variances.

The rest of the paper is organized as follows. The sample selection model is outlined in Section 2. The econometric specification is set out in Section 3. In Section 3, I also discuss the estimation procedure and methods used. The empirical results along with a comparison of the performance of the different models are reported in Section 4. Finally, Section 5 presents the conclusions and a summary of this study.

2. Sample selection model

Assume that the farms use a multi-output joint production technology with some degree of specialization in the production of two jointly pro-

[1] Estimation issues related to use of rotating panel data are discussed in Biorn (1981), Biorn and Jansen (1983), Nijman et al. (1991), Kumbhakar and Heshmati (1991) and Heshmati (1994) and Heshmati et al. (1994).

ducible outputs, crop and non-crop. Let the stochastic model be expressed as:

$$Y_{it}^{*} = X_{it}'\beta + \epsilon_{it} \quad (1.a)$$

$$\epsilon_{it} = u_i + w_{it} \quad (1.b)$$

$$D_{it}^{*} = X_{it}'\beta + S_{it}'\zeta + \eta_{it} \quad (1.c)$$

in which $i = 1, 2, \ldots, N \quad t = t_i, t_i + 1, \ldots, T_i$ where Y_{it}^{*} is the output of crop, and D_{it}^{*} is a crop response indicator; the subscripts i and t denote, respectively, the production units and time periods; t_i and T_i are the first and the last time periods the ith farm is observed, respectively; X and S are matrices of explanatory variables, β and ζ are the vectors of the unknown parameters to be estimated; ϵ_{it} and η_{it} are error terms distributed i.i.d. $N(0,\sigma_e^2)$ and i.i.d. $N(0,\sigma_\eta^2)$, respectively; ϵ_{it} is composed of two components: u_i is the farm-specific effect, and w_{it} is the statistical noise.

Y_{it}^{*} is not directly observable, Y_{it} is its observable counterpart; Y_{it}^{*} is observed only if it is greater than a threshold, denoted by $Y_{\min}$. $Y_{\min}$ can be interpreted as the minimum level of revenue required for the farm to be considered as specialized in the production of crops. The following selection rule is used:

$$\begin{aligned} Y_{it} &= Y_{it}^{*} \quad \text{if } D_{it}^{*} > 0 \quad \text{or equivalently } Y_{it}^{*} > Y_{\min} \\ Y_{it} &= 0 \quad \text{otherwise} \end{aligned} \quad (2)$$

If we take the sample of crop producers only, the observation of Y_{it} is not random and will depend on ϵ_{it}. The decision whether to produce a crop or not is not random. It is made by individual farms. Thus, the application of regression models to the data by discarding the observation at the threshold will result in biased and inconsistent estimators of β. Estimation of the production function (1.a) must be done subject to the selection rule (2). The sample selection model consists of a selection equation governing the probability of observing the dependent variable and a regression equation based on observable observations only. The observability of Y_{it} is thus governed by a separate probit function (1.c), where $D_{it} = 1$ if $D_{it}^{*} > 0$, else $D_{it} = 0$.

In analyzing production relationships, the inputs used in the production process are not determinants of the probability of producing crops. The vector of explanatory variables in the selection equation consists of two subsets. One subset contains elements overlapping those entering the regression equation, X_{it}. The other subset of the explanatory variables entering the selection equation, S_{it}, does not enter the conditional expectation of Y_{it}.

In panel data literature, the estimation of the model according to the structure defined above has been developed in two directions. First, the fixed effects (FE) model, where u_i is assumed to be fixed and in general correlated with the regressors. Second, the random effects model (RE), where u_i is assumed to be random and uncorrelated with the regressors (see Hsiao, 1986) [2].

The main issues discussed frequently regarding random or fixed treatment of the effects are the efficiency, unbiasedness and consistency of the estimates. In empirical applications the RE models are frequently chosen. The main argument for this choice is that it allows a reduction in the number of parameters to only two, the mean and variance. Assuming that u_i is random also allows the inclusion of time invariant variables which vanish in FE models after the within transformation.

In this study, I use a random effect formulation. Generally, it would be desirable, if possible, to include the farm-specific and time-specific effects in the production function. However, variables reflecting managerial differences, weather conditions, etc are either measurable but have been ignored due to the lack of information or are not observable and consequently impossible to include in the estimation.

Consistent estimates of the model parameters can be obtained with Heckman's two-stage method. The first step involves estimation of the selection equation with probit so as to obtain

[2] Recent developments in the econometrics of panel data is surveyed by Baltagi and Raj (1992). The methods used in the estimation of limited dependent variable models with panel data related to the problems of FE vs. RE is surveyed by Maddala (1987). Further discussion of the estimation of fixed and random effects models with selectivity bias are found in Verbeek (1990) and Zabel (1992).

consistent parameter estimates. The estimates are used to estimate the Mill's Ratio, $\hat{MR}_{it} = \phi(\cdot)/\Phi(\cdot)$. To correct the selection bias MR is introduced in the second step as an extra explanatory variable in the regression or production function over non-zero observations. Thus, consistent estimates of (1.a) can be obtained if the following relation is estimated:

$$Y_{it} = X'_{it}\beta + \psi \hat{MR}_{it} + \epsilon_{it} \tag{3}$$

by OLS. The model in (3) differs from the model in (1.a) by the inclusion of the correction factor, $\hat{MR}_{it}$. The estimation method used in the second step is the Generalized Least Squares technique (GLS). GLS will result in more efficient parameter estimates provided there is no correlation between the effects and the X variables.

The estimates of the standard errors are biased and inconsistent. They can under- or overestimate their correct counterparts (Heckman, 1979; Greene, 1981)). The simplest way to estimate consistently the variance-covariance matrix is to use the White's (1980) robust estimator expressed as:

$$V\hat{\gamma} = (Z'Z)^{-1}Z'AZ(Z'Z)^{-1} \tag{4}$$

where $Z = (X, \hat{MR})$ is $N \times k + 1$ matrix of explanatory variables, $\gamma = (\beta, \psi)$ is $k + 1$ vector of unknown parameters. The diagonal matrix A is replaced by a matrix with diagonal elements (see Amemiya, 1985, p. 370):

$$\left[Y_{it} - X'_{it}\hat{\beta} - \hat{\psi}\hat{MR}_{it}\right]^2$$

3. Econometric model

We assume that the objective of a crop producer is to maximize profit. The farm follows a two-stage decision process where the input choice decision is made prior to the output decision. Thus, maximizing profit with given input and output prices is equivalent to maximizing output (see Kumbhakar and Hjalmarsson, 1993) [3]. To avoid strong a priori restrictions on technology, a flexible functional form, translog, is chosen (see Christensen et al., 1973). Since the Cobb-Douglas (CD) function is nested within the translog, the CD specification will be tested for. Thus, the production technology of the Swedish crop producers is represented by:

$$\begin{aligned} y_{it} = {} & \beta_0 + \Sigma_j \beta_j x_{jit} + \beta_t t + \psi \hat{MR}_{it} \\ & + \frac{1}{2}\left[\Sigma_j \Sigma_k \beta_{jk} x_{jit} x_{kit} + \beta_{tt} t^2\right] \\ & + \Sigma_j \beta_{jt} x_{jit} t + e_{it} \end{aligned} \tag{5.a}$$

$$e_{it} = u_i + w_{it} \tag{5.b}$$

where y is the log of the output of crops, x is the log of inputs defined as previously. The βs and ψ are parameters of the model to be estimated. We include time (t) as one of the explanatory variables representing the rate of exogenous technical change.

Returns to scale is measured by the elasticity of output with regard to a proportionate change in all inputs (the directional elasticity of the production function) and is equal to the sum of marginal elasticities, i.e. the elasticities of output with regard to the different inputs (see Forsund and Hjalmarsson, 1987, pp. 83–84):

$$\text{RTS} = \Sigma_j E_j \quad j = 1,2,\ldots,k \tag{6}$$

where

$$E_j = \partial y_{it}/\partial x_{jit} = \beta_j + \Sigma_k \beta_{jk} x_{kit} + \beta_{jt} t \tag{7}$$

If RTS is greater than, equal to, or less than one, then the corresponding returns to scale are increasing, constant, or decreasing.

The rate of technical change is conventionally defined as the partial derivative of the production function with regard to time, i.e.

$$E_t = \partial y_{it}/\partial t = \beta_t + \beta_{tt} t + \Sigma_j \beta_{jt} x_{jit} \tag{8}$$

The rate of technical change, E_t, can further be decomposed additively into pure technical change $(\beta_t + \beta_{tt} t)$ and non-neutral technical change $(\Sigma_j \beta_{jt} x_{jit})$ components. Technical change is defined as non-neutral if the passage of time affects the marginal rate of technical substitution between inputs.

[3] This is consistent with the fact that input and output prices are exogenous to the farms which is a reasonable assumption.

The modeling of heteroscedasticity may differ according to the way the farm-specific variances are defined. One alternative is to estimate a farm-specific variance for each farm. Treating u_i and w_{it} as random, using the Baltagi and Griffin (1988) approach, the following distributional assumptions on the heteroscedastic error components are imposed

(i) u_i is i.i.d. $N(0, \sigma_{u_i}^2)$
(ii) w_{it} is i.i.d. $N(0, \sigma_w^2)$ and
(iii) u_i and w_{it} are independent of each other and of the x variables.

Stacking the time-series observations for the ith farm in vector form and with the above distributional assumptions, the variance-covariance matrix of ϵ_i is

$$\Omega_i = E(\epsilon_i' \epsilon_i) = \sigma_{u_i}^2 J_{q_i} + \sigma_w^2 \mathrm{I}_{q_i} \tag{9}$$

Ω is a block diagonal matrix where q_i is the number of times the ith farm is observed; I_{q_i} is an identity matrix of order $q_i \times q_i$, and J_{q_i} is a $q_i \times q_i$ matrix with all elements equal to one. Thus the inverse of Ω_i is:

$$\Omega_i^{-1} = 1/\sigma_w^2 \left[\mathrm{I}_{q_i} - \left(\sigma_{u_i}^2 / \left(q_i \sigma_{u_i}^2 + \sigma_w^2 \right) \right) \mathrm{J}_{q_i} \right] \tag{10}$$

The GLS estimates of β are equivalent to the least square estimates when the following transformations are applied to the data:

$$\tilde{y}_{it} = y_{it} - \alpha_i \bar{y}_i \text{ and } \tilde{x}_{it} = x_{it} - \alpha_i \bar{x}_i \tag{11}$$

where $\bar{y}_i = T^{-1} \Sigma_t y_{it}$, $\bar{x}_i = T^{-1} \Sigma_t x_{it}$ and $\alpha_i = [1 - (\sigma_w / (q_i \sigma_{u_i}^2 + \sigma_w^2)^{0.5})]$

The model in (5.a) in vector form is rewritten as:

$$\tilde{y}_i = \tilde{Z}_i \beta + \tilde{e}_i \tag{12}$$

To accomplish the above data transformations, estimates of the unknown variance components $\sigma_{u_i}^2$ and σ_w^2 are to be obtained first. A two-step GLS estimation procedure is used. In the first step consistent estimates of the variance components are obtained. In the second step the estimated variance components are used to transform the data and perform the least squares method to the above transformed data.

The overall estimation procedure has the following steps:

(a) Mill's Ratio, MR_{it}, is unknown and must be estimated. We estimate the selection equation with probit to obtain a consistent estimate of the MR_{it}. MR_{it} is then introduced as an extra explanatory variable in the production function.

(b) Regress the within mean transformed y_{it} on the within mean transformed x_{it} and MR_{it} to get the within parameter estimates and the mean square error which is an unbiased and consistent estimator of σ_w^2.

(c) Ignore the individual farm effect and obtain the OLS residuals, $\hat{e}_{it}$, without any transformation and estimate $\mathrm{Var}(e_{it}) = \lambda_i^2 = \sigma_{u_i}^2 + \sigma_w^2$ from $\hat{\lambda}_i^2 = \Sigma_t [\hat{e}_{it}^2 / (q_i - k)]$ for each farm which is unbiased and consistent. Since we have a short panel with a large number of parameters (k), the expression $(q_i - k)$ is replaced with 1.

(d) Estimates of the variances $\sigma_{u_i}^2$ and $\sigma_{e_i}^2$ are obtained as $\hat{\sigma}_{u_i}^2 = \hat{\lambda}_i^2 - \hat{\sigma}_w^2$ and $\hat{\sigma}_{e_i}^2 = q_i \hat{\sigma}_{u_i}^2 + \hat{\sigma}_w^2$ using steps (b) and (c), and then calculate the transformation parameter $\hat{\alpha}_i$ for each farm. Since the estimate of $\hat{\sigma}_{u_i}^2$ may be negative, we specify $\hat{\alpha}_i = 0$ if $\hat{\sigma}_{u_i}^2 < 0$.

(e) Given the α_i,s calculated in step (d), transform the data as $\tilde{y}_{it} = y_{it} - \hat{\alpha}_i \bar{y}_i$ and $\tilde{x}_{it} = x_{it} - \hat{\alpha}_i \bar{x}_i$. Regress $\tilde{y}_{it}$ on $\tilde{x}_{it}$ and MR_{it} by using OLS to get GLS estimates of the parameters of the model [4]. The GLS is a weighted combination of the OLS and within estimators. If $\hat{\alpha}_i$ proves to be equal to zero, then the model collapses to OLS and if $\hat{\alpha}_i = 1$, it collapses to within estimator.

(f) The estimates of standard errors are biased and inconsistent and deviate from the correct asymptotic standard errors. Consistent estimates of standard errors are obtained using the White's robust estimator (White, 1980) as is done in (4).

In order to avoid or reduce the occurrence of negative variances, some modifications will be introduced where instead of estimating farm-specific variances we estimate group-specific variances for each homogeneous group of farms. The distributional assumptions and the estimation

[4] It is possible to use an iteration procedure using the GLS parameters to estimate new residuals, λ_i^2, $\sigma_{u_i}^2$, $\sigma_{e_i}^2$ and α_i until convergence is obtained. This step was not fully concluded due to the problem of large memory requirement.

procedure are the same as in the previous case except for minor differences in the calculation of the components of the group-specific variances described in the steps (c) and (d). The subscript i in $\sigma_{u_i}^2$ and $\sigma_{e_i}^2$ and α_i are replaced by g indicating groups of farms. The estimates of $\hat{\sigma}_u^2$, $\hat{\sigma}_e^2$ and $\hat{\alpha}$ are not only group-specific but they also vary within the same group due to the variation in the number of time periods the farms are observed. Since we have more than one farm in each group, there is no need to replace the expression in step (c), $(q_i - k)$, with 1. The new expression is $(\Sigma_i q_i - k)$ or $(n_g \bar{q}_g - k)$, where n_g is the number of farms in group g and $\bar{q}_g$ is the average years of observation.

4. Empirical results

The models specified in Sections 2 and 3 are estimated on the rotating panel data from Swedish crop producers, as described in the Appendix. The source of the data is an annual national survey of the economic conditions of family farms (JEU) during the period 1976 to 1988. The input categories used in the production of crops are seed, fertilizer, energy, cash expenditure, net rental cost, labor, traction power, user cost of capital and land. A summary of the statistics of the variables is given in Table 1.

The probit estimates of the selection equations are given in Table 2. The GLS estimates of the

Table 1
Summary statistics of the variables

Variable	Definition	Mean	SD
(A) Variables included in the selection equation (N = 10 611)			
SX_l	Share of farm land	0.7423	0.2768
SX_d	Share of drained farm land	0.2586	0.3356
X_a	Age of farmer	47.3828	10.2768
X_o	Off-farm income	25 679.0614	41 053.5821
X_t	Time trend	6.7736	3.6946
PSE_c	% producer subsidy, crop	0.3497	0.1183
PSE_d	% producer subsidy, dairy	0.5104	0.0484
(B) Variables induded in the regression equation (N = 3077)			
SY_{ca}	Crop share of output	0.8597	0.1691
Y_c	Production of crop	168112.6225	136 157.8095
Y_d	Production of dairy	34 657.5302	60 732.9314
Y_a	Aggregate output	203 236.8698	165 318.8161
X_s	Seed	12 283.0834	13 167.1140
X_f	Fertilizer	29 111.2118	20 082.1445
X_e	Energy	16 443.5668	11 428.6835
X_c	Cash expenditure	19 601.7417	24 731.5630
X_r	Net rental cost	13 673.1113	21 410.3942
X_w	Labor	57 530.5405	42 730.7507
X_p	Traction power	480.6580	1 020.7829
X_k	User cost of capital	85 009.5975	66 005.5904
X_l	Farmland	48.0050	21.5159
X_t	Time trend	6.8811	3.5790
F	No. of farms observed	1 034.0000	
N	No. of observation	3 077.0000	
q_i	No. of times observed	3.8047	1.0497
MR	Mill's Ratio	0.8481	0.4114

Dummy variables not included in the table are as follows: (a) Production type: production of milk (D_m), beef (D_b), and pork (D_p). (b) Production area: area P_1 (high) to P_8 (low), ranked by fertility of land. (c) Regional location: high, medium and low productive regions. (d) Farm size: small, medium and large sizes.

parameters associated with the sample selection and the conventional production function models, along with the corrected asymptotic standard errors, defined in (4), are given in Table 4. The estimated standard errors underestimate their correct counterparts by 57% in the sample selection model and by 59% in the conventional model.

A Cobb-Douglas versus a translog functional form, both including MR, was tested using a F-test (see Johnston, 1984, p. 189). The test is based on the residual sum of squares calculated using the GLS parameter estimates. The resulting F-statistic value of 8.73 (critical value $F(55, 3009) = 1.47$) indicates that the restricted Cobb-Douglas was rejected at a 1% level of significance, in favor of the translog functional form. A similar test based on the translog specification where the parameter of MR is restricted to zero is found to be significant at 5% level of significance, which means that selectivity bias is a problem.

Since we have many parameters to estimate, one could expect multi-collinearity to be a problem. Most data sets exhibit some degree of multi-collinearity. A simple measure of its degree can be obtained by regressing each of the explanatory variables on the remaining explanatory variables. The R^2 obtained can then be taken as a measure of the degree of multi-collinearity. The various values of R^2 were as follows: seed (0.12), fertilizer (0.15), energy (0.35), cash (0.41), net rental cost (0.14), labor (0.36), traction power (0.06), capital (0.44) and land (0.47) indicating

Table 2
Probit parameter estimates [a]

	Milk		Beef		Pork		Crop	
Var	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
α_0	1.1573	0.2102	0.7512	0.3838	−3.2585	0.3000	0.4283	0.4049
SX_1	2.0999	0.4194	3.4079	0.6041	0.5287	0.0524	7.1515	0.8290
SX_1^2	−3.5348	0.3016	−4.2555	0.4265			−5.6718	0.7511
SX_d			0.2008	0.1501	0.2423	0.0424		
SX_d^2			−0.3795	0.1588				
X_a			0.0245	0.0123	0.0567	0.0109		
X_a^2			−0.0003	0.0001	−0.0006	0.0001		
X_t			−0.0349	0.0051	−0.0554	0.0042	−0.0459	0.0085
X_o	−8.5E − 6	4.0E − 6	−4.9E − 6	3.5E − 7	−1.2E − 6	3.3E − 7	−1.9E − 6	7.6E − 7
PSE_c							0.5167	0.1498
PSE_d	0.5728	0.3023	0.6044	0.3663	1.2867	0.3223		
D_m [b]							−3.0942	0.3282
D_b							−1.2156	0.2977
D_p							−1.2552	0.3712
P_1	−1.0331	0.0763	−0.1942	0.0833	1.6003	0.0797	0.2213	0.2385
P_2	−0.4279	0.0752	0.3584	0.0864	1.3216	0.0788	0.3177	0.1888
P_3	−0.5579	0.0758	0.1831	0.0865	0.6908	0.0815	0.2222	0.1329
P_4	−0.5513	0.0762	−0.0007	0.0840	0.4413	0.0807	0.1353	0.1124
P_5	−0.0911	0.0848	0.6377	0.1051	0.8340	0.0852	0.2112	0.1457
P_6	−0.3373	0.0882	−0.0382	0.0988	0.0712	0.0966	−0.0704	0.1122
P_8	0.5374	0.0974	0.7839	0.1125	0.1928	0.1004	0.1632	0.1412
N_o [c]	5562		8013		3808		7534	
N_1	5049		2598		6803		3077	
N	10611		10611		10611		10611	
Log L	−4853.2648		−4283.0275		−5910.1549		−4641.2211	

[a] The dependent variable is 1 if the farm is a producer of milk, beef, pork, or crop else zero.

[b] The dummy variables D_m, D_b and D_p in the crop model are considered to be endogenous and are replaced by their calculated predicted probability values.

[c] Number of producers (N_1), non-producers (N_o) and observations (N).

that multi-collinearity is not high and consequently not a problem in the present study (see Kmenta, 1986, p. 439).

The correlation coefficient among the explanatory variables was found to be in the interval 0.00–0.50. It was computed using the Pearson product-moment and Spearman's rank-order correlation procedures. An inspection of the significance probabilities associated with the correlations indicates no significant correlation among the variables.

A test of the regularity conditions was also performed. First, the marginal elasticities of each input with regard to output were calculated at each point. The percentage frequency of positive marginal productivities is as follows: seed (91.0), fertilizer (97.5), energy (73.0), cash (93.2), net rental cost (72.2), labor (95.8), traction power (63.3), capital (94.6) and land (89.4). Second, the concavity of the production function is checked by testing whether the matrix $M = B - \mathrm{diag}(\alpha) + \alpha\alpha'$ is negative semidefinite, where $y = \alpha_0 + \alpha' x + \frac{1}{2}Bx$ is the translog production function. The determinant of the matrix M was found to be negative semidefinite ($-2.68\mathrm{E}-7$) indicating diminishing marginal productivity.

4.1. Selection equation

The selection equation is estimated to correct for selectivity bias and to provide some measure of the probability of producing crops. As can be seen from Eq. (1.b), the determinants of probability are divided into two subsets. The first, S_{it}, includes variables that enter the selection equation only. These include variables or conditions that are important to the manager of the farm at the pre-production stage, in making decisions whether to produce crops or alternative products. The second subset, X_{it}, consists of variables entering the production function, mainly inputs, some of which are constrained to a certain level (e.g. land) and variables important in terms of allocation when producing multiple outputs.

The selection equation specified includes the following variables. The share of farm land, the share of farm land with drainage, age of the farmer, time trend, off-farm income, producer subsidy share of crop and dairy prices, dummy variables representing production areas and production type dummies representing production of milk, beef and pork.

The decision whether to produce a certain type of product is usually made by the individual farms, e.g. production of milk, beef and pork. However, the producers decisions are influenced by factors not under the control of the farmers, e.g. price policy, credit policy, quality of land, type of farm buildings and the climate. These factors make the decisions weakly exogenous. In order to avoid misspecification, first we estimate separately three probit models where the dependent variables are 1 if a farm has positive production of milk (D_{m}), beef (D_{b}) or pork (D_{p}) else zero. The parameter estimates associated with these models are used to calculate the predicted probabilities that D_{m}, D_{b} and D_{p} are 1 denoted by $\hat{D}_{\mathrm{m}}$, $\hat{D}_{\mathrm{b}}$ and $\hat{D}_{\mathrm{p}}$. These calculated predicted probabilities are then included in the specification of the crop probit model.

The issue whether the above production type dummy variables should be considered endogenous or exogenous is tested using the procedure suggested by Hausman (1978). The χ^2 statistic computed is 23.67. The critical value at the 1% significance level and 2 degrees of freedom is 9.21. This is an indication that the above production type dummies are endogenous.

The estimated parameters of the four (milk, beef, pork and crop) probit models, using the maximum likelihood method, are reported in Table 2. Application of Pearsons and likelihood-ratio tests for the goodness of fit of the crop model yields chi-square values of 11582 and 9282. The models' predictive performance are satisfactory. The percentage of correctly predicted crop producer farms (i.e. predicted probability > 0.50) is about 60% (see Table 3) [5]. The farms classified as crop producers are characterized as farms which, on average, have large farm land with a

[5] The percentage predicted probability > 0.50 in the alternative specification where the production type dummy variables D_{m}, D_{b} and D_{p} are considered as exogenous is much higher about 77%.

Table 3
Calculated predicted probabilities of crop producers

Probability interval	0	%	1	%
0.00 – 0.20	4849	64.36	299	9.72
0.21 – 0.40	869	11.53	411	13.36
0.41 – 0.60	1411	18.73	367	44.43
0.61 – 0.80	392	5.20	956	31.07
0.81 – 1.00	13	0.17	44	1.43
Prob. ≤ 0.50	6332	84.05	1231	40.01
Prob. > 0.50	1202	15.95	1846	59.99

high share of drainage, owned by a young part-time working farmer and located in the more fertile production areas, P_1 to P_4. About 39% and 28% of these farms also have joint production of crops and beef or crops and pork, respectively. About 16% produce a mixture of crops, beef and pork.

From Table 2 we can see that the coefficient of the share of farm land (SX_l) is highly significant and as expected has a positive sign. On other hand, the square of SX_l has a significant coefficient, but negative sign, indicating a decreasing marginal probability of being a crop producer. The coefficient of the time trend (X_t) is negative and significant. This can be interpreted as showing that the probability of having crops as a main line of production is declining over time. The off-farm income (X_o) has a negative and significant effect. This indicates that though specialization in production of crops for family farms relies on and is integrated with part-time off-farm work, an increase in the level X_o will nevertheless result in loss of useful information with negative impacts on the efficiency of farms.

The coefficient of the producer subsidies share of the crop price (PSE_c) is, as expected, positive and significant. A higher PSE_c is positively related to the propensity of producing crops. The predicted production type dummy variables milk (D_m), beef (D_b) and pork (D_p) are all significant and negatively related to the probability of being a crop producer. Three of the seven dummies (P_7 excluded), that represent production areas (P_1 to P_8) ranked by fertility of land, are significant. The coefficients associated with the age of the farmer (X_a), the share of farm land with drainage (SX_d) and the producer subsidies share of dairy price (PSE_d) were insignificant and were finally excluded from the specification.

The Mill's ratio (MR), evaluated for each observation using the parameter estimates from the crop probit model, was included as one of the explanatory variables in the production function. From Table 4 we can see that the coefficient of MR is negative and significant at the 5% level of significance, indicating presence of selectivity bias. Thus not accounting for selection bias will yield biased and inconsistent estimators. It should be noted, however, that the use of MR adjustment to account for sample selectivity is only an approximation. Empirical findings show that different estimation methods applied to the same data set might yield different estimates of the selection effect (see Olsen, 1982, and Little, 1985).

4.2. Variance components

One important feature of the model outlined in this study is the introduction of farm heterogeneity into the production function. We follow the generalized error component model with heteroscedastic disturbances proposed by Baltagi and Griffin (1988). The serious disadvantage of this approach is that the procedure used to estimate the variance components may result in negative estimates of the farm-specific variances. In this paper, I specify heterogeneity by defining groups of farms with similar farming conditions. Estimation of group-specific variances instead of farm-specific variances reduces the frequency of the negative variances. One variance (σ_u^2) is estimated for each homogeneous group of farms and one common white noise variance (σ_w^2), both are used to calculate the total variance (σ_e^2). These variance components are then used to calculate the transformation parameter (α). The heteroscedastic variance component σ_u^2 makes α group-specific, varying in the interval zero to one. An inspection of the values showed that only 0.1% (i.e., three observations out of 3077) of the group-specific variances are negative [6].

[6] An inspection of the frequency of negative farm-specific variances in a case where one σ_u^2 is estimated for each farm showed that 18.1% of the σ_u^2 were negative.

Table 4
Generalized least squares parameter estimates [a]

Parameter	Sample selection model (SSM)		Conventional regression model (CRM)	
	Estimate	SE [b]	Estimate	SE
β_0	-0.0753	0.0530	-0.1139	0.0466
β_s	0.0887	0.0726	0.0881	0.0724
β_f	0.2901	0.1901	0.2884	0.1899
β_e	0.1856	0.2660	0.1778	0.2666
β_c	-0.2888	0.2314	-0.2872	0.2324
β_r	0.0616	0.0556	0.0635	0.0556
β_w	0.0877	0.2508	0.1039	0.2518
β_p	-0.1051	0.0690	-0.1040	0.0690
β_k	-0.3084	0.2945	-0.2990	0.2954
β_t	-0.1578	0.0799	-0.1606	0.0803
β_l	4.3049	0.9584	4.2589	0.9630
β_{ss}	0.0057	0.0014	0.0057	0.0014
β_{ff}	0.0074	0.0022	0.0075	0.0022
β_{ee}	0.0011	0.0053	0.0012	0.0053
β_{cc}	0.0117	0.0037	0.0118	0.0037
β_{rr}	0.0068	0.0013	0.0069	0.0013
β_{ww}	0.0311	0.0071	0.0305	0.0072
β_{pp}	-0.0008	0.0019	-0.0008	0.0019
β_{kk}	0.0296	0.0170	0.0302	0.0170
β_{tt}	-0.0025	0.0009	-0.0024	0.0009
β_{ll}	0.0096	0.0873	0.0074	0.0881
β_{sf}	-0.0020	0.0022	-0.0019	0.0022
β_{se}	-0.0170	0.0054	-0.0172	0.0053
β_{sc}	0.0027	0.0028	0.0027	0.0028
β_{sr}	-0.0017	0.0009	-0.0017	0.0009
β_{sw}	0.0086	0.0077	0.0086	0.0077
β_{sp}	-0.0005	0.0011	-0.0005	0.0011
β_{sk}	-0.0059	0.0057	-0.0058	0.0058
β_{st}	0.0015	0.0012	0.0015	0.0012
β_{sl}	-0.0004	0.0115	-0.0004	0.0115
β_{fe}	-0.0010	0.0077	-0.0012	0.0078
β_{fc}	-0.0028	0.0040	-0.0028	0.0040
β_{fr}	0.0008	0.0017	0.0008	0.0017
β_{fw}	-0.0007	0.0162	-0.0006	0.0161
β_{fp}	0.0017	0.0018	0.0016	0.0018
β_{fk}	-0.0312	0.0118	-0.0310	0.0118
β_{ft}	0.0013	0.0022	0.0013	0.0022
β_{fl}	0.0066	0.0160	0.0062	0.0160
β_{ec}	0.0258	0.0136	0.0256	0.0136
β_{er}	-0.0003	0.0025	-0.0002	0.0025
β_{ew}	-0.0184	0.0231	-0.0175	0.0232
β_{ep}	0.0040	0.0052	0.0042	0.0052
β_{ek}	-0.0049	0.0171	-0.0045	0.0171
β_{et}	-0.0026	0.0047	-0.0026	0.0047
β_{el}	-0.0033	0.0444	-0.0043	0.0446
β_{cr}	-0.0046	0.0030	-0.0047	0.0030
β_{cw}	0.0033	0.0249	0.0022	0.0252
β_{cp}	-0.0013	0.0037	-0.0012	0.0037
β_{ck}	0.0205	0.0173	-0.0204	0.0175
β_{ct}	0.0124	0.0044	0.0124	0.0044
β_{cl}	0.0248	0.0372	0.0276	0.0373
β_{rw}	-0.0096	0.0058	-0.0097	0.0059

Table 4 (continued)

Parameter	Sample selection model (SSM)		Conventional regression model (CRM)	
	Estimate	SE [b]	Estimate	SE
β_{rp}	−0.0011	0.0008	−0.0011	0.0008
β_{rk}	0.0070	0.0044	0.0068	0.0044
β_{rt}	0.0005	0.0007	0.0005	0.0007
β_{rl}	−0.0081	0.0076	−0.0080	0.0077
β_{wp}	0.0140	0.0075	0.0138	0.0075
β_{wk}	0.0539	0.0295	0.0516	0.0296
β_{wt}	0.0053	0.0080	0.0057	0.0080
β_{wl}	−0.2801	0.0969	−0.2757	0.0976
β_{pk}	−0.0069	0.0050	−0.0071	0.0050
β_{pt}	0.0008	0.0008	0.0008	0.0008
β_{pl}	0.0006	0.0104	0.0009	0.0104
β_{kt}	0.0153	0.0046	0.0154	0.0046
β_{kl}	−0.0777	0.0510	−0.0777	0.0510
β_{tl}	−0.0497	0.0104	−0.0504	0.0104
β_{h}	0.0468	0.0463	0.0622	0.0447
β_{m}	0.0039	0.0488	0.0096	0.0481
ψ	−0.0307	0.0136		
MSE	0.1049		0.1048	
R^2_{adj}	0.9067		0.9057	

[a] Based on 1034 farms each observed on the average 3.8047 years, i.e. a total of 3077 observations.
[b] Corrected standard errors.
Glossary of variables: s, seed; f, fertilizer; e, energy; c, cash expenditure; r, net rental cost; h, m, regional dummies; w, labor; p, traction power; k, user cost of capital; t, time trend; l, land.

The existence of a large number of farms with large cross-sectional differences suggests that heteroscedasticity could be a problem and motivates estimation of the α-parameters for each homogeneous group of farms. Application of Barlett's test (see Kmenta, 1986, p. 297) for the null hypothesis of homoscedasticity, i.e. $\sigma_{u_1}^2 = \sigma_{u_2}^2 = \ldots = \sigma_{u_{21}}^2$, against the alternative hypothesis of heteroscedasticity yields the chi-square values of 735.84. The critical value of χ^2 with 20 degrees of freedom is 37.57 at the 1% level of significance [7]. This allows us to reject homoscedasticity at the 1% level of significance.

The group-specific variances, σ_u^2 (excluding one with zero value) varying from 0.25 to 5.07, with an overall sample mean of 0.95 and standard deviation of 0.70, are reported in Table 5. The common white noise variance, σ_w^2, is 0.09. The transformation parameter α ranges from 0.69 to 0.93 with a mean of 0.80 and a standard deviation of 0.08. For purposes of comparison, I also report the means of the variance components and α by years, regions and farm sizes in Table 5. When looking at the distribution over time, σ_u^2, σ_e^2 and α are found to be increasing from 1976 to 1983. They reach their highest values in 1983 and then decrease during the following years to the lowest levels in the final year (1988). The sizes of σ_u^2, σ_e^2 and α are found to be negatively related to the productivity of land and farm size. Small farms and farms located in less productive regions are found to be more heterogeneous.

4.3. Input elasticities

Since the coefficients of the translog production function do not have any direct interpretation, I calculate the elasticity of output with respect to each of the inputs as defined in (7).

[7] The number of cell combinations of three sizes and eight production areas is 24 groups of farms. No farms were observed within three cells. The degree of freedom is $(G-1)=20$.

These elasticities are both farm- and time-specific and are used in drawing inferences regarding the allocation of resources by farm, within the sample and over time. The elasticities and the measures of returns to scale, evaluated at the mean, for each year, region and farm size and based on the

Table 5
Variance components by groups of farm, year, location and size [a]

	N	q_i	σ_w^2	σ_u^2	σ_e^2	α
G_{1s}	179	3.6536	0.0955	0.3397	1.3368	0.7261
G_{1m}	304	3.8388	0.0955	0.2499	1.0547	0.6947
G_{1l}	410	3.8854	0.0955	0.3185	1.3332	0.7235
G_{2s}	93	3.9462	0.0955	0.6484	2.6542	0.8054
G_{2m}	89	3.5169	0.0955	0.4185	1.5673	0.7487
G_{2l}	167	3.8982	0.0955	0.3519	1.4673	0.7390
G_{3s}	197	3.7614	0.0955	1.1284	4.3398	0.8484
G_{3m}	244	4.0738	0.0955	1.3582	5.6283	0.8656
G_{3l}	302	3.7020	0.0955	1.2165	4.5991	0.8525
G_{4s}	176	3.7670	0.0955	1.6209	6.2016	0.8729
G_{4m}	238	3.5672	0.0955	2.0641	7.4586	0.8846
G_{4l}	335	3.7522	0.0955	0.9476	3.6512	0.8336
G_{5s}	41	3.7317	0.0955	1.0484	4.0080	0.8435
G_{5m}	48	3.6458	0.0955	0.3910	1.5211	0.7470
G_{6s}	67	4.1343	0.0955	1.7959	7.5204	0.8831
G_{6m}	68	3.8529	0.0955	0.7510	2.9892	0.8188
G_{6l}	3	2.0000	0.0955	0.0000	0.0955	0.0000
G_{7s}	35	3.6857	0.0955	0.7717	2.9399	0.8186
G_{7m}	31	4.5161	0.0955	1.0737	4.9445	0.8560
G_{8s}	23	4.3043	0.0955	5.0679	21.9093	0.9321
G_{8m}	27	3.6667	0.0955	3.0662	11.3381	0.9066
1976	202	3.2822	0.0955	0.9122	3.0137	0.7823
1977	259	3.4015	0.0955	0.9894	3.4109	0.7936
1978	256	3.8242	0.0955	0.9276	3.6300	0.7967
1979	238	4.0168	0.0955	0.9562	3.9236	0.8102
1980	221	3.9910	0.0955	0.9865	4.0137	0.8100
1981	252	4.0595	0.0955	0.9697	4.0376	0.8111
1982	240	4.0417	0.0955	0.9931	4.1722	0.8140
1983	249	4.0522	0.0955	1.0673	4.4238	0.8193
1984	258	4.1124	0.0955	0.9610	3.9981	0.8140
1985	277	3.9495	0.0955	0.9345	3.7651	0.8087
1986	276	3.7971	0.0955	0.9236	3.6133	0.8000
1987	221	3.4072	0.0955	0.9047	3.2418	0.7844
1988	128	3.0313	0.0955	0.7640	2.3827	0.7538
Regional location						
High	2734	3.7919	0.0955	0.8920	3.4546	0.7981
Medium	227	3.8458	0.0955	1.0271	4.1619	0.8162
Low	116	4.0259	0.0955	2.2383	9.1915	0.8716
Farm size						
Small	811	3.8015	0.0955	1.1536	4.5716	0.8255
Medium	1049	3.8122	0.0955	1.0694	4.1413	0.8028
Large	1217	3.8003	0.0955	0.7183	2.7970	0.7862
Overall sample mean						
SSM	3077	3.8047	0.0955	0.9527	3.7231	0.8022
CRM	3077	3.8047	0.0955	0.9595	3.7498	0.8036

[a] G.., group of small (s), medium (m) and large (l) farms located in production areas P_1–P_8. N, No. of observations; q_i, No. of times observed. Variances: constant (σ_w^2), heterogenous (σ_u^2) and total (σ_e^2); α, transformation parameter. SSM, Sample selection model. CRM, Conventional regression model.

parameters estimated for the sample selection model, are reported in Table 6.

In general, the size and development of these elasticities reflect important structural features of Swedish grain farming. Largest in magnitude is the land elasticity with a sample mean of 0.34, a large standard deviation (0.21) and a strong decreasing trend moving the average value from 0.59 in 1976 to 0.07 in 1988. This trend reflects the existence of a large crop surplus and the subsequent reduction in the size of land. This has in turn increased the supply of low cost land, inducing a reduction in the rate of return to farming. Thus, in spite of the variations, land has on average become a much less constraining factor on production. The land elasticity is found to be negatively related to the size of farms.

Traction power provides the other extreme, with virtually constant elasticity at 0.01 (and many negative individual elasticities). This is a strong indication of tax-induced over-mechanization in Swedish farming – a result which is consistent with a lot of anecdotal evidence.

The seed elasticity is on the average 0.05 and is increasing over time. The differences in use of seed across farms of different sizes and located in different regions are found to be quite small. The mean value is lower for the large farms, in comparison to the small and medium sizes, and it is somewhat higher in the northern regions.

The fertilizer elasticity is somewhat larger than the seed elasticity. It is on the average 0.07 and fluctuating around the value of the overall mean. We observe minor differences in the utilization of fertilizer across the different farm sizes and regions. Large farms located in high productive regions use more fertilizer per unit of land. This indicates that this type of farms use land more intensively.

The energy elasticity has a mean value of 0.05 with large standard deviations. It is increasing during 1976 to 1978. The outbreak of the second

Table 6
Input elasticities and returns to scale

Year	Elasticity of output with respect to									Returns to scale	
	Seed	Fertil.	Energy	Cash	Rent	Labor	T. power	Capital	Land	RTS_1	RTS_2
1976	0.0420	0.0671	0.0434	0.0299	0.0656	0.2062	0.0101	0.1067	0.5948	0.5710	1.1658
1977	0.0425	0.0678	0.0544	0.0401	0.0653	0.2017	0.0107	0.1203	0.5509	0.6028	1.1537
1978	0.0431	0.0661	0.0583	0.0493	0.0669	0.2091	0.0101	0.1405	0.4950	0.6434	1.1384
1979	0.0427	0.0688	0.0541	0.0641	0.0671	0.2026	0.0104	0.1369	0.4757	0.6467	1.1224
1980	0.0486	0.0676	0.0420	0.0729	0.0674	0.2130	0.0106	0.1487	0.4194	0.6708	1.0902
1981	0.0464	0.0661	0.0423	0.0783	0.0672	0.2358	0.0096	0.1685	0.3404	0.7142	1.0546
1982	0.0477	0.0656	0.0443	0.0917	0.0681	0.2486	0.0103	0.1879	0.2909	0.7642	1.0551
1983	0.0495	0.0681	0.0463	0.1040	0.0692	0.2343	0.0089	0.1971	0.2551	0.7774	1.0325
1984	0.0476	0.0621	0.0408	0.1187	0.0722	0.2800	0.0086	0.2217	0.1968	0.8517	1.0485
1985	0.0529	0.0670	0.0381	0.1384	0.0709	0.2673	0.0092	0.2287	0.1706	0.8725	1.0431
1986	0.0515	0.0666	0.0445	0.1610	0.0705	0.2673	0.0101	0.2476	0.1300	0.9191	1.0491
1987	0.0554	0.0667	0.0404	0.1739	0.0715	0.2626	0.0103	0.2607	0.1037	0.9415	1.0452
1988	0.0549	0.0645	0.0292	0.1948	0.0743	0.2681	0.0104	0.2683	0.0730	0.9645	1.0375
Regional location											
High	0.0479	0.0658	0.0412	0.1051	0.0696	0.2416	0.0097	0.1881	0.3325	0.7689	1.1014
Medium	0.0461	0.0742	0.0670	0.0871	0.0655	0.2141	0.0098	0.1728	0.4167	0.7366	1.1533
Low	0.0553	0.0695	0.0914	0.0735	0.0507	0.2174	0.0133	0.2328	0.3953	0.8039	1.1992
Farm size											
Small	0.0527	0.0717	0.0505	0.0808	0.0660	0.2593	0.0111	0.1802	0.4055	0.7723	1.1778
Medium	0.0492	0.0678	0.0477	0.0961	0.0676	0.2461	0.0103	0.1941	0.3390	0.7789	1.1179
Large	0.0440	0.0620	0.0396	0.1207	0.0714	0.2180	0.0086	0.1896	0.2972	0.7539	1.0511
Overall sample mean											
SSM	0.0480	0.0665	0.0454	0.1028	0.0689	0.2386	0.0099	0.1887	0.3419	0.7688	1.1107
CRM	0.0602	0.0786	0.0465	0.1280	0.0726	0.2412	0.0103	0.1756	0.4031	0.8130	1.2161

SSM, Sample selection model. CRM, Conventional regression model.

oil crisis in 1979 introduced some changes in the energy consumption behavior of the farms. The elasticity decreases during 1979 to 1980 and thereafter increases until 1983. There is a sharp decline in the size of the elasticity in the final year of 1988. Small farms located in the northern parts of Sweden use, on average, more energy per unit of output than other farms.

The cash expenditure elasticity is relatively large with a sample mean of 0.10 and standard deviation of 0.06. It shows a strong increase from 0.03 in 1976 to 0.19 in 1988. Moreover, there are substantial variations across different locations and sizes. Large farms and farms located in highly productive regions are more cash intensive than small farms and farms located in the less fertile areas.

The sample mean net rental cost elasticity is 0.07 and is increasing over time. Large farms located in fertile areas have somewhat larger elasticity than the smaller ones. Complicated production processes and large quantities of output require access to storage spaces and expensive machinery other than those owned by the large farms.

The labor elasticity is the second largest in magnitude. The sample mean value is about 0.24. It increases during the period 1976 to 1984 but is somewhat lower, and fluctuates around 0.26, during the remaining period. This is a reflection of increased mechanization. As expected, there are large variations across regions and farm sizes. The labor elasticity is found to be negatively related to the size of the farm, but positively related to the fertility of the land. Farms of the same size employ different levels of labor intensity. The differences are generated by the variations in the types of crops produced and their variability in labor requirements.

The capital elasticity is fairly high and increasing, starting at 0.11 in 1976 and ending at 0.27 in 1988. The elasticities for farms located in high and medium productivity areas are smaller than those for small farms in low productivity regions. This is due to the over-mechanized production process. The latter seems to mainly to relate to tractor power. Small farms use less capital per unit of output and more efficiently compared to other sizes. This type of farm replaces capital less frequently.

For comparison purposes, I have calculated these elasticities using the conventional model which does not account for selection bias. The mean sample elasticities of output with respect to seed, fertilizer, energy, cash expenditure, net rental cost, labor and traction power and land are higher while those of the user cost of capital are lower than those calculated for the sample selection model.

4.4. Returns to scale

The estimates of returns to scale (RTS) defined in (6), as the elasticity of output with regard to a proportionate change in all inputs, are given in Table 6. Two measures of returns to scale are defined. In the first case, RTS_1, the variable land is not included while in the second, RTS_2, land is also included.

The sample mean of RTS_1 is found to be 0.77 with a standard deviation of 0.11. RTS_1 is increasing continuously over time. Keeping land compact, crop farms are characterized by decreasing returns to scale. RTS_1 is gradually increasing during the sample period, from 0.57 in 1976 to 0.96 in 1988. It is somewhat higher for the small and medium size farms, located in the low fertile areas.

By the second definition, the sample mean of returns to scale is 1.11 with a standard deviation of 0.06. RTS_2 is also found to be, on average, greater than one, indicating increasing returns to scale, during 1976 to 1979. RTS_2 has declined over time from 1.17 in 1976 to 1.04 in 1988. This is caused by the sharp decline in the returns to land. The size of RTS_2 is low for large farms located in the highly fertile areas and higher for smaller farms located in areas with high fertility.

The following is the interpretation of the results obtained. At the margin, land has become a less scarce resource in the sense that its marginal productivity has decreased. At the same time the returns to scale of other inputs have increased, which is totally consistent with the fact that land has become gradually less of a binding constraint on production in Swedish agriculture. Together, however, the elasticity of scale has decreased towards constant returns to scale, which is also consistent with the fact that land is not a very binding restriction on the production possibilities, mainly due to the existence of a large surplus and the reduction in the total area of land utilized.

The main difference between the sample selection model and the conventional model appears in the scale properties – in the marginal elasticities as well as in the elasticity of scale. According to the both models there is overall increasing returns to scale. The sum of differences correspond to 4.4% by RTS_1 and 10.5% by RTS_2 measures.

4.5. Rate of technical change and its decomposition

Technical change is represented by a simple time trend in the production function. The use of a flexible functional form and introduction of quadratic terms in time and interaction of time with the inputs allows for non-neutrality of technical change.

Estimates of the rate of technical change defined in (8), are obtained as the partial derivative of the production function with regard to time. The sample mean values show technical regress during the period of study, at the rate of 0.03 per year. The rate of technical change is further decomposed into pure and non-neutral components. These measures are reported in Table 7. The sample mean of the pure and non-neutral components are −0.25 and 0.22, respectively. The pure component contributes negatively to the rate of technical change while the contribution from the non-neutral component is found to be positive. Considered over time the elasticity was negative indicating technical regress, but at a decreasing rate. The overall rate declined continuously from the value of zero in 1976 to −0.06 in the final year 1988. The same decreasing pattern was evident for the pure component. The non-neutral component increases over time.

We observe some variations in the behavior of the technical change and its non-neutral component for different sizes and locations. Small farms and farms located in fertile regions have, on average, experienced a lower rate of technical regress. The large negative pure component could, to some extent, be explained by the changes in agricultural policy caused by environmental concerns. The changes have sought to limit the use of fertilizer and pesticides per unit of land by introducing more restrictions and taxes on the use of these inputs. The highly positive non-neutral component is an indication that changes have taken place in the composition of inputs that are used in crop production. The reduction in the size of land employed in the production of crops by laying off less fertile land has been important in this respect.

Table 7
Decompostition of the rate of technical change

Year	Pure	Non-neutral	Overall
1976	−0.2183	0.2180	−0.0002
1977	−0.2236	0.2136	−0.0100
1978	−0.2290	0.2154	−0.0136
1979	−0.2344	0.2138	−0.0206
1980	−0.2397	0.2152	−0.0245
1981	−0.2451	0.2195	−0.0256
1982	−0.2505	0.2250	−0.0255
1983	−0.2558	0.2205	−0.0353
1984	−0.2612	0.2269	−0.0343
1985	−0.2666	0.2255	−0.0410
1986	−0.2719	0.2243	−0.0476
1987	−0.2773	0.2205	−0.0568
1988	−0.2827	0.2209	−0.0618
Regional location			
High	−0.2504	0.2209	−0.0295
Medium	−0.2462	0.2137	−0.0325
Low	−0.2446	0.2125	−0.0321
Farm size			
Small	−0.2473	0.2274	−0.0200
Medium	−0.2498	0.2219	−0.0279
Large	−0.2515	0.2136	−0.0379
Overall sample mean			
SSM	−0.2498	0.2200	−0.0298
CRM	−0.1941	0.1644	−0.0297

SSM, Sample selection model. CRM, Conventional regression model (CRM).

5. Conclusions

In this paper we have discussed the estimation of production functions when sample selection bias due to non-randomness of the sample can be expected. A sample selection model consisting of a selection equation and a production function was estimated using Heckman's two-stage method. In the estimation of the production function, where the underlying technology was represented by a translog functional form, we assumed a random effects model with heteroscedastic variances. The model was estimated using the Generalized Least Squares method. In the empirical part we utilized a large rotating panel data set from Swedish crop producers over the period of 1976–1988. The major findings and conclusions derived from this study are:

First, a correct specification of the determinants of the probability of being a crop producer is limited by data observability and accessibility. This affects the explanatory power of the probit model, the evaluation of Mill's Ratio (*MR*) and the detection of the possible presence of selectivity bias. The probit model was estimated using the maximum likelihood method. The models predictive performance has been satisfactory. The coefficient of the *MR* in the production model is found to be negative and significant, indicating the presence of selectivity bias.

Second, introduction of heteroscedasticity by estimating group-specific variances shows to be useful. The groups of farms are defined based on a combination of farm sizes and locations. The mean distribution of group-specific variances has increased during the period before 1983 and declined in the following years. Small farms located in less productive regions are found to be more heterogeneous. The procedure used here to estimate the variance components did only result in 0.1% negative estimates of the group-specific variances.

Third, the elasticity of output with respect to each of the inputs was calculated. These elasticities are both farm- and time-specific. The mean energy, seed, fertilizer and rental cost elasticities, ranked by size lie between 0.04 and 0.07. The mean cash and capital elasticities are 0.10 and 0.19, respectively. The labor elasticity is the second largest, about 0.24. The cash, labor and capital elasticities are increasing over time. The elasticity with respect to land is the largest in magnitude. The mean value is about 0.34 and is decreasing over time.

Fourth, two measures of returns to scale are defined. The sample mean of returns to scale where land is not included in the calculation is 0.77. It shows that crop farms are characterized by decreasing returns to scale. When land is included, the sample mean is 1.11. It is larger than one during 1976 to 1988. The returns to scale are declining over time. This is caused by a sharp decline in the returns to land. Land has become a less scarce resource and gradually less of a binding constraint on production.

Fifth, the sample mean value of technical change shows technical regress during the period of study at the rate of 0.03 per year. It is declining from the value of zero in 1976 to -0.06 in 1988. The measure of technical change is further decomposed into its pure and non-neutral components. The means for the pure and non-neutral components are -0.25 and 0.22, respectively. Small farms and farms located in fertile regions have on the average experienced a lower rate of technical regress.

Finally, in comparison between the sample selection and the conventional model, the empirical results show that introduction of heteroscedasticity and accounting for selection in the production relationship to be important. The impact of the selectivity bias considered on the results obtained in terms of scale properties are found to be significant. A significant coefficient of the *MR* is an indication of the presence of selectivity bias. However, the estimates of the selection effects are sensitive to the way the farm effects are treated and the choice of estimation methods. In comparison between farm- and group-specific formulations of the heteroscedasticity, the group-specific formulation is preferred. It results in very few negative variances.

6. Appendix: Data

6.1. Description of the rotating sample

The data used in this study are part of an annual national survey of the economic conditions of agriculture (JEU), carried out by Statistics Sweden (SCB). The annual survey includes about 1,000 small and medium-sized family farms drawn from a population of 30 000 to 35 000 farms during the period 1976 to 1988. These farms have arable land consisting of 20 to 100 hectares in the plain districts and 20 to 50 hectares in the forest and northern districts. A stratified random sample of the farms was obtained by SCB using the Farm Register in agriculture and forestry as the sampling frame. The stratification of farms with respect to the production area was in direct proportion to the total number of farms within the area, cattle units and the area of arable land.

Each chosen farm is expected to be part of the survey for a total of 4 years with a regeneration rate of 25% per year. The main reason for applying a rotating sampling design by SCB is to reduce the degree of non-response and the desire to maintain the property of the sample representation of the true population. The annual degree of non-response has kept unchanged at 10–12% of the sample. During the period under investigation, 1976–1988, there were major changes in sample representation of the population due to changes that occurred in the structure of production and the population of the farms.

The distribution of land by size shows that the percentage of small farms has declined over time whereas the percentage of medium and large farms has increased. The area under cultivation and the number of farms has declined by 3.6% and 22.9%, respectively. These changes are mainly caused by rapid policy induced structural change. The idea of rotation has been helpful in capturing changes in the structure of agriculture and taking them into account in the annual stratified sampling plan.

6.2. Definition of variables

The observed farms in my sample do not all produce the same outputs nor do they use the same inputs. I have classified the farms' production activities and resources used into several output and input categories, common to all farms under consideration. In this paper, I am interested only in the production of crops. I use three measures of output to define a farm as a crop producer, Y_c, Y_d and Y_a ($Y_a = Y_c + Y_d$), as the total income generated from the production of crops, dairy and from an aggregate output, respectively. I define a farm as a crop producer during a period if the share of crop of the aggregate output is larger than 50%. The share of crops in the farm's aggregate output in the conditioned sample is on average, 86% (see Table 1). Y_c, Y_d and Y_a are measured in Swedish currency (SEK) converted to 1980 prices using the producer price index for crop, dairy and an aggregate of crops and dairy products.

In the present study, I use seed, fertilizer, energy, cash expenditure, net rental cost, labor, traction power, user cost of capital and land as inputs in the production of crops.

Seed (X_s) is one of the main components of the farm's expenditure defined as the total expenditure on the purchase of seed. Fertilizer (X_f) is the aggregate value of the plant nutrient utilized at the farm. Cash expenditure (X_c) is the aggregate value of cash expenditure items purchased and used mainly in the production of crops. X_c consists of costs induced by the use of pesticides, fodder preservative, cleaning, education and health service, cost of hiring labor, consultation service, marketing, communication and transportation. X_s, X_f and X_c are measured in SEK and transformed to constant 1980 prices using the agricultural requisites cost price index.

Energy (X_e) is the aggregate value of the farm's consumption of energy including electricity and fuel. X_e is measured in SEK and transformed to constant 1980 prices using an aggregate cost price index for fuel and electricity.

Net rental cost (X_r) is the difference between total rental cost and total rental income of the farm. Rental income and costs includes income, respective costs generated from renting farm buildings, land and machines. We define X_r as a difference because net rental costs reflect the available net input to the farm. X_r is measured in SEK and transformed to 1980 prices using the service cost price index in agriculture.

The labor variable (X_w) is the total cost of family and hired labor used exclusively in the production of crop products. Labor includes all crop production activities such as planning, ploughing, sowing, spraying plant protection and nutrient and harvesting. Labor is measured in SEK and transformed to constant 1980 prices using the cost price index for labor in agriculture.

Traction power (X_p) is the value of aggregate traction power used at the farm excluding those related to the forest activities. The capital variable (X_k) is user cost of capital equipment including depreciation, maintenance, insurance and net interest rate costs. X_k covers the capital equipment of machinery, inventory, farm buildings and land improvement. The rate of depreciation applied to machinery equipment was between 14 and 17% respectively. Different rates were used depending on the size of the farm and the differences in the farm's intensity of capital use. A rate of 3.7% was used for farm buildings and 11% was used for inventories. X_p and X_k are measured in SEK and transformed to constant 1980 prices using a cost price index for capital equipment in agriculture.

There are two types of land. Farming land (X_l) covers arable land and pasture land (i.e. the area used for pasture) both measured in hectares. The land variable used in the production function is only aggregate farming land.

In addition, we consider off-farm income (X_o), age of the farmer (X_a), a time variable for the observation year (X_t), the share of farm land (SX_l), the share of farm land with covered drainage (SX_d), production type dummies indicating if a farm, in addition to crops, also produces milk (D_m), beef (D_b) or pork (D_p), location dummies (P) and percentage producer subsidy (PSE) as explanatory variables in the selection equation. A summary of the statistics of all the variables is given in Table 1.

Off-farm income (X_o) is a measure of income generated from the off-farm activities. X_o includes both the farmers and his wife's income from non-farm work such as part-time work, seasonal piece work and other activities than production of crop, dairy products and forest. X_o is measured in SEK and transformed to constant 1980 prices using the consumer price index. Producer Subsidy Equivalent (PSE) is a measure of money transfer to the farms at the existing production level. PSE consists of the market price support, i.e. the difference between the world market and domestic prices of agricultural products, plus direct payments and other budget payments less taxes paid by producers. The PSE for crops (PSE_c) and dairy products (PSE_d) are given as the producer subsidy share of crops and dairy prices.

Since the data are related to farms of different sizes, in terms of hectares, located in different regions within Sweden, regional dummies are used to reflect differences in production behavior with respect to location. A farm's geographical location is classified by Statistics Sweden using two levels, based on farming conditions and land productivity. First, location is differentiated by 8 production areas (P_1 to P_8), used in the estimation of the selection equation. Second, at a more aggregate level, location is classified into three major regions: high (P_1–P_3), medium (P_4–P_6), and low (P_7–P_8) productive regions. The farm size consists of three size classes based on arable land in hectares: small (20–30), medium (30–50), and large (50–100). The regional location and farm size variables are used in the classification and presentation of the results. A combination of the different production areas and size classes are also used in the specification of group-specific heteroscedasticity.

Acknowledgements

The author gratefully acknowledges financial support from SJFR and would like to thank Anders Klevmarken, Subal Kumbhakar, Lennart Hjalmarsson, D.J.H. Capel and two anonymous referees for their valuable comments and suggestions.

References

Amemiya, T., 1985. Advanced Econometrics. Basil Blackwell, Oxford, UK.

Baltagi, B.H. and Griffin, J.M., 1988. A generalized error component model with heteroscedastic disturbances. Int. Econ. Rev., 29: 745–753.

Baltagi, B.H. and Raj, B., 1992. A survey of recent theoretical developments in the econometrics of panel data. Emp. Econ., 17: 85–109.

Biorn, E., 1981. Estimating economic relations from incomplete cross-section/time-series data. J. Econometrics, 16: 221–236.

Biorn, E. and Jansen, E.S., 1983. Individual effects in a system of demand functions. Scand. J. Econ., 85: 461–483.

Christensen, L.R., Jorgenson, D.W. and Lau, L.J., 1973. Transcendental logarithmic production frontiers. Rev. Econ. Stat., 55: 28–45.

Forsund, F.R. and Hjalmarsson, L., 1987. Analysis of Industrial Structure. A Putty-Clay Approach. IUI, Almqvist & Wiksell International, Stockholm.

Greene, W.H., 1981. Sample selection bias as a specification error: Comment. Econometrica, 49: 795–798.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica, 46: 1251–1271.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica, 47: 153–161.

Heshmati, A., 1994. Estimating technical efficiency, productivity growth and selectivity bias using rotating panel data: an application to Swedish agriculture. Ph.D. thesis, Department of Economics, Univ. Gothenburg, Sweden.

Heshmati, A., Kumbhakar, S.C. and Hjalmarsson, L., 1994. Efficiency of the Swedish pork industry: a farm level study using rotating panel data 1976–1988. Eur. J. Oper. Res. (forthcoming).

Hsiao, C., 1986. Analysis of Panel Data. Cambridge University Press, UK.

Johnston, J., 1984. Econometric Methods (3rd Edition. McGraw-Hill, New York.

Killingsworth, M.R. and Heckman, J.J., 1986. Female labor supply: a survey. Chapter 2 in: O. Ashenfelter and R. Layard R. (Editors) Handbook of Labor Economics 1, pp. 103–204.

Kmenta, J., 1986. Elements of Econometrics (2nd Edition) Macmillan, New York.

Kumbhakar, S.C. and Heshmati, A., 1991. Efficiency measurement using rotating panel data: an application to Swedish dairy farms 1976–1988. Dep. Econ. Memo 155, Univ. Gothenburg, 31 pp.

Kumbhakar, S.C. and Hjalmarsson, L., 1993. Technical efficiency and technical progress in Swedish dairy farms. Chapter 9 in: H. Fried, C.A.K. Lovell and S. Schmidt (Editors) The Measurement of Productive Efficiency Techniques and Applications. Oxford University Press, New York, pp. 256–270.

Little, R.J.A., 1985. A note about models for selectivity bias. Econometrica, 53: 1469–1474.

Maddala, G.S., 1987. Limited dependent variable models using panel data. J. Human Resour., 22: 307–338.

Nijman, T., Verbeek, M. and Van Soest, A., 1991. The efficiency of rotating-panel designs in an analysis-of-variance model. J. Econometrics, 49: 373–399.

Olsen, R.J., 1982. Distributional tests for selectivity bias and more robust likelihood estimator. Int. Econ. Rev., 23: 223–240.

Verbeek, M., 1990. On the estimation of a fixed effects model with selectivity bias. Econ. Lett., 34: 267–270.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48: 817–838.

Zabel, J.E., 1992. Estimating fixed and random effects models with selectivity. Econ. Lett., 40: 269–272.