ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Lönnrotinkatu 4 B, 00120 Helsinki 12, Finland, tel. 601322

Keskusteluaiheita
Discussion papers

Pekka Lastikka

ESTIMATING HETEROSCEDASTIC REGRESSION MODELS: SOME NEW METHODS AND THEIR APPLICATIONS

A revised version

No. 125 22.11.1982

A revised version of the paper presented at the European Meeting of the Econometric Society in Dublin, September 5-9, 1982. (ETLA's Discussion Paper No. 105).

ESTIMATING HETEROSCEDASTIC REGRESSION MODELS: SOME NEW METHODS AND THEIR APPLICATIONS

Pekka Lastikka*

The Research Institute of the Finnish Economy,

SF-00120 Helsinki 12, Finland

The study is concerned with the estimation of regression models for Finnish industrial workers' wages from heteroscedastic cross-sectional data. A model is constructed and estimated for the disturbance variance in wage models. New methods are suggested for constructing and estimating such models. The mathematical form of the model for the disturbance variance is carefully chosen. In estimating that model a very simple autoregressive model for wages is utilized. Separate models for the disturbance variance are estimated for geographical regions and for branches of industry. These models are used for estimating the following:

1) the weights used in estimating the wage models from cross-sectional data
2) the variance of the logarithm of individual workers' wage from aggregate cross-sectional data
3) the interdependence of workers' wages in variously-sized geographical regions and branches of industry, making use of the generalized intra-class correlation coefficient. A method is developed by means of which the "wage transfer" effect can be quantitatively measured in aggregated cross-sectional units when the population is finite.

1. Introduction

This paper considers the estimation of wage models from cross-sectional data where the observations differ in size and are not independent. The

* An earlier version of this paper was presented at the European Meeting of the Econometric Society in Dublin, September 5-9, 1982. I wish to thank the participants for valuable comments, and my particular thanks are due to Professor T. Kloek. Moreover, I wish to express my deepest gratitude to my teacher, Professor L. Törnqvist, for numerous stimulating discussions and for his encouragement for this work. Finally I wish to thank Jaakko Railo, M.A., for translating parts of the Finnish text into English and for checking my English in some parts of the text.

main interest is in the heteroscedasticity of the wage models, and I focus on the problem concerning the weights that should be used in estimating regression models for industrial workers' average hourly wages with so few theoretical assumptions as possible. This estimation problem is solved by estimating a separate model for the disturbance variance of the wage models - an idea which is not new in itself [Glejser (1969), Goldfeld-Quandt (1972), Harvey (1974,1976), Rutemiller-Bowers (1968)]. In contrast to earlier studies, this paper is interested especially in the principles according to which a model for the disturbance variance should be constructed. A very flexible function to be used as a model for the disturbance variance is suggested. The method of estimating that model is a new one and is based on a few short time series of wages for variously-sized geographical regions and branches of industry.

The data relates to industrial workers in Finland. The main interest is in the cross-sectional data related to 170 commuting regions and 22 branches of industry in 1970. The commuting areas are formed by dividing Finland into 170 mutually exclusive variously-sized areas, in such a way that each area forms a geographically connected whole. The branches of industry are different in size and mutually exclusive, too, and together they form the whole country's industry.

Separate models for the disturbance variance is estimated for geographical regions and for industry branches. These models proved very informative and useful. Apart from their use in estimating the weights needed for cross-sectional wage models they can also be used for some other purposes proposed in chapter 3.

2. The model of wages and the model of the disturbance variance

2.1. The model of the logarithm of industrial workers' wages

Let us consider the population F of all industrial workers in Finland in the time period t[1]. The whole population is divided successively by some procedure into sets i ($i \in F$),which are mutually exclusive in the same division of the population.Let us study one of such divisions.In that division $i=1,\ldots,I$. Let the number of all industrial workers ν in aggregate units i be N_i and in the whole population N_F. Let $w_{i\nu}$ stand for the hourly wage paid to the industrial worker ν in aggregate unit i. To explain the logarithm of the wage $w_{i\nu}$ we introduce the model

$$\ln w_{i\nu} = \alpha + \sum_k \beta_k \ln x_{i\nu k} + \varepsilon_{i\nu}, \qquad \begin{array}{l}(i=1,\ldots,I),\\ (\nu=1,\ldots,N_i),\\ (k=1,\ldots,K),\end{array} \tag{1}$$

where $\ln x_{i\nu k}$ is the value of the k:th explanatory variable and $\varepsilon_{i\nu}$ is the disturbance. Let α be a constant such that $\sum_i p_i' \sum_\nu p_{i\nu} \varepsilon_{i\nu} = 0$, where the weights p_i' and p_{iv} will be regarded as non-stochastic variables and $\sum_i p_i' = \sum_\nu p_{i\nu} = 1$. Let the β_k:s be those constants which would be obtained if model (1) could be estimated from the whole population by OLS.

Let us consider the average hourly wage in the aggregate unit i, which is $\bar{w}_{i\cdot} = \sum_\nu p_{i\nu} w_{i\nu}$. The logarithm of this wage is

$$\ln \bar{w}_{i\cdot} = \ln \sum_\nu p_{i\nu} w_{i\nu} \,. \tag{2}$$

[1] The subscript t will be omitted because the following considerations relate to the same period t.

In appendix A it has been shown that a regression model at the aggregate level for the variable (2) can be evaluated:

$$\ln \bar{w}_{i\cdot} = \alpha' + \sum_k \beta_k \overline{\ln x}_{i\cdot k} + \varepsilon'_{i\cdot}\,, \tag{3}$$

where

$$\alpha' = \alpha + [\sum_i p'_i \sum_\nu p_{i\nu}(\frac{1}{2}\dot{w}^2_{i\nu} + \frac{1}{6}\dot{w}^3_{i\nu} + \ldots)]\,, \tag{4}$$

$$\overline{\ln x}_{i\cdot k} = \sum_\nu p_{i\nu} \ln x_{i\nu k}\,, \tag{5}$$

and

$$\varepsilon'_{i\cdot} = (\sum_\nu p_{i\nu}\varepsilon_{i\nu} + \frac{1}{2}\sum_\nu p_{i\nu}\dot{w}^2_{i\nu} + \frac{1}{6}\sum_\nu p_{i\nu}\dot{w}^3_{i\nu} + \ldots) - (\alpha' - \alpha). \tag{6}$$

In equations (4) and (6) $\dot{w}_{i\nu} = \ln w_{i\nu} - \sum_\nu p_{i\nu} \ln w_{i\nu}$. In model (3) $\sum_i p'_i\varepsilon'_{i\cdot} = 0$ owing to the α':s and $\varepsilon'_{i\cdot}$:s manner of construction[2].

In this paper a model of type (3) is estimated from two different sets of cross-sectional aggregate data: 1) data for geographical regions (commuting regions) and 2) data for branches of industry. Since the statistical units associated with these kinds of data differ in size, in terms of the number of workers, particular attention in estimating the regression models of type (3) will be devoted to rendering the observations on the various regions and branches of industry comparable through weighting. For this purpose a model is estimated for the disturbance variance of model (3).

[2] In the text model (3) is also sometimes referred to when its estimated counterpart is meant. This should not cause any confusion.

Although the observations in the regional and the industry-branch divisions are mutually dependent, the independence assumption of observations is used as a working hypothesis in estimating the models (3) since no reliable estimates for the correlations of observations are available[3]. The intention is to estimate model (3) mainly by empirical means. An effort will be made to avoid theoretical, restrictive assumptions the correctness of which could be considered questionable in the empirical data.

2.2. The model of the disturbance variance

An effort is made to construct a model for the disturbance variance of model (3) in variously-sized aggregate units. The important point is that the aggregate units are not collected by random sampling and, thus, no variance formula, derived by mathematical methods, is available. The disturbance variances, in this study, relate to distributions that result when aggregate units are formed from a finite population in accordance with some procedure other than random sampling. In the case of geographical regions the aggregate units each form a connected geographical whole and are mutually exclusive in the same regional division (as is the case with the commuting regions). In the case of industry branches the aggregate units are formed on an institutional basis and are mutually

[3] I have investigated the effect of this assumption on regression models and statistics making use of cross-sectional data for administrative labour-force districts. These districts are much larger than the commuting regions; in 1970 Finland was divided into 11 administrative labour-force districts. My investigation suggested i.a., that the variance of regression coefficient can be quite sensitive to the independence assumption of observations. I compared the variances of regression coefficient estimated under the assumptions that the observations were independent and that they were dependent; in the latter case I made use of estimated correlations between observations, i.e. $\binom{11}{2} = 55$ estimated correlations. These correlations were estimated utilizing time series from the residuals of models (16)-(17). The results of these experiments are reported in my forthcoming study.

exclusive, too. The variances of disturbance' distributions which result in collecting information by applying of principles this kind are not known theoretically. I try to estimate those by empirical methods.

In the model of the disturbance variance, the size of the aggregate unit, relative to the whole population, was chosen as the only explanatory variable, even though other variables descriptive of the characteristics of the aggregate unit could also have been used as explanatory variables. The relative size of the aggregate unit was denoted by p and it was dealt with as a continuous variable ($p \in [0,1]$). A continuous function of p can then be chosen as the explanatory model of the disturbance variance. The size of the disturbance variance associated with an aggregate unit of size p will be denoted by σ_p^2.

Certain conditions, considered desirable, may be imposed on the explanatory model of σ_p^2. The variance σ_p^2 should be non-negative and finite. As in the case of random sampling, the further desideratum is imposed here that the expression of σ_p^2 should be capable of being decomposed into two factors, one of which is a constant, denoted by σ_0^2 (which corresponds to the value $p = 0$), the other being a function dependent on the size of the sample. Let θ_p^2 stand for this function. The function θ_p^2 is postulated to be continuous, to possess derivatives of the first order and to be monotonically decreasing in the interval $[0,1]$.

One desideratum for the function θ_p^2 is obtained from the properties of the disturbances $\varepsilon_{i.}'$ of model (3). Let ε_p' be the disturbance associated with an aggregate unit of size p and let ε_q' be the disturbance associated with an aggregate unit of size q ($q = 1 - \rho$). Then, because $\sum_i p_i'\varepsilon_i' = 0$ in (3), we have

$$p\varepsilon_p' + q\varepsilon_q' = 0 \,, \tag{7}$$

whence, further,

$$p\varepsilon_p' = -q\varepsilon_q' \tag{8}$$

and, by squaring (8) ,

$$p^2(\varepsilon_p')^2 = q^2(\varepsilon_q')^2 \tag{9}$$

Taking the mathematical expectation, on both sides of (9), over all the aggregate units of size p and q that can be formed, by successive divisions, from the population[4] we get, after re-arranging the factors

$$\frac{\sigma_p^2}{\sigma_q^2} = \frac{q^2}{p^2} \,, \tag{10}$$

where

$$\sigma_p^2 = E_p(\varepsilon_p')^2 \text{ and } \sigma_q^2 = E_q(\varepsilon_q')^2 \,.$$

If $\theta_p^2 = q^2 f(p,q)$ and if $f(p,q)$ is a symmetric function, or $f(p,q) = f(q,p)$, the condition (10) will be satisfied.

[4] In the case of geographical regions the aggregate units of size p and q each form a connected geographical whole and are mutually exclusive in the same regional division of the population. In the case of industrial branches they are mutually exclusive, too, in the same division of the population.

On the basis of the above, the following desiderata are imposed on the explanatory model of the disturbance variance σ_p^2

1) $0 \leq \sigma_p^2 < \infty$ for all $p \in [0,1]$.

2) $\sigma_p^2 = \sigma_0^2 \theta_p^2$; σ_0^2 being a constant and $0 \leq \theta_p^2 \leq 1$; $\theta_0^2 = 1$, $\theta_1^2 = 0$.

3) The function θ_p^2 is continuous, has derivatives of first order and decreases monotonically in the interval $p \in [0,1]$; $-\infty < \frac{d\theta_p^2}{dp} \leq 0$.

4) $\theta_p^2 = q^2 f(p,q)$, where $f(p,q) = f(q,p)$.

Several functions satisfying the desiderata 1) - 4) can be found. In the present study,

$$\theta_p^2 = q^2(1+pq)^2(1+\lambda p^2 q^2)^{-\delta}, \qquad (\lambda > 0), \tag{11}$$

was settled on after many trials.

Then,

$$\sigma_p^2 = e^{\gamma} q^2(1+pq)^2(1+\lambda p^2 q^2)^{-\delta}, \tag{12}$$

where $e^{\gamma} = \sigma_0^2$.

It is easily seen that desideratum 1 is met in the case of function (12). For all $p,q \in [0,1]$, $\sigma_p^2 \geq 0$ and $\sigma_p^2 < \infty$ hold true, provided that the parameters γ, δ and λ are finite.

Desideratum 2 is also met by (12). It could first be recalled that

$\sigma_p^2 = \sigma_0^2\theta_p^2$. In addition, $\theta_p^2 \geq 0$. Furthermore, $\theta_p^2 \leq 1$, because the maximum value of $q^2(1+pq)^2$ in the interval $p \in [0,1]$ is 1 (p being then equal to 0) and because the maximum value of $(1+\lambda p^2q^2)^{-\delta}$ in the same interval is also 1 ($\delta \geq 0$). In the case of the function (11), $\theta_0^2 = 1$ and $\theta_1^2 = 0$ hold true, in addition.

Also (11) is continuous and has a derivative for all values $p \in [0,1]$. It can be shown (appendix B) that $-\infty < \frac{d\theta_p^2}{dp} \leq 0$, when $p \in [0,1]$,

if

$$0 \leq \delta \leq 1 \tag{13}$$

and

$$0 < \lambda < \infty. \tag{14}$$

If conditions (13) and (14) are met, desideratum 3 is satisfied by function (11) when $p \in [0,1]$.

Desideratum 4 is also satisfied by function (11) since the function θ_p^2 is of the type $\theta_p^2 = q^2f(p,q)$, where

$$f(p,q) = f(q,p) = (1+pq)^2(1+\lambda p^2q^2)^{-\delta}.$$

2.3. Estimation of the disturbance variance

Transformation to logarithms in both sides of (12), replacement of σ_p^2 by its estimate s_p^2 and introduction of the disturbance ξ_p into formula (12) yield the model

$$\ln s_p^2 = 2 \ln q + 2 \ln (1 + pq) + \gamma - \delta \ln (1 + \lambda p^2 q^2) + \xi_p \,. \qquad (15)$$

Model (15) can be estimated from the observational data if estimates of $\ln s_p^2$ for various values of p are available. The finding of reasonable substitutes for p and q is more simple.

In the present study, the estimates of $\ln s_p^2$ are based on the residual variances of the wage models for $\ln \bar{w}_{i.}$ and $\ln \bar{w}_{.j}$ obtained from time series data[5]. Regarding such models of $\ln \bar{w}_{i.}$ and $\ln \bar{w}_{.j}$ it is justified to presuppose that the regression coefficients of those models are constants, since the parameters of model (3), except the coefficient α' of the constant, are constants[6]. In addition, it can be presupposed that the disturbances of the time series' models of $\ln \bar{w}_{i.}$ and $\ln \bar{w}_{.j}$ approach zero as the size of the aggregate unit approaches the size of the population. This condition corresponds to the property $\varepsilon'_{F.} = 0$ of model (3).

The variance s_p^2 involved in (15) was estimated in this study as follows. From the time-series data, 1960-1971, for variously sized regions i and various branches of industry j, the residual variances for wage models

$$\ln \bar{w}_{i.t} = \ln \bar{w}_{i.t-1} + (\ln \bar{w}_{..t} - \ln \bar{w}_{..t-1}) + \bar{e}_{i.t} \,, \qquad (16)$$

and $(i=1,\ldots,38),(t=1,\ldots,12),$

$$\ln \bar{w}_{.jt} = \ln \bar{w}_{.j(t-1)} + (\ln \bar{w}_{..t} - \ln \bar{w}_{..(t-1)}) + \bar{e}_{.jt} \,, \qquad (17)$$

$(j=1,\ldots,22), \ (t=1,\ldots,12)$,

[5] The subscript i henceforth refers to geographical regions and the subscript j to branches of industry.

[6] In successive divisions of the population the coefficient α' may have different values.

were computed from the formulas

$$s_t^2(\bar{e}_{i\cdot t}) = \frac{1}{T-1} \sum_{t=1}^{T} (\bar{e}_{i\cdot t} - \bar{\bar{e}}_{i\cdot\cdot})^2 \tag{18}$$

and

$$s_t^2(\bar{e}_{\cdot jt}) = \frac{1}{T-1} \sum_{t} (\bar{e}_{\cdot jt} - \bar{\bar{e}}_{\cdot j\cdot})^2 \tag{19}$$

respectively[7].

The desiderata imposed on the cross-sectional models of $\ln\bar{w}_{i\cdot}$ and $\ln\bar{w}_{\cdot j}$ are satisfied by models (16) - (17): the coefficients do not depend on the size of the aggregate unit, the parameters of models (16) - (17) can be considered to equal unity in absolute value, and the residuals $\bar{e}_{i\cdot t}$ and $\bar{e}_{\cdot jt}$ approach zero as the size of the aggregate unit (region i or industrial branch j) approaches the size of the whole Population.

The use of the residual variances (18) - (19) as estimated from time series related to variously sized areas or industrial branches, in estimating the disturbance variance σ_p^2 is based on the working hypothesis that the residual variance estimated from time series data for variously sized aggregate units behaves approximately as the disturbance variance of model (3) as the size of the aggregate unit changes. This hypothesis rests on the view that the time-series variance and cross-section variance associated with wages are of the same type. The cross-sectional variation observed in wages at any given point in time can be considered

[7] The dot in the place of the subscript j and i in the formulas (16)-(17) respectively indicates all branches of industry (in (16)) or the whole of Finland (in (17)). In equations (18) - (19) $\bar{\bar{e}}_{i\cdot\cdot} = \frac{1}{T} \sum_{t=1}^{T} \bar{e}_{i\cdot t}$ and $\bar{\bar{e}}_{\cdot j\cdot} = \frac{1}{T} \sum_{t=1}^{T} \bar{e}_{\cdot jt}$ respectively.

to be due to a variety of factors which have changed and been at work in the course of time. The structure of the labour force of any one region, for instance, which can be regarded as one of the central factors capable of explaining regional wage levels, can be considered to be a result of such a course of historical development. Thus, the regional variation present in cross-sectional data on wages can be regarded as a result of the time-series variation shown by wages.

Although the working hypthesis is introduced that the residual variances of models (16) and (3) as well as models (17) and (3) respectively decrease at approximately equal rates as the size of the aggregate unit (the region in the case of the first pair of models and the branch of industry in the case of the second pair of models) increases, the residual variances of models (16) - (17) related to time-series data, can be expected to be smaller in absolute value than the residual variances of model (3) associated with cross-sectional data of the corresponding size. This does not matter in the estimation of model (3), since, for that purpose, it is necessary to know only the relative weights rather than the absolute weights of the observations.

2.4. Estimation of the model (15) in the case of geographical regions

Time series of industrial workers' average hourly wages in the years 1960-1971 were formed for 38 variously sized regions, which each formed a geographically connected whole, just as did the 170 commuting regions and the 11 administrative labour-force districts (Footnote 3). The 38 regions were mainly regions other than the commuting regions and the administrative labour-force districts, but they were constructed according to the same principles as these regions. From this observational data, the residual variances of model (16) were estimated for the various regions and the logarithms of these estimates were used as the dependent variable in model (15). The size variables p and q were measured in terms of the relative numbers of industrial workers.

Model (15) was estimated by the OLS method, weighting the observations by unity, in such a way that the parameter λ was given different values. The model of which the multiple correlation coefficient was the largest was chosen as the final model, from among the models involving the different values of λ tried out. The estimation results for model (15) are set out in Table 1, which reveals that for model 9 the multiple correlation coefficient attained its maximum value when λ was given the value $\lambda = 3 \cdot 10^6$. For that model $\delta = 0.408$, so that model 9 can be accepted in the sense of formula (13).

A scatter diagram of the observations on the regressand $\ln s^2_{p_i}$ and the regressor p_i of the models given in Table 1 is presented in Figure 1. In addition, the graph of the function

$$\ln s^2_p \approx 2 \ln q + 2 \ln(1 + pq) - 4.963 - 0.408 \ln(1 + 3 \cdot 10^6 p^2 q^2) \qquad (20)$$

corresponding to the model 9 given in Table 1 and the graph of the function

$$\ln s^2_p = -4.963 + \ln \frac{1}{359608 \cdot p} + \ln q \ , \quad (p \in [\frac{1}{359608}, 1] \text{ and} \qquad (21)$$
$$q = 1 - p),$$

are represented in Figure 1.

Function (21) is related to that imagined situation in which each of the 38 regions used as units of observation is interpreted as a random sample drawn without replacement from the population formed by the industrial workers in the whole country[8].

Figure 1 shows that the variance formula pertaining to the case of random sampling without replacement is badly suitable for the explanation of the residual variances related to the 38 regions constituting the observational data. As the relative size p of the region increases, function (20) will decrease definitely more slowly than function (21).

2.5. Some empirical results

If for a model of type (3) it is true that

$$\sigma^2(\varepsilon'_{i.}) = \frac{\sigma_0^2}{\omega_i^2}, \tag{23}$$

[8] Consider a random sample of N_i^{wr} workers drawn (without replacement) from the finite population formed by industrial workers. As is well known, the variance of the mean of a variable associated with this kind of sample is

$$\sigma^2 = \frac{\sigma_0^2}{N_i^{wr}} (1 - \frac{N_i^{wr} - 1}{N_.^{wr} - 1}) ,$$

where σ_0^2 is the variance of the variable concerned in the total population. Denoting

$$\sigma_0^2 = 0.00699 \; (= e^{-4.963}) , \quad p = \frac{N_i^{wr}}{N_.^{wr}} \text{ and } N_.^{wr} = 359608 \text{ we get}$$

$$\sigma^2 \approx \frac{0.00699}{359608\,p}\, q , \tag{22}$$

where q = 1 - p. Transforming to logarithms on both sides in (22),an approximation corresponding to the function (21) is obtained.

where ω_i stands for a non-stochastic weight variable, then for a model

$$\omega_i \ln \bar{w}_{i\cdot} = \omega_i \alpha' + \sum_k \beta_k (\omega_i \overline{\ln x}_{i\cdot k}) + \omega_i \varepsilon'_i \,. \tag{24}$$

it is true that

$$\sigma^2(\omega_i \varepsilon'_{i\cdot}) = \omega_i^2 \sigma^2(\varepsilon'_{i\cdot}) = \sigma_0^2 \,. \tag{25}$$

Some estimation results concerning models of type (3) are given in Table 2[9]. The models in this table were estimated by OLS after multiplying the original observations with the weights

$$\omega_i = \sqrt{[1 + 3\cdot 10^6 (p_i q_i)^2]^{0.408} / q_i (1 + p_i q_i)} \,, \tag{26}$$

where $p_i = \frac{N^{wr}_{i\cdot 70}}{403393}$ and $q_i = 1 - p_i$ ($N^{wr}_{i\cdot 70}$ standing for the number of industrial workers in the commuting region i in 1970. In 1970 there was 403393 industrial workers in whole Finland). The weights (26) are equal to $\sqrt{e^{-4.963}} = \hat{\sigma}_0$ divided by the standard deviation $\hat{\sigma}_p$ based on model 9 in Table 1. The transformed residual of models of type (24) ought to be now homoscedastic.

The homoscedasticity assumption relating to transformed models of type (24) was studied, with respect to the size of regions, by a wide variety of methods which made use of the empirical residuals of the models in Table 2. All the results suggested that heteroscedasticity was no longer perceptible in the transformed residuals.

[9] The dependent variable of the models in Table 2 was $100 \ln \bar{w}_{i\cdot}$, where $\bar{w}_{i\cdot}$ is industrial workers' average hourly wage in the region i. The independent variables were designed to measure industrial workers' characteristics, the type of their work performance and certain characteristics of the industrial establishments and regions.

One method, which I applied, is based on the idea of calculating estimates of the disturbance variance of model (24) both from small regions and from large regions. If these estimates are approximately the same, the homoscedasticity assumption of the model of type (24) can be accepted.

The regions for which the weight ω_i was in the interval $\omega_i \in (1, 1.0027]$ were classified as "small", and those for which it was in the interval $\omega_i \in [4.096, 7.765]$ were classified as large. Each of the two groups came to contain 10 regions. The observations for the Helsinki commuting region were excluded because in most models the variance of the residual $\bar{e}_{i.}$ related to this region considerably differed from the variances of the residuals $\bar{e}_{i.}$ related to other commuting regions[10]. The variances were computed with respect to both "0" (the theoretical mean of the disturbance $\varepsilon'_{i.}$) and the mean of the residuals themselves. The results are set out in Table 3.

The variance ratios for small and large regions of models 1-11 of Table 2 were, in the case of most models, comparatively close to unity. The fact whether the variance of a variable was calculated with respect to its own mean (the formula for the variance s_2^2) or with respect to zero (the formula for the variance s_1^2) did not greatly affect the variance ratios.

[10] It is well known that the theoretical variances of empirical residuals can be quite different in different observations also in the case of independent observations (Draper-Smith (1981) pp. 151-152). The theoretical variances of empirical residuals depend in general on the theoretical variance of disturbances, on independent variables and on correlations between the observations. The general formulas for the theoretical variances of empirical residuals are derived by the author in a forthcoming study.

If the variance ratios given in Table 3 were F-distributed, none of the values of the ratios would be statistically significant at the 1 % level and only one ($s_{2s}^2/s_{2\ell}^2$ = 3.23 for model 11) would be significant at the 5 % level. Despite the fact that the variance ratios cannot be considered F distributed, the disturbances of most theoretical models corresponding to the models in Table 2 can be regarded as homoscedastic. This applies at least to the models where the variance ratios $s_{1s}^2/s_{1\ell}^2$ and $s_{2s}^2/s_{2\ell}^2$ are close to unity.

2.6. Estimation of the model (15) in the case of industry branches

A model of the disturbance variance for the branches of Finnish industry was also estimated. This was used to estimate models of industrial workers' average hourly wage rate from the cross-sectional data concerning all branches of Finnish industry in 1970. These models were of type (3). The industry branch division represents a type of division of Finnish industrial workers' population into mutually exclusive sets, different from the division into commuting regions. Because the 22 branches of industry were of unequal size in terms of the number of workers, the same econometric problem arose as in the case of geographical areas: what weights should be used in estimating wage models from data concerning all branches of Finnish industry in 1970? This problem was solved in the same way as in the case of geographical regions. First the dependent variable of a model of type (15) was constructed utilizing model (17) and its residual variance (19). Then the model of type (15) was estimated by OLS for all 22 branches of Finnish industry weighting the observations by unity, in such a way that the parameter λ was given different values. Model 7 in Table 4, which is

$$\ln s_p^2 \approx 2 \ln q + 2 \ln(1+pq) - 7.93 - 0.917 \ln (1+400\, p^2q^2) , \qquad (27)$$

was finally chosen because, for this model, the coefficient of multiple correlation was highest (See Figure 2)[11]. On the basis of this model the weights

$$\omega_j = \sqrt{[1+400(p_jq_j)^2]^{0.917}/q_j(1+p_jq_j)} , \qquad (28)$$

where

$$p_j = \frac{N^{wr}_{\cdot j70}}{403393} \quad \text{and} \quad q_j = 1 - p_j , \qquad (29)$$

were constructed ($N^{wr}_{\cdot j70}$ is the number of industrial workers in branch j in 1970). Weights (28) were used in the same way as in the case of geographical regions to estimate wage models for industrial workers from data concerning all branches of Finnish industry in 1970.

2.7. Further remarks on the model of the disturbance variance

The suggested method concerning the construction of the model of the disturbance variance and its estimation seems very attractive both empirically and theoretically. The function of the disturbance variance (12) is very flexible and suitable for a variety purposes where variance formulas derived by mathematical methods cannot be used and where the desiderata 1-4 (in chapter 2.2.) should be satisfied. In different kinds of aggregate data, where the observation aggregates are constructed in accordance with different principles, the rate of decrease of the

[11] In model (27) $\delta = 0.917$, which is acceptable in the sense of formula (13).

disturbance variance can be very different as it is in the geographical data (Figure 1) and in the data concerning the branches of industry (Figure 2). The function (12) allows the disturbance variance to decrease at very many different rates depending on the size of the parameters λ and δ. The observational material can be used to estimate these parameters.

The method of estimating the model of the disturbance variance is very simple to apply. Only rather short time series data from the variables $\ln \bar{w}_{i\cdot}$ and $\ln \bar{w}_{\cdot j}$, which were the dependent variables of the "main" models of type (3), and very simple autoregressive models of types (16)-(17) are needed. The error variance of models (16)-(17) can be estimated from formulas (18)-(19) by weighting observations with unity weights because the size of the regions and industrial branches is approximately constant over relatively short time periods. In contrast to many earlier studies, the only explanatory variable of the variance model (15) is the size (p) of the aggregate unit. This kind of model was found to behave quite satisfactorily.

Methods proposed earlier for estimating a model for the disturbance variance are based on the residuals $\bar{e}'_{i\cdot}$ of the "main" model of type (3). These residuals are estimated by OLS with unity weights and an explanatory model is estimated for the squares or the absolute values of these residuals (Park (1966), Glejser (1969), Goldfeld-Quandt (1972), Amemiya (1977)). One of the shortcomings of these methods is that, in estimating the original explanatory model of type (3), in order to determine the residuals $\bar{e}'_{i\cdot}$, the observations are from the outset weighted with wrong weights in cases where the homoscedasticity assumption concerning the disturbances of the model of type (3) does not hold true and the correct weights of the observations are not known. Harvey (1976) suggests the application of the maximum likelihood method, by means of which it is possible to estimate simultaneously the original explanatory model (type (3)) and the unknown parameter involved in the multiplicative formula of the disturbance variance. However, when this method is used, it is necessary to assume already in the estimation stage that the disturbances of the original model of type (3) are normally distributed and stochastically independent of one another. These assumptions are very strong and there are apparently numerous studies where they are hardly true.

3. Some further applications of the disturbance variance models

With the disturbance variance models it is possible to calculate the weights of geographical regions and branches of industry in estimating models for wages from cross-sectional data. Making use of the disturbance variance models it is possible further to

1) estimate the variance of the logarithm of individual workers' average hourly wage rate
2) estimate numerically the interdependence of individual workers' wages in geographical regions and branches of industry differing in size. This can be done employing the generalized intra-class correlation coefficient and the disturbance variance models.

3.1. The standard deviation of the logarithm of individual workers' wage

The models of the average hourly wage rate were estimated from cross-sectional data for geographical regions and branches of industry in the form (24). That is, the original observations were first multiplied by weights (26) in the case of geographical regions and by weights (28) in the case of the branches of industry and models of type (24) were then estimated by OLS with unity weights for the transformed observations. The residual variance of those wage models of type (24),where the only explanatory variable was the weight ω_i (or ω_j) can be interpreted as the estimate of the variance of the logarithm of individual workers' hourly wage rate according to (25). The estimates from different sets of observations are given in Table 5.

As can be seen from Table 5 the estimate $\hat{\sigma}_0$ calculated from data for the branches of industry is very close to the estimate obtained from data of the Confederation of Finnish Employers. This result suggests that, by suitable methods, very accurate information can also be obtained from secondary data which is originally produced for administrative purposes and which is not at all optimal from the point of view of the study.

3.2. The estimation of interdependence of wages

Let us consider the following decomposition

$$(\frac{1}{n}\sum_i \varepsilon_i)^2 = \bar{\varepsilon}^2 = \frac{1}{n^2}(\Sigma\, \varepsilon_i^2 + \sum_{i \neq j}\Sigma\, \varepsilon_i \varepsilon_j) , \qquad (30)$$

where ε_i is the disturbance of a regression model and n is the sample size. Let E be the operator of expectation and let us write, as a matter of notation,

$$E(\bar{\varepsilon}^2) = \sigma^2_{\bar{\varepsilon}} \, , \tag{31a}$$

$$E(\varepsilon_i^2) = \sigma^2_{\varepsilon_i} \, , \tag{31b}$$

$$\rho_{ih} = \frac{\underset{i \neq h}{E(\varepsilon_i \varepsilon_h)}}{\sigma^2_{\varepsilon_i}} \text{ and} \tag{31c}$$

$$\bar{\rho} = \frac{1}{n(n-1)} \sum_{i \neq h} \sum \rho_{ih} \, . \tag{31d}$$

Taking expectations on both sides of (30), utilizing notations (31a-d) we get

$$\bar{\rho} = \frac{\dfrac{\sigma^2_{\bar{\varepsilon}}}{\sigma^2_{\varepsilon_i}} - \dfrac{1}{n}}{1 - \dfrac{1}{n}} \, . \tag{32}$$

Formula (32) defines the average intra-class correlation coefficient $\bar{\rho}$. Leo Törnqvist has suggested the generalization of (32) to cases where the sizes of research units are continuous variables. Instead of (32) he suggests the formula

$$\rho(p_1,p_2) = (\frac{\sigma^2_{p_2}}{\sigma^2_{p_1}} - \frac{p_1}{p_2})/(1 - \frac{p_1}{p_2}) \, , \quad (p_1 < p_2) \, , \tag{33}$$

where the sample size is denoted by p_2 and the size of the measure region (inter region) by p_1. The variance $\sigma^2_{p_2}$ corresponds to $\sigma^2_{\bar{\varepsilon}}$ and the variance $\sigma^2_{p_1}$ corresponds to $\sigma^2_{\varepsilon_i}$ in formula (32) [12)]

[12)] $\sigma^2_{p_2}$ may be considered the variance of the disturbance $\bar{\varepsilon}$ over all geographical regions of size p_2, which are obtained by successive division of the population into mutually exclusive sets. These geographical regions each form a connected geographical whole. In the case of the branches of industry $\sigma^2_{p_2}$ may be considered the variance of the disturbance $\bar{\varepsilon}$ over all "industry branches" of size p_2, which are obtained by successive division of the population into mutually exclusive "industry branches", real or hypothetical. Similar interpretations can be given to $\sigma^2_{p_2}$.

From empirical data it is possible to compute an estimate of the intra-class correlation coefficient defined by (33). The variances $\sigma^2_{p_2}$ and $\sigma^2_{p_1}$ can then be replaced by estimates which can be computed from the formulas (20) and (27) constructed in the preceding sections.

Formulas (20) and (27) are based on time series data. If the disturbance variance of the wage models associated with cross-sectional data decreases as fast as the disturbance variance of the time-series models of the same variable as the size of the aggregate unit increases - which assumption was made in constructing the weights of observations to be used in estimating the wage models from cross-sectional data - then the model of the intra-class correlation coefficient constructed in the above manner can also be considered to describe the intra-class correlation coefficient of the disturbances of the wage models related to cross-sectional data.

It may be of particular interest to examine the estimate $\hat{\rho}(p_1 = 0, p_2) = \hat{\rho}(0, p_2)$, which can be considered to describe approximately the intra-class correlation coefficient of the disturbances related to individual workers in aggregate units of size p_2. In Figure 3, the intra-class correlation coefficients $\hat{\rho}_I(0,p_2)$ and $\hat{\rho}_J(0,p_2)$ related to geographical regions and branches of industry respectively are represented as functions of p_2.

Figure 3 shows that, when p_2 increases, the intra-class correlation coefficient $\hat{\rho}_J(0,p_2)$ for the branches of industry approaches zero much more slowly than does the intra-class correlation coefficient $\hat{\rho}_I(0,p_2)$ for geographical areas. This result suggests that, in any two areas of the same size - which may be either geographical areas or "industrial branch" areas, which also form geographical wholes - the disturbances of models for industrial workers' individual level average hourly wage rates are more similar to each other in the case of the branches of industry than in the case of geographical areas.

With the method outlined in this section it is possible to estimate quantitatively the dependence of wages in different types of sets and in sets differing in size, which may be geographical areas or branches of industry or some other sets utilizing a generalized intra-class correlation coefficient and the models for the disturbance variances.

4. Concluding remarks

The focus in this paper is on the problem of estimating wage models from aggregated cross-sectional data which is heteroscedastic. The important idea was that the cross-sectional units, where the main point of interest laid, were constructed with procedures other than random sampling. In such cases, what can be said about the behaviour of wage models' disturbance variances in aggregated cross-sectional units differing in size? This question, to which an answer was found by constructing and estimating models for disturbance variances, may be of great importance for those who try to extract more information out of, e.g., official statistics where the aggregation of micro-units has already been done in accordance with certain principles. Some results given in chapter 3.1. suggested that by utilizing the model of the disturbance variance extremely accurate individual level information could be obtained from highly aggregated data.

The criteria which I used in judging the goodness or badness of my method were mainly empirical. The results in chapter 2.5. concerning the realism of the homoscedasticity assumption related to the transformed model (24) showed that no heteroscedasticity was in evidence in the estimated residuals after applying the method suggested in this paper. Results of this kind suggest that the disturbance variance model and its application (chapter 3) may be applicable in a wide variety of studies where the econometric problems are similar to those in this study.

The following question may be raised: is the proposed new estimation method for heteroscedastic models then "better" than other methods suggested earlier? This question can be answered as follows. My method is better in that it is very simple to apply and requires fewer unrealistic theoretical assumptions than the methods proposed earlier (see chapter 2.7.). If we wanted to study whether my method will produce better estimates in some well-defined statistical sense than the methods suggested earlier, a common model should be specified and the competing estimation methods could be compared in that frame. Obviously this would lead to very complicated mathematics if in the specification of the model we rejected, most of the customary assumptions of the disturbance. In any case, as I see it, this is a very interesting problem for further research.

Appendix A

The model of the logarithm of industrial workers' wages

Let us consider the model for the logarithm of the wage $w_{i\nu}$ of worker ν in aggregate unit i

$$\ln w_{i\nu} = \alpha + \sum_k \beta_k \ln x_{i\nu k} + \varepsilon_{i\nu}, \qquad \begin{array}{l}(i=1,\ldots,I),\\(\nu=1,\ldots,N_i),\\(k=1,\ldots,K),\end{array} \tag{1}$$

and the logarithm of the average hourly wage in the aggregate unit i

$$\ln \bar{w}_{i.} = \ln \sum_\nu p_{i\nu} w_{i\nu}. \tag{2}$$

In (2) $\sum_\nu p_{i\nu} = 1$. Applying Törnqvist's (1936) formulas we get[13)]

[13)] Y. Vartia (1976) has derived Törnqvist's formulas as follows. Let us consider weighted moment means $({}_\alpha W_0^1)^\alpha$ and geometric means ${}_0W_0^1$ of wage ratios defined by

$$({}_\alpha W_0^1)^\alpha = \Sigma\, p_\nu (w_\nu^1/w_\nu^0)^\alpha = \Sigma\, p_\nu e^{\alpha \ln (w_\nu^1/w_\nu^0)}, \tag{A.1}$$

$$\ln({}_0W_0^1) = \Sigma\, p_\nu \ln (w_\nu^1/w_\nu^0), \tag{A.2}$$

where $p_\nu \geq 0$ and $\Sigma\, p_\nu = 1$.
Dividing every term of (A.1) by $({}_0W_0^1)^\alpha$ we get

$$\begin{aligned}({}_\alpha W_0^1/{}_0W_0^1)^\alpha &= \Sigma\, p_\nu e^{\alpha \ln (w_\nu^1/w_\nu^0({}_0W_0^1))}\\ &= \Sigma\, p_\nu e^{\alpha \dot{w}_\nu},\end{aligned} \tag{A.3}$$

where $\dot{w}_\nu = \ln (w_\nu^1/w_\nu^0) - \ln ({}_0W_0^1)$. By expanding (A.3) to a power series of α we get for all values of $\dot{w}_\nu$:s

$$({}_\alpha W_0^1/{}_0W_0^1)^\alpha = 1 + \frac{\alpha^2}{2!} \Sigma\, p_\nu \dot{w}_\nu^2 + \frac{\alpha^3}{3!} \Sigma\, p_\nu \dot{w}_\nu^3 + \ldots. \tag{A.4}$$

Taking logarithms, dividing both sides of (A.4) by α and rearranging the terms we get:

$$\ln ({}_\alpha W_0^1) = \ln ({}_0W_0^1) + \frac{\alpha}{2!} \Sigma\, p_\nu \dot{w}_\nu^2 + \frac{\alpha^2}{3!} \Sigma\, p_\nu \dot{w}_\nu^3. \tag{A.5}$$

Specifying $\alpha = 1$ and $w_\nu^0 = 1$ we get the formula (A.6).

$$\ln \bar{w}_{i\cdot} = \sum_{\nu} p_{i\nu} \ln w_{i\nu} + \frac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^{2}_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^{3}_{i\nu} + \dots , \tag{A.6}$$

where

$$\dot{w}_{i\nu} = \ln w_{i\nu} - \sum_{\nu} p_{i\nu} \ln w_{i\nu} .$$

Substituting $\ln w_{i\nu}$ from (1) to (A.6) we get

$$\ln \bar{w}_{i\cdot} = \alpha + \sum_{k} \beta_k \overline{\ln x}_{i\cdot k} + \sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \frac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^{2}_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^{3}_{i\nu} + \dots , \tag{A.7}$$

where

$$\overline{\ln x}_{i\cdot k} = \sum_{\nu} p_{i\nu} \ln x_{i\nu k} . \tag{5}$$

Let us define α' as the sum of α and the average of the residual of the model (A.7)

$$\begin{aligned} \alpha' &= \alpha + \Big(\sum_{i} p'_{i} \sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \frac{1}{2} \sum_{i} p'_{i} \sum_{\nu} p_{i\nu} \dot{w}^{2}_{i\nu} + \frac{1}{6} \sum_{i} p'_{i} \sum_{\nu} p_{i\nu} \dot{w}^{3}_{i\nu}\Big) \\ &= \alpha + \Big[\sum_{i} p'_{i} \sum_{\nu} p_{i\nu} \Big(\frac{1}{2} \dot{w}^{2}_{i\nu} + \frac{1}{6} \dot{w}^{3}_{i\nu} + \dots\Big)\Big] , \end{aligned} \tag{4}$$

since $\sum_{i} p'_{i} \sum_{\nu} p_{i\nu} \varepsilon_{i\nu} = 0$.

Let us define $\varepsilon'_{i\cdot}$ as the difference of the residual of model (A.7) and its average

$$\begin{aligned} \varepsilon'_{i\cdot} &= \Big(\sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \frac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^{2}_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^{3}_{i\nu} + \dots\Big) \\ &\quad - \Big[\sum_{i} p'_{i} \sum_{\nu} p_{i\nu} \Big(\frac{1}{2} \dot{w}^{2}_{i\nu} + \frac{1}{6} \dot{w}^{3}_{i\,\nu} + \dots\Big)\Big] \\ &= \Big(\sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \frac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^{2}_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^{3}_{i\nu} + \dots\Big) - (\alpha' - \alpha) . \end{aligned} \tag{6}$$

Then at the aggregate level we obtain a model for $\ln \bar{w}_{i\cdot}$.

$$\ln \bar{w}_{i\cdot} = \alpha' + \sum_k \beta_k \overline{\ln x}_{i\cdot k} + \varepsilon'_{i\cdot} , \tag{3}$$

where $\sum_i p'_i \varepsilon'_{i\cdot} = 0$ because of the way the α':s and $\varepsilon'_{i\cdot}$:s have been constructed.

Appendix B

The proof of monotonicity of the function (11)

Let us consider the function $\frac{d\theta_p^2}{dp}$. We get

$$\frac{d\theta_p^2}{dp} = \theta_p^2 \left(\frac{-2p(1+3q)}{q(1+pq)} + \frac{2\delta\lambda pq(2p-1)}{1+\lambda p^2 q^2}\right) . \tag{B.1}$$

The term $\frac{-2p(1+3q)}{q(1+pq)} \leq 0$, when $0 \leq p \leq 1$ and $\frac{2\delta\lambda pq(2p-1)}{1+\lambda p^2q^2} \leq 0$ when $0 \leq p \leq \frac{1}{2}$. Therefore, $\frac{d\theta_p^2}{dp} \leq 0$, when $0 \leq p \leq \frac{1}{2}$.

In the interval $\frac{1}{2} < p < 1$, $\frac{d\theta_p^2}{dp} \leq 0$, if

$$\delta \leq \frac{p(1+3q)(1+\lambda p^2q^2)}{q(1+pq)\lambda pq(2p-1)} = \frac{(1+3q)(\frac{1}{\lambda}+p^2q^2)}{q^2(1+pq)(2p-1)} . \tag{B.2}$$

Since, in the interval $\frac{1}{2} < p < 1$,

$$\frac{(1+3q)p^2}{(1+pq)(2p-1)} < \frac{(1+3q)(\frac{1}{\lambda}+p^2q^2)}{q^2(1+pq)(2p-1)} , \qquad (0 < \lambda < \infty) , \tag{B.3}$$

then the inequality (B.2) holds true in the same interval if

$$\delta \leq \frac{(1+3q)p^2}{(1+pq)(2p-1)} . \tag{B.4}$$

The smallest value of the right-hand side of (B.4) in the interval $1/2 \leq p \leq 1$ is 1, and thus, in the interval $1/2 < p < 1$ $\frac{d\theta_p^2}{dp} \leq 0$ if

$$0 \leq \delta \leq 1 \tag{13}$$

and

$$0 < \lambda < \infty . \tag{14}$$

In addition, $-\infty < \frac{d\theta_p^2}{dp}$ in the interval $0 \leq p \leq 1$ if $|\delta\lambda| < \infty$.

References

Amemiya, T., 1977, A note on a heteroscedastic model, Journal of Econometrics 6, 365-370.

Draper, N.R. and Smith, H., 1981, Applied regression analysis, 2nd ed. (Wiley, New York).

Glejser, H., 1969, A new test for heteroscedasticity, Journal of the American Statistical Association 64, 316-323.

Goldfeld, S.M. and Quandt, R.E., 1972, Nonlinear methods in econometrics, Ch. 3: Analyses of heteroscedasticity (North-Holland, Amsterdam) 78-123.

Harvey, A.C., 1974, Estimation of parameters in a heteroscedastic regression model, Paper presented at the European Meeting of the Econometric Society, Grenoble, Sept.

Harvey, A.C., 1976, Estimating regression models with multiplicative heteroscedasticity, Econometrica 44, 461-465.

Lastikka, P., Relative differences in industrial workers' average hourly wages between various geographical regions and branches of industry in Finland, 1960-1971, A forthcoming study (Research Institute of the Finnish Economy, Helsinki).

Park, R.E., 1966, Estimation with heteroscedastic error terms, Econometrica 34, 888.

Rutemiller, H.C. and Bowers, D.A., 1968, Estimation in a heteroscedastic regression model, Journal of the American Statistical Association 63, 552-557.

Törnqvist, L., 1936, Levnadskostnadsindexerna i Finland och Sverige, deras tillförlitlighet och jämförbarhet, Ekonomiska Samfundets Tidskrift 37, 59-93.

Vartia, Y., 1976, Fisher's five-tined fork and other quantum theories of index number, Discussion paper no. 4 (Research Institute of the Finnish Economy, Helsinki).

Table 1. The estimation results for model (15). Data related to 38 geographically connected areas[a].

Number of model	Regression coefficients and t-values[b]					Error terms standard deviation	Coefficient of multiple correlation
	λ	$\hat{\gamma}$	t	$\hat{\delta}$	t		
1	1	-6.829	26.73	62.140	3.88	1.4611	0.5913
2	10	-6.802	26.85	7.809	4.06	1.4410	0.6062
3	10^2	-6.696	27.32	1.935	4.76	1.3636	0.6585
4	10^3	-6.465	28.16	0.972	6.19	1.2117	0.7435
5	10^4	-6.039	29.40	0.694	8.86	0.9758	0.8426
6	10^5	-5.509	32.31	0.555	13.44	0.7096	0.9201
7	10^6	-5.132	33.14	0.445	16.86	0.5834	0.9467
8	2.10^6	-5.029	31.93	0.421	17.12	0.5757	0.9482
9*	3.10^6	-4.963	30.94	0.408	17.14	0.5762	0.9483
10	4.10^6	-4.912	30.12	0.401	17.11	0.5759	0.9481
11	10^7	-4.721	27.11	0.382	16.93	0.5815	0.9471
12	10^8	-4.046	18.99	0.365	16.58	0.5922	0.9451

[a] The dependent variable of model (15) was constructed by using formula (18) and the independent variable p_i was computed from the formula $p_i = \bar{N}^{wr}_{i..}/\bar{N}^{wr}_{...}$ where $\bar{N}^{wr}_{i..}$ stands for the average number of industrial workers in region i in 1960-1971 ($q_i = 1-p_i$) and $\bar{N}^{wr}_{...}$ (= 359608) stands for the corresponding variable in the whole country in 1960-1971.
Error terms standard deviation was computed from the formula $s(u_{p_i}) = \sqrt{\frac{1}{36}\sum_{i=1}^{38} u^2_{p_i}}$, where u_p is the residual of model (15). The coefficient of multiple correlation was computed from the formula

$$R = \sqrt{1 - \frac{s^2(u_{p_i})}{s^2(\ln s^2_{p_i})}}, \quad \text{where} \quad s^2(\ln s^2_{p_i}) = \frac{1}{37}\sum_{i=1}^{38}(\ln s^2_{p_i} - \overline{\ln s^2_{p.}}) \text{ and } \overline{\ln s^2_{p.}} = \frac{1}{38}\sum_{i=1}^{38}\ln s^2_{p_i}.$$

[b] The absolute value of regression coefficient divided by its standard deviation calculated under the assumption that the observations are independent.

Table 2. Models for industrial workers' average hourly wage rate (%). Data related to 170 commuting regions in 1970. Models have been estimated by using the weights (26)[a].

Explanatory variable and its t-value[b] / Number of model	1	2	3	4	5	6	7	8	9	10	11
Constant	164.402	-129.765	-138.821	-124.066	-112.077	-111.077	-103.148	-98.727	-84.340	-86.721	-86.607
t-value	154.17	7.40	10.23	9.78	8.71	8.60	8.06	7.65	6.23	6.47	6.50
Calculated hourly wage rate (%)		1.758	1.473	1.405	1.330	1.336	1.318	1.325	1.366	1.338	1.343
t-value		16.78	17.31	17.83	16.39	16.56	16.76	16.96	17.59	17.27	17.41
Regional index of female industrial workers						-4.597	-5.880	-6.535	-7.396	-6.981	-7.452
t-value						1.81	2.34	2.60	2.99	2.85	3.04
Regional index of salaried industrial employees				18.123	16.894	17.928	15.056	13.173	10.319	11.220	9.665
t-value				5.77	5.45	5.73	4.74	3.99	3.06	3.35	2.79
Indicator of additional education (%)			5.726	3.679	3.289	3.268	3.251	3.196	1.471	2.239	2.535
t-value			10.71	6.09	5.43	5.43	5.56	5.50	1.79	2.57	2.86
Share of leftists					14.729	17.102	16.050	10.274	11.151	13.898	15.377
t-value					2.96	3.35	3.22	1.78	1.97	2.44	2.68
Regional index of productivity (%)							0.059	0.056	0.054	0.051	0.055
t-value							3.21	3.00	2.99	2.89	3.06
Regional index of the average size of establishments (%)								0.026	0.035	0.033	0.029
t-value								1.92	2.58	2.48	2.16
Regional consumer price index (%)									0.649	0.677	0.687
t-value									2.91	3.08	3.14
Industrial concentration in three biggest branches (%)										0.058	0.077
t-value										2.37	2.85
Unemployment rate (%)											-0.020
t-value											1.64
Degrees of freedom	169	168	167	166	165	164	163	162	161	160	159
Error terms standard deviation	32.587	19.977	15.426	14.122	13.802	13.709	13.336	13.228	12.932	12.752	12.686
Coefficient of multinle correlation	0	0.790	0.881	0.901	0.906	0.907	0.912	0.914	0.918	0.920	0.921

a) Log-percentages are defined by the operator 100 ln (   ) and denoted by the symbol (%).

The model k (k=1,...,11) was estimated by minimizing $\sum_i (\omega_i e_{ik})^2$, where ω_i is computed from formula (26).

Error terms standard deviation was computed from the formula

$$s(e_k) = \sqrt{\frac{1}{n-m} \sum_{i=1}^{n} (\omega_i e_{ik})^2} ,$$

where n stands for the number of regions and m for the number of explanatory variables.

Coefficient of multiple correlation was computed from the formula

$$R_k = \sqrt{1 - \frac{s^2(e_k)}{s^2(y)}},$$

where $s^2(y) = \frac{1}{n-1} \sum_i \omega_i^2 (y_i - \bar{y})^2$ and $\bar{y} = \frac{\sum_i \omega_i^2 y_i}{\sum_i \omega_i^2}$.

Table 3. The estimated variances from small ($1 < \omega_i \leq 1.0027$) and large ($4.096 \leq \omega_i \leq 7.765$) regions computed from the residuals of models 1-11 in Table 2, the number of observations and the ratio between variances computed from small and large regions[a)]

The number of model in Table 2	Small regions			Large regions				
	s^2_{1s}	s^2_{2s}	Number of observations	$s^2_{1\ell}$	$s^2_{2\ell}$	Number of observations	$s^2_{1s}/s^2_{1\ell}$	$s^2_{2s}/s^2_{2\ell}$
1	595.42	482.23	10	1150.98	1080.37	10	0.52	0.45
2	483.21	465.61	10	463.16	350.86	10	1.04	1.33
3	335.99	368.96	10	346.76	381.33	10	0.97	0.97
4	290.06	316.94	10	268.06	291.03	10	1.08	1.09
5	298.70	320.21	10	257.02	243.04	10	1.16	1.32
6	299.21	321.56	10	282.61	268.75	10	1.06	1.20
7	347.32	364.75	10	263.10	256.09	10	1.32	1.42
8	362.98	346.29	10	265.33	219.32	10	1.37	1.58
9	374.29	368.99	10	195.59	157.16	10	1.91	2.35
10	339.62	352.02	10	151.72	128.23	10	2.24	2.75
11	374.07	384.09	10	132.68	118.85	10	2.82	3.23

[a)] The variances s_1^2 and s_2^2 have been computed from the formulas $s_1^2 = \frac{1}{n}\sum_i e_i^2$ and

$$s_2^2 = \frac{1}{n-1}\sum_i (e_i - \bar{e})^2, \quad \text{where } \bar{e} = \frac{1}{n}\sum_i e_i, \ (i = 1,\dots,n).$$

The index s refers to small regions and the index ℓ to large regions.

Table 4. The estimation results for model (15). Data related to 22 branches of industry (two-digit ISIC[a]) subdivision)[b].

Number of model	Regression coefficients and t-values[c]					Error terms standard deviation	Coefficient of multiple correlation
	λ	$\hat{\gamma}$	t	$\hat{\delta}$	t		
1	1	-8.022	40.00	165.775	3.24	.6923	.6289
2	10	-8.018	39.85	17.146	3.24	.6921	.6292
3	50	-8.003	39.26	3.911	3.25	.6913	.6303
4	100	-7.988	38.65	2.235	3.26	.6905	.6314
5	200	-7.964	37.68	1.375	3.27	.6896	.6326
6	300	-7.945	36.90	1.074	3.28	.68924	.6331
7*	400	-7.930	36.25	0.917	3.28	.68919	.6332
8	500	-7.917	35.70	0.819	3.28	.6893	.6330
9	700	-7.896	34.78	0.699	3.27	.6899	.6322
10	1000	-7.871	33.71	0.602	3.26	.6910	.6307

a) International Standard Industrial Classification of All Economic Activities.

b) The dependent variable of model (15) was constructed by using formula (19) and the independent variables p_j and q_j were computed from formula (29).

The error terms standard deviation was computed from the formula

$$s(u_{p_j}) = \sqrt{\frac{1}{20} \sum_{j=1}^{22} u^2_{p_j}} \,,$$

where u_{p_j} is the residual of model (15). The coefficient of multiple correlation was computed from the formula

$$R = \sqrt{1 - \frac{s^2(u_{p_j})}{s^2(\ln s^2_{p_j})}} \,, \text{ where } s^2(\ln s^2_{p_j}) = \frac{1}{21} \sum_{j=1}^{22} (\ln s^2_{p_j} - \overline{\ln s^2_{p.}})^2 \text{ and } \overline{\ln s^2_{p.}} = \frac{1}{22} \sum_{j=1}^{22} \ln s^2_{p_j} .$$

c) See footnote b of Table 1.

Table 5. Estimates of the standard deviation $\hat{\sigma}_0$ of the logarithm of individual workers' hourly wage rate calculated from different sets of observations(%)[a].

Set of observations	$\hat{\sigma}_0$
Confederation of Finnish Employers (218 000 observations)[b]	23.12
Commuting regions (170 regions)	32.59
Branches of industry (22 branches)	23.02

[a] If the standard deviation of the variable $\ln w_{i\nu}$ is $\hat{\sigma}_0$, then the standard deviation of the variable $\ln w_{i\nu}$ in log-percentages is $100 \cdot \hat{\sigma}_0$. (See footnote a) of Table 2).

[b] Using the statistics collected by the Confederation of Finnish Employers, the estimate $\hat{\sigma}_0$ was calculated from observations for the 3rd quarter of 1976. The material included 218 000 workers in all, which is about half the total number of industrial workers in Finland. The estimates $\hat{\sigma}_0$ calculated from the data collected by the Confederation of Finnish Employers were very stable from 1970 to 1976, so that it is possible to assume, as a working hypothesis, that in 1970 the standard deviation of the logarithm of individual workers' hourly wage rate was in the whole population about 23.12 (%).

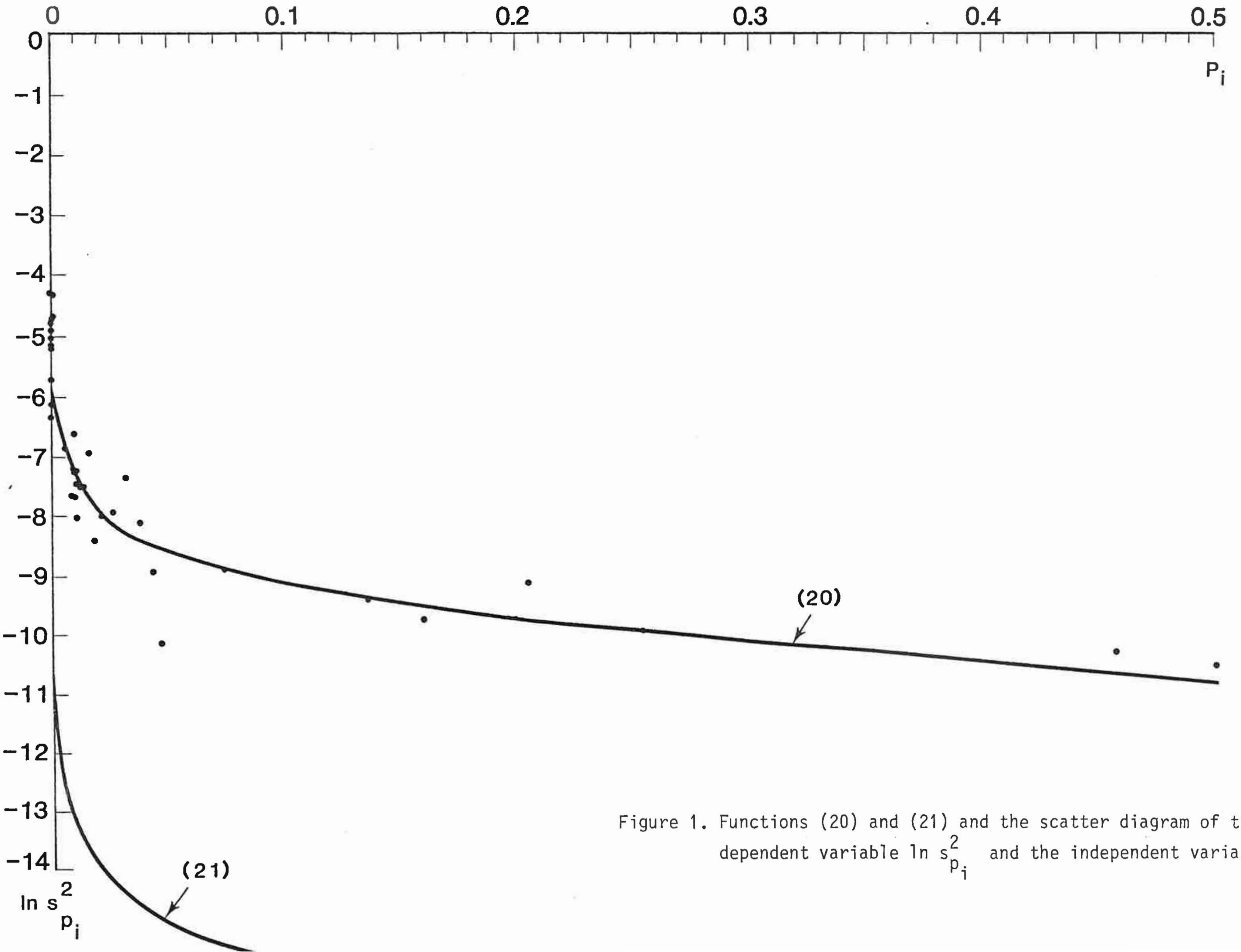


Figure 1. Functions (20) and (21) and the scatter diagram of the dependent variable $\ln s^2_{p_i}$ and the independent variable p_i.

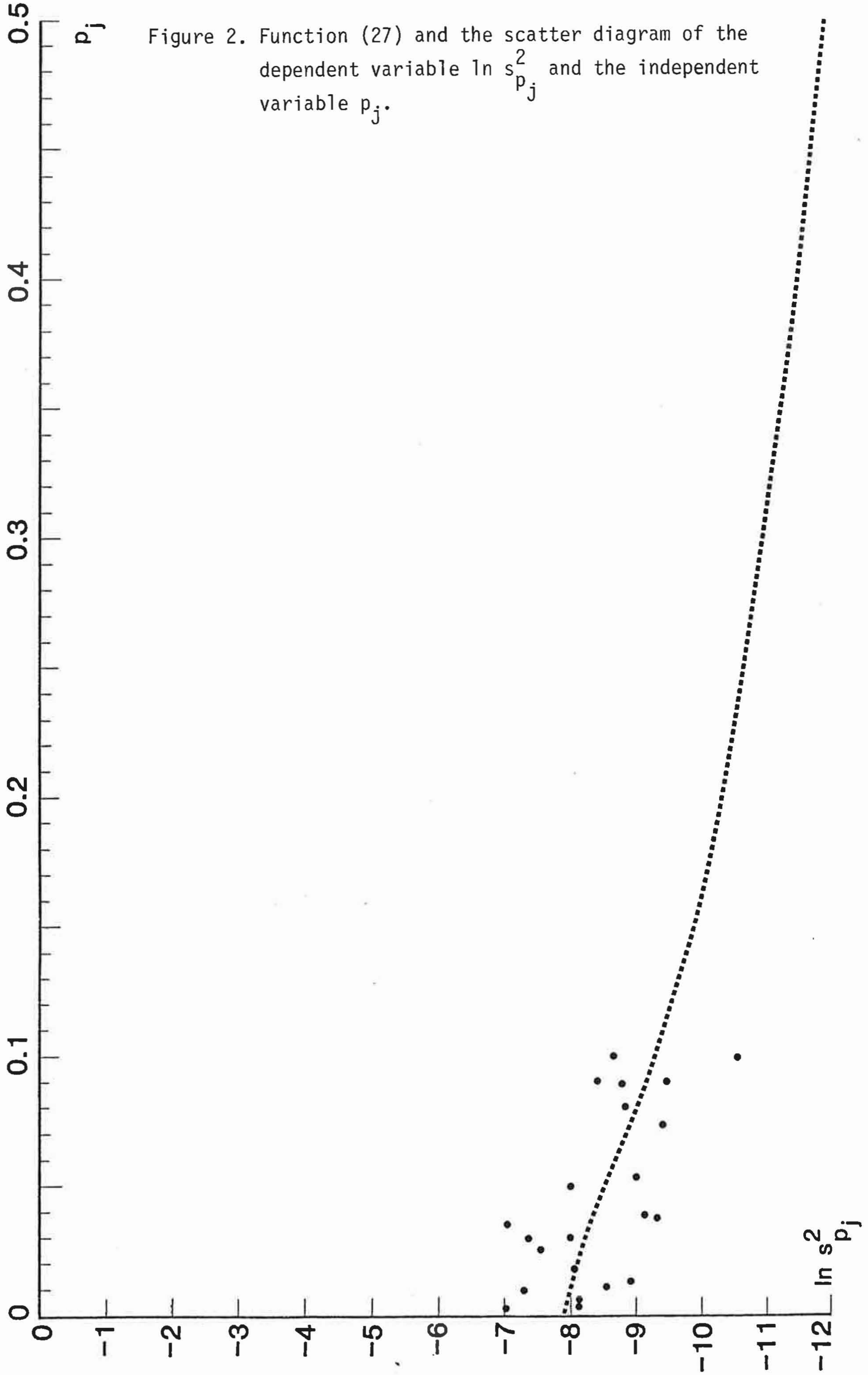
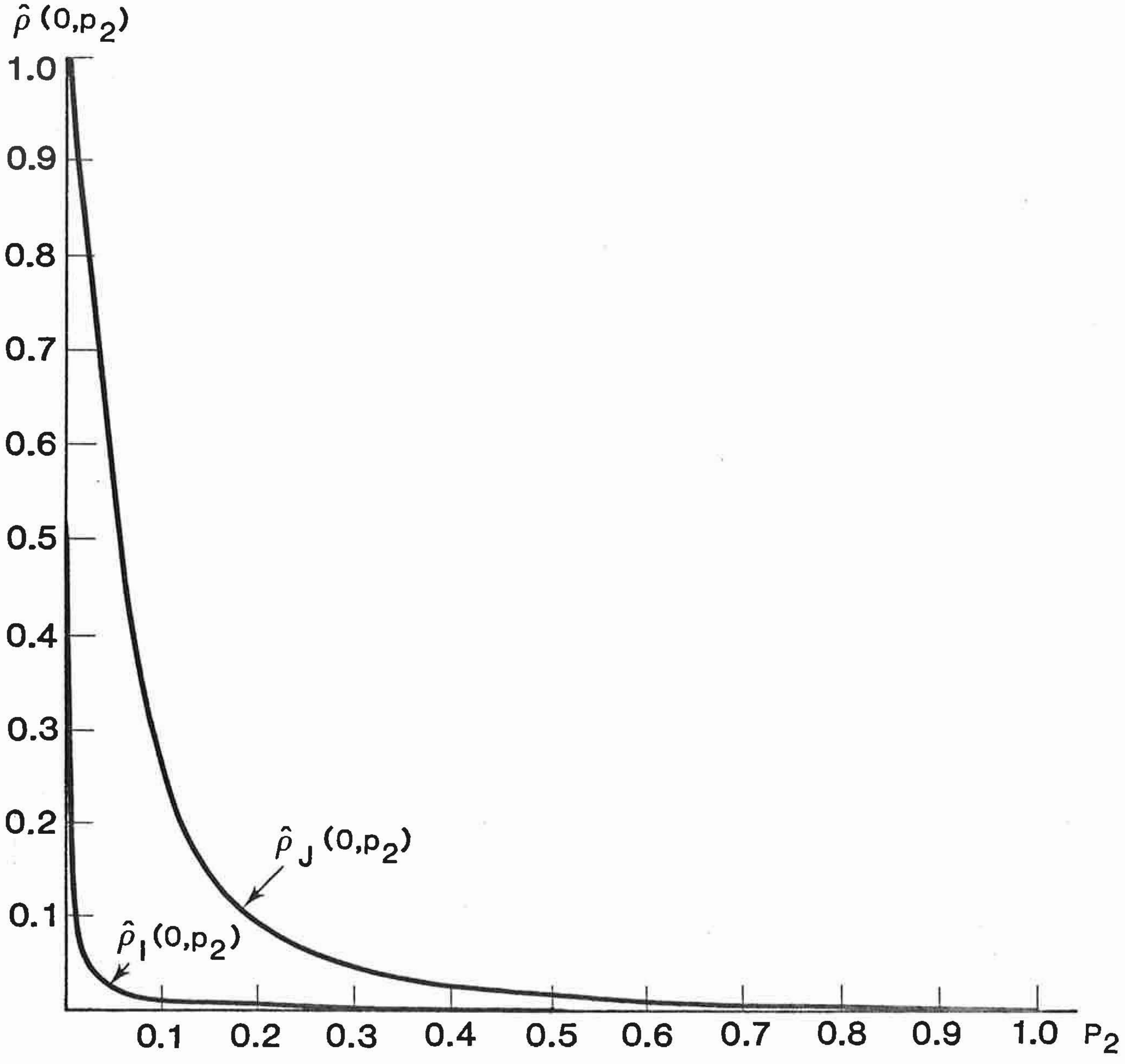


Figure 2. Function (27) and the scatter diagram of the dependent variable $\ln s^2_{p_j}$ and the independent variable p_j.

Figure 3. The intra-class correlation coefficients related to connected geographical regions ($\hat{\rho}_I(0,p_2)$) and branches of industry ($\hat{\rho}_J(0,p_2)$)[a].



[a] The intra-class correlation coefficients have been computed from formula (33) utilizing thereby formulas (20) and (27) in the case of geographical regions and branches of industry respectively.