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**

No 105 1.4.1982

A paper presented at the European Meeting of the Econometric Society in Dublin, September 5-9,1982.

*The Research Institute of the Finnish Economy, Helsinki, Finland.

**This paper is an abridged version of a forthcoming doctoral thesis at the University of Helsinki concerning "Relative Differences in Industrial Workers' Average Hourly Wages between Various Geographical Regions and Branches of Industry in Finland, 1960-1971".

Acknowledgements. I wish to express my deepest gratitude to my teacher, Professor Leo Törnqvist, for numerous stimulating discussions and encouraging this work. 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.

ABSTRACT FORM

Author(s):
Pekka Lastikka*

Institute:
The Research Institute of the Finnish Economy, Helsinki, Finland

Title:
Estimating heteroscedastic regression models: some new methods and their applications

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

The models for the error variance are used for the following estimations:

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 geographical regions and industrial branches of different size, utilizing the generalized intra-class correlation coefficient. A method is developed by means of which the "wage transfer" effect can be quantitatively measured from cross-sectional data when the population is finite.

1. Introduction

This paper concerns the estimation of wage models from cross-sectional data where the observations are of different size and they are not independent. The main point of interest is in the heteroscedasticity of the wage models and I focus on the problem which weights 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 error terms variance of the wage models, which idea is not new in itself. [Glejser (1969), Goldfeld-Quandt (1965), Harvey (1974,1976), Rutemiller-Bowers (1968)]. In contrast to earlier studies there is in this paper a special interest in the principles in which way a model to the error terms variance should be constructed. The method to estimate the dependent variable in that model is a new one and bases on some short time series of wages from geographical regions and industrial branches of different size.

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

Separate models for the error terms variance is estimated for geographical regions and for industrial branches. These models turned out to be very informative and useful. Except their use to estimate weights needed for cross-sectional wage models they can also be used to some other purposes proposed in chapter 4.

2. The explanatory model of wages and the model of the residual variance

2.1 The explanatory model of the logarithm industrial workers' wages

Let $w_{i\nu}$ stand for the hourly wage paid to the industrial worker ν employed in the region i for the time period t (the subscript t will be omitted because the following considerations relate to the same period t). Let i belong to the whole Finland and industrial worker ν to the set N_i of all industrial workers in the region i. To explain the logarithm of the wage $w_{i\nu}$ in the population of all industrial workers in the whole Finland N_F we introduce the model

$$(2.1) \qquad \ln w_{i\nu} = \alpha + \sum_k \beta_k \ln x_{i\nu k} + \varepsilon_{i\nu} , \qquad \begin{matrix} k \in K \\ i \in F \\ \nu \in N_i \end{matrix}$$

where $\ln x_{i\nu k}$ is the value of the k:th explanatory variable belonging to the set K, and $\varepsilon_{i\nu}$ is the residual. Let α be a constant such that $\sum_i \sum_\nu p_{i\nu} \varepsilon_{i\nu} = 0$ and let the β_k:s be those constants which would be obtained if model (2.1) could be estimated from the whole population by OLS. The weights $p_{i\nu}$ will be regarded as non-stochastic variables and $\sum_\nu p_{i\nu} = p_{i.} = 1$.

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

$$(2.2) \qquad \ln \bar{w}_{i.} = \ln \sum_\nu p_{i\nu} w_{i\nu} .$$

In appendix 1 it has been shown that a regression model at the regional level for the variable (2.2) can be evaluated:

$$(2.3) \qquad \ln \bar{w}_{i\cdot} = \alpha' + \sum_k \beta_k \overline{\ln x}_{i\cdot k} + \varepsilon'_{i\cdot} \ ,$$

where

$$(2.4)^{1)} \qquad \alpha' = \alpha + [\Sigma\Sigma\Sigma p_{i\nu}(\frac{1}{2}\dot{w}^2_{i\nu} + \frac{1}{6}\dot{w}^3_{i\nu} + \ldots)]$$

and

$$(2.5) \qquad \overline{\ln x}_{i\cdot k} = \sum_\nu p_{i\nu} \ln x_{i\nu k}$$

and

$$(2.6) \qquad \varepsilon'_{i\cdot} = (\sum_\nu p_{i\nu}\varepsilon_{i\nu} + \frac{1}{2}\sum_\nu p_{i\nu}\dot{w}^2_{i\nu} + \frac{1}{6}\sum_\nu p_{i\nu}\dot{w}^3_{i\nu} + \ldots) - (\alpha' - \alpha).$$

In the model (2.3) $E\varepsilon'_{i\cdot} = \sum_i p_{i\cdot}\,\varepsilon'_{i\cdot} = \sum_i \varepsilon'_{i\cdot} = 0$ owing to the α':s and $\varepsilon'_{i\cdot}$:s manner of construction.

In this paper a model of type (2.3) is estimated from sets of cross-sectional data for geographical areas (commuting regions) and industrial branches. Since the statistical units associated with the regional and the industrial -branch divisions differ in size, in terms of the number of workers, particular attention in estimating the regression models of the type (2.3) will be devoted to rendering the observations on the various regions and branches comparable through weighting. For this purpose a model is estimated for the residual variance of model (2.3).

---

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

Although the observations are mutually dependent, the independence assumption of observations is used as a working hypothesis in estimating the models (2.3) because estimates for the correlations of observations were not available.[1] The intention is to estimate model (2.3) mainly by empirical means. An effort will be made to avoid theoretical, restricting assumptions the correctness of which could be considered questionable in the empirical data.

2.2. The explanatory model of the residual variance

An effort is made to construct a model for the theoretical (expected) residual variance of model (2.3). In the explanatory model of the residual variance, the size of the region, relative to the whole population, was chosen as the only explanatory variable, even though other variables descriptive of the characteristics of the regions could also be used as explanatory variables. The relative size p of the region 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 residual variance. The size of the expected residual variance associated with a geographically connected area of size p will be denoted by σ_p^2 .

---

1) I have studied the effect of this assumption on regression models and statistics utilizing thereby cross-sectional data concerning administrative labour-force districts. These districts are much bigger regions than commuting regions. In 1970 Finland was divided into 11 administrative labour-force districts. These studies suggested, i.a., that the variance of regression coefficient can be quite sensitive to the independence assumption of observations. I compared thereby the variances of regression coefficient estimated under the assumptions that the observations were independent and that the observations were dependent, in which case I utilized estimated correlations between observations, i. e. $\binom{11}{2} = 55$ estimated correlations.

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 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 residuals $\varepsilon_{i.}'$ of model (2.3). Let ε_p' be the residual associated with a region of size p and let ε_q' be the residual associated with a region of size q ($q = 1 - \rho$). From the properties of the mean, and taking into account that the residual associated with the whole country is $\varepsilon_1' = 0$, we have

$$(2.7) \qquad p\varepsilon_p' + q\varepsilon_q' = \varepsilon_1' = 0$$

whence, further,

$$(2.8) \qquad p\varepsilon_p' = -q\varepsilon_q'$$

and, by squaring (2.8),

$$(2.9) \qquad p^2(\varepsilon_p')^2 = q^2(\varepsilon_q')^2 \ .$$

Taking the mathematical expectation, on both sides of (2.9), over all the areas of size p and q belonging to the population that is formed of regions that form a connected geographical whole and are mutually exclusive in the same regional division we get, after re-arranging the factors,

$$(2.10) \qquad \frac{\sigma_p^2}{\sigma_q^2} = \frac{q^2}{p^2}$$

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 (2.10) will be satisfied.

On the basis of the above, the following desiderata are imposed on the explanatory model of the residual 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_θ^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,

$$(2.11) \qquad \theta_p^2 = q^2(1+pq)^2 e^{-\delta \ln(1+\lambda p^2 q^2)}, \qquad (\lambda>0)$$

was settled on after many trials.

Then,

$$(2.12) \quad \sigma_p^2 = q^2(1+pq)^2 e^{\gamma - \delta \ln(1+\lambda p^2 q^2)},$$

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

It is easily seen that desideratum 1 is met in the case of function (2.12). 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 (2.12). 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 $e^{-\delta \ln(1+\lambda p^2 q^2)}$ in the same interval is also 1 ($\delta \geq 0$). In the case of the function (2.11), $\theta_0^2 = 1$ and $\theta_1^2 = 0$ hold true, in addition.

Also (2.11) is continuous and has a derivative for all values $p \in [0,1]$. It can be shown (appendix 2) that $\frac{d\theta_p^2}{dp} \leq 0$, when $p \in [0,1]$,

if

$$(2.13) \quad 0 \leq \delta \leq 1 \qquad \text{and}$$

$$(2.14) \quad 0 < \lambda < \infty .$$

In addition, $\frac{d\theta_p^2}{dp} > -\infty$ in the interval $0 \leq p \leq 1$ if $|\delta\lambda| < \infty$. If conditions (2.13) and (2.14) are met, desideratum 3 is satisfied by function (2.11) in the interval $0 \leq p \leq 1$.

Desideratum 4 is also satisfied by function (2.11) since the function θ_p^2 is of the type $\theta_p^2 = q^2 f(p,q)$, where

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

2.3. Estimation of the residual variance

Transformation to logarithms (ln) in both sides of (2.12), replacement of σ_p^2 by its estimate s_p^2, computed from the observational data, and introduction of the error term ξ_p into formula (2.12) yield the model

$$(2.15) \quad \ln s_p^2 = 2\ln q + 2\ln(1+pq) + \gamma - \delta \ln(1+\lambda p^2 q^2) + \xi_p$$

Model (2.15) can be estimated from the observational data if estimates of $\ln s_p^2$ for various values of p are available.

In the present study, the estimates of $\ln s_p^2$ are based on the estimation of the residual variance of the explanatory model of $\ln \bar{w}_{i.}$ from time series data. Regarding such a model of $\ln \bar{w}_{i.}$ it is justified to presuppose that the parameters of this model do not depend on p, since the parameters of model (2.3), except the coefficient, α', of the constant do not depend on the region. In addition, it can be presupposed that the residual of the explanatory model of $\ln \bar{w}_{i.}$ approaches zero as the size of the region approaches the population. This condition corresponds to the property $E\varepsilon'_{i.} = \sum_i \varepsilon'_{i.} = 0$ of model (2.3).

The variance s_p^2 involved in (2.15) was estimated in this study as follows. From the time-series data for the various branches of industry j and variously sized regions i, the residual variance $s_t^2(\bar{e}_{i\cdot t})$ for model

$$(2.16) \quad \ln\bar{w}_{i\cdot t} = \ln\bar{w}_{i\cdot t-1} + (\ln \bar{w}_{\cdot\cdot t} - \ln\bar{w}_{\cdot\cdot t-1}) + \bar{e}_{i\cdot t}$$ (where · is in the place of index j)

was computed from the formula

$$(2.17) \quad s_t^2(\bar{e}_{i\cdot t}) = \frac{1}{T-1}\sum_{t=1}^{T}(\bar{e}_{i\cdot t} - \bar{e}_{i\cdot\cdot})^2.$$

The desiderata imposed on the explanatory model of $\ln\bar{w}_{i\cdot}$ are satisfied by model (2.16): its coefficients do not depend on the size of the region (the parameters of model (2.16)) can be considered to equal unity in absolute value) and the residual $\bar{e}_{i\cdot t}$ of the model approaches zero as the size of the region i approaches the whole country.

The use of the residual variance $s_t^2(\bar{e}_{i\cdot t})$, as estimated from time series related to variously sized areas in accordance with formula (2.17), in estimating the residual variance σ_p^2 is based on the working hypothesis that the residual variance estimated from time series data for variously sized regions behaves approximately as the residual variance of model (2.3) as the size of the region 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 inter-region variation observed in wages at any given point in time can be considered 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 hypothesis is introduced that the residual variances of models (2.16) and (2.3) decrease at approximately equal rates as the size of the region decreases, the residual variances of model (2.16), related to time-series data, can be expected to be smaller in absolute value than the residual variances of model (2.3) associated with cross-sectional data for regions of the corresponding size. This does not matter in the estimation of model (2.3), since, for that purpose, it is necessary to know only the relative weights rather than the absolute weights of the observations.

Time series of industrial workers' average hourly wages in the years 1960-1971 were formed for 38 variously sized regions, which formed a geographically connected whole, just as did the commuting regions and the administrative labour-force districts. The 38 regions were mainly regions other than the commuting regions and the administrative labour-force districts. From this observational data, the residual variances of model (2.16) were estimated for the various regions.

2.4. Estimation of the explanatory model (2.15) of the residual variance

Model (2.15) 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 (2.15) are set out in Table (2.1). As appears from Table (2.1) the multiple correlation coefficient attained its maximum value when λ was given the value $\lambda = 3 \cdot 10^6$ (Model 9).

A scatter diagram of the observations on the regressand $\ln s_{pi}^2$ and the regressor p_i of the models given in Table (2.1) is presented in Chart (2.1). In addition, the graph of the function

$$(2.18) \qquad \ln s_p^2 \approx 2 \ln q + 2 \ln(1 + pq) - 4.963 - 0.408 \ln(1 + 3 \cdot 10^6 p^2 q^2)$$

corresponding to the Model 9 given in Table (2.1) and the graph of the function

$$(2.19) \qquad \ln s_p^2 = -4.963 + \ln \frac{1}{359608 \cdot p} + \ln q, \quad (p \in [\frac{1}{359608}, 1] \text{ ja } q = 1 - p)$$

are represented in Chart (2.1).

Function (2.19) is related to that imagined situation in which the residuals $\varepsilon_{i.}^{!}$ of model (2.3) are means computed from a random sample. In that case, each of the 38 regions used as units of observation is interpreted as a random sample drawn without replacement from the population

Table 2.1. The estimation results for model (2.15). Data composed of 38 geographically connected areas.[1)]

Number of model	Regression coefficients and t-values					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

1) The dependent variable of model (2.15) was constructed utilizing formula (2.17) 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 and $\bar{N}^{wr}_{...}$ stands for the corresponding variable in the whole country in 1960-1971 ($q_i = 1-p_i$).

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 empirical residual of model (2.15). The coefficient of multiple correlation was computed from the formula $R = \sqrt{1 - \frac{s^2(u_{p_i})}{s^2(\ln s^2_{p_i})}}$, where $s^2(\ln s^2_{p_i}) = \frac{1}{37}\sum_{i=1}^{38}(\ln s^2_{p_i} - \overline{\ln s^2_{p.}})$ and $\overline{\ln s^2_{p.}} = \frac{1}{38}\sum_{i=1}^{38} \ln s^2_{p_i}$.

formed by the industrial workers in the whole country.[1)]

Chart (2.1) 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, the residual variance associated with the regions obtained by dividing up the population will decrease definitely more slowly than the residual variance connected with random sampling.

---

1) 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

$$\sigma^2 = \frac{\sigma_0^2}{N_i^{wr}} \left(1 - \frac{N_i^{wr} - 1}{N^{wr} - 1}\right),$$

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

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

$$(2.20) \quad \sigma^2 \approx \frac{0.00699}{359608\, p}\, q,$$

where $q = 1 - p$. Transforming to logarithms on both sides in (2.20), an approximation corresponding to the function (2.19) is obtained.

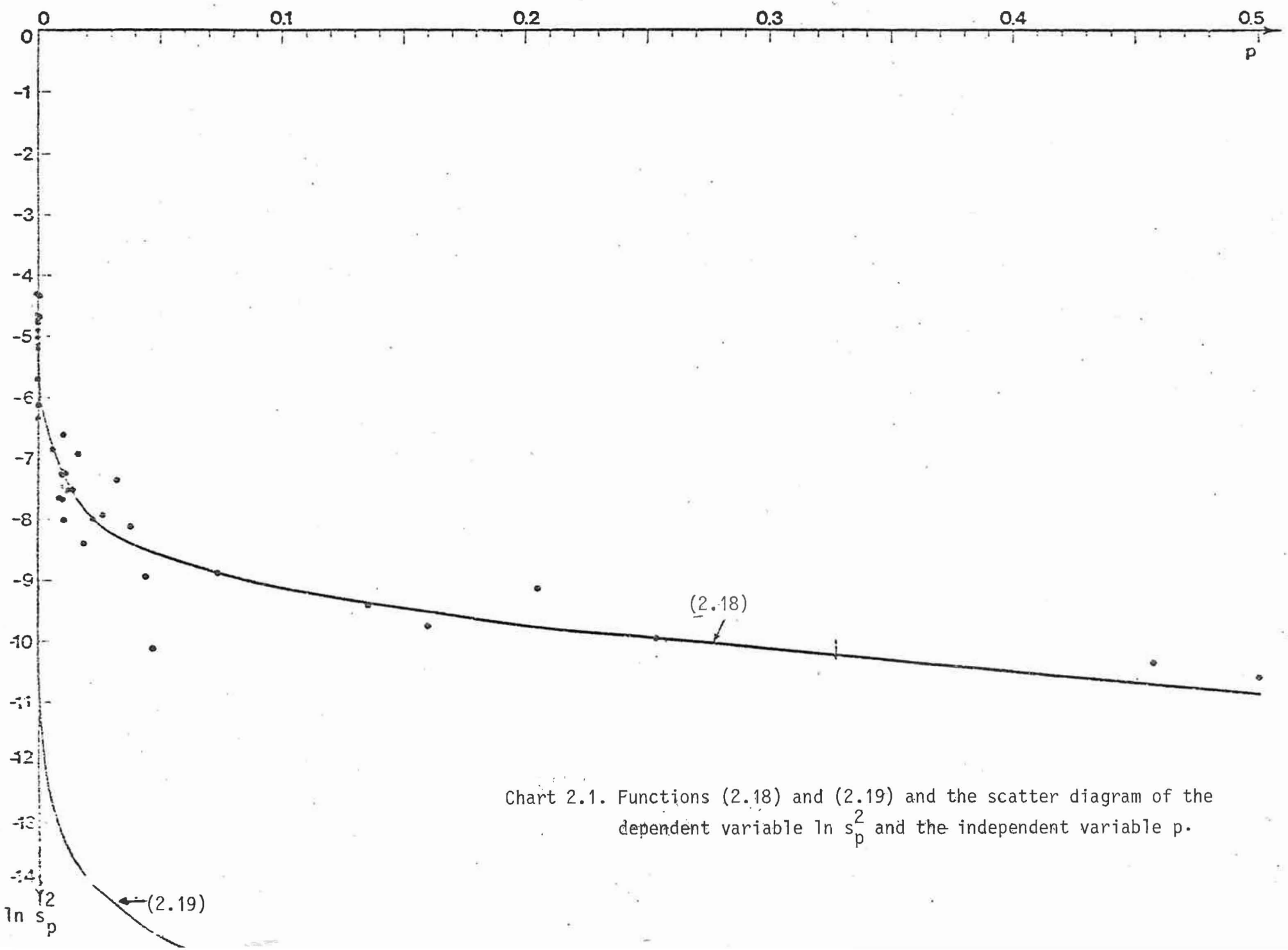


Chart 2.1. Functions (2.18) and (2.19) and the scatter diagram of the dependent variable ln s_p^2 and the independent variable p.

2.5. Remarks on the model of the residual variance and its estimation compared to earlier studies

The suggested method concerning the construction of the model of the residual variance and its estimation seems to be very attractive both empirically and theoretically. To find out sensible weights for estimating a model of type (2.3) from cross-sectional data only quite short time series data from the dependent variable $\ln \bar{w}_{i.}$ and a very simple autoregressive model of type (2.16) are needed. The error variance of this model can be estimated from formula (2.17) by weighting observations with unity weights because the size of the regions is approximately constant over relatively short time periods.

In contrast to many earlier studies, the only explanatory variable of the variance model (2.12) is the size (p) of the region. This kind of model turned out to work quite satisfactory.

Methods proposed earlier for estimating a model for the error variance are based on the residuals $\bar{e}'_{i.}$ of the original explanatory model of type (2.3). These residuals are estimated by OLS with unity weights and an explanatory model is estimated for the squares or absolute values of these residuals (Park (1966), Glejser (1969), Goldfeld-Quandt (1972), Hildreth-Houch (1968), Anemiya (1977)). One of the shortcomings of these methods is that, in estimating the original explanatory model of type (2.3) (in order to determine the residuals $\bar{e}'_{i.}$), the observations are from the outset weighted with wrong weights in cases where the homoscedasticity assumption concerning the theoretical residuals of the explanatory model of type (2.3) 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 (2.3)) and the unknown parameter involved in the multiplicative formula of the residual variance. However, when this method is used, it is necessary to assume that the theoretical residual terms of the original model of type (2.3) are normally distributed and stochastically independent of one another.

These assumption are very strong in many studies. In this study, where the population of industrial workers is finite and the observations have been formed by dividing whole Finland into mutually exclusive areas, the assumption of independent observations is obviously not correct.

2.6. Some empirical results

If for a model of type (2.3) it is true that

$$(2.21) \quad \sigma^2(\varepsilon'_{i\cdot}) = \frac{\sigma_0^2}{\omega_i^2},$$

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

$$(2.22) \quad \omega_i \ln \bar{w}_{i\cdot} = \omega_i \alpha' + \sum_k \beta_k \, (\omega_i \, \overline{\ln x}_{i\cdot k}) + \omega_i \varepsilon'_{i\cdot}$$

it is true that

$$(2.23) \quad \sigma^2(\omega_i \varepsilon'_{i\cdot}) = \omega_i^2 \, \sigma^2(\varepsilon'_{i\cdot}) = \sigma_0^2 \, .$$

Some estimation results concerning models of type (2.3) are given in Table 2.2.[1] The models in this table were estimated by OLS after multiplying the original observations with weights

$$(2.24) \qquad \omega_i = \sqrt{[1+3\cdot10^6(p_iq_i)^2]^{0.408}}/q_i(1+p_iq_i)\,,$$

where $p_i = \frac{N^{wr}_{i\cdot70}}{403393}$ and $q_i = 1-p_i$ ($N^{wr}_{i\cdot70}$ standing for the number of industrial workers in the commuting region i in 1970). Weights (2.24) are equal to $\sqrt{e^{-4.963}} = \hat{\sigma}_0$ divided by the standard deviation $\hat{\sigma}_p$ based on model (9) in Table 2.1. The transformed residual of models of type (2.3) ought to be now homoscedastic.

The homoscedasticity assumption relating to models of type (2.3) was tested, with respect to the size of regions, by different methods which utilized the empirical residuals of models in Table 2.2. The results suggested that the homoscedasticity assumption of the residuals of model (2.3) can be accepted.

One method, which I applied, bases on the idea of calculating estimates of the residual variance of model (2.3) both from small regions and from large regions. If these estimates are approximately the same, the homoscedasticity assumption of the original model of type (2.3) can be accepted.

---

1) The dependent variable of the models in Table 2.2 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.

Table 2.2. Models for industrial workers' Average Hourly Wage Rate (%). Data composed of 170 commuting regions in 1970. Models have been estimated by using weights (2.24)[1)]

Explanatory variable and its t-value	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 multiple correlation		0	0.790	0.881	0.901	0.906	0.912	0.914	0.918	0.920	0.921	

1) 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 the formula (2.24).

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 and $\bar{y} = \frac{\sum_i \omega_i^2 y_i}{\sum_i \omega_i^2}$.

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 ω_i 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 was 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.[1] The variances were computed with respect to both "0" (the theoretical mean of the residual $\varepsilon'_{i.}$) and the mean of the residuals themselves. The results are set out in Table 2.3.

The variance ratios for small and large regions of models 1-11 of Table 2.2 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.

If the variance ratios given in Table 2.3 were F-distributed, none of the values of the ratios would be statistically significant at the 1 % level and only one ($s_{2s}^2/s_{21}^2 = 3.23$ for model 11) would be significant at the 5 % level. Despite the fact that the variance ratios cannot be considered F distributed the residuals of most theoretical models corresponding to the models in Table 2.2 can be regarded as homoscedastic. This applies at least to the models in which the variance ratios s_{1s}^2/s_{11}^2 and s_{2s}^2/s_{21}^2 are close to unity.

---

1. 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 (1967) pp. 93-94). The theoretical variances of empirical residuals depend in general on the theoretical variance of theoretical residuals, 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 the forthcoming doctoral thesis.

Table 2.3: 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 2.2, the number of observations and the ratio between variances computed from small and large regions[1].

The number of model in Table 2.2	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

1) The variances s^2_1 and s^2_2 have been computed from the formulas $s^2_1 = \frac{1}{n}\sum_i e_i^2$

$$s^2_2 = \frac{1}{n-1}\sum_i (e_i - \bar{e})^2, \quad \text{where } \bar{e} = \frac{1}{n}\sum_i e_i \quad (i = 1,\ldots,n)$$

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

3. Models of the residual variance for all branches of Finnish industry

A model of the residual variance for Finnish industrial branches 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. The industry branch division represents another type of division of Finnish industry workers' population into mutually exclusive sets than does 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: which 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 a model of type (2.15) was estimated for all branches of Finnish industry. The dependent variable of this model was constructed utilizing the model

$$\ln \bar{w}_{.jt} = \ln \bar{w}_{.j(t-1)} + (\ln \bar{w}_{..t} - \ln \bar{w}_{..(t-1)}) + \bar{e}_{.jt}. \tag{3.1}$$

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

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

and its error terms variance

$$s_t^2(\bar{e}_{.jt}) = \frac{1}{T-1} \sum_t (\bar{e}_{.jt} - \bar{e}_{.j.})^2 \tag{3.2}$$

In formulas (3.1) - (3.2) the index j indicates the 22 industry branches and the index t indicates the years 1960-1971.

The model for the residual variance (2.15) was estimated by OLS, weighting the observations by unity, in such a way that the parameter λ was given different values. Model 7 in Table 3.1, which is

$$(3.3) \qquad \ln s_p^2 \approx 2 \ln q + 2 \ln(1+pq) - 7.93 - 0.917 \ln(1+400\, p^2q^2)$$

was finally chosen because for this model the coefficient of multiple correlation was highest (Chart 3.1). On the basis of this model the weights

$$(3.4) \qquad \omega_j = \sqrt{[1+400(p_jq_j)^2]^{0.917}}/q_j(1+p_jq_j)$$

were constructed[1]. These weights were used in the same way as in the case of geographical regions to estimate models for the average hourly wage rate by industrial workers from data concerning all branches (22) of Finnish industry in 1970.

4. Some further applications of the residual variance models for geographical regions and branches of industry

With the residual variance models it is possible to calculate the weights of geographical regions and industrial branches in estimating models

---

1)
$$(3.5) \qquad p_j = \frac{N^{wr}_{.j70}}{403393}$$

and $q_j = 1 - p_j$, $N^{wr}_{.j70}$ being the number of industrial workers in branch j in 1970.

Table 3.1. The estimation results for model (2.15). Data composed of 22 branches of industry (two-digit ISIC subdivision)[1)]

Number of model	Regression coefficients and t-values					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

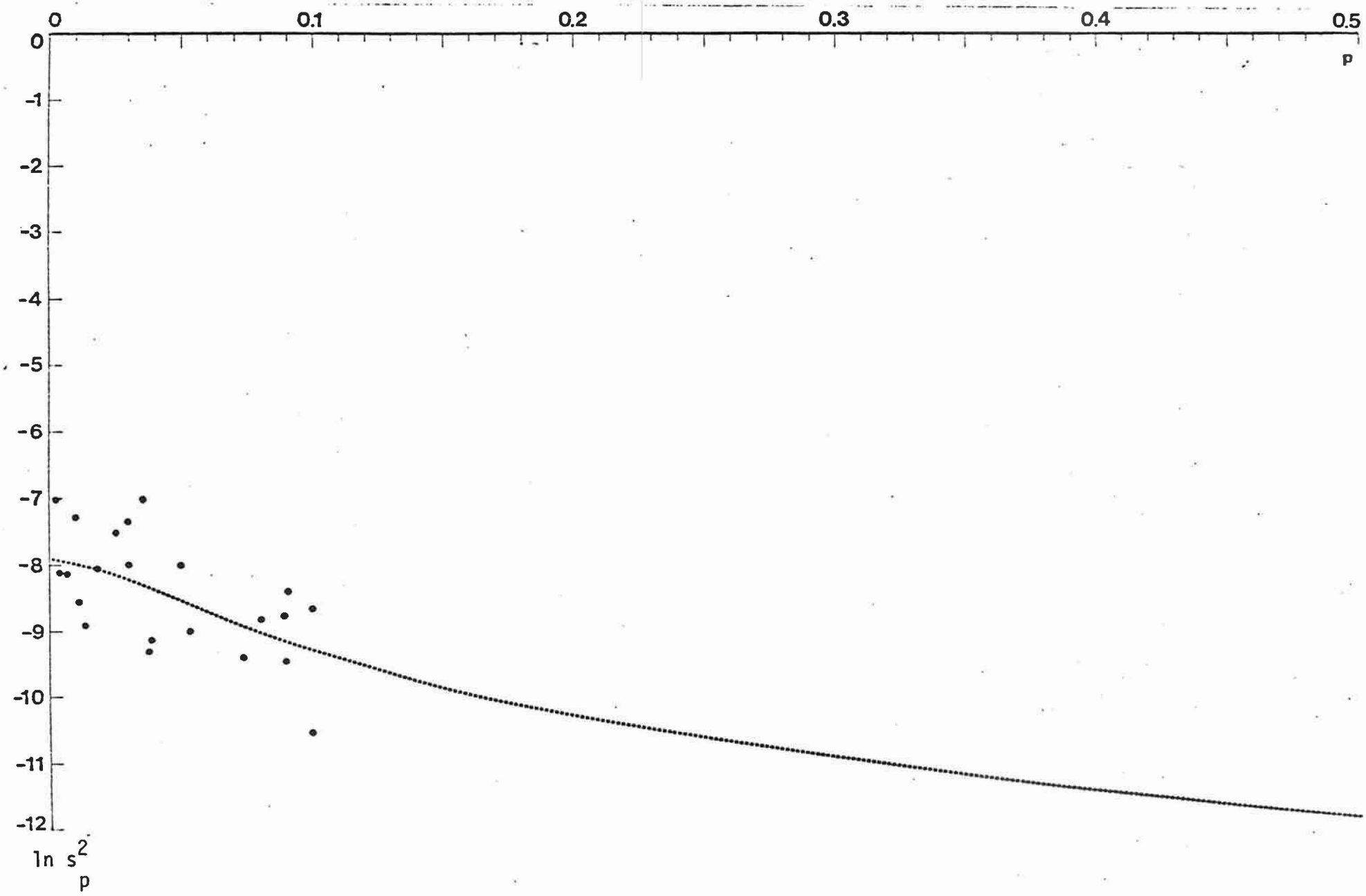
1) The dependent variable of model (2.15) was constructed utilizing formula (3.2) and the independent variable p_j was computed from formula (3.5) ($q_j=1-p_j$).
The error terms standard deviation was computed from the formula

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

where u_{p_j} is the empirical residual of model (2.15). The coefficient of multiple correlation was computed from the formula

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

Chart 3.1. Function (3.3) and the scatter diagram of the dependent variable $\ln s_p^2$ and the independent variable p.



for wages from this cross-sectional data. Making use of the residual 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 industrial branches of different size. This can be done utilizing the generalized intra-class correlation coefficient and the residual variance models.

4.1 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 industrial branches in form (2.22). That is, the original observations were first multiplied with weights (2.24) in the case of geographical regions and with weights (3.4) in the case of industrial branches and models of type (2.22) were then estimated by OLS with unity weights for the transformed observations. The residual variance of those wage models of type (2.22) where the only explanatory variable was the weight ω_i can be interpreted as the estimate of the variance of the logarithm of individual workers' hourly wage rate according to (2.23). The estimates from different sets of observations are given in Table 4.1.

Table 4.1: Estimates of the standard deviation $\hat{\sigma}_0$ of the logarithm of individual workers' hourly wage rate calculated from different sets of observations (%)[1)]

Set of observations	$\hat{\sigma}_0$
Confederation of Finnish Employers (218 000 observations) [2)]	23.12
Commuting regions (170 regions)	32.59
Industrial branches (22 industrial branches)	23.02

As can be seen from Table 4.1 the estimate $\hat{\sigma}_0$ calculated from data for industrial branches is very near to the estimate obtained from data of the Confederation of Finnish Employers. This result suggests that by suitable methods very accurate information can be obtained also 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.

---

1) If the standard deviation of the variable $\ln w_{iv}$ is $\hat{\sigma}_0$, then the standard deviation of variable $\ln w_{iv}$ in log-percentages is $100 \cdot \hat{\sigma}_0$.

2. From the statistics collected by the Confederation of Finnish Employers the estimate $\hat{\sigma}_0$ was calculated from the observations of the 3rd quarter of 1976. The material included totally 218 000 workers in all, which is about a half of the total industrial workers in Finland. The estimates $\hat{\sigma}_0$ calculated from the data collected by the Confederation of Finnish Employers have been 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 (%).

4.2 The estimation of interdependence between wages in different types of areas of different size using the generalized intra-class correlation coefficient

Let us consider the following decomposition

$$(4.1) \qquad \left(\frac{1}{n}\sum_i \varepsilon_i\right)^2 = \bar{\varepsilon}^2 = \frac{1}{n^2}\left(\sum \varepsilon_i^2 + \sum_{i\neq j}\sum \varepsilon_i \varepsilon_j\right)$$

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

$$(4.2\ a\text{-}d) \quad E(\bar{\varepsilon}^2) = \sigma^2_{\bar{\varepsilon}}$$

$$E(\varepsilon_i^2) = \sigma^2_{\varepsilon_i}$$

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

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

Taking expectations on both sides of (4.1), utilizing notations (4.2 a-d) we get

$$(4.3) \qquad \bar{\rho} = \frac{\dfrac{\sigma^2_{\bar{\varepsilon}}}{\sigma^2_{\varepsilon_i}} - \dfrac{1}{n}}{1 - \dfrac{1}{n}}$$

Formula (4.3) defines the average intra-class correlation coefficient $\bar{\rho}$. Prof. Leo Törnqvist has suggested the generalization of (4.3) to cases where the sizes of research units are continuous variables. Instead of (4.3) he suggests the formula

$$(4.4) \qquad \rho(p_1,p_2) = \left(\frac{\sigma^2_{p_2}}{\sigma^2_{p_1}} - \frac{p_1}{p_2}\right)/\left(1 - \frac{p_1}{p_2}\right) \qquad (p_1 < p_2).$$

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 (4.3).[1)]

From empirical data it is possible to compute an estimate of the intra-class correlation coefficient defined by (4.4). The parameters $\sigma^2_{p_2}$ and $\sigma^2_{p_1}$ can then be replaced by estimates which can be computed from the formulas (2.18) and (3.3) constructed in the preceding sections.

Formulas (2.18) and (3.3) are based on time series data. If the residual variance of the explanatory models of the average hourly wage rate associated with cross-sectional data diminishes as fast as the residual variance of the time-series models of the same variable when the size of the region increases - which assumption was made in constructing, by means of explanatory models of the residual variance based on time series data, weights of observations to be used in estimating the explanatory model of the average hourly wage rate from cross-sectional data - then the model of the intra-class correlation coefficient constructed in the above-described manner can also be considered to describe the intra-class correlation coefficient of the residual terms of the explanatory models of the average hourly wage rate related to cross-sectional data.

---

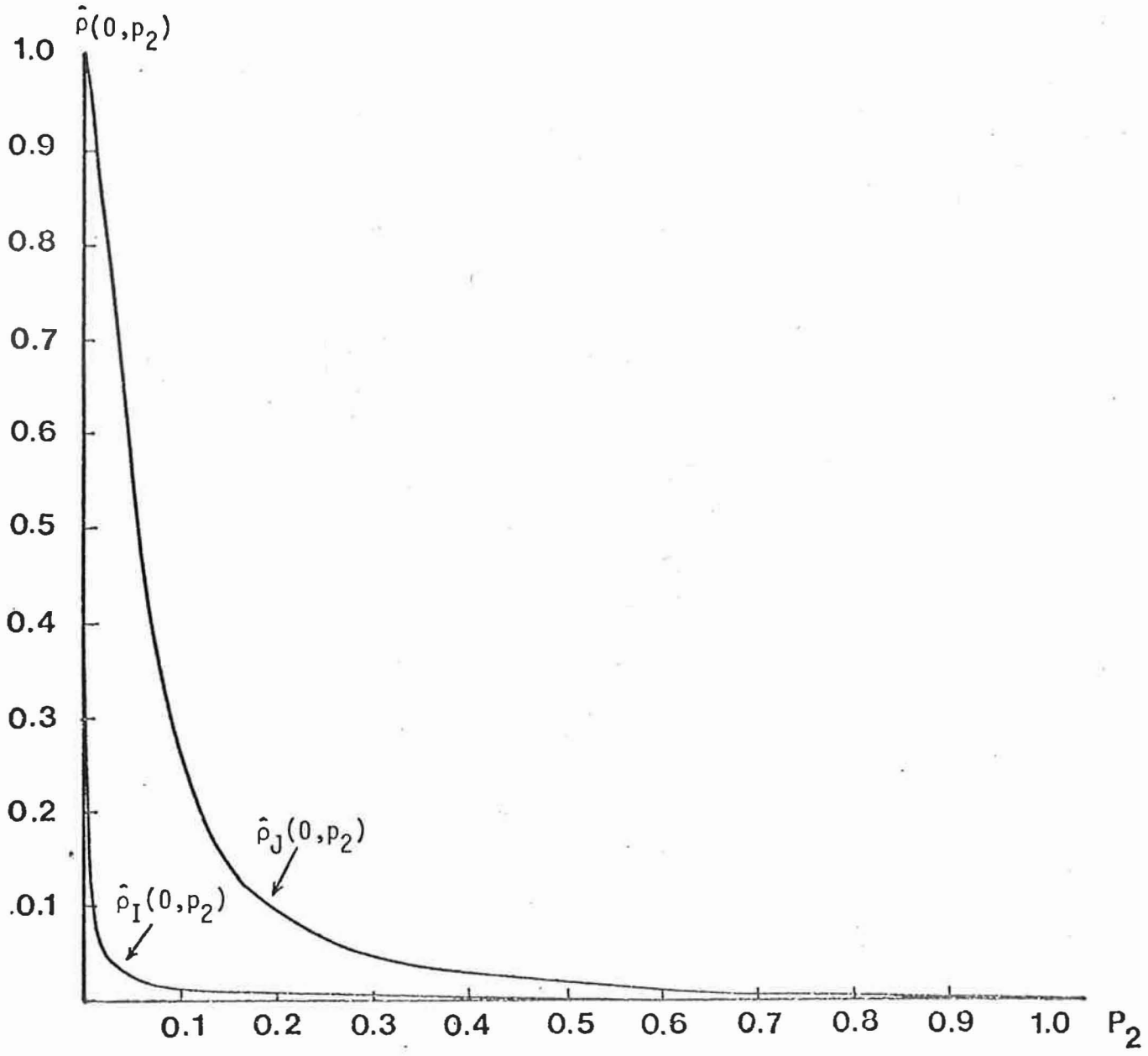
1) $\sigma^2_{p_2}$ may be considered the expected variance of the residuals $\bar{\varepsilon}$ over all geographical regions of size p_2 that form a connected geographical whole and are mutually excluded. In the case of industry branches $\sigma^2_{p_2}$ may be considered the expected variance of the residuals ε over all "industry branches" of size p_2 which can be formed from the population. These branches can be real or hypotethical. Similar interpretations can be given to $\sigma^2_{p_1}$.

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 approximately describe the intra-class correlation coefficient of the residual terms related to individual workers in regions of size p_2. In Chart 4.1, the intra-class correlation coefficient $\hat{\rho}_I(0,p_2)$ and $\hat{\rho}_J(0,p_2)$ related to connected geographical regions and branches of industry respectively are represented as functions of p_2.

From Chart 4.1 it appears 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 connected geographical areas. This result suggests that, in any two areas of the same size which may be either connected geographical areas or "industrial branch" areas, which also form geographical wholes - the residual terms of explanatory models or 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 geographically connected areas.

If the residuals in models (2.16) and (3.1) are interpreted to represent approximately wage drift, the results in Chart 4.1 suggest that, in the case of Finnish industrial workers, the interdependence of wage drift is stronger within branches of industry than within a connected geographical area of the same size. The interpretation of the residuals of models (2.16) and (3.1) as approximately describing wage drift is quite realistic because the percentual wage increases in 1960-1971 were in fact very similar in Finnish industry. If this were the case the residuals in models (2.16) and (3.1) would measure the difference in wage drift between region i and the whole country (model (2.16)) and the difference in wage drift between branch of industry j and the whole country (model (3.1)).

Chart 4.1. The intra-class correlation coefficients related to connected geographical regions ($\hat{\rho}_I(0,p_2)$) and branches of industry ($\hat{\rho}_J(0,p_2)$).[1]



[1] The intra-class correlation coefficients have been computed from the formula (4.4) utilizing thereby formulas (2.18) and (3.3) in the case of geographical regions and branches of industry respectively.

With the method outlined in this section it is possible to estimate quantitatively the dependence of wages in different type and of different size of sets, which can be geographical areas or industrial branches or some other sets utilizing generalized intra-class correlation coefficient and the models for the residual variance.

REFERENCES

ANEMIYA, T., A Note on a Heteroscedastic Model, Journal of Econometrics, November 1977, 365-370.

DRAPER, N. and SMITH, H., Applied Regression Analysis, John Wiley & Sons, Inc., New York 1967.

GLEJSER H., A New Test for Heteroscedasticity, Journal of the American Statistical Association, Vol. 64, 1969, 316-323.

GOLDFELD, S.M, and QUANDT, R.E., Some Tests for Homoscedasticity. Journal of the American Statistical Association, Vol. 60, 1965, 539-547.

GOLDFELD, S.M. and QUANDT, R.E., Nonlinear Methods in Econometrics, Ch. 3: Analyses of heteroscedasticity, North-Holland, Amsterdam 1972, 78-123.

HARVEY, A.C., Estimation of Parameters in a Heteroscedastic Regression Model. Paper presented at the European Meeting of the Econometric Society, Grenoble, September 1974.

HARVEY, A.C., Estimating Regression Models with Multiplicative Heteroscedasticity, Econometrica, May 1976, 461-466.

LASTIKKA, PEKKA, Relative Differences in Industrial Workers' Average Hourly Wages between Various Geographical Regions and Branches of Industry in Finland, 1960-1971. A forthcoming doctoral thesis, The Research Institute of the Finnish Economy, Series A 8.

PARK, R.E., Estimation with Heteroscedastic Error Terms, Econometrica, October 1966, 888.

TÖRNQVIST, Leo, Levnadskostnadsindexerna i Finland och Sverige, deras tillförlitlighet och jämförbarhet, Ekonomiska Samfundets Tidskrift, Vol. 37, Helsingfors 1936, 59-93.

VARTIA, YRJÖ O., Fisher's Five-tined Fork and Other Quantum Theories of Index Number, The Research Institute of the Finnish Economy, Discussion papers No 4, 10.8.1976.

Appendix 1

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

$$(2.1) \qquad \ln w_{i\nu} = \alpha + \sum_k \beta_k \ln x_{i\nu k} + \varepsilon_{i\nu} , \qquad \begin{matrix} k \in K \\ i \in F \\ \nu \in N_i \end{matrix}$$

and the logarithm of the average hourly wage in the region i

$$(2.2) \qquad \ln \bar{w}_{i.} = \ln \sum_\nu p_{i\nu} w_{i\nu} .$$

In (2.2) $\sum_\nu p_{i\nu} = p_{i.} = 1$. Applying Törnqvist's (1936) formulas we get[1)]

---

1) 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

$$(1) \qquad ({}_\alpha W_0^1)^\alpha = \sum p_\nu (w_\nu^1/w_\nu^0)^\alpha = \sum p_\nu e^{\alpha \ln (w_\nu^1/w_\nu^0)}$$

$$(2) \qquad \ln({}_0W_0^1) = \sum p_\nu \ln (w_\nu^1/w_\nu^0) ,$$

where $p_\nu \geq 0$ and $\sum p_\nu = 1$.

Dividing every term of (1) by $({}_0W_0^1)^\alpha$ we get

$$(3) \qquad ({}_\alpha W_0^1/{}_0W_0^1)^\alpha = \sum p_\nu e^{\alpha \ln (w_\nu^1/w_\nu^0({}_0W_0^1))}$$

$$= \sum p_\nu e^{\alpha \dot{w}_\nu}$$

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

$$(4) \qquad ({}_\alpha W_0^1/{}_0W_0^1)^\alpha = 1 + \frac{\alpha^2}{2!} \sum p_\nu \dot{w}_\nu^2 + \frac{\alpha^3}{3!} \sum p_\nu \dot{w}_\nu^3 + \ldots$$

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

$$(5) \qquad \ln ({}_\alpha W_0^1) = \ln ({}_0W_0^1) + \frac{\alpha}{2!} \sum p_\nu \dot{w}_\nu^2 + \frac{\alpha^2}{3!} \sum p_\nu \dot{w}_\nu^3$$

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

$$(6) \qquad \ln \bar{w}_{i.} = \sum_{\nu} p_{i\nu} \ln w_{i\nu} + \tfrac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^2_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^3_{i\nu} + \ldots$$

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

Substituting $\ln w_{i\nu}$ from (2.1) to (6) we get

$$(7) \qquad \ln \bar{w}_{i.} = \alpha + \sum_{k} \beta_k \overline{\ln x}_{i.k} + \sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \tfrac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^2_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^3_{i\nu} + \ldots,$$

where

$$(2.5) \qquad \overline{\ln x}_{i.k} = \sum_{\nu} p_{i\nu} \ln x_{i\nu k} .$$

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

$$(2.4) \qquad \alpha' = \alpha + \left(\sum_{i} \sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \frac{1}{2} \sum_{i} \sum_{\nu} p_{i\nu} \dot{w}^2_{i\nu} + \frac{1}{6} \sum_{i} \sum_{\nu} p_{i\nu} \dot{w}^3_{i\nu} \right)$$

$$= \alpha + \left[\sum_{i} \sum_{\nu} p_{i\nu} \left(\frac{1}{2} \dot{w}^2_{i\nu} + \frac{1}{6} \dot{w}^3_{i\nu} + \ldots \right) \right]$$

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

$$(2.6) \qquad \varepsilon'_{i.} = \left(\sum_{\nu} p_{i\nu} \varepsilon_{i\nu} + \frac{1}{2} \sum_{\nu} p_{i\nu} \dot{w}^2_{i\nu} + \frac{1}{6} \sum_{\nu} p_{i\nu} \dot{w}^3_{i\nu} + \ldots \right)$$

$$- \left[\sum_{i} \sum_{\nu} p_{i\nu} \left(\frac{1}{2} \dot{w}^2_{i\nu} + \frac{1}{6} \dot{w}^3_{i\nu} + \ldots \right) \right]$$

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

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

$$(2.3) \qquad \ln \bar{w}_{i.} = \alpha' + \sum_{k} \beta_k \overline{\ln x}_{i.k} + \varepsilon'_{i.} ,$$

where $E\varepsilon'_{i.} = \sum_{i} p_{i.} \varepsilon'_{i.} = \sum_{i} \varepsilon'_{i.} = 0$ because of the α':s and $\varepsilon'_{i.}$:s manner of construction.

Appendix 2 The Proof of monotonicity of the function (2.11)

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

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

The term

$$\frac{-2p(1+3q)}{q(1+pq)} \leq 0, \text{ when } p \in [0,1] \text{ and } \frac{2\delta\lambda pq(2p-1)}{1+\lambda p^2 q^2} \leq 0$$

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

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

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

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

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

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

The smallest value of the right-hand side of (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

(2.13) $0 \leq \delta \leq 1$ and

(2.14) $0 < \lambda < \infty$.

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