KM 21(1986)
pag 133-142

REGRESSION EFFECTS IN TABULATING FROM PANEL DATA

Huib van de Stadt*
Tom Wansbeek*

Abstract

Tabulation of income changes between two years by income classes based on panel data may suffer from regression to the mean. For a correct tabulation, knowledge of the data generation process is required. In general a good procedure is tabulation by the average income over the two years.

1. Introduction

Over the last decade, the increasing availability of longitudinal data from socio-economic panels has given an enormous impetus to the social sciences. In a number of cases inspired by the well-known 'Panel Study of Income Dynamics' of the University of Michigan, started in 1968 but still ongoing, panels are conducted in various countries in order to collect and analyze information on socio-economic variables that influence welfare. Among these, income plays of course a prominent role.

Although an impressive array of statistical and econometric techniques has been developed to handle panel data, a small but important problem has been left unexplored to the best of our knowledge, viz. how to tabulate income change by income class. There, one runs the risk of

* Huib van de Stadt: Netherlands Central Bureau of Statistics, P.O. Box 959, 2270 AZ Voorburg, tel. 070-694341, Tom Wansbeek: Econometrics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, tel. 050-633810. The views expressed in this paper are those of the authors and do not necessarily reflect to policies of the Netherlands Central Bureau of Statistics. We thank Martin Fase, Peter Kooiman, Frank van der Pol, Jules Theeuwes and a referee for their comments on this paper, and Paul van Batenburg, Jan de Leeuw and Ivo Molenaar for guiding us to some of the references.

getting results that may be misleading due to 'regression to the mean'. In this note we look at this problem and suggest a solution. It will appear that a simple provision can be made to reduce the misrepresentation problem considerably, but that a really satisfactory solution requires profound knowledge of the data generation process.

In section 2 we illustrate the problem by a simple example and propose a criterium for unbiased tabulation. In section 3 the analysis is repeated for a more general model, and section 4 concludes with some comments on the data generation process.

2. Regression to the mean

Regression to the mean may occur when we want to tabulate the change in some variable, measured by means of a panel, by that variable itself. An example is tabulating income changes by income. Table 1, based on the data used by Keller et al. (1985) in their study of real income changes of households in the Netherlands, provides an illustration.

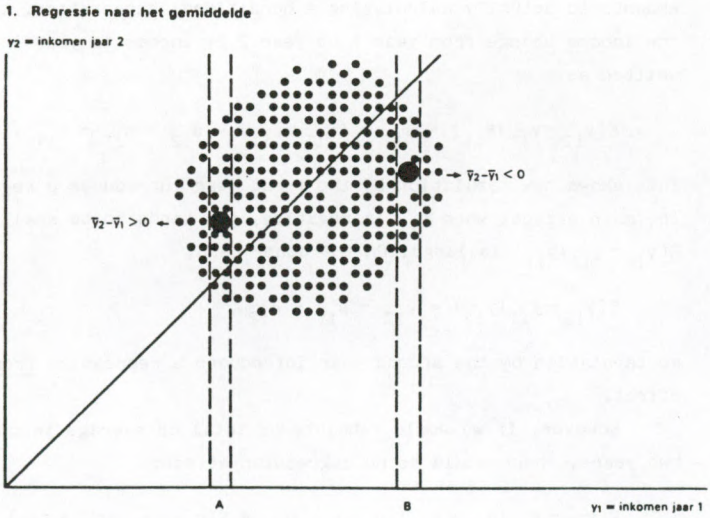
The first column of this table, based on tabulation by 1979 income, suggests a sizeable decrease in income inequality from 1979 to 1981. Individuals with the lowest incomes in 1979 did nearly 9% per year better than those with the highest incomes (viz. +3.6% versus -5.3%). The second column, however, where the same observations are tabulated by 1981 income, suggests a sizeable increase in income inequality. So the conclusions drawn from the table are strongly influenced by the way the data are tabulated.

The situation is also illustrated by figure 1, which gives a hypothetical income distribution in two years for the members of a panel. The distribution is symmetric around the 45 degree line, so income inequality has not changed between both years. It is easily seen that for this distribution the mean income change for low incomes in year 1 (group A) is positive, whereas it is negative for high incomes (group B). If, however, we would have tabulated according to income in the second year, we would get the converse result: an income increase for high income groups and an income decrease for low income groups.

Table 1. Median change in real income per year, 1979-1981, the Netherlands.

Income class		Classification according to		
		1979 income	1981 income	average income 1979 and 1981
first	10%-group	3.6	-11.3	-1.9
first	25%-group	-0.9	-5.3	-2.4
second	25%-group	-2.3	-3.2	-2.9
third	25%-group	-2.8	-1.4	-2.1
fourth	25%-group	-4.8	-0.1	-2.2
tenth	10%-group	-5.3	-0.1	-2.7

Source: Keller et al. (1985) and supplementary computations.



Instances of tables based on longitudinal data where the authors note that they suffer from regression to the mean can be found in the literature (e.g., Schiller (1977) and Park et al. (1983)). However, in these papers no further analysis is added. There also are some more theoretical writings on the subject (e.g. a few introductory pages in Goldstein (1979), pp. 119 ff.; the comprehensive paper by Nesselroade et al. (1980); and a follow-up on the latter by Labouvie (1982)), but these do not explicitly deal with tabulation. The Nesselroade et al. paper correctly stresses knowledge of the data generation process as a prerequisite for avoiding regression to the mean, and it is this aspect that we will elaborate upon below.

Suppose that the income generation process is

$$y_{it} = \mu_{it} + \varepsilon_{it} \tag{2.1}$$

with y_{it} log-income in year t, ε_{it} an error term with zero expectation, constant variance and uncorrelated over time and between invididuals and μ_{it} a non-stochastic function of exogenous variables (education, experience, "permanent income"). Since tabulation by some variable amounts to actually calculating a conditional expectation, tabulating the income change from year 1 to year 2 by income in year 1 can be written as

$$E(y_{i2}-y_{i1}|y_{i1}) = \mu_{i2} - (\mu_{i1}+\varepsilon_{i1}) = \mu_{i2} - \mu_{i1} - \varepsilon_{i1} . \tag{2.2}$$

This shows how tabulation by the first year introduces a regression to the mean effect: when ε_{i1} is negative, y_{i1} tends to be small whereas $E(y_{i2}-y_{i1}|y_{i1})$ is large. On the other hand,

$$E(y_{i2}-y_{i1}|y_{i2}) = \mu_{i2} - \mu_{i1} + \varepsilon_{i2} , \tag{2.3}$$

so tabulation by the second year introduces a regression from the mean effect.

However, if we would tabulate by total or average income over the two years, there would be no regression effect:

$$\begin{aligned} E(y_{i2}-y_{i1}|y_{i2}+y_{i1}) &= \mu_{i2} - \mu_{i1} + E(\varepsilon_{i2}-\varepsilon_{i1}|\varepsilon_{i1}+\varepsilon_{i2}) \\ &= \mu_{i2} - \mu_{i1} , \end{aligned} \tag{2.4}$$

for reasons of symmetry. So when the data are generated according to (2.1), tabulation by total or average income in both years is a solution for the problem. This is illustrated in the third column of table 1.

An important aspect of this analysis is of course how a 'solution' should be defined. It is defined here as a tabulation which does not depend on the stochastic part of the data generation process, which in this case implies that the conditional expectation should not depend on ε_{it} for t = 1,2. This is true for (2.4), but not for (2.2) or (2.3). Because the conditioning only influences the stochastic part of the process, the requirement that the tabulation does not depend on the stochastic part of the process is equivalent to the requirement that the conditional expectation is equal to the unconditional expectation, which $\mu_{i2} - \mu_{i1}$ for the process described by (2.1).

Tabulation by average income is not a solution for some other processes. An example is the autoregressive process of the form

$$y_{it} = y_{it-1} + \varepsilon_{it}, \qquad t = 1,2,\ldots, \tag{2.5}$$

with ε_{it} independent of y_{it-1} and y_{i0} fixed. Now $E(y_{i1} - y_{i0}|y_{i0}) = E(y_{i2} - y_{i1}|y_{i1}) = 0$, so tabulation by the first year is correct, but for example $E(y_{i1} - y_{i0}|y_{i0} + y_{i1}) = \varepsilon_{i1}$, so tabulation by average income introduces regression from the mean.

Both examples in this section illustrate the link between tabulation and data generation. We will formalize this below.

3. A general formulation

Let the income generation process be described by

$$y_{it} = \mu_{it} + u_{it}, \qquad t = 1,2,\ldots, \tag{3.1}$$

$$u_{it} = \beta u_{it-1} + \varepsilon_{it}, \qquad t = 1,2,\ldots, \tag{3.2}$$

where μ_{it} is again some non-stochastic function of exogenous variables, u_{it} an autoregressive error term with autoregression parameter β ($0 \leq \beta \leq 1$) and ε_{it} a normally distributed error term with zero expectation, constant variance and uncorrelated over time and between individuals. The process starts with a fixed u_{i0}. We want to tabulate

the change in income by some linear combination of y_{it-1} and y_{it}, so the relevant expression is

$$E(y_{it} - y_{it-1} \mid ay_{it-1} + (1-a)y_{it}) \ , \tag{3.3}$$

where a is some constant to be determined. Substituting (3.1) into (3.3) gives

$$\mu_{it} - \mu_{it-1} + E(u_{it} - u_{it-1} \mid au_{it-1} + (1-a)u_{it}) \ , \tag{3.4}$$

and next substituting

$$u_{it} = \beta^t u_{i0} + \sum_{s=1}^{t} \beta^{t-s} \varepsilon_{is} \tag{3.5}$$

into (3.4) yields

$$\mu_{it} - \mu_{it-1} + (\beta^t - \beta^{t-1})u_{i0} + E(\sum_{s=1}^{t} \beta^{t-s} \varepsilon_{is} - \sum_{s=1}^{t-1} \beta^{t-1-s} \varepsilon_{is} \mid a \sum_{s=1}^{t-1} \beta^{t-1-s} \varepsilon_{is} + (1-a) \sum_{s=1}^{t} \beta^{t-s} \varepsilon_{is}) \tag{3.6}$$

We are interested in the value of a for which (3.6), the conditional expectation, is equal to the unconditional expectation. Therefore, the expectation term in (3.6) should be set equal to zero. This amounts to

$$E(\varepsilon_{it} + \sum_{s=1}^{t-1} (\beta^{t-s} - \beta^{t-1-s})\varepsilon_{is} \mid (1-a)\varepsilon_{it} + \sum_{s=1}^{t-1} (a\beta^{t-1-s} + (1-a)\beta^{t-s})\varepsilon_{is}) = 0 \ . \tag{3.7}$$

From standard multivariate normal theory we know that

$$E(\delta_1 \mid \sum_j \gamma_j \delta_j) = E(\delta_1) + \frac{\operatorname{cov}(\delta_1, \sum_j \gamma_j \delta_j)}{\operatorname{var}(\sum_j \gamma_j \delta_j)} (\sum_j \gamma_j \delta_j - E(\sum_j \gamma_j \delta_j)) = \frac{\gamma_1}{\sum_j \gamma_j^2} \sum_j \gamma_j \delta_j \tag{3.8}$$

for arbitrary constants γ_j and normally, independently and identically distributed δ_j with zero expectation. Using this result and taking into account that the expectation of a weighted sum is equal to the weighted sum of the expectations, (3.7) holds if

$$(1-a) + \sum_{s=1}^{t-1}(\beta^{t-s} - \beta^{t-1-s})(a\beta^{t-1-s} + (1-a)\beta^{t-s}) = 0$$

$$\Leftrightarrow \quad (1-a) + (\beta-1)(a+(1-a)\beta)\sum_{s=1}^{t-1}\beta^{2t-2-2s} = 0$$

$$\Leftrightarrow \quad (1-a) + (\beta-1)(a+(1-a)\beta)\frac{(1-\beta^{2t-2})}{(1-\beta^{2})} = 0$$

$$\Leftrightarrow \quad (1-a)(1+\beta) - (a+\beta-a\beta)(1-\beta^{2t-2}) = 0$$

$$\Leftrightarrow \quad a = \frac{1+\beta^{2t-1}}{2+\beta^{2t-1}-\beta^{2t-2}} \; . \qquad (3.9)$$

In this derivation we have assumed $\beta \neq 1$; if $\beta = 1$ the final result in (3.9) is obviously $a = 1$.

Several interesting observations can be made on the basis of this equation. First, note that for the process given by (2.1), the autoregression parameter β is equal to zero, which leads to $a = \frac{1}{2}$. So in the absence of autoregression one should classify according to average income in both years.

The process given by (2.5) is also a special case of (3.1) and (3.2). Now $\mu_{it} = 0$ and $\beta = 1$, resulting in $a = 1$. So in this situation one should classify according to first year income.

If $0 \leqq \beta < 1$, then $\frac{1}{2} \leqq a < 1$. However, if t goes to infinity, a approaches $\frac{1}{2}$ very fast. For example, if $\beta = \frac{1}{2}$ then for $t = 2,3,4,5$, a equals 0.60, 0.52, 0.51 and 0.50, respectively. For other values of β the pattern is similar. So in most situations, especially if the process has started a number of periods ago, one should take $a = \frac{1}{2}$, which comes down to classification according to average income. In practice we tabulate data pertaining to many different individuals of different ages and hence different t's. As long as $\beta \neq 1$ the conclusion remains of course the same.

An important but hitherto neglected role in this analysis is played by the error term at the start of the process, u_{i0}. It is assumed fixed for convenience, but in many situations it would be more appropriate to consider it as a stochastic variable too. A possible formulation for $\beta \neq 1$ is to specify u_{i0} as an error term and to impose stationarity on the process:

$$\text{var}(u_{i0}) = \text{var}(\varepsilon_{it})/(1-\beta^2) \ , \tag{3.10}$$

because then

$$\text{var}(u_{it}) = \ldots = \text{var}(u_{i1}) = \beta^2 \text{var}(u_{i0}) + \text{var}(\varepsilon_{i1})$$

$$= \text{var}(u_{i0}) \ . \tag{3.11}$$

It can be shown that for this specification of u_{i0}, equality of conditional and unconditional expectation implies $a = \frac{1}{2}$, irrespective the value of β and t (as long as $\beta \neq 1$). So for stationary processes, one should tabulate according to average income. This result once again stresses the importance of a correct understanding of the data generating process.

Another important assumption in the analysis is the assumption of constant variance of ε_{it} over time. If this assumption is relaxed, the expression for a in (3.9) becomes also a function of the variance of ε_{it} at the different moments in time, so in this situation tabulation will be somewhat more complicated. However, if the fluctuations in the variance are minor, (3.9) will offer a good approximation in most practical situations.

4. Concluding remarks

In this note we have shown the importance of knowing the data generation process in tabulating panel data. In the important case of tabulating income changes by income the (straightforward) classification by the value in the first year gives rise to regression to the mean, whereas classification according to the average value is correct for almost all values of β. Consequently, if one has no specific reason to believe that the data are generated by an autoregressive process with $\beta = 1$, the choice of the average value as the classification variable therefore seems most appropriate.

An important question is of course: what is the true process in the case of income? Much work on this has been done by Lillard (1978) and Lillard and Willis (1978), based on data from the Panel Study of Income Dynamics. Their equation is approximately equal to ours, with μ_{it} specified as

$$\mu_{it} = \alpha' x_{it} + \delta_i + \gamma_t \,, \qquad (4.1)$$

where x_{it} stands for a vector of exogenous variables, δ_i for a random individual effect and γ_t for a fixed time effect. The fact that δ_i is random does not change our analysis because it cancels out in the computation of $y_{it} - y_{it-1}$.

Their estimate of the autoregression parameter β depends heavily on the model specification. For the different specifications it takes values between 0.35 and 0.83. However, for all specifications it is significantly different from one, so according to our analysis tabulation by average income would be the most appropriate. These findings contrast with an earlier analysis by Fase (1971), who analyzed age-income profiles by means of a model with $\beta = 1$ and obtained reasonably good results. However, the hypothesis $\beta = 1$ was not tested. It should be noted that the relevant economic theory (on-the-job training in the context of the human capital theory; e.g., Theeuwes et al., 1985) is not rich enough to generate an a priori idea on the issue whether $\beta = 1$ or $\beta \neq 1$.

Things are different as to the development over time of the size of firms. In the spirit of the pioneering work of Gibrat (1931), it is sometimes postulated that the relative size change is independent of the size of the firm. So $\beta = 1$ and one should tabulate by the first year's size. If this is not true (see e.g. Jovanovic (1982) for a recent analysis and references), this tabulation will be biased as a result of regression to the mean. This sheds some new light on the sometimes proposed idea that the growth of employment is due to small firms (e.g. Birch, 1979).

References

Birch, D.L., 1979, The job generation process (Program on Neighborhood and Regional Change, MIT, Cambridge MA).

Fase, M.M.G., 1971, On the estimation of lifetime income. Journal of the American Statistical Assocation 66, pp. 686-692.

Gibrat, R., 1931, Les inégalités économiques (Sirey, Paris).

Goldstein, H., 1979, The design and analysis of longitudinal studies (Academic Press, London).

Jovanovic, B., 1982, Selection and the evolution of industry. Econometrica 50, pp. 649-670.

Keller, W.J., A. ten Cate, A.J. Hundepool and H. van de Stadt, 1985, Real income changes of households in the Netherlands, 1977-1983. Proceedings of the ISI conference, Amsterdam.

Labouvie, E.W., 1982, The concept of change and regression to the mean. Psychological Bulletin 92, pp. 251-257.

Lillard, L.A., 1978, Estimation of permanent and transitory response functions in panel data: a dynamic labor supply model. Annales de l'INSEE 30-31, pp. 367-395.

Lillard, L.A. and R.J. Willis, 1978, Dynamic aspects of earnings mobility, Econometrica 46, pp. 985-1012.

Nesselroade, J.R., S.M. Stigler and P.B. Baltes, 1980, Regression toward the mean and the study of change. Psychological Bulletin 88, pp. 622-637.

Park, R.E., B.M. Mitchell. B.M. Wetzel and J.H. Alleman, 1983, Charging for local telephone calls. Journal of Econometrics 22, pp. 339-364.

Schiller, B.R., 1977, Relative earnings mobility in the United States. American Economic Review 67, pp. 926-941.

Theeuwes, J., C.C. Koopmans, R. van Opstal and H. van Reijn, 1985, Estimation of optimal human capital accumulation parameters. European Economic Review 29, 233-257.