Journal of Official Statistics
Vol. 6. No. 3, 1990, pp. 311–317

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

Regression Effects in Tabulating From Panel Data

Huib van de Stadt[1] *and Tom Wansbeek*[1]

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 panel surveys has given an enormous impetus to the social sciences. Panels are conducted in various countries to collect and analyze information on the socio-economic variables that influence welfare. One of the most important variables in this respect is income.

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

In Section 2 the problem is illustrated by a simple example. In Sections 3 and 4 we propose a criterion for unbiased tabulation and analyze this criterion under different data generation models. Section 5 concludes.

2. Regression to the Mean

Regression to the mean can occur when the changes in a variable, measured by means of longitudinal data, are tabulated by that variable itself. One example is the tabulation of income changes by income. Table 1,

[1] Netherlands Central Bureau of Statistics (CBS) and Groningen University, respectively. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the CBS. We thank Martin Fase, Peter Kooiman, Frank van de Pol, Geert Ridder, Jules Theeuwes, the editor and three referees for their helpful comments, and Jan de Leeuw and Ivo Molenaar for guiding us to some of the references.

Table 1. Median change in real income, 1979–1981, the Netherlands

Income class	Classification according to		
	1979 income	1981 income	average income 1979 and 1981
	% per year		
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. (1987) and supplementary computations

based on the data used by Keller, Ten Cate, Hundepool, and Van de Stadt (1987) in their study of real income changes of households in the Netherlands, illustrates this problem. The first column of this table suggests a sizeable decrease in income inequality from 1979 to 1981. Individuals with the lowest incomes in 1979 did nearly 9 percentage points per year better than those with the highest incomes (+3.6 vs. −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 depend on the way the data are tabulated.

Regression to the mean is problematic because it is a "natural property" of longitudinal data and a nuisance simultaneously. Depending on what one would like to show, one can either use the first or the second column of Table 1 and reach seemingly contradictory conclusions about the development of income inequality. The final column of Table 1 gives the result from a classification by average income, and these figures look more satisfactory.

Examples of tabulated longitudinal data where regression to the mean is acknowledged but not analyzed exist in the literature (Schiller 1977; Park, Mitchell, Wetzel, and Alleman 1983). There are also theoretical writings on the subject, e.g., a brief introduction in Goldstein (1979), the comprehensive discussion by Nesselroade, Stigler, and Baltes (1980), a follow-up on the latter by Labouvie (1982), and an insightful analysis by Das and Mulder (1983), who stress that outside the multinormal context "regression to the mode" is a better name. These papers, though, do not explicitly deal with tabulation. Nesselroade et al. correctly stress that knowledge of the data generating process is a prerequisite for avoiding regression to the mean, and it is this aspect that we now develop.

3. The Basic Ideas

The problem of regression to the mean is actually a problem of endogenous selection: observations are selected on the basis of values of the variable to be analyzed (the endogenous variable). For example, when we have two variables y_1 and y_2, selection of observations with a low value of y_1 will automatically lead to a high expected value of $y_2 - y_1$ (since y_1 has a minus sign in that expression). In other contexts, endogenous selection has also been subject to analysis. A

well-known example is Heckman (1979), who suggests an adjustment for a regression equation when the selection of observations for the regression depends on the error term of the regression. Another example is Ten Cate (1986), who analyzes regression when the design of the sample depends on the endogenous variable of the regression.

In this paper we also try to solve the problems associated with endogenous selection. We do this by specifying a general stochastic structure for the income generation process and analyzing the characteristics of this structure. Let z_{it} be the (latent) "true" log-income of individual i in year t. Instead of z_{it} we observe y_{it}, which we assume is generated by the following process

$$y_{it} = z_{it} + \varepsilon_{it}, \quad t = 1, 2, \ldots, \qquad (3.1)$$

where ε_{it} is an error term. We start with the assumption that ε_{it} is normally distributed with zero expectation, constant variance, and independent over time and between individuals. It is important to stress that ε_{it} captures both real (but stochastic) fluctuations in annual income (for example, due to temporary illness or other unusual circumstances) and measurement errors.

The important question is now: What is the best way to tabulate income change against income? The answer to this question depends on what one wants to know. We discuss three possibilities.

a. One is interested in the income change of a person with observed income in a specific income class in the first year. In that case ε_{i1} should not be considered random and the first column of Table 1 (tabulation by first year income) should be used.

b. One is interested in the development of income inequality. In that case one can simply look at the two marginal distributions and compare a measure of income inequality in the first and the second year. It is not necessary to use panel data.

c. One is interested in $z_{i2} - z_{i1}$, that is, the change in income apart from random fluctuations. We show that in this case tabulation by income in the first year leads to misleading results. If one is interested in the change of the structural part of the income generating process (3.1), a different kind of tabulation is necessary.

In our opinion most users of statistics are more interested in (c) than in (a). That is, it is more interesting to know the change in "true" income than the change in observed income. For this reason we concentrate on (c) and analyze this in greater detail to determine the most appropriate tabulation.

In Table 1 regression to the mean is analyzed by tabulating income changes. We now approximate "tabulating" by "calculating a conditional expectation." Tabulating the income change from year 1 to year 2 by income in year 1 (as has been done in Table 1) can be written as the following conditional expectation

$$\begin{aligned} \mathrm{E}(y_{i2} - y_{i1} | y_{i1}) &= \mathrm{E}(y_{i2} | y_{i1}) - y_{i1} \\ &= \mathrm{E}(y_{i2}) - z_{i1} - \varepsilon_{i1} \\ &= z_{i2} - z_{i1} - \varepsilon_{i1}. \end{aligned} \qquad (3.2)$$

This shows how tabulation by income in the first year introduces a regression to the mean effect. If y_{i1} is relatively low, ε_{i1} is probably negative and the expectation of $y_{i2} - y_{i1}$ given y_{i1} is high. Hence a low income in the first year tends to be associated with a relatively large income increase, and similarly a high income in the first year tends to be associated with a relatively low value of $y_{i2} - y_{i1}$. This is exactly the effect that can be observed in the first column of Table 1.

For tabulation by the second-year income, the equation is

$$E(y_{i2} - y_{i1} | y_{i2}) = z_{i2} - z_{i1} + \varepsilon_{i2}. \quad (3.3)$$

A high income in the second year tends to be associated with a high value of $y_{i2} - y_{i1}$, and vice versa. This "regression from the mean" can be observed in the second column of Table 1.

Under this model, there are no regression effects if we tabulate by average income over the two years (or, equivalently, by total income over both years)

$$\begin{aligned} E(y_{i2} - y_{i1} | y_{i1} + y_{i2}) &= z_{i2} - z_{i1} \\ &+ E(\varepsilon_{i2} | \varepsilon_{i1} + \varepsilon_{i2}) - E(\varepsilon_{i1} | \varepsilon_{i1} + \varepsilon_{i2}) \\ &= z_{i2} - z_{i1} \end{aligned} \quad (3.4)$$

since ε_{i1} and ε_{i2} are identically distributed. This tabulation is used in the third column of Table 1.

However, this does not hold for some other processes. An example is the autoregressive process of the form

$$y_{it} = y_{it-1} + \varepsilon_{it} \quad \text{for } t = 1, 2, \ldots, \quad (3.5)$$

with ε_{it} independent of y_{it-1}, and y_{i0} fixed. This process might seem unlikely because it implies increasing income inequality over time, but note that for birth cohorts increasing inequality with rising age is not unreasonable. Now $E(y_{i1} - y_{i0} | y_{i0}) = 0$, so tabulation by the first year is correct. In this situation the two effects "regression to the mean" and "increasing inequality" cancel out each other. Since $E(y_{i1} - y_{i0} | y_{i0} + y_{i1}) = \varepsilon_{i1}$, tabulation by average income introduces regression from the mean. We conclude that process (3.5) requires another form of tabulation than that of process (3.1).

4. The General Solution

Both examples in the previous section illustrate the link between the tabulation and the data generating process. We formulate this in a more general framework in this section. Let the income generating process again be described by

$$y_{it} = z_{it} + \varepsilon_{it}, \quad t = 1, 2, \ldots, \quad (4.1)$$

but ε_{it} now denotes a normal error term which might be correlated over time and which does not necessarily have a constant variance over time. We want to tabulate the change in income by some linear combination of y_{it-1} and y_{it}, so the expression is

$$E(y_{it} - y_{it-1} | a y_{it-1} + (1 - a) y_{it}) \quad (4.2)$$

where a is some constant to be determined. Substitution of (4.1) into (4.2) and standard multivariate normal theory give

$$\begin{aligned} &E(y_{it} - y_{it-1} | a y_{it-1} + (1 - a) y_{it}) \\ &= z_{it} - z_{it-1} + E(\varepsilon_{it} - \varepsilon_{it-1} | a\varepsilon_{it-1} \\ &\quad + (1 - a)\varepsilon_{it}) = z_{it} - z_{it-1} \\ &\quad + \frac{\operatorname{cov}(\varepsilon_{it} - \varepsilon_{it-1}, a\varepsilon_{it-1} + (1 - a)\varepsilon_{it})}{\operatorname{var}(a\varepsilon_{it-1} + (1 - a)\varepsilon_{it})} \\ &\quad \times (a\varepsilon_{it-1} + (1 - a)\varepsilon_{it}). \end{aligned} \quad (4.3)$$

For unbiased tabulation (4.3) should be equal to $z_{it} - z_{it-1}$, hence

$$\begin{aligned} &\operatorname{cov}(\varepsilon_{it} - \varepsilon_{it-1}, a\varepsilon_{it-1} + (1 - a)\varepsilon_{it}) = 0 \\ &\Leftrightarrow (1 - a) \operatorname{var}(\varepsilon_{it}) + (-a) \operatorname{var}(\varepsilon_{it-1}) + (2a - 1) \operatorname{cov}(\varepsilon_{it-1}, \varepsilon_{it}) = 0 \\ &\Leftrightarrow a = \frac{\operatorname{var}(\varepsilon_{it}) - \operatorname{cov}(\varepsilon_{it-1}, \varepsilon_{it})}{\operatorname{var}(\varepsilon_{it}) + \operatorname{var}(\varepsilon_{it-1}) - 2 \operatorname{cov}(\varepsilon_{it-1}, \varepsilon_{it})}. \end{aligned} \quad (4.4)$$

Note that $a = \frac{1}{2}$ if the variance of ε_{it} is constant. For other processes, the result for a is more complicated. An interesting example is the autoregressive process

$$\varepsilon_{it} = \beta\varepsilon_{it-1} + \delta_{it}, \quad t = 1, 2, \ldots, \tag{4.5}$$

where ε_{it} is an autoregressive error term with autoregression parameter β $(0 \leqslant \beta \leqslant 1)$ and δ_{it} a normally distributed error term with zero expectation, constant variance, and independent over time and between individuals. The process starts with a fixed ε_{i0}. Using

$$\text{var}\,(\varepsilon_{it}) = \sum_{s=1}^{t} \beta^{2t-2s}\,\text{var}\,(\delta), \tag{4.6}$$

$$\text{cov}\,(\varepsilon_{it-1}, \varepsilon_{it}) = \sum_{s=1}^{t-1} \beta^{2t-2s-1}\,\text{var}\,(\delta), \tag{4.7}$$

it can be shown that (4.4) is

$$a = \frac{1 + \beta^{2t-1}}{2 + \beta^{2t-1} - \beta^{2t-2}}. \tag{4.8}$$

Equation (4.8) determines a in the case of the autoregressive process. Note first that for the process given by (3.1), the autoregression parameter β equals 0, which leads to $a = \frac{1}{2}$. We draw the same conclusion as in the previous section; in the absence of autoregression one should classify according to average income in both years. If there is autoregression with $\beta = 1$, then $a = 1$. So in this situation one should classify according to first year income.

The value of a in (4.8) is between $\frac{1}{2}$ and 1 for every β between 0 and 1. If t approaches infinity, a approaches $\frac{1}{2}$ very fast. For example, if $\beta = \frac{1}{2}$ then for $t = 2$, 3, 4, and 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 was started a number of periods ago, one should take $a = \frac{1}{2}$, which results in classification according to average income. Since in most situations we want to classify a large number of individuals of different ages and different t, average income is a good choice as long as $\beta \neq 1$.

The error term at the start of the process, ε_{i0}, is also important, despite the neglect it has received so far in the discussion. This error term is assumed fixed for convenience, but in many situations it would be more appropriate to consider it a random variable too. A possible formulation for $\beta \neq 1$ is to specify ε_{i0} as an error term and to impose stationarity on the process

$$\text{var}\,(\varepsilon_{i0}) = \text{var}\,(\delta_{it})/(1 - \beta^2). \tag{4.9}$$

This process is stationary because

$$\begin{aligned} \text{var}\,(\varepsilon_{i1}) &= \beta^2\,\text{var}\,(\varepsilon_{i0}) + \text{var}\,(\delta_{i1}) \\ &= \text{var}\,(\varepsilon_{i0}) \end{aligned} \tag{4.10}$$

resulting in a constant var (ε_{it}). It can be shown that in this case $a = \frac{1}{2}$ should be chosen, regardless of the values of β and t (as long as $\beta \neq 1$). For stationary processes one should classify 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 that of a constant variance of δ_{it} over time. If this assumption is relaxed, the expression for a in (4.8) also is a function of the variance of δ_{it} at different points in time. So in this situation, correct classification is more complicated. However, if the fluctuations in the variance are minor, classification by average income offers a good approximation in practice.

5. Concluding Remarks

We have seen the importance of knowing the data generation process if we want to

classify individuals according to income class. If $\beta < 1$, it is usually safe to classify according to average income. An important question is now: What is the true process in the case of income? A more specific question is: What is the most likely value of β?

Much empirical work on the income generating process has been done by Lillard (1978) and Lillard and Willis (1978), based on the Panel Study of Income Dynamics from the University of Michigan. Their equation is approximately equal to our autoregression model, with z_{it} specified as

$$z_{it} = \boldsymbol{\alpha}' \boldsymbol{x}_{it} + \theta_i + \gamma_t, \qquad (5.1)$$

where $\boldsymbol{\alpha}$ denotes a vector of regression coefficients, $\mathbf{x}_{it}$ a vector of exogenous variables, θ_i a random individual effect, and γ_t a fixed time effect (see MaCurdy 1982, for a more general approach). The fact that θ_i is random does not change our analysis because it cancels out in the computation of $y_{it} - y_{it-1}$.

Their estimation of the autoregression parameter depends heavily on the specification of the model and assumes values between 0.35 and 0.83 for the different specifications. However, their estimate is always significantly different from one. So according to our analysis, classification by average income would be the most appropriate. This result contrasts with an earlier analysis by Fase (1971), who analyzed age-income profiles by means of a model with $\beta = 1$. He obtained reasonably good results, but he did not test the hypothesis $\beta = 1$. It should also be noted that the relevant microeconomic theory (for example, the human capital theory) is not adequate to decide whether $\beta = 1$ or $\beta \neq 1$ (see Theeuwes, Koopmans, Van Opstal, and Van Reijn 1985).

6. References

Das, P. and Mulder, P.G.H. (1983). Regression to the Mode. Statistica Neerlandica, 37, 15–20.

Fase, M.M.G. (1971). On the Estimation of Life Time Income. Journal of the American Statistical Association, 66, 686–692.

Goldstein, H. (1979). The Design and Analysis of Longitudinal Studies. London: Academic Press.

Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47, 153–162.

Keller, W.J., Ten Cate, A., Hundepool, A.J., and Van de Stadt, H. (1987). Real Income Changes of Households in the Netherlands, 1977–1983. The Review of Income and Wealth, 33, 257–271.

Labouvie, E.W. (1982). The Concept of Change and Regression to the Mean. Psychological Bulletin, 92, 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, 367–395.

Lillard, L.A. and Willis, R.J. (1978). Dynamic Aspects of Earnings Mobility. Econometrica, 46, 985–1012.

MaCurdy, T.E. (1982). The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis. Journal of Econometrics, 18, 83–114.

Nesselroade, J.R., Stigler, S.M., and Baltes, P.B. (1980). Regression Toward the Mean and the Study of Change. Psychological Bulletin, 88, 622–637.

Park, R.E., Mitchell, B.M., Wetzel, B.M., and Alleman, J.H. (1983). Charging for Local Telephone Calls. Journal of Econometrics, 22, 339–364.

Schiller, B. R. (1977). Relative Earnings

Mobility in the United States. American Economic Review, 67, 926–941.

Ten Cate, A. (1986). Regression Analysis Using Survey Data With Endogenous Design. Survey Methodology, 12, 121–138.

Theeuwes, J., Koopmans, C.C., Van Opstal, R., and Van Reijn, H. (1985). Estimation of Optimal Human Capital Accumulation Parameters. European Economic Review, 29, 233–257.

Received April 1987
Revised April 1990