Journal of Economics and Econometrics Vol. 54, No.1, 2011 pp. 1-6

ISSN 2032-9652 E-ISSN 2032-9660

Using "shares" vs. "log of shares" in fixed-effect estimations

CHRISTER GERDES[†]

ABSTRACT

This paper looks at potential implications emerging from including "shares" as a control variable in fixed effect estimations. By shares we refer to the ratio of a sum of units over another, such as the share of immigrants in a city or school. As will be shown in this paper, a logarithmic transformation of shares has methodological merits compared to the use of shares defined as mere ratios.

JEL Classification: C23, J10.
Keywords: Scale dependency, consistency, spurious significance.

[†]Swedish Institute for Social Research (SOFI), Stockholm University, SE-106 91, Stockholm, Sweden. E-mail: christer.gerdes@sofi.su.se, Tel: +46 (8) 6747815.

1 INTRODUCTION

Occasionally one aims to examine variables that refer to a share (used here synonymous with odds, proportion or ratio) of some sort. This could be the share of unemployed in different regions, the share of women within the board of public companies, or the share of persons of foreign origin in a state or municipality, just to mention a few examples. In empirical research one habitually includes such kind of variable by its simplest form, i.e. just by taking the ratio of A to B. Sometimes, however, shares occur by their logarithmic transformation, i.e. log(A/B). The tendency of using a linear rather than a log-linear approach likely follows from convenience in use. However, for a number of reasons the linear measure could fall short of standard consistency requirements, as we intend to show in this paper.

In the following section the methodological derivation underlying the claims made here will be explained. The last section concludes.

2 FIXED-EFFECT MODELING

The main feature of standard fixed-effect estimation in a panel data setting is its focus on a variable's relative outcome to its mean value over time. That is, for the purpose of identifying coefficient estimates this approach merely utilises the within variation of a variable over time. This can be seen by the following way of notation (see for example Verbeek (2000), p. 313):

$$y_{it} - \overline{y}_{i.} = \beta'(x_{it} - \overline{x}_{i.}) + (\varepsilon_{it} - \overline{\varepsilon}_{i.}), \quad \text{where } \varepsilon_{it} \sim IID\left(0, \sigma_{\varepsilon}^2\right) \tag{1}$$

Here x_{it} are time varying control variables in region i at time t (for the purpose of the paper these variables include at least one variable denoting a share of some sort), while y_{it} denotes the according dependent variable. The coefficient vector β is estimated by conducting ordinary least squares estimations (OLS) on the demeaned variable. Similarly, in a log-linear setting one would have the following expression:[1]

[1]For ease of notation we here refer to the case where all explanatory variables enter the model in logarithms, but for the purpose of argument it does not matter how other right hand variables other than the "share"-variable(s) are treated.

$$\ln y_{it} - \overline{\ln(y_{i.})} = \beta' \left(\ln(x_{it}) - \overline{\ln(x_{i.})} \right) + (\varepsilon_{it} - \overline{\varepsilon}_{i.}) \quad (2)$$

Another way of achieving fixed-effect estimations works by including dummies in line with the following notation

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}, \text{ where } \varepsilon_{it} \sim IID\left(0, \sigma_{\varepsilon}^2\right) \quad (3)$$

As before, x_{it} are time varying control variables, but now in addition a dummy variable for the respective entity of observations (e.g. US states, municipalities or schools) are included, denoted by α_i. Frequently this way of formalising the model is referred to as "Least Squares Dummy Variable" (LSDV) approach. It can be shown that both approaches will lead to the same coefficient and standard deviation estimates; see for example Greene (2003), Chapter 3.3. That is to say, using (1) or (3) will result in equal regression estimates $\hat{\beta}$. Such similarity implies that even studies that use an approach of controlling for time constant effects by means of including dummy variables essentially are utilising within differences over time as their tool in identifying $\hat{\beta}$. The latter aspect highlights why fixed-effect estimators frequently are called "within estimators" as they suppress variation in the cross-sectional dimension.

2.1 Fixed-effect regressions with shares

Start by denoting a share in a given period as S_{it}, where $S_{it} = \frac{a_{it}}{b_{it}}$. In line with the notation in (1) the within variation of the share S_{it} can be written as $S_{ik} - \overline{S}_{i.} = S_{ik} - \frac{S_{is} + S_{i2} + ...S_{iT}}{T} = S_{ik} - \sum_{t=1}^{T} S_{it} \, 1/T$, where t is a time index, ranging from 1 to T, and $k \in \{1,...T\}$. To facilitate the presentation we will denote $\sum_{t=1}^{T} S_{it} \, 1/T$ as ΦS_i, referring to Φ as the "arithmetic mean value operator" that is applied on a sequence of shares $\{S_{i1}, S_{i2}...S_{iT}\}$.

Similarly, in a log-linear setting one has the following

$$\ln(S_{ik}) - \frac{\ln(S_{i1}) + \ln(S_{i2}) + \ldots \ln(S_{iT})}{T} = \ln(S_{ik}) - \frac{1}{T}\ln(S_{i1}S_{i2}\ldots S_{iT})$$
$$= \ln(S_{ik}) - \ln\left(\prod_{t=1}^{T}(S_{it})^{1/T}\right)$$

Subsequently we will denote $\prod_{t=1}^{T}(S_{it})^{1/T}$ as ΔS_i, saying that Δ is the "geometric mean value operator".

Focusing on the linear case to start with, one can restate the within estimator as

$$S_{ik} - \Phi S_i = \frac{a_{ik}}{b_{ik}} - \frac{\Phi a_i}{\Phi b_i}\left(\frac{\frac{1}{T}\sum b_{it}}{\frac{1}{T}\sum a_{it}}\frac{1}{T}\sum\frac{a_{it}}{b_{it}}\right), \tag{4}$$

where $\Phi a_i = \sum_{t=1}^{T} a_{it}\frac{1}{T}$ and $\Phi b_i = \sum_{t=1}^{T} b_{it}\frac{1}{T}$.

The last factor in expression (4), i.e., $\frac{1}{T}\sum\frac{a_{it}}{b_{it}}\frac{\sum b_{it}}{\sum a_{it}}$ is a statistic relating the "mean of ratios" times to the inverse of "the ratios of means". Simply for ease of notation we will call this term Pi[2]

Using the Pi notation, (4) can be rewritten $S_{ik} - \Phi S_i = \frac{a_{ik}}{b_{ik}} - \frac{\Phi a_i}{\Phi b_i}Pi$.

Dividing by Φa_i and multiplication with b_{ik} results in

$$S_{ik} - \Phi S_i = \left[\frac{a_{ik}}{\Phi a_i} - \frac{b_{ik}}{\Phi b_i}Pi\right]\frac{\Phi a_i}{b_{ik}} \tag{5}$$

This expression says that the within variation in the share S_i with respect to time in a fixed-effect setting is the weighted (!) difference in

---

[2]Letting t go to infinity Pi becomes $E(a/b)\left[E(a)/E(b)\right]^{-1}$. A standard result in statistics holds that the expectation of a ratio does not equal the ratio of expectations, i.e. $E(a/b) \neq E(a)/E(b)$. In certain situations equality applies; that is the case if (and only if) $Cov(a/b, b) = 0$, see Heijmans (1999). Sometimes equality is said to hold as a close approximation, see Angrist and Pischke (2008; 207).

the relative size of a_{ik} and b_{ik} with respect to their respective arithmetic mean values.

The implications of such a result might become clearer when one compares the above expression with the one attained with the set up in the log-linear case. One can rewrite the within estimator in log shares as follows:

$$\ln(S_{ik}) - \ln(\Delta S_i) = \ln\left(\frac{a_{ik}}{b_{ik}}\right) - \ln\left(\frac{\Delta a_i}{\Delta b_i}\right) \tag{6}$$

The equality holds simply because of $S_{it} = \frac{a_{it}}{b_{it}}$ so that

$$\ln(\Delta S_i) = \ln(\Delta \frac{a_i}{b_i}) = \ln\left(\prod\left(\frac{a_{it}}{b_{it}}\right)^{1/T}\right) = \ln\left(\prod\frac{(a_{it})^{1/T}}{(b_{it})^{1/T}}\right) = \ln\left(\frac{\prod(a_{it})^{1/T}}{\prod(b_{it})^{1/T}}\right) = \ln\left(\frac{\Delta a_i}{\Delta b_i}\right)$$

The right hand side of equation (6) can then be rephrased as

$$\ln\left(\frac{a_{ik}}{b_{ik}}\right) - \ln\left(\frac{\Delta a_i}{\Delta b_i}\right) = \ln(a_{ik}) - \ln(b_{ik}) - \left[\ln(\Delta a_i) - \ln(\Delta b_i)\right] = \ln\left(\frac{a_{ik}}{\Delta a_i}\right) - \ln\left(\frac{b_{ik}}{\Delta b_i}\right) \Leftrightarrow$$

$$\ln(S_{ik}) - \ln(\Delta S_i) = \ln\left(\frac{a_{ik}}{\Delta a_i}\right) - \ln\left(\frac{b_{ik}}{\Delta b_i}\right) \tag{7}$$

The last expression specifies the within variation (with respect to its geometric mean over time) in the logarithmic share S_{ik} as the difference in the according relative size of a_{ik} and b_{ik} with respect to their respective geometrical mean values. Comparing the linear estimator in (5) and the log-linear estimator in (7), the main difference is that the latter does not apply a weighting by $\Phi a_i / b_{ik}$. While the population indicator b_{ik} is varying over time, the numerator Φa_i is constant over the whole time period for each i.

2 DISCUSSION AND CONCLUSIONS

This paper has shown that in fixed-effect estimations the linear estimator weights changes in shares by its denominator. This result opposes the common view that shares are independent from the actual population size, i.e. that shares are not subject to scale. Stated

differently, in a fixed-effect framework a linear share indicator is scale dependent. This implies that implicit weighting of the share variable in the linear setting implies scope for spurious correlation between the share and the dependent variable.

The choice between using a log-linear or linear approach should be anchored in accordance with a number of considerations, both theoretical and empirical. In empirical research there often is no structural model available to base the model to be estimated on, so that the decision on using shares (ratios) in a linear or a log-linear way becomes intrinsically ad hoc. In such situation, the recommendation emerging from this paper would be to consider a logarithmic transformation of shares as default choice rather than to use a simple ratio.

REFERENCES

Angrist, J. and S. Pischke. 2008. Mostly Harmless Econometrics: An Empiricists' Companion. Princeton University Press, Princeton, NJ.

Greene, W. H. 2003. Econometric Analysis. (Fifth Edition). Upper Saddle River, NJ: Pearson Education.

Heijmans, R. 1999. "When does the expectation of a ratio equal the ratio of expectations?" Statistical Papers 40: 107-115.

Verbeek, M. 2000. A Guide to Modern Econometrics. John Wiley & Sons, Ltd, Chichester.

Wooldridge, J. M. 2002. Econometric Analysis of Cross Section and Panel Data, Cambridge: MIT Press.