DISCUSSION PAPER SERIES

IZA DP No. 5171

Using "Shares" vs. "Log of Shares" in Fixed-Effect Estimations

Christer Gerdes

September 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Using “Shares” vs. “Log of Shares” in Fixed-Effect Estimations

Christer Gerdes
SOFI, Stockholm University
and IZA

Discussion Paper No. 5171
September 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5171
September 2010

ABSTRACT

Using “Shares” vs. “Log of Shares” in Fixed-Effect Estimations

This paper looks at potential implications emerging from including “shares” as a control variable in fixed effect estimations. By shares I 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 some methodological merits as compared to the use of shares defined as mere ratios. In certain empirical settings the use of the latter might result in coefficient estimates that, spuriously, are statistically significant more often than they should.

JEL Classification: C23, C29, J10

Keywords: consistency, Törnqvist index, symmetry, spurious significance

Corresponding author:

Christer Gerdes
Swedish Institute for Social Research
Stockholm University
SE-106 91 Stockholm
Sweden
E-mail: Christer.Gerdes@sofi.su.se

1 Introduction

Occasionally one aims to examine variables that refer to a *share* (used here synonymous with a *ratio* or a *proportion*) 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, municipality, or school, 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 I intend to show in this paper. To be more exact, here I will focus on different aspects that emerge from incorporating shares as control variables in fixed-effect regression estimation. The overriding question of this paper is the following: What are the methodological implications of conducting fixed-effect estimations with variables stating *shares* in its linear form, in comparison with using its logarithmic transformation, i.e., the *logarithm of shares*?

For some scholars such question might look like an issue of marginal relevance. To others, especially those dealing with issues regarding outcomes emerging on some aggregated level, e.g. the country, state or municipality level, such questions are in no way far-fetched, as ratios or percentage shares frequently are of particular interest. For example, a well known study by Husted and Kenny (1997) includes the percent of black and elderly within US states in fixed-effects regression estimations, where the dependent variable is state government spending.

In the following section the methodological derivation underlying the claims made here will be explained. This is followed by a discussion as to how consistency assumptions of

coefficient estimates could be violated by the choice of estimator, while the subsequent section provides results from estimations on simulated data to test for empirical implications. 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 utilizes 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):

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

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]

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

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

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

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) are included, denoted by α_i. Frequently this way of formalizing the model is referred to as "Least Squares Dummy Variable" (LSDV)

[1] For ease of notation I 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.

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 *(i)* or *(iii)* 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 utilizing 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.

Fixed-effect regressions with shares

Start by denoting a share in a given period as S_{it}, where $S_{it} = \frac{a_{it}}{b_{it}}$.[2] In line with the notation in *(i)* the within variation of the share S_{it} can be written as $S_{ik} - \bar{S}_{i.} = S_{ik} - \frac{S_{is} + S_{i2} + ...S_{iT}}{T} = S_{ik} - \sum_{t=1}^{T} S_{it} \, 1/T$, where *t* is a time index, ranging from 1 to *T*, and $k \in \{1,...T\}$. To facilitate the presentation I 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

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

[2] Subsequently I will refer to b_{it} as "population". Depending on the research question, the population might include a_{it}, such that $b_{it} = a_{it} + c_{it}$, with c_{it} denoting "others". For the argument of this section such difference in defining b_{it} is of no relevance. However, on the margin it could play a role for the consistency argument addressed in the next section.

Subsequently I 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

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

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

The last factor in expression *(iv)*, i.e., $1/T\sum \frac{a_{it}}{b_{it}} \frac{\sum b_{it}}{\sum a_{it}}$ is a statistic relating the "mean of ratios" times to the inverse of "the ratios of means". Simply for ease of notation I will call this term *Pi*.[3]

Using the *Pi* notation, *(iv)* 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

(v) $$S_{ik} - \Phi S_i = \left[\frac{a_{ik}}{\Phi a_i} - \frac{b_{ik}}{\Phi b_i} Pi\right]\frac{\Phi a_i}{b_{ik}} \qquad \text{¤}$$

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

[3] 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 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

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

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

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

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

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

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

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 *(v)* and the log-linear estimator in *(vii)*, the main difference is that the latter does not apply a weighting by $\frac{\Phi a_i}{b_{ik}}$. While the population indicator b_{ik} is varying over time, the numerator Φa_i is constant over the whole time period for each *i*.

Next the argument will be addressed more formally. For that purpose I will connect to a paper by Törnqvist et al. (1985). In their study the authors look at "indicators of relative differences of a variable /…/ measured on a ratio scale"; see Törnqvist et al. (1985), p.1. To facilitate a

comparison with their work, I introduce a new notation, $\Omega_a \in \left\{ \frac{a_{ik}}{\Phi a_i}, \frac{a_{ik}}{\Delta a_i} \right\}$ and $\Omega_b \in \left\{ \frac{b_{ik}}{\Phi b_i}, \frac{b_{ik}}{\Delta b_i} \right\}$. Accordingly Ω_a states the relative size of a_{ik} in one time period in relation to either its arithmetic or geometric mean over time. An analogous interpretation holds for Ω_b, simply by addressing b_i instead of a_i.[4]

Using that notation and rewriting the right hand side of equation *(v)* and *(vii)* results in

(viii) $\left[\Omega_a - \Omega_b Pi\right] \frac{\Phi a_i}{b_{ik}}$ and

(ix) $\ln(\Omega_a) - \ln(\Omega_b)$ respectively.

As a first step, it will be shown that the linear estimator *(viii)* is an indicator of "relative differences" – albeit a trembling one – in line with the definition given in Törnqvist et al. (1985). As their paper already considered the log-linear setting akin to the one in *(ix)* – which also emerged to be their chosen estimator of reference – I only aim to discuss the linear case.

Definition: $C(\Omega_a, \Omega_b)$ is an indicator of relative differences given that

1) $C(\Omega_a, \Omega_b) = 0$ iff $\Omega_a = \Omega_b$ (iff ≡ "if and only if")
2) $C(\Omega_a, \Omega_b) > 0$ iff $\Omega_a > \Omega_b$

   $C(\Omega_a, \Omega_b) < 0$ iff $\Omega_a < \Omega_b$
3) $C(\Omega_a, \Omega_b)$ is a continuous and increasing function of Ω_a when Ω_b is fixed
4) $\forall l : l > 0 \rightarrow C(l\Omega_a, l\Omega_b) = C(\Omega_a, \Omega_b)$

[4] For the argument of the paper, any difference in numeric values between the arithmetic and the geometric mean is of no significant importance.

The last condition says "that the values of an indicator of relative difference must be independent of the unit of measurement", p. 43. Stated differently, it says that function $C(,.,)$ is homogenous of degree zero. Here $C(.,.)$ is used to indicate a real valued function that takes Ω_a and Ω_b as its arguments.

Looking at condition 1), one immediately realizes that *Pi* has to be equal to one in *(viii)* as otherwise the first condition will not be met. Depending on the actual empirical conditions, the factor *Pi* might take on various values, but here I will assume that *Pi* is about equal to one, so that the linear estimator *(viii)* can be approximated by $C(.,.) \approx \tilde{C}(.,.) = [\Omega_a - \Omega_b]\frac{\Phi a_i}{b_{ik}}$.

The conditions under point 2) are satisfied (provided that *Pi* is sufficiently close to one), i.e., $\Omega_a > \Omega_b \Rightarrow \Omega_a > Pi\Omega_b$ and $\Omega_a < \Omega_b \Rightarrow \Omega_a < Pi\Omega_b$.

Condition 3) can be shown to hold by differentiating $C(\Omega_a, \Omega_b)$ with respect to Ω_a:

$\frac{\partial C(.,.)}{\partial \Omega_a} = \frac{\Phi a_i}{b_{ik}} > 0$, saying that after an increase in Ω_a the functional value of the (continuous) function $C(,.,)$ will increase monotonically, given Ω_b (and *Pi*) fixed.

Finally, accuracy of condition 4) can be shown by utilizing that

$$\Omega_b = \frac{b_{ik}}{\Phi b_i} \Leftrightarrow b_{ik} = \Omega_b \Phi b_i \Rightarrow l\Omega_b = \frac{lb_{ik}}{\Phi b_i} \Leftrightarrow lb_{ik} = l\Omega_b \Phi b_i$$

Hence, one achieves the following equality

$$C(l\Omega_a, l\Omega_b) = [l\Omega_a - l\Omega_b Pi]\frac{\Phi a_i}{lb_{ik}} = l[\Omega_a - \Omega_b Pi]\frac{\Phi a_i}{lb_{ik}} = [\Omega_a - \Omega_b Pi]\frac{\Phi a_i}{b_{ik}} = C(\Omega_a, \Omega_b)$$

Setting $l = 1/\Omega_b$ (and *Pi* =1) one also has

$\tilde{C}(l\Omega_a, l\Omega_b) = \tilde{C}\left(\Omega_a/\Omega_b, 1\right) = \left[\Omega_a/\Omega_b - 1\right]\frac{\Phi a_i \Omega_b}{a_{ik}} = \tilde{H}\left(\Omega_a/\Omega_b\right)$, i.e., a (continuous)

function $\tilde{H}$, taking the ratio $\left(\Omega_a/\Omega_b\right)$ as its argument.

To sum up, even in the linear setting the conditions stated by Törnqvist et al. (1985) are met, given that *Pi* is approximately equal to one. As a next step, I now turn to discussing the advantage of using the logarithmic estimator as in *(ix)* over the linear estimator in *(viii)*.

According to Törnqvist et al. (1985), the key factor to highlight regards an estimator being *symmetric*. In particular, symmetry can be defined as follows

$H\left(\Omega_a/\Omega_b\right) = -H\left(\Omega_b/\Omega_a\right)$.

which is the standard definition for a symmetric function in two variables, i.e., $f(y,x) = -f(x,y)$.

In the log linear case symmetry applies trivially by logarithm rules, i.e.,

$\ln\left(\Omega_a/\Omega_b\right) = -\ln\left(\Omega_a/\Omega_b\right)$.

In the linear case, however, equality will not apply, i.e.,

$\tilde{H}\left(\Omega_a/\Omega_b\right) = \left[\Omega_a/\Omega_b - 1\right]\frac{\Phi a_i}{b_{ik}}\Omega_b \neq -\left[\Omega_b/\Omega_a - 1\right]\frac{\Phi b_i}{a_{ik}}\Omega_a = -\tilde{H}\left(\Omega_b/\Omega_a\right)$.

Proposition 1: Only the log-linear measure of relative differences in fixed-effect estimations is symmetric

To provide an intuitive understanding of *non-symmetry* in accordance with the Törnqvist et al. (1985) set up, note the following example: assuming a share of 4 percentage points in the

initial period, and a share of 5 percentage points in the subsequent period, the *relative change* can be measured by relating either to the value for the former, saying that the share increased by 25% (i.e., 5/4 -1). On the other hand, using as base period the latter, this would imply that the share initially was 20% lower (i.e., 4/5 -1) than it was in the subsequent period. Hence, depending on the chosen base level, one will receive different results, meaning that an "ordinary" linear percentage estimator is not symmetric. This is in contrast to the log-linear estimator, where ln(5/4) = -ln(4/5), saying that the relative change is exactly the same in absolute values. Of course, in terms of absolute changes both measures are symmetric, i.e. |5-4|=|4-5| and |ln(5)-ln(4)|=|ln(4)-ln(5)| respectively.[5]

Extension to first-difference estimations

A related estimation concept in panel based estimations to fixed-effects is the "first-difference" approach.[6] The results achieved above for the fixed-effect setting can easily be extended to first-difference estimations as shown below.

Proposition 2: Only the log-linear measure of relative differences in first-difference estimations is symmetric

Proof: First rewriting the linear estimator according to its first difference analogue results in:

$$S_{ik} - \Phi S_i = S_{ik} - \left(\frac{S_{ik} + S_{ik-1}}{2} \right) = \frac{S_{ik} - S_{ik-1}}{2} = \frac{1}{2} \left[\frac{a_{ik}}{b_{ik}} - \frac{a_{ik-1}}{b_{ik-1}} \right] = \frac{1}{2} \left[\frac{a_{ik}}{a_{ik-1}} - \frac{b_{ik}}{b_{ik-1}} \right] \frac{a_{ik-1}}{b_{ik}}$$

[5] One of the few papers referring to Törnqvist et al. (1985) is Ashenfelter and Greenstone (2004). Applying a difference-in-difference framework they emphasize the advantage of the "ln difference approach", see p. 248. A field of application where the Törnqvist et al. (1985) paper has received more attention regards price or quantity indexation; see for example Armstrong (2001) and Reinsdorf et al. (2002).

[6] It can be shown that in the two-period case first-difference and fixed effect estimation result in identical coefficient estimates and standard deviations. (See for example Wooldridge, 2003, pp. 67-68)

Applying the same reformulation in a log-linear setting result in:

$$\ln(S_{ik}) - \Phi \ln(S_i) = \frac{\ln(S_{ik}) - \ln(S_{ik-1})}{2} = 1/2\left[\ln\left(\frac{a_{ik}}{b_{ik}}\right) - \ln\left(\frac{a_{ik-1}}{b_{ik-1}}\right)\right] = 1/2\left[\ln\left(\frac{a_{ik}}{a_{ik-1}}\right) - \ln\left(\frac{b_{ik}}{b_{ik-1}}\right)\right]$$

Slightly changing the notation for Ω_a and Ω_b such that

$\Omega_a = \frac{a_{ik}}{a_{ik-1}}$ and $\Omega_b = \frac{b_{ik}}{b_{ik-1}}$, respectively, one receives

$$S_{ik} - \Phi S_i = 1/2\left[\Omega_a - \Omega_b\right]\frac{a_{ik-1}}{b_{ik}}$$

$$\ln(S_{ik}) - \Phi \ln(S_i) = 1/2\left[\ln(\Omega_a) - \ln(\Omega_b)\right]$$

This means that even in the first-difference setting the properties of the estimator are similar as in the fixed-effect setting, meaning symmetry applies in the log-linear but not in the linear setting. ■

Note that in the first-differences case there is not factor *Pi* "disturbing" the estimator. The lack of symmetry solely refers to $\left[\Omega_a - \Omega_b\right]\frac{a_{ik-1}}{b_{ik}} \neq -\left[\Omega_b - \Omega_a\right]\frac{b_{ik-1}}{a_{ik}}$.

(Non-)Consistency of estimates due to weighting in the linear set-up

Departing from equation *(i)* and using the notation in Wooldridge (2002), p. 269, the regression coefficients are estimated by: $\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_i'\ddot{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_i'\ddot{y}_i\right)$

Here $\ddot{x}_i = (x_{it} - \bar{x}_{i.})$ and $\ddot{y}_i = (y_{it} - \bar{y}_{i.})$, *t =1, 2… T* are the respective time periods and *i =1, 2,…N* are index numbers for each observation, capturing the cross sectional dimension, e.g.

US-states as in the Husted and Kenny (1997) paper. Consistency in coefficient estimations is defined as $\lim_{n\to\infty} E\left[\hat{\beta}_{FE}\right] = \beta_{FE}$

It says that the expected value of the estimated coefficient vector $\hat{\beta}_{FE}$ should be equal to β_{FE} when the sample size goes to infinity.

How could estimator $\hat{\beta}_{FE}$ become non-consistent? The most immediate argument relates to bias occurring because of a correlation between $\ddot{x}_i$ and the error term $\ddot{\varepsilon}_i$. To see this more clearly, note that one can rewrite the estimator as follows

$$\hat{\beta}_{FE} = \beta_{FE} + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_i'\ddot{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{x}_i'\ddot{\varepsilon}_i\right).$$

Given that the consistency assumption is violated, i.e. $E(\ddot{x}_i'\ddot{\varepsilon}_i) \neq 0$, then $\hat{\beta}_{FE}$ will be biased, as it will not converge to the true value of β_{FE}. According to the preceding section, a linear estimator of a share variable implies weighting by "population". For several reasons "population" could be correlated with the error term. Stated differently, the coefficient estimate of a share variable could become statistically significant due to that "population" b_i and outcome variable $\ddot{y}_i$ are correlated. More formally this can be formulated as follows. Let $\ddot{\varepsilon}_i = g(b_i) + \ddot{v}_i$ such that $E(\ddot{x}_i'\ddot{v}_i) = 0$ and $g(.)$ be some function taking b_i as its argument. In case $Cov\left(b; g(b_i)\right) \neq 0$ then $\hat{\beta}_{FE}\ddot{S} \neq \beta_{FE}\ddot{S}$, i.e. the coefficient for the linear share variable would be biased due to omitted variable bias.

3 Simulations and Estimations

The results presented above have been inherently theoretical. They should be complemented by studies on real data and/or simulation to measure the actual implications for applied research. One attempt of doing this is provided in this section. For that purpose a data set is

created with 50 time periods and 100 “states”. The share variable is constructed by letting the denominator (i.e. the “population” denoted b_i above) be a function of $\bar{y}_{i.}$, i.e. by design the denominator will be correlated with the time-demeaned dependent variable (see Appendix A for details on the design of variables and the estimation set up). To secure similar preconditions, the expected value of the demeaned share variable is approximately zero in both settings (see Appendix B). The purpose is to test the following predictions from the theoretical section above:

First: The linear share estimator will be correlated with the dependent variable due to the implicit weighting by “population”, while this is not the case when using the logarithm of a share. This implies that there is scope for spuriously significant coefficient estimates in the first, but not in the second setting.

Second: Letting the numerator (denoted a_i above) be correlated with $\ddot{y}_i$ should to a lesser extent result in spuriously significant regression estimates (in the linear setting). This is due to the fact that the shape of the weighting factor $\frac{\Phi a_i}{b_{ik}}$ implies that the denominator b_{ik} is more likely to pick up changes in the error term over time than what should be the case with the numerator Φa_i, which is constant for each unit *i*. Subsequently instead of the denominator the numerator is defined to be a function of $\bar{y}_{i.}$.

Regression estimations are conducted using both the linear and the log-linear share variable respectively, while the various appearances of the dependent variable are following exactly the same data-generating process in both settings. Thus, the models estimated read

$\ddot{y}_i = \ddot{x}_i'\beta + \ddot{\varepsilon}_i$ and $\ddot{y}_i = \ln(\ddot{x}_i')\beta + \ddot{\varepsilon}_i$, respectively.

The dependent variables are constructed in a random manner by using autoregressive parameters (subsequently denoted *k*) of size 0; 0.1; 0.2; ...1.0; respectively, utilizing various

grades of stationary in the dependent variable. Randomization allows creating outcome variables that follow a normal distribution, such that from the outset the variance of the error term is normally distributed, a feature that will translate to the standard errors of the coefficient estimates. Figure 1 shows one realization of the dependent variable after a random draw of an AR(1) process for each of the considered parameter *k*.[7] Conversely, variation of the share variables over time follow a simple random walk. This implies that dependent and independent variable are following different trends, so that the identification of coefficients, standard errors, and, consequently, statistical significance should not be the result of spurious correlation. Due to the overall random design of the share variable the null-hypothesis of a zero-effect should not be rejected more than is determined by the chosen level of significance (with deviations in the range of predictable statistical error).

Table 1 and Table 2 present the number of times the regression estimates have been found to be significantly different from zero after utilizing a five percent level of significance. Accordingly these numbers should be significant in about five out of one hundred estimations. The tables account for the numbers after two thousand iterations of estimations have been evaluated, accordingly one should expect there to be approximately 100 statistically significant estimates when testing for 5 percent level of significance. A number very different from 100 would indicate that the estimated model is not consistent, i.e. it violates the assumption that $\sqrt{N}\hat{\beta}$ is asymptotically normal distributed. Given 2000 independent draws, each with the same probability distribution, one can settle the lower and upper boundary

[7] Given the regression model in *(i)*, introducing an AR(1) process in the dependent variable leads to $y_{it} - \bar{y}_{i.} = \beta'(x_{it} - \bar{x}_{i.}) + k(y_{t-1} - \bar{y}_{i.}) + (\varepsilon_{it} - \bar{\varepsilon}_{i.})$. If one does not explicitly account for the dynamic structure the error term would become $\ddot{v}_{it} = k\ddot{y}_{i,t-1} + \ddot{\varepsilon}_{it}$, saying that $\hat{\beta}_{FE}$ will become bias if $E(\ddot{x}_{it}\ddot{v}_{it}) \neq 0$.

values of a two-sided 95% confidence interval for testing H_0: $p = 100/2000 = 0.05$ to be 81 and 119 respectively.[8]

As becomes clear from the first row in Table 1 are the numbers about 200 when the denominator is a function of $\bar{y}_{i.}$. Looking at Table 2 one can see that the log share variable is doing a much better job, lying in the range of the proposed consistency level. The different realizations of the autoregressive process by means of increasing *k* in the dependent variable (see Figure 1) describe a range of trajectories, being linear, concave or exponential, respectively. If the superior results for the log-linear estimator would just be caused by a more appropriate association of the variance covariance structure of the dependent variables and the share-variable, one should expect to find at least some sensitivity as to the distribution of test statistics in the different form of appearances of the dependent variable. As that apparently does not seem to be the case one can conclude that it is the actual structure of the estimator per se that is causing the diverging patterns of the respective share variable.

According to the predictions of the theoretical section the scope for spurious significance of share coefficients should be larger in case the denominator (the "population") is correlated with the outcome variable. The second row in Table 1 and Table 2 reports the number of cases where the p-value of the t-statistic becomes smaller than .05. The numbers are in

[8] More precisely, the confidence interval is given by $\left[x/2000 - 1.96 * \sqrt{\frac{0.05(1-0.05)}{2000}},\ x/2000 + 1.96 * \sqrt{\frac{0.05(1-0.05)}{2000}} \right]$, i.e., assuming a normal distribution given by $N\left[p_0, p_0(1-p_0)/n \right]$. The applied standard errors weighting scheme, based on estimations using the "cluster"-command in *Stata*, results in estimates that are robust to both heteroskedasticity in a cross sectional dimension and to serial correlation. For a discussion on such and related issues, see Kézdi (2004) and Bertrand et al. (2004).

accordance with statistical expectations for the log-linear setting, and approximately so in the linear setting. The latter point suggests that estimates are less prone to be affected when the correlation goes between the numerator in a share-variable and the dependent variable, than when it is the denominator.

The probable interpretation of this exercise goes as follows: given that both the dependent variable and the (denominator of the) share variable are determined by factors that relate to a common base, such as population size, there is an underlying risk of attaining non-reliable statistics by using a linear share variable, while there is no such risk in a log-linear setting. According to the numbers achieved from the present exercise, the effect can be rather sizable, rejecting the null-hypothesis of a zero-effect ten times out of hundred while testing on a five percent level of significance, which should be significant only five times out of hundred. Stating differently, in the latter case there is greater risk of Type-I errors, i.e. rejecting a null-hypothesis of a zero-effect. Of course, the way the variables are generated is inherently ad-hoc, so one cannot draw general conclusion on the overall scope of disturbance that could emerge in general. They might be smaller, but could also be larger.

4 Conclusion

This paper has shown that the linear estimator (inversely) weights changes in shares by its denominator. Relying on the work by Törnqvist et al. (1985), it can be shown that the linear estimator is non-symmetric. The 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 is determined by the particular research question under study and the data examined. Accordingly, one should not take the results presented in this paper as strict advice to use a log-linear approach anytime one

includes “share”-variables in a fixed-effect estimation framework. Indeed, the choice should be anchored in accordance with a number of considerations, both theoretical and empirical. Among others, using logs changes the reading of coefficient estimates. Anyway, 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 rather ad hoc. In such situation, the recommendation emerging from this paper would be to consider a logarithmic transformation of shares as ones default choice rather than to use a simple ratio.

References

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

Armstrong, K. G. (2001), "What impact does the choice of formula have on international comparisons?" *Canadian Journal of Economics*, 34, 697–718.

Ashenfelter, O., and Greenstone, M. (2004), "Using Mandated Speed Limits to Measure the Value of a Statistical Life," *Journal of Political Economy*, 112, 226–267.

Banerjee, Abhijit V., and Esther Duflo. 2004. "Inequality and Growth: What Can the Data Say?" *Journal of Economic Growth*, 8, 267–299.

Bertrand, M., E. Duflo and S. Mullainathan. 2004. "How much should we trust Differences-in-Differences Estimates?", *Quarterly Journal of Economics*, 119, 249–275.

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.

Husted, T. A., and L. W. Kenny, L. W. (1997), "The Effect of the Expansion of the Voting Franchise on the Size of Government," *Journal of Political Economy*, 105, 54–82.

Kézdi, Gabor. 2004. "Robust Standard Error Estimation in Fixed-Effects Panel Models", *Hungarian Statistical Review*, 9, 96–116.

Reinsdorf, Marshall B., Diewert E., and Ehemann C. (2002), "Additive Decompositions for the Fisher, Törnqvist and Geometric Mean Indexes," *Journal of Economic and Social Measurement*, 28, 51–61.

Törnqvist, L, Pentti, V., and Vartia Y. O. (1985), "How Should Relative Changes Be Measured?" *The American Statistician*, 39, 43–46.

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.

Wooldridge, J M. (2003), *Solution manual and supplementary materials for Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.

Table 1. Counting the number of significant coefficient estimates on a five percent level for "share of immigrants" as explanatory variable. Randomized dependent and independent variable; Results after 2000 iterations when using *linear shares.*

$y_t = k*y_{t-1} + e_t$; $e_t \sim N(0,1)$	k=0	k=0.1	k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7	k=0.8	k=0.9	k=1.0
Denominator function of $\bar{y}_{i.}$	204	215	217	208	192	224	234	202	232	260	206
Numerator function of $\bar{y}_{i.}$	115	142	125	120	131	117	126	120	117	121	98

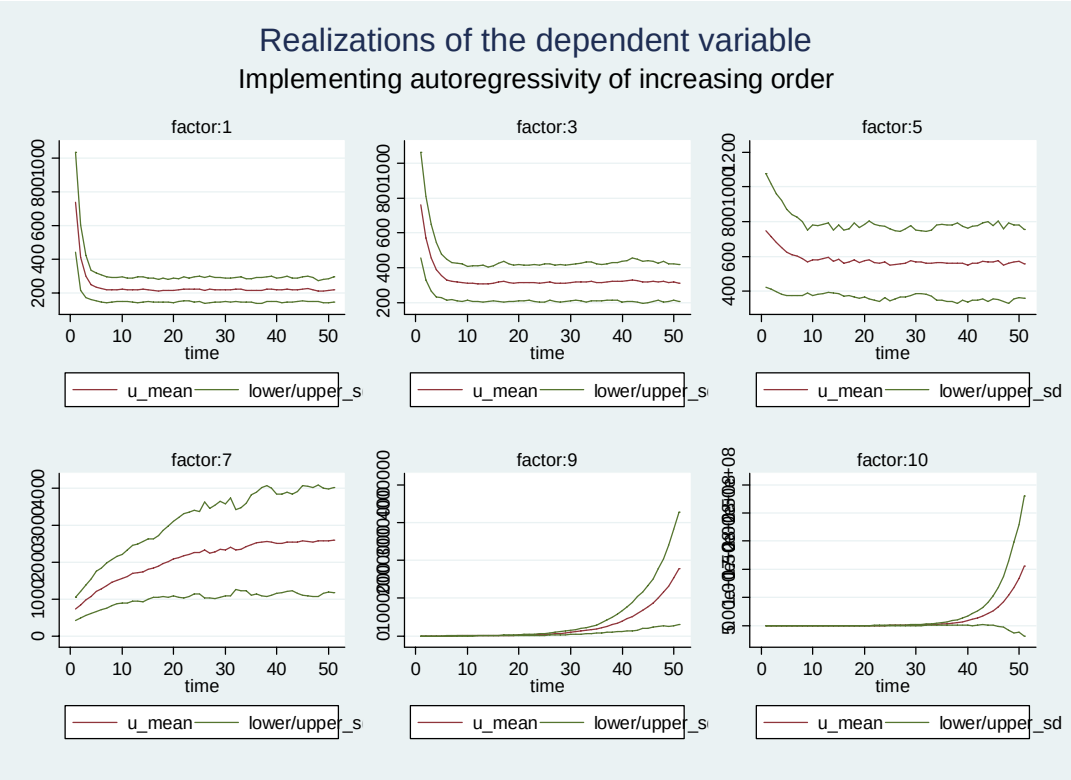
Notes: 50 time periods and 100 "states". Adjusted standard errors with respect to state clusters.

Table 2. Counting the number of significant coefficient estimates on a five percent level for "share of immigrants" as explanatory variable. Randomized dependent and independent variable; Results after 2000 iterations when using *log-linear shares*.

$y_t = k*y_{t-1} + e_t$; $e_t \sim N(0,1)$	k=0	k=0.1	k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7	k=0.8	k=0.9	k=1.0
Denominator function of $\bar{y}_{i.}$	99	108	101	84	104	102	103	103	94	100	99
Numerator function of $\bar{y}_{i.}$	106	89	98	104	102	91	98	107	106	99	120

Notes: 50 time periods and 100 "states". Adjusted standard errors with respect to state clusters.

Figure 1


Realizations of the dependent variable
Implementing autoregressivity of increasing order
factor:1
factor:3
factor:5
factor:7
factor:9
factor:10
time
u_mean
lower/upper_sd

Appendix A

```
drop _all
set obs `N'
*******
foreach k of numlist 0 1 2 3 4 5 6 7 8 9 10     {
gen pv`k'=.
                                                }
************************
foreach k of numlist 0 1 2 3 4 5 6 7 8 9 10     {

      local j=0
            while `j' < `N'               {
            local j=`j'+1
            preserve

clear

foreach i of numlist 1950/2000      {
      clear
      range Z 100 400 100
      range X 200 800 100

gen municipnr=_n
gen artal=`i'
sort artal municipnr
save arbetsdata`i', replace
                                    }

use arbetsdata1950, clear
foreach i of numlist 1951/2000      {
append using arbetsdata`i'
                                    }

tab artal, gen(d)
foreach i of numlist 1/51     {
local l=`i'+1949
rename d`i' d`l'
                              }

sort municipnr artal

* Introducing uniform random component for “Population size” applying an
* AR(1) process without drift and trend
replace Z=Z*(uniform()+1)  if artal==1950
replace Z=Z[_n-1]  + invnormal(uniform()) if municipnr[_n]==municipnr[_n-1]
& artal~=1950

replace X=X*(uniform()+1)  if artal==1950

* Defining the dependent variable by applying an autocorrelation structure
* of increasing order
gen w = X + invnormal(uniform()) if artal==1950
replace w=(`k'/10)*w[_n-1] + invnormal(uniform()) if w[_n-1]~=. &
municipnr[_n]==municipnr[_n-1] & artal~=1950

* Defining the year-average of the dependent variable
by municipnr, sort: egen w_mean=mean(w)
```

```
* Defining a variable that combines the year-average dependent variable and
* the randomly generated variable Z
gen Z_w_mean=Z*w_mean

* Defining a constant that will be used in the definition of the share-
* variable
gen constant=10+10*uniform()

* Defining the share variable. Here the denominator, indicated b, by design
* is a function of the time average of the dependent variable. Subsequently
* the algorithm for generating a and b are alternated to test if there are
* any differences emerging from the position in the share variable, i.e. if
* there are differences occurring from the position as denominator or
* numerator
gen a=  constant
gen b= (constant + uniform()-0.5)*Z_w_mean

gen share_rand=a/b
gen lnshare_rand=ln(a/b)

* Defining “log of population” to be included in the estimations as another
* covariate in the log setting
gen lnbef=ln(b)

* Running fixed-effect regression
xtreg w d1951-d2000 `ln'b `ln'share_rand, fe i(municipnr)cluster(municipnr)

gen V= _se[`ln'share_rand]    /* get standard-error for share variable */
gen b= _b[`ln'share_rand]     /* get coefficient for share variable */
     gen tv=b/V        /* the "t"-ratio */
     scalar pv`k' = 2 * ttail(e(df_r), abs(tv))    /* the p-value  */


restore

replace pv`k'=scalar(pv`k') in `j'  /* set pv to the p-value for the ith
simulation */

      save data_pv_`label'_`N'_`ln', replace

                                          }

                                                }
```

Appendix B

The linear estimator

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

$$= \left[\frac{cons}{\Phi cons} - \frac{(cons + uniform() - 0.5) * Z_{ik} \bar{y}_i}{\Phi((cons + uniform() - 0.5) * Z_{ik} \bar{y}_i)} Pi \right] \frac{\Phi cons}{(cons + uniform() - 0.5) * Z_{ik} \bar{y}_i}$$

$$= \left[\frac{cons}{\Phi cons} - \frac{(cons2) * Z_{ik} \bar{y}_i}{\Phi((cons2) * Z_{ik} \bar{y}_i)} Pi \right] \frac{\Phi cons}{(cons2) * Z_{ik} \bar{y}_i} = \left[\frac{cons}{\Phi cons} - \frac{(cons2) * Z_{ik} \bar{y}_i}{\Phi(cons2) * \Phi(Z_{ik} \bar{y}_i)} Pi \right] \frac{\Phi cons}{(cons2) * Z_{ik} \bar{y}_i}$$

$$= \left[\frac{cons}{cons} - \frac{(cons2) * Z_{ik} \bar{y}_i}{(cons2) * \bar{y}_i * \Phi Z_{ik}} Pi \right] \frac{cons}{cons2 * Z_{ik} \bar{y}_i} = \left[1 - \frac{Z_{ik}}{\Phi Z_{ik}} Pi \right] \frac{cons}{cons2 * Z_{ik} \bar{y}_i}$$

Here cons and cons2 denote two constants (*cons2=cons+uniform()-0.5*) with $E(cons2) = E(cons) = cons$. By design $E(Z_{ik}) = \Phi Z_{ik}$ so that the last equation becomes about zero when the sample size goes to infinity and assuming that *Pi*≈1

The log-linear estimator

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

$$= \ln\left(\frac{cons}{\Delta cons} \right) - \ln\left(\frac{(cons + \text{uniform}() - 0.5) * Z_{ik} \bar{y}_i}{\Delta((cons + \text{uniform}() - 0.5) * Z_{ik} \bar{y}_i)} \right)$$

$$= \ln\left(\frac{cons}{\Delta cons} \right) - \ln\left(\frac{(cons2) * Z_{ik} \bar{y}_i}{\Delta((cons2) * \Delta Z_{ik} \bar{y}_i)} \right) = \ln\left(\frac{cons}{\Delta cons} \right) - \ln\left(\frac{(cons2)}{\Delta(cons2)} \right) - \ln\left(\frac{Z_{ik}}{\Delta Z_{ik}} \right) - \ln\left(\frac{\bar{y}_i}{\bar{y}_i} \right)$$

$$= \ln\left(\frac{cons}{cons} \right) - \ln\left(\frac{(cons2)}{(cons2)} \right) - \ln\left(\frac{Z_{ik}}{\Delta Z_{ik}} \right) - \ln\left(\frac{\bar{y}_i}{\bar{y}_i} \right) = \ln(1) - \ln(1) - \ln(\frac{Z_{ik}}{\Delta Z_{ik}}) - \ln(1)$$

As above, $E(Z_{ik}) = \Phi Z_{ik}$, when the sample size goes to infinity, so that $\ln(\frac{Z_{ik}}{\Delta Z_{ik}})$ can be approximated by $\ln(\frac{\Phi Z_{ik}}{\Delta Z_{ik}})$. As a general rule, the arithmetic mean is larger than the geometric mean, which implies $\ln(\frac{\Phi Z_{ik}}{\Delta Z_{ik}}) \geq \ln(1)$. However, as long as $Z_{i1} \approx Z_{i2} .. \approx Z_{iT}$ even $\Phi Z_{ik} \approx \Delta Z_{ik}$, such that the last part of the equality in the log-linear approach is approximately $\ln(1) - \ln(1) - \ln(1) - \ln(1) = 0$