Chain Error as a function of Seasonal Variation

Vartia, Yrjö, University of Helsinki,
Suoperä, Antti, Statistics Finland,
Nieminen, Kristiina, Statistics Finland,
Markkanen, Hannele, Statistics Finland[1].

Abstract

In this study, we examine statistically the dependence between *Seasonal Variation* of consumed values and the *ChainErrors* of corresponding excellent indices in different subgroups *Ak*.

First, cyclic seasonal variation of values is calculated by simple regression analysis and the ChainError is calculated by the Multi Period Identity Test. Secondly, *Quadratic Means QM* of these two variables (or dimensions) are used in our analysis. Two quite obvious properties of the variation of quadratic mean should be specified: First, Mean of Absolute Values (MAV) varies roughly in proportion to single absolute values. Secondly, Quadratic Mean (QM) varies roughly like MAV, because for moderate changes $QM(x) \approx MAV(x)$, although $QM(x) \geq MAV(x)$ for any variable x.

The Quadratic Means of cyclic seasonal variation of values and ChainError (difference between base and chain strategies) both show variation found in typical months. The dependence between these two quadratic means is shown in the paper by *simple regression* analysis. We show that there is a *very strong statistically significant dependency* between Quadratic Means of Chain Errors and Quadratic Means of values in the seasonal index. Our main empirical findings are the following: '*Never use any construction strategy that is somehow connected with the chain strategy*'.

Our test data is a scanner data from one big Finnish retail trade chain which includes monthly information of unit prices, quantities and values form January 2014 to December 2018, and has more than 20 000 homogeneous commodities that are comparable in quality.

[1] Satu Montonen has also participated, but is at the moment on maternity leave.

1. Introduction

We use a scanner data from one big Finnish retail trade chain. The test data set contains monthly data from five years starting from January 2014 and ending to December 2018. The classification of test data is based on cartesian product of coicop7 commodity groups (151 groups) and GTIN commodity identifiers. For each GTIN, the data set contains price, quantity and value information and the data can be called as complete micro data. This data includes more than 20 000 commodities that are comparable in quality in all time periods.

Our research question in this study is "Does the seasonal variation of values cause ChainError to the index series constructed by chain strategy?" To answer this question we need:

1. Excellent index number formulas (Vartia & Suoperä, 2017, 2018).
2. Index series constructed with both base and chain strategies. In both strategies the base period is defined to be previous year normalized as average month. We get four blocks of index series to years 2015, 2016, 2017 and 2018.
3. The Multi Period Identity Test (MPIT) (Walsh, C. M., 1901, 1921), described empirically in Vartia, Suoperä, Nieminen & Montonen (2018) and the Quadratic Mean (QM) of the monthly ChainErrors (CE) derived from the MPIT in log-scale.
4. The Seasonal Index to estimate systematic seasonal variation of values and the Quadratic Mean (QM) of the monthly seasonal components of the seasonal index in log-scale.
5. Regression of the Quadratic Mean of ChainError on the Quadratic Mean of Seasonal Index.

The basic index number formulas (Laspeyres (L), log-Laspeyres (l), Harmonic-Laspeyres Lh), Palgrave (Pl), Log-Paasche (p) and Paasche (P)) are *contingently biased* and may never be used for complete micro dataset (Vartia & Suoperä, 2017, 2018). Therefore we perform analysis using the following excellent index number formulas: Stuvel (S), Törnqvist (T), Montgomery-Vartia (MV), Sato-Vartia (SV), Walsh-Vartia (WV) and Fisher (F).

The items one to two in the list above are already familiar to most of the index number statisticians, but three and four are not. We show in chapters two, three and four how ChainErrors, Seasonal Index and their Quadratic Means are derived. In chapter five the empirical relation of the Quadratic Means is demonstrated. Chapter six concludes.

Our benchmark index series is efficient base strategy that is free of the ChainError: This strategy is easily applied with the excellent index number formulas. Our proposal is based on the following links $Year(t-1) \rightarrow Year(t).m$. It compares all months m of the current year $Year(t).m$ with the (normed) previous year $Year(t-1)$. With this method there is no need to differentiate commodities according to seasonal variation – all commodities may be treated equally in index calculation. This strategy with excellent index number formula is hard or perhaps impossible to beat. Detailed analysis for our benchmark base strategy and for the chain strategy is presented in Appendix 2.

Our aim is to explore if the seasonal variations in the commodity group induce differences in the base and chain indices calculated by excellent index number formulas. More precisely, does the largeness of the seasonal components in the value series, as measured by its Quadratic Mean (QM) per month during the observation period, reflect itself in the largeness of ChainErrors (CE) derived by Multi Period Identity Test (MPIT) (Walsh, 1901, 1921; see Vartia, Suoperä, Nieminen & Montonen, 2018).

2. Definition of Quadratic Mean

Both Seasonal Index and ChainError vary around zero and we are interested how much they deviate from zero. To measure the mutual dependence between seasonal variation and ChainError, we think, it is most essential to use Quadratic Means for that. It is the **focus of the paper**.

The Quadratic Mean (also called the root mean square) is a type of average. It is used mostly in the physical sciences referencing the *"square root of the mean squared deviation of a signal from a given baseline or fit"* (Wolfram, 2019). Quadratic Mean statistic for vector *x* of *n* observations is defined as

(1) $$QM(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Taking the square root eliminates much of the huge variation of the *squares*, and the resulting output, namely the mean, $QM(x)$, is of the same overall size, but always larger than the Mean Absolute Value (MAV) of these signed numbers, that is $QM(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \geq MAV(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i \mid \geq 0.$

Quadratic Mean is used because it has 'better' mathematical properties and easier to understand statistically. The basic properties of the Quadratic Mean are demonstrated in Appendix 1.

3. Measuring seasonality

The seasonal variation (relative stationary variation of the month's values from the trend) is most easily calculated by regression analysis using logarithms of values $y^{t.m} = logv^{t.m}$. This is done for every 151 subset Ak of commodities during a *balanced* period having in our data the years 2014-2018, which as a balanced time series contains all the 12 months during each year. We have $y^{t.m} = logv^{t.m}$ as the dependent variable to be explained, while the independent variables are time $t = t.m$ and its square (a quadratic time trend) and monthly dummies. Only 11 monthly dummies or indicators can be used and one month must be arbitrarily chosen as the reference month. It does not matter which of the months is chosen as the reference because the differences in the (logarithmic) seasonal indices are invariant of the choice.

The estimates of the dummy coefficients may be deduced to seasonal components ($s^m, m = 1,2, \dots 12$ for all subsets Ak). The seasonal coefficients are normally small values around zero either up- or downwards summing up to zero for all subsets Ak.

Table 1 shows some examples of cyclically behaving seasonal components, $s^m, m = 1, 2, \dots, 12$ and their t-test statistics. The seasonal component and t-test statistic are symmetric in signs: if the t-test statistic is negative, the seasonal component is also negative, and vice versa. If the seasonal component is negative, it has been sold less than the average would indicate. Quite surprisingly, all these commodity groups have also very serious Chain Errors in all blocks that is for years 2015, 2016, 2017 and 2018. Diewert and Fox (2017) describe that if chain bias is 6-8%, it is significant.

Table 1: Some coicop7 commodity groups, their seasonal components and t-test statistics in log-percentages.

		Month											
coicop7		1	2	3	4	5	6	7	8	9	10	11	12
Sweet pastry	s^m	-0,104	0,071	0,115	0,082	0,318	-0,008	-0,091	-0,060	-0,173	-0,052	0,074	-0,173
	t	-4,967	3,393	5,509	3,899	15,164	-0,376	-4,325	-2,875	-8,250	-2,459	3,540	-8,253
Crisp bread	s^m	-0,299	-0,219	-0,029	-0,138	0,072	0,104	0,095	0,056	-0,073	0,015	0,221	0,195
	t	-12,986	-9,519	-1,268	-6,014	3,128	4,531	4,135	2,431	-3,179	0,648	9,604	8,489
Beef top side	s^m	0,179	-0,048	0,211	0,185	-0,293	-0,407	-0,596	-0,411	0,009	0,132	0,547	0,492
	t	3,235	-0,859	3,812	3,336	-5,297	-7,358	-10,766	-7,417	0,169	2,387	9,876	8,881
Filet of beef	s^m	-0,180	-0,296	-0,004	-0,044	0,095	0,224	0,134	0,113	-0,083	-0,092	0,124	0,009
	t	-4,618	-7,615	-0,099	-1,138	2,448	5,761	3,453	2,896	-2,137	-2,353	3,174	0,227
Beef strips	s^m	0,193	0,065	0,207	0,001	-0,155	-0,316	-0,341	-0,145	0,020	0,129	0,211	0,131
	t	6,812	2,303	7,294	0,039	-5,472	-11,145	-12,035	-5,122	0,708	4,538	7,444	4,636
Pork tenderloin	s^m	-0,205	-0,346	-0,082	0,043	0,364	0,491	0,390	0,237	-0,233	-0,200	-0,107	-0,352
	t	-4,775	-8,076	-1,918	1,004	8,491	11,446	9,091	5,528	-5,425	-4,658	-2,496	-8,213
Pork strips	s^m	0,121	0,005	0,179	-0,046	-0,160	-0,272	-0,257	-0,007	0,107	0,196	0,195	-0,061
	t	4,935	0,204	7,330	-1,903	-6,534	-11,118	-10,528	-0,287	4,382	8,042	7,986	-2,508
Pork joint	s^m	-0,079	-0,152	0,071	-0,030	-0,156	0,014	-0,083	-0,165	0,028	0,147	0,276	0,127
	t	-2,056	-3,985	1,867	-0,788	-4,073	0,369	-2,162	-4,317	0,732	3,858	7,229	3,325
Pork sirloin	s^m	-0,375	-0,341	-0,070	-0,017	0,251	0,419	0,351	0,183	-0,040	-0,054	-0,017	-0,290
	t	-9,664	-8,783	-1,807	-0,430	6,456	10,794	9,041	4,715	-1,036	-1,394	-0,427	-7,465
Cucumber	s^m	0,109	0,042	0,117	-0,046	-0,052	-0,061	-0,109	0,122	-0,123	-0,109	0,136	-0,026
	t	4,727	1,816	5,085	-1,984	-2,267	-2,676	-4,758	5,306	-5,335	-4,726	5,941	-1,129

When seasonal components are calculated for all of the 151 coicop7 groups, especially January, February and December come up most often having quite large up- or downward seasonal components with significant t-test statistics. In groups displayed in Table 1, many commodities have exceptionally high or low seasonal components in May, June and July. Figure 1 shows an example of seasonal variation for the commodity group ’01.2.2 Pork’ and its coicop7 groups.

Figure 1: Seasonal components for the commodity group ’01.2.2 Pork’.

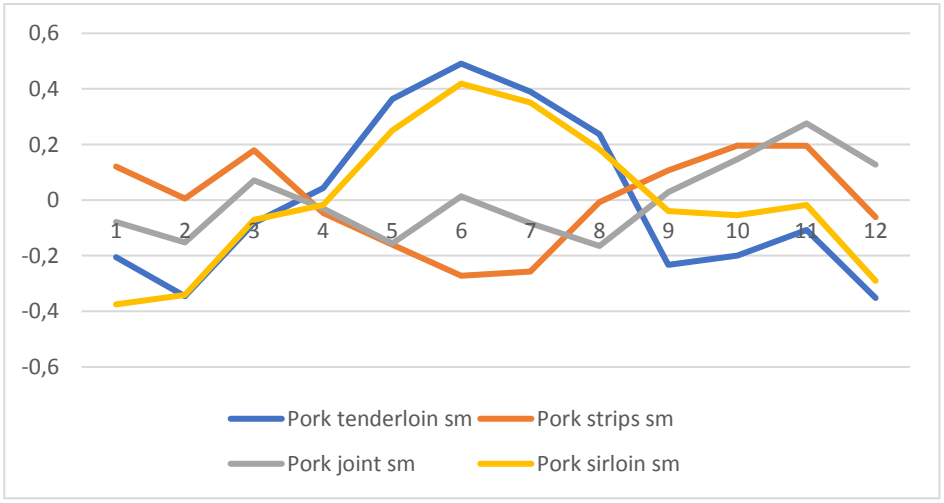


Figure 1 above gives some examples of cyclic seasonal variation of values. In practice, almost all commodity groups have different profiles telling us that the seasonal variation of values is not so nicely behaving as commonly believed. The profiles of seasonal variations may be for example downward descending, upward increasing, up- or downward concave curves, they may have the shape of saw blade or some mix of them.

When prices and especially quantities vary highly between months (bouncing effect), it may cause problems in calculations. In this study, we use values because these are natural part of the index numbers and are essential in our hypothesis. Same analysis could be performed with quantities as well.

4. Measuring Chain Error

The base and chain strategies based on the excellent index number formulas satisfying the Time Reversal Test (TR) are used for defining the Multi Period Identity Test (MPIT). The MPIT *reveals that the chain error occurs when an index does not return to unity when prices in the current period return to their levels in the base period*.

We test excellent formulas comparing chain and base strategies for each index number formula separately. We simply calculate the base and chain indices for any price index number formula *P* satisfying the time reversal test. The base period in both strategies is the previous year normalized as average month. Because the direct price-link or binary compilations have no circular or chain error, then if chained indices for any time path deviates from corresponding direct price-link, the chain strategy includes chain error surely.

In this study we do not use the basic form of the MPIT (see Vartia, Suoperä, Nieminen & Montonen, 2018), but its logarithmic difference, that is

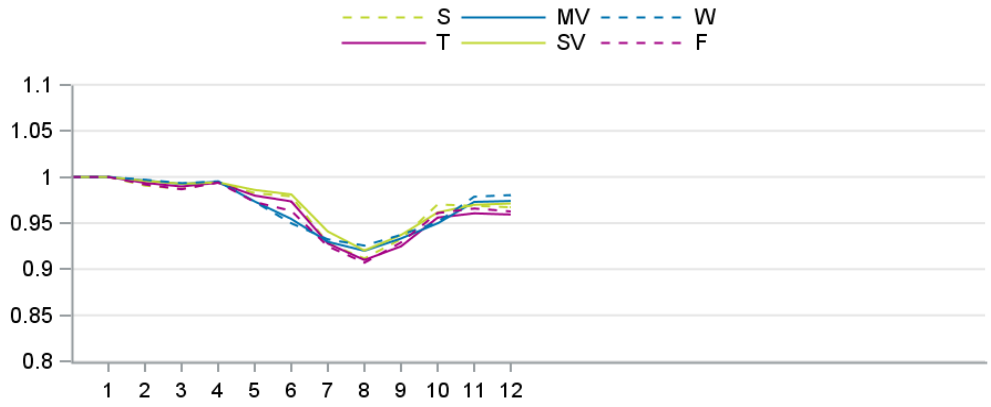
(2) $$ChainError(P, Period) = (z^{t.m}) = \left(logP_{Base}^{t.m} - logP_{Chain}^{t.m}\right), \text{ where } t.m \in Period$$

where *P* is an index number formula belonging to the family of excellent index numbers (Vartia & Suoperä, 2017, 2018). Equation (2) defines the ChainError (CE) used in this study as the relative (actually logarithmic) difference of the index series calculated using the chain strategy compared to its values calculated using the base strategy. The CE varies around zero and gets data contingently positive or negative values, if the chain index exceeds (is lower) the base index.

In the $ChainError$ the first year 2014 is used as the **base of our index computation strategy**, which means the $ChainError$ is calculated only for the periods $t.m$, where $m = 1, 2, \dots, 12$ and $t = 2015, 2016, 2017, 2018$. Therefore, the ChainErrors are calculated only for the time series starting from 2015.1 and ending to 2018.12 . As an example, for June in 2017 we have $z^{2017.6} = logP_{Base}^{2017.6} - logP_{Chain}^{2017.6}$ and for the whole year 2017 we have a piece of the time series $\left(z^{2017.m}\right) = \left(logP_{Base}^{2017.m} - logP_{Chain}^{2017.m}\right) = \left(z^{2017.1}, z^{2017.2}, \dots, z^{2017.12}\right)$.

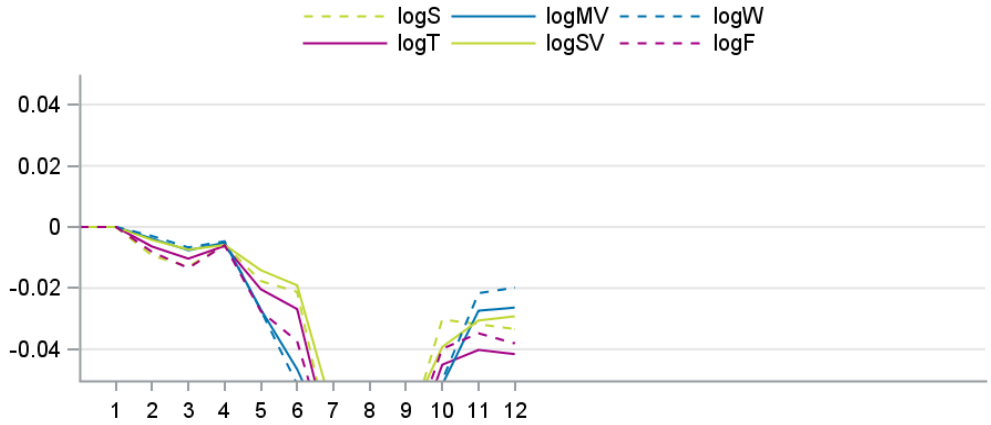
We give two graphical examples of that. Figure 2 presents the MPIT's for Stuvel (S), Törnqvist (T), Montgomery-Vartia (MV), Sato-Vartia (SV), Walsh-Vartia (W) and Fisher (F) for the commodity group '01.1.7.1.3.2 Cucumber' and Figure 3 its log-transformation (i.e. equation (2)).

Figure 2: The MPIT for selected excellent index number formulas for commodity group '01.1.7.1.3.2 Cucumber'.



Because the index does not return to unity, ChainError occurs. This can be seen in both figures.

Figure 3: Logarithmic difference of the MPIT for selected excellent index number formulas for commodity group '01.1.7.1.3.2 Cucumber'.



We have 151 'coicop7 commodity groups' and four blocks of the MPIT's and their log-transformations (i.e. for years 2015, 2016, 2017 and 2018) - so we have 604 similar pairs of figures. In most cases the MPIT's are close to one (i.e. in log-scale close to zero), but sometimes deviate very strongly from it. The y-scale is compressed to +/- 4 log-% in Figure 3, so that the commodities with high ChainError can be distinguished.

5. The Quadratic Means for the Chain error and the Seasonal components

In chapters three and four we defined seasonal components, $s^{t.m}, m = 1,2, \dots ,12$ and chain error components $\left(z^{t.1}, z^{t.2}, \dots , z^{t.12}\right)$ for t = 2015, 2016, 2017, 2018 for all subsets Ak. Now we define the Quadratic Means of them for all subsets Ak, that is for 151 commodity groups (see some basic properties of the Quadratic Mean in Appendix 1).

5.1 The Quadratic Means for Chain error

In the Quadratic Mean of $ChainError(P, Period)$ for index number formula *P*, all the signed individual components $z^{t.m} = logP_{Base}^{t.m} - logP_{Chain}^{t.m}$ of its input time series are denoted generally as the following vector or more concretely, a time series:

(3) $$ChainError(P, Period) = (z^{t.m}) = \left(logP_{Base}^{t.m} - logP_{Chain}^{t.m}\right), \text{ where } t.m \in Period$$

and the QM of them are

(4) $$QM\ of\ ChainError(P, Period) = QM(z^{t.m}) = \sqrt{\frac{1}{T}\sum_{t.m\ \epsilon\ Period}(z^{t.m})^2}.$$

For example, when the MPIT's deviate only harmlessly from one for all $t.m \in Period$, the CE components in (3) are close to zero and the QM in (4) gives small positive values close to zero. Note, that the components $z^{t.m}$ are typically small numbers, but sometimes can vary quite greatly around zero. Then their squares $(z^{t.m})^2$ become all positive (or very rarely null), but small numbers, say $z = 10^{-3}$ even become much smaller $z\text{^}2 = 10^{-6}$, but large numbers, say $z = 0{,}2$ or $z = 0{,}5$ stay large as their squares are $z\text{^}2 = 0{,}04$ and $z\text{^}2 = 0{,}25$. Thus, the relative size or relative variation of numbers grows much in squaring. When we take an average over these squares and the square root of it, we end at the Quadratic Mean of the original signed numbers.

As we already have told, the first year 2014 is used as the **base of our index computation strategy**, which means the ChainError is calculated only for the periods $t.m$, where $m = 1, 2, \dots, 12$ and $t = 2015, 2016, 2017, 2018$. Therefore, the ChainErrors are calculated only for the time series starting from 2015.1 and ending to 2018.12 . It has 4*12 = 48 months (not 60 = 5*12 months) so the Quadratic Mean of this time series is finally

$$QM\ of\ ChainError(P, Period) = QM\left(z^{t.m}\right)$$
$$= \sqrt{\frac{1}{48}\left[\sum_{2015.m=1}^{12}(z^{2015.m})^2 + \cdots + \sum_{2018.m=1}^{12}(z^{2018.m})^2\right]}$$

The calculation of the QM of chain error for all subsets Ak may be calculated using quite simply three steps:

1. Square all signed log-differences $z^{t.m}$ over the whole period $z^{2015.1} - z^{2018.12}$ of 48 observations.
2. Calculate means of squared CE i.e. $MS\left(z^{t.m}\right)$ $(= \left[QM\left(z^{t.m}\right)\right]^2)$.
3. Finally take square root of $MS\left(z^{t.m}\right)$ to get $QM\left(z^{t.m}\right) = QM\ of\ ChainError(P, Period)$.

Steps one to three have been applied to excellent index number formulas *P* – for Stuvel (S), Törnqvist (T), Montgomery-Vartia (MV), Sato-Vartia (SV), Walsh-Vartia (W) and Fisher (F). All calculations have been programmed with SAS.

The small squares $(z^{t.m})^2$ have only small influence in $QM(z^{t.m})$, but the influence of large squares is stronger. Typically $QM(z^{t.m})$ gets small or very small positive values, when it is calculated for different subgroups Ak of consumption. However, sometimes it abruptly gets very large values, sometimes perhaps even 1 000 times higher than the typical small values. In statistical language, the distribution of $QM(z^{t.m})$

over our 151 subgroups of commodities get values or lies always on the positive side of the real line, mostly near the zero, but is very skewed to the right.

5.2 The Quadratic Means for Seasonal components

The seasonal components, $s^{t.m}$ behave cyclically for t = 2014, 2015, 2016, 2017, 2018 such that, $s^{2014.m} = s^{2015.m} = \cdots = s^{2018.m}$, for $m = 1,2,\ldots,12$. Thus, it is not necessary to calculate the QM of seasonal components over years 2014 to 2018 – only one year and 12 seasonal components of it is enough (for example 2015.m, m = 1,…,12). Now the Quadratic Mean of seasonal components reduces to

$$QM\colon (s^{t.m} = seasonal\ index\ \ expressed\ in\ log - values) =$$

$$QM\colon (Season, Ak, Period) = QM\left(s^{2015.m}\right) = \sqrt{\frac{1}{12}\left[\sum_{2015.m=1}^{12}\left(s^{2015.m}\right)^2\right]} \quad (5)$$

The seasonal components sum up to zero, that is $\sum_{t.m=1}^{12} s^{t.m} = 0$, and it is easy to see, that the quadratic mean of seasonal components coincide with square root of variance of seasonal components. When seasonal components are small/large numbers around zero, then the Quadratic Mean of them is small/large number.

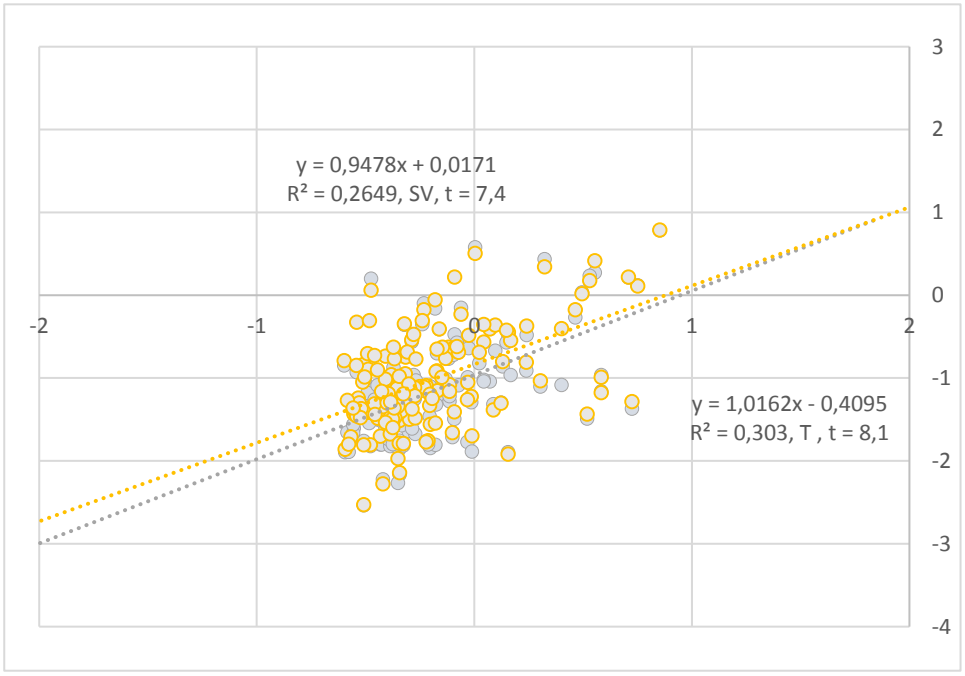
The same analysis holds for the QM of the logarithmic seasonal indices in the value time series, as before (i.e. QM of CE). Also, as before, the distribution of $QM(s^{t.m})$ lies always on the positive side of the real line, mostly near the origo, but is also very skewed to the right. Actually, the maximum values are about 100 times larger than a typical (rather small) values, which means large skewness to the right.

The Quadratic Mean of seasonal components is estimated with three similar steps as in the case of the QM of chain error.

6 Empirical Results

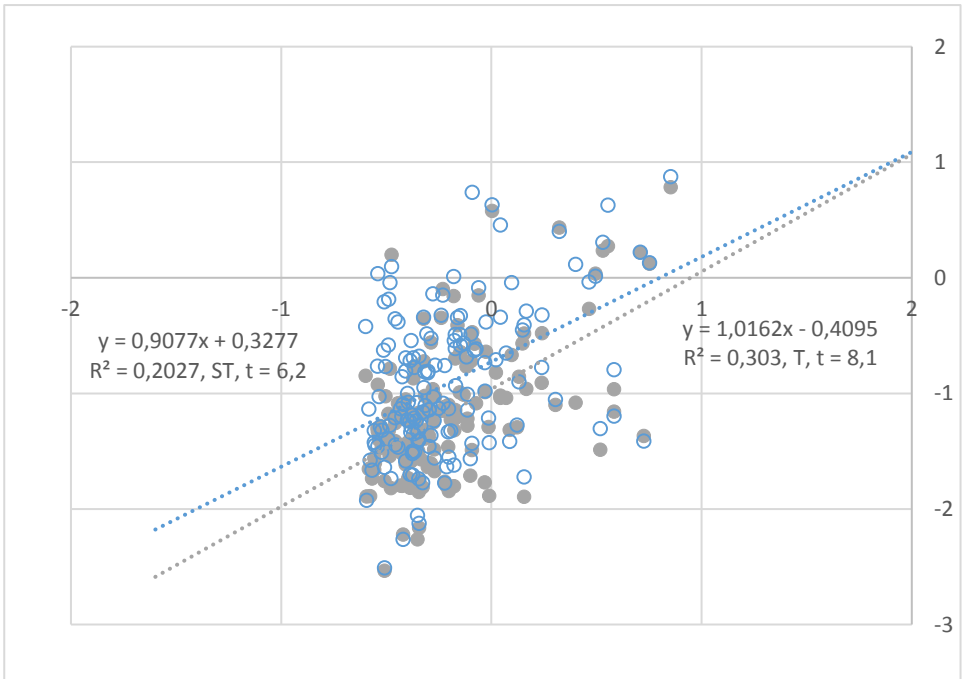
To get some sense in the analysis, we need to remove the extreme skewness of the QM variables. Therefore we have to take logarithms (here logs of base 10, log10) of both, the QM of ChainError and QM of seasonal components to reveal the relative changes in their values. This produces figures having double logarithmic coordinates, where the extreme variation is now condensed to a more manageable scale. After taking the logarithms, we need to perform the OLS regression of the logarithmic Quadratic Mean of ChainError according to the logarithmic Quadratic Mean of the seasonal component to show their mutual dependence. The OLS regression produces unbiased but inefficient estimators. This is not a problem since the results are already efficient enough for statistical purposes.

Figure 4: *LogQM of ChainError (SV,Ak,Period)* and *logQM of ChainError (T,Ak ,Period)* according to *logQM* of the Seasonal component for the same *Ak* and *Period*.



In figures 4 and 5, the seasonal component is displayed in the horizontal axis and the magnitude of Chain Error is shown in the vertical axis. Both axis are on log10-scale which makes it double-log10-scale. This makes the figures easier to interpret. Sato-Vartia (SV) is displayed with yellow dots and Törnqvist (T) with grey dots. The coefficients of determination and t-statistics are highly statistically significant. Below is the same picture with Törnqvist and Stuvel index formulas.

Figure 5*: LogQM of ChainError(SV,Ak,Period) and logQM of ChainError(T,Ak,Period)* according to logQM of Seasonal for the same Ak and *Period*. Hollow points correspond to *T*.



In the figures 4 and 5 above, the x-scale observation varies from -2 to 2, which means that the largest original values of QM: $(Season, Ak, Period)$ are roughly 70 times higher than the smallest ones. In the y-scale, the observations vary roughly between log10-values from -2 to +2. This means large variation in the original scale: the large values of QM: $CE(logMV, Period)$ are 1 000 times larger than the small ones.

The two regressions, one for SV =Sato-Vartia and the other for T = Törnqvist, are almost identical in this double-logarithmic scale. They evidently make excellent sense with R^2-values 26% and 30%, which according to standard t-test are extremely significant statistically. These two figures essentially solves the question of our article: *relative or logarithmic differences in the largeness of seasonal variation in the log-values of time series in various 151 subgroups, clearly have average positive effects on the log-values of largeness of Chain Errors (= relative differences between index numbers produced by chain or, alternatively, by the base method). It does not matter, which of the excellent index number formula is used, because this choice affects results only slightly*. As evidence of this in figure 4, Sato-Vartia and Törnqvist indices differ from each other only "slightly", compared to the large differences between different subset Ak (our 151 points in the figure).

Note also the correct understanding of the 'random variation' around the regression lines. In the middle of the figure, roughly 50% of the points are both above and below the regression lines. This is just how according to the statistical theory should happen. What does this mean? Both the deviations in the log10-scale are typically 2 log-units. This means 10-fold or 1/10-fold in the original scale. Observations 10 times above the average can be seen more clearly in the original scale while the 1/10-fold cannot be distinguished as they are too small or too near the x-axis to be seen.

The large skewness of the distributions of these two quadratic means (from one subgroup Ak to another) is really extraordinary and requires carefully adjusted statistical and mathematical methods. No standard methods applied to their original scales are suitable or more clearly stated: standard methods such as regression analysis on their original absolute scales is clearly "forbidden". However, statistical programs would do the senseless job of calculating such regressions without *"explosion*". In Appendix 4 we show some examples of these "forbidden" estimation methods.

7 Conclusions

ILO manual (2004, p. 393-410) classifies commodities into *normal, weakly and strongly seasonal ones* and suggests **different treatment** for them. In this study, we do not make any distinction between commodities. As noticed, we derive cyclically behaving seasonal indices of values and log-differences of the MPIT's (i.e. ChainErrors in log-scale) for all 151 coicop7 commodity groups without categorization of commodities into seasonal and non-seasonal ones.

We derive by these two statistics – seasonal index and chain error – and the quadratic means of them. We show empirically with these quadratic means that *relative or logarithmic differences in the largeness of seasonal variation in the log-values of time series in various 151 subgroups, clearly have average positive effects on the log-values of largeness of ChainErrors.*

Practically this means that all construction strategies of index series, which are based somehow on chaining[2], should all be avoided or *actually forbidden*. For example, the simple chain strategy using weekly,

[2] The actual reasonable number of these strategies for one year is somewhat less than 40 000 000. (The exact number for the 12 months of any year is 11! = 39 916 800.)

monthly and quarterly links should never be used. Similarly, the multilateral RYGEKS presented in Ivancic, Diewert & Fox (2011, p. 33 equation 9), is a method that cannot be recommended as actual method for official statistics. It is just 'equation 9-type' updating of index series simply by multiplication, that necessarily causes chain error. Compare also Vartia, Suoperä, Nieminen & Montonen, 2018.

This study is based on *our special strategy* combined with *excellent index number formulas* as benchmark method. Our bilateral strategy uses previous year normalized to average month as its base period. This base period describes representative consumption (in the relevant year) and this holds all commodities – irrespective whether they are normal, weakly or strongly seasonal! Direct *price-links* of our strategy use only binary comparisons between base period of years length and all 12 observation months. Price-links should be based on *"flexible basket approach"* which correctly reflects consumers' expenditure patterns in all binary comparisons. Artificial distinctions between normal, weakly and strongly seasonal commodities are not needed.

Our natural and simple bilateral strategy has three important properties:

(i) Our strategy removes all problems caused by chaining (i.e. *multiplications* needed in chain index). The only practical index series, that are totally free from *ChainError CE*, are simple versions of our strategy.

(ii) Our strategy treats all months of every year *equally*. All other strategies contain at least one link implying a multiplication. And we have proved, that this multiplication necessarily causes (data contingent) ChainError.

(iii) Our strategy treats weakly seasonal, strongly seasonal and non-seasonal commodities totally symmetrically. They all have their proper contributions to the overall *CPI* .

This message, we want to share with you.

References

Diewert and Fox (2017): "Substitution bias in multilateral strategies for CPI construction using scanner data", Ottawa group, 2017

Eurostat (2018): Harmonised Index of Consumer Prices (HICP): Methodological Manual

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, Peter Hill (ed.), Geneva: International Labour Office.

Ivancic, L., Diewert, E. and Fox, K. J. (2011): "Scanner data, time aggregation and the construction of price indexes", Journal of Econometrics 161 p. 24-35.

Johansen, I., Nygaard, R. (2011): "Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages", Ottawa Group, 2011

Nieminen, K. and Montonen, S (2018), "The foundation of index calculation".

Vartia, Yrjö (2018), 'Good choices in index number production'.

Vartia, Yrjö and Suoperä, Antti (2018): "Contingently biased, permanently biased and excellent index numbers for complete micro data".

Vartia, Y., Suoperä, A., Nieminen, K. and Montonen, S. (2018a), "Circular Error in Price Index Numbers Based on Scanner Data. Preliminary Interpretations".

Vartia, Y., Suoperä, A., Nieminen, K. and Montonen, S. (2018b), "The Algebra of GEKS and Its Chain Error".

Walsh, C.M. (1901), *The Measurement of General Exchange Value*, New York: Macmillan and Co.

Walsh, C. M. (1921), "Discussion", *Journal of the American Statistical Association* 17, 537-544.

Wolfram (2019) : Weisstein, Eric W. "Root-Mean-Square." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/Root-Mean-Square.html.

Appendix 1: Some important properties of the Quadratic Mean

The basic properties of $QM(x)$:

1. If all the values of x are non-negative, then $QM(x)$ lies always between smallest and largest of them.
2. For any non-negative multiplier c and even for any real (not necessarily non-negative) variable, the $QM(x)$ is *homogeneous*: For any $c \geq 0$, $QM(cx) = cQM(x)$. This applies trivially for $c = 0$. This shows how change of units affect $QM(x)$. Naturally, as you should see.
3. The order in which the values of x are expressed is irrelevant for $QM(x)$. It is order independent or invariant in *permutations* ψ of the vector x: $QM(\psi x) = QM(x)$. Such a function is also called as symmetric in (the components of) the variable x.
4. For non-negative variables x, $QM(x)$ is always greater or equal to the ordinary arithmetic mean $AM(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \ = \bar{x}$ expressed as $0 \leq min(x) \leq AM(x) = \bar{x} \leq QM(x) \leq max(x)$. All these numbers must be non-negative, though normally they all are positive. This inequality requires rather sophisticated mathematics in its proof. Without the non-negativity condition, no such equality holds, because e.g. $max(x)$ can be negative, while $0 \leq QM(x)$ always by squaring the components.
5. Perhaps a helpful relation, which hold also for any real variables x, whose components may attain also negative (or even only negative) values, is the following. Consider the MAV or Mean of Absolut Values denoted and defined by $MAV(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i|$, where $|x_i|$ = the non-negative absolute value of the possibly negative x_i. The inequality in 4 actually implies the following inequality between QM and MAV. For all real (positive, negative or null) components x_i of the variable $x = (x_1 , x_2 , \dots , x_n)$, the following inequality always holds: $0 \leq min(|x|) \leq MAV(x) \leq QM(x) \leq max(|x|)$. This is very helpful when interpreting the positive values produced by $QM(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \geq 0$.
6. Note that the inputs $x = (x_1 , x_2 , \dots , x_n)$ of $MAV(x)$ and $QM(x)$ may well all be negative, $x \in R^n$, unlike the vector of absolute values, for which $|x| = (|x_1|, |x_2|, , \dots , |x_n|), \in R_+^n$, The inputs in $MAV(x)$ and $QM(x)$ are not the positive absolute values or the squares, but these vectors $x = (x_1 , x_2 , \dots , x_n)$ with possibly and usually negative components. This may sound odd, but an important point is that the absolute valuing and squaring *happens within the functions* $MAV(x)$ and $QM(x)$. Although e.g. the equalities $MAV(x) = MAV(|x|)$ and $QM(x) = QM(|x|)$ hold trivially, it is conceptually important and is hidden in our notation, that both these functions use as their arguments the real (not the non-negative) arguments. Mathematically, they are functions of type: $MAV: R^n \rightarrow R_+^n$ and $QM: R^n \rightarrow R_+^n$. Note how clumsy the *alternative notation* $QM(|x|) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|x_i|)^2}$ would be. Notations like concepts should be as elegant and natural as possible.

These are typical properties of all means, which shows the Quadratic Mean $QM(x)$ deserves its name and it really is a mean, a special mean of its special type! Note, that our important paper Vartia & Suoperä (2018) based its results heavily on general moment means (of which QM is a special case) and on log-changes.

Quadratic Mean multiplied by $\sqrt{n}$ is the Euclidean Length of a vector:

$$l(x) = \sqrt{n} * QM(x) = \sqrt{\sum_{i=1}^{n} x_i^2}$$

and its square is the (Pythagorean) square length of a vector, which is mathematically much easier to manipulate than its length:

$$l^2(x) = \sum_{i=1}^{n} x_i^2 = n * QM(x)^2.$$

This is how our $QM(x)$ is connected not only to statistics and to general mathematic, but also to the geometry of the common n-dimensional Euclidean space.

Appendix 2: Current base-based strategy and formula applied for scanner data in Finland

It is known fact that the value of a price index number depends on three things: data in question, the strategy used (base, chain or rather a mixture of them) and on the index number formula. These should fit together and give maximally reliable results. So we had to find the most suitable composition of the formula and strategy.

Scanner-data opens up new possibilities to change traditional practices because:

- In data having more than 100 000 homogenous commodities, quality errors virtually vanishes
- Information concerning value and quantities are available up from daily or weekly level
- Quantity and value information in addition to price information allows to use the excellent formulas
    - Different index number formulas (excellent and contingently biased) are easily compared
- New strategies, not just pure base or pure chain strategies, may be used

During recent years we tested various combinations of alternative construction strategies and the index number formulas.

Based on the research work, international research reports and all our test results, following decisions were made and implemented to the production in the beginning of year 2019 concerning the scanner-data:

- The Törnqvist index number formula is applied
- The base strategy is used with the normalized average month of previous year as a base period
- Only false registrations like missing or erroneous classification, erroneous product label, negative prices and quantities are filtered out
- The price relation is calculated by each item identified with the GTIN-code
- Elementary aggregates are composed by using the excellent index number formula. In Finland, the elementary aggregate level is by time, region and coicop-7.
- The scanner-data elementary aggregates are integrated together with the traditionally collected and processed elementary aggregates using enterprise-specific weights
- New items are taken into account at the update of the next base period.
- Items that do not have price or quantity for year t are deleted from calculations.
- Annual chaining is used for merging together index series having different base periods

Appendix 3: Definition of the chain error in the case of Montgomery-Vartia

We define the chain error as

$$CE(\Delta_{Base\to Chain} logMV, A_k, Period)_{year(t-1)}^{year(t).m} = \log\widetilde{MV}_{year(t-1)}^{year(t).m} - logMV_{year(t-1)}^{year(t).m}$$

where the MV-base and chain indices are calculated for the commodities in subgroup A_k (not denoted in the RHS symbol) and for every month $m = 1, \dots, 12$ of every block of calendar $year(t), t = 2015, \dots, 2018$ in our current $Period$.

In our recommended base strategy, which is also a special case of GEKS (see Vartia, Suoperä, Nieminen & Montonen, 2018), the base indices $MV_0^{year.m}$ are comparing directly the months *t* of any year $year(0).m$ with the average month of the previous year $year(-1)$. Its links are, therefore, very simple:

$year(-1) \to year(0).m$, $m = 1, \dots$,12 and annual links: $year(t) \to year(t+1)$

and the base index using MV-index is

$$logMV_{year(t-1)}^{year(t).m}$$

for any $m = 1, \dots$,12 and any $year(t)$ where in our present data $t = 2015, \dots$,2018. The year $t = 2014$ must be used as a starting value in this recursive calculation. Annual changes are calculated naturally using only annual data

$logMV_{year(t-1)}^{year(t)}$ and we set $logMV_{year(t-1).12}^{year(t).1} = logMV_{year(t-1)}^{year(t)} + logMV_{year(t)}^{year(t).1}$.

This means that annual data is treated like the 13th month, which raises the level of the next year indices to the appropriate level. This is best presented using a table.

The whole index series is now defined by adding relevant log-changes:

$logMV_0^{year(t).m} = \sum_{t=2015}^{2018} logMV_{year(t-1)}^{year(t).m}$, with

MV-index based on this intuitively evident strategy is in log-form $logMV_0^{year.m}$, where the time variable $t = year.m$ gets values $m = 1, \dots$,12 within every calendar year-block of the $Period$. In our data we have five calendar years 2014 - 2018, from which the year 2014 forms the initial point 0 of our calculations and the first month is 2015.1. For all the months of 2015 calculate $logMV_0^{year.m}$ and then the same procedure repeats for 2016.m. In these, the average month of 2015 forms the basis. We have produced in this way four annual blocks of 12 months, together 4*12 = 48 observations, where every month appears exactly 4 times. For these 48 observations, we compare the outcomes of the base and chain indices, calculated by the same formula (in our example MV).

On the other hand, the (pure) chain strategy is based on 4 blocks of 12 months, where for every year *t* we start from the comparison

$year(t-1) \to year(t).1$, which is the same as in the base strategy.

The later links compare consecutive months, because $\log\widetilde{MV}_0^{year.m}$ is the pure chain index:

$$year(t).(m-1) \to year(t).m, m = 2, \dots, 12.$$

Then the year-block changes and we start anew using (9). For every year-block, we have for the chain index

$$\widetilde{MV}_{year(t-1)}^{year(t).1} = MV_{year(t-1)}^{year(t).1} \text{ and } \widetilde{MV}_{year(t-1)}^{year(t).m} = \widetilde{MV}_{year(t-1)}^{year(t).(m-1)} MV_{year(t).(m-1)}^{year(t).m},$$

where $m = 2, \dots, 12$. This is much easier in logarithms:

$$log\widetilde{MV}_{year(t-1)}^{year(t).1} = logMV_{year(t-1)}^{year(t).1} \text{ and}$$

$$\widetilde{logMV}_{year(t-1)}^{year(t).m} = log\widetilde{MV}_{year(t-1)}^{year(t).(m-1)} + logMV_{year(t).(m-1)}^{year(t).m} =$$

$$= \sum_{k=2}^{m-1} logMV_{year(t).(k-1)}^{year(t).k} + logMV_{year(t).(m-1)}^{year(t).m} =$$

$$= \sum_{k=2}^{m} logMV_{year(t).(k-1)}^{year(t).k}.$$

In logarithms, the pure chain strategy means just summing log-changes from the previous month, $logMV_{year(t).(k-1)}^{year(t).k}$, where the index number formula is MV, but expressed in logarithms $logMV_{m-1}^{m}$. Of course, within the months *m* of every calendar year. In the change of the year, start the strategy anew, by calculating the changes from the average month of the previous year. Note, that we *do not* propose the chain strategy, but the base strategy.

Now we are ready to define the important concept Chain Error CE. It is simply the vector between the chain and base vectors:

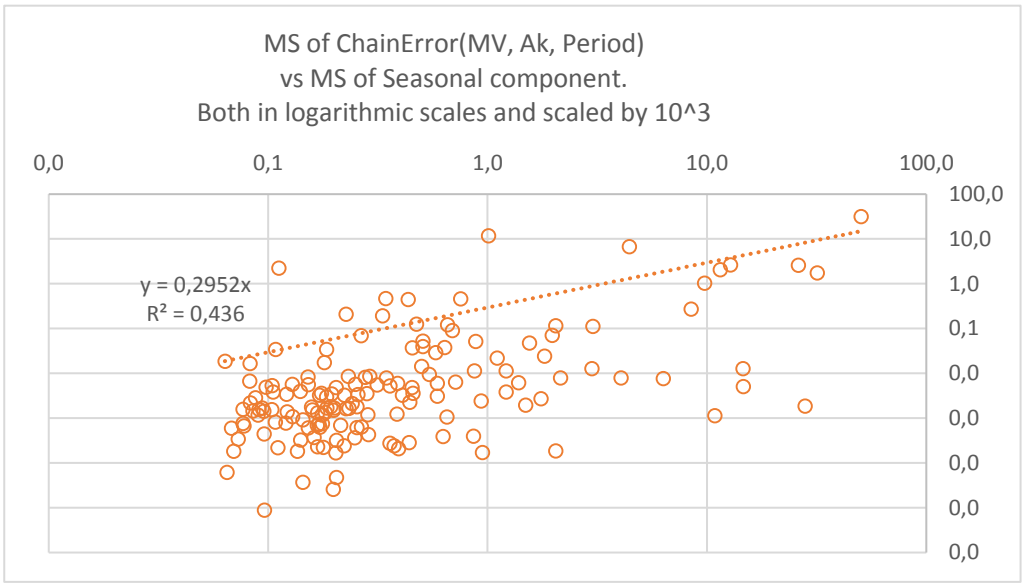
$$\Delta log\pi^{t.m} = \widetilde{logMV}_{year(t-1)}^{year(t).m} - logMV_{year(t-1)}^{year(t).m}$$

$$CE(logMV, A_k, Period) = (\Delta log\pi, A_k, Period),$$

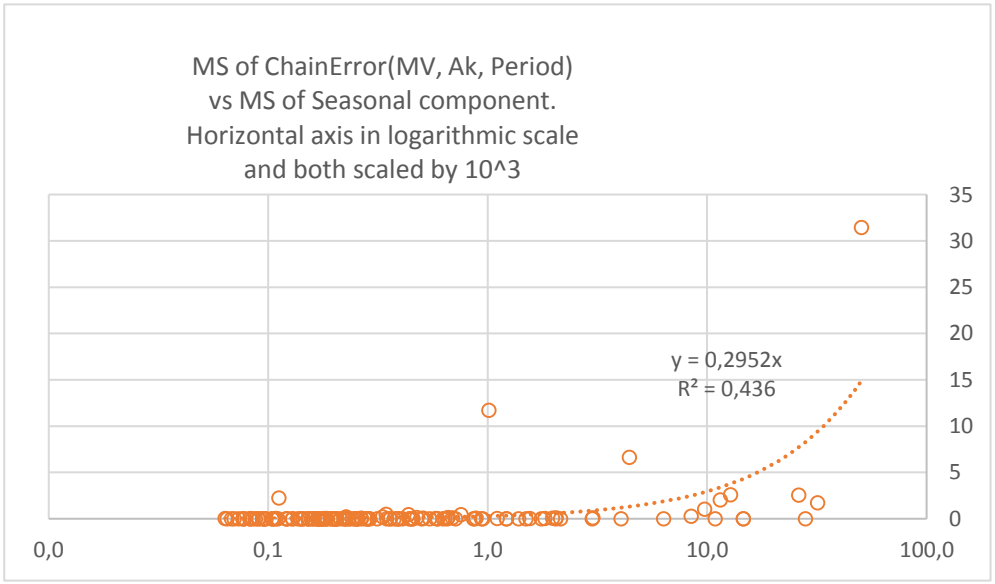
where also the background data $(A_k, Period)$ is included in the notation

Appendix 4: Some unsuitable estimation methods

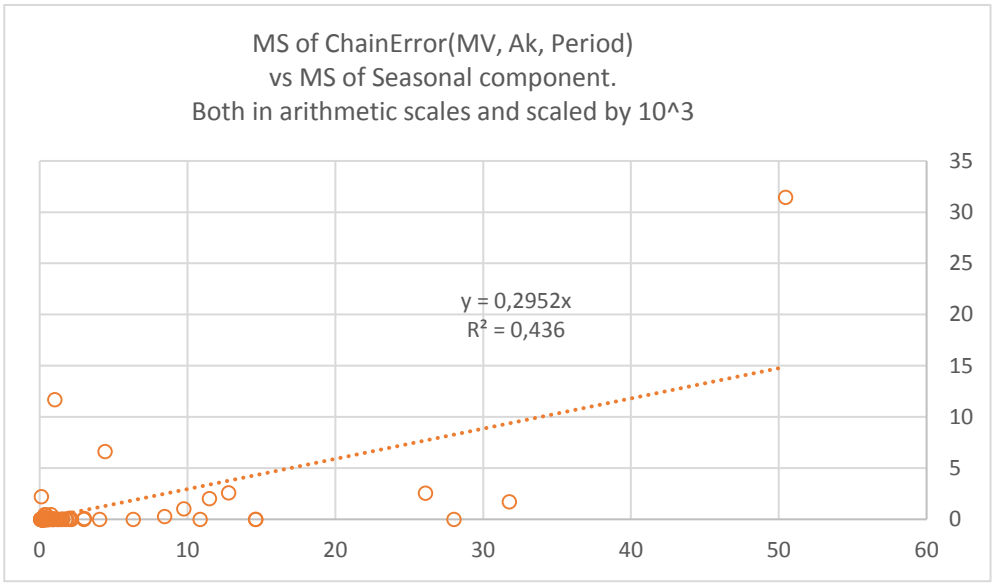
Here we show some examples of “forbidden” estimation methods, where the scale of variables are not admissible. The following figure is almost the same as in chapter 5, but now we calculate MS = Mean of Squares instead of the QM = Quadratic Means. Note the change in the scales (both logarithmic in this figure), where variation has increased by 1000. Also, there is a slight change in the regression (forced to go through origo in both), including its R^2 = 0,436. Not even R^2 is invariant in this transformation. This shows that there is something wrong in it.



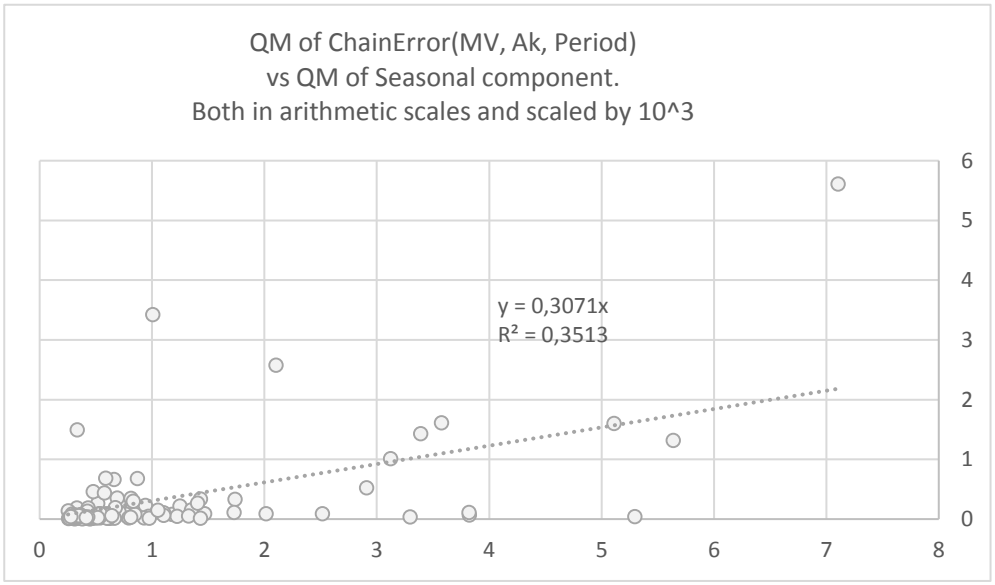
This is how the figure comes and looks if only the x-axis is logarithmic.



And if both variables are in the original absolute scale. This is the space where the regression through the origo was calculated. The space is not at all the correct or proper space for meaningful or best calculations of regressions.



Note the effect of using QM instead of MS. QM is better, easier to understand, in original units.



In the figures, both the coordinate variables have been scaled by a large constant. For example, scaling could be by 10^3, so to get small integer values for the both scales. This corresponds to the usual practice, when small share of a disease, say 0,00013 in the population, is communicated to the public. It is transformed to 1,3 cases per 10 000 inhabitants.