Variance of the generalized regression estimator under measurement error

Jan van den Brakel[1,2] and John Michiels[1]

1 Statistics Netherlands, 2 Maastricht University

Paper presented at 17th Workshop on LFS Methodology, Neuchatel 25-26 April 2024

This project is financed by Eurostat under grant agreement 2023-NL-LFS, nr. 101139082

1 Introduction

The purpose for the Dutch Labour Force Survey (DLFS) is to publish monthly, quarterly and annual figures about the employed and unemployed labour force in the Netherlands. Monthly publication tables are compiled with a multivariate structural time series model (STM), while quarterly and annual figures are predominantly produced with the more commonly used generalized regression (GREG) estimator. To enforce consistency between monthly and quarterly figures, the weighting scheme for quarterly figures contains a component that is based on the monthly labour force estimates. The variance of the GREG estimator ignores the uncertainty of including components in the weighting scheme that are observed with measurement error. This interim report describes the first results of a project that aims to develop a variance approximation for the GREG estimator that accounts for the additional uncertainty that is a result of using weighting components that are observed with error.

The paper is organized as follows. In Section 2 the DLFS is being described, followed by a presentation of the STM for monthly figures in section 3. In Section 4 an analytical expression for the variance of the GREG estimator that accounts for uncertainty in the population totals of the weighting scheme is proposed. In Section 4 an outline of the research project is described.

2 Dutch Labour Force Survey

Before 2000, the DLFS was designed as a cross-sectional survey. Since October 1999, the DLFS has been conducted as a rotating panel design. Until the redesign in 2010, data in the first wave were collected by means of computer assisted personal interviewing (CAPI). Respondents were re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing (CATI). During these re-interviews, a condensed questionnaire was used to establish changes in the labour market position of the respondents.

In 2010, a major redesign for the DLFS started. The main objective of this redesign was to reduce the administration costs of this survey. This is accomplished by changing the data collection in the first wave from CAPI to a mixed data collection mode using CAPI and CATI. Households with a listed telephone number are interviewed by telephone, the remaining households are interviewed face-to-face. To make CATI data collection in the first wave feasible, the questionnaire for the first wave needed to be abridged since a telephone interview, according to data collection literature, should not take longer than 15 to 20 minutes. Therefore, parts of the questionnaire were transferred from the first to the second or the third wave. In 2012, a second major redesign of the DLFS took place. Data collection changed to a sequential mixed-mode design that starts with Web interviewing.

The new survey design of the DLFS, that was first implemented in 2021, is different in several ways. The sample design changed from a stratified two-stage cluster sample of households to a stratified two-stage cluster sample of persons. The target population is people from 15 to 89 years old living in private households. Under the new design, samples will be drawn on a weekly instead of a monthly basis.

3 Time series model for monthly estimates

Since June 2010, Statistics Netherlands uses a multivariate structural time series model (STM) for the production of monthly labour force figures. The DLFS is based on a rotating panel design. Each month a new sample enters the panel. This sample is observed five times at quarterly intervals. After the fifth interview round, respondents leave the panel. The sample observed for the jth time is further shortly denoted as the jth wave. As a result of the rotation scheme, each month data are collected in five independent samples, i.e. the sample of the first wave that enters the panel for the first time, the sample of the second wave that entered the panel three months ago and that is observed for the second time, the sample of the third wave that entered the panel six months ago and is observed for the third time, etc. Let $\hat{y}_t^{[j]}$ denote the general regression (GREG) estimator (see Särndal et al. (1992)) for an unknown population parameter in month t, based on the sample that is observed for the jth time. As a result, each month, five GREG estimates are observed that can be collected in a five dimensional vector, say $\hat{\boldsymbol{y}}_t = (\hat{y}_t^{[1]}, \dots, \hat{y}_t^{[5]})'$. From this, a five dimensional time series can be constructed, which is the input of the following STS model:

$$\hat{\boldsymbol{y}}_t = \mathbf{1}_{[5]}\theta_t + \boldsymbol{\lambda}_t + \boldsymbol{\Delta}_t^1\boldsymbol{\gamma}^1 + \boldsymbol{\Delta}_t^2\boldsymbol{\gamma}^2 + \boldsymbol{\Delta}_t^3\boldsymbol{\gamma}^3 + \boldsymbol{\delta}_t^{COV}\gamma^{COV} + \boldsymbol{\varepsilon}_t. \qquad (1)$$

This is an extension of the model proposed by Pfeffermann (1991). The components in STS model (1) can be motivated as follows. In the first component θ_t denotes the unknown population parameter and $\mathbf{1}_{[5]}$ a five dimensional column vector with each element equal to one. This component states that $\hat{\boldsymbol{y}}_t$ contain five GREG estimates for the population parameter in month t. The population parameter is modelled with a so-called basic STM, i.e.

$$\theta_t = L_t + S_t + I_t, \qquad (2)$$

with L_t a time-varying or dynamic trend model for the low frequency variation in the series of the population parameter, S_t a dynamic seasonal model for the monthly effects in the series and I_t a white noise component for the unexplained variation of the population parameter. For L_t the so-called smooth trend model and for S_t the trigonometric seasonal model are used, see Durbin and Koopman 2012, Ch. 3 for details.

The second component in (1), i.e. $\boldsymbol{\lambda}_t$, models the rotation group bias (RGB) induced by the rotating panel design. In this application it is assumed that the first wave is free from RGB and thus gives the most reliable estimates for θ_t, see Van den Brakel and Krieg (2009) for a motivation. The other four components contain random walks, denoted $\lambda_t^{[j]}$ ($j = 2, \dots, 5$), and model the systematic difference between the first wave and the four follow-up waves. The third, fourth and fifth component model the discontinuities in the input series that are the result of three major survey redesigns that took place in 2010, 2012, and 2021 respectively. The sixth component in (1) contains a correction for the loss of CAPI respondents in the first wave during the lockdown of the corona crisis in 2020 and 2021. See Van den Brakel et al. (2022) for details.

The last component in (1) is a time series model for the survey errors that accommodate heteroscedasticity due to e.g. varying sample sizes over time and serial correlation which is a result of the partial sample overlap of the rotating panel design. The sampling errors are stacked in a five dimensional vector $\boldsymbol{\varepsilon}_t = (\varepsilon_t^{[1]}, \varepsilon_t^{[2]}, \varepsilon_t^{[3]}, \varepsilon_t^{[4]}, \varepsilon_t^{[5]})'$. To account for heteroscedasticity, the sampling errors are scaled with the standard errors of the GREG estimates of the input series, i.e. $\varepsilon_t^{[j]} = \sqrt{var(\hat{y}_t^{[j]})}\tilde{\varepsilon}_t^{[j]}$. The standard errors of the GREG estimates are estimated from the survey data. The scaled sampling error for the first wave, i.e. $\tilde{\varepsilon}_t^{[1]}$, is a normally and independently distributed error term that is not correlated with past observations, since the first wave is observed for the first time. The scaled sampling errors of the follow-up waves are modeled with an AR(1) model to accommodate serial correlation with past observations. See Van den Brakel and Krieg (2015) for details.

Model (1) can be expressed in the so-called state space representation. Subsequently the Kalman filter is applied to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2012). The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman et al. (2008).

Population parameters estimated by the time series model are the unemployed labour force, employed labour force and the total labour force. These three parameters are estimated at the national level and a break down in six domains that is based on the cross classification of gender and age in three classes. Parameters of interest are the trend (L_t) and the signal. The latter is defined as the trend plus the seasonal component ($L_t + S_t$).

4 Analytic approximation of the variance of the GREG estimator under measurement error

An analytic approximation for the variance of the GREG estimator that accounts for the additional uncertainty of incorporating components in the weighting scheme that are subject to measurement error is obtained as follows. In a first step the correction term of the regression estimator is split in a term for which the true population totals are known, say $\boldsymbol{t}_a$ (where subscript a stands for administration), and a term for which the true population totals are estimated, say $\tilde{\boldsymbol{t}}_m$ (where subscript m stands for model estimate). This results in:

$$\hat{t}_y^R = \hat{t}_y^\pi + \widehat{\boldsymbol{\beta}}'(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^\pi) = \hat{t}_y^\pi + \widehat{\boldsymbol{\beta}}'_a(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^\pi) + \widehat{\boldsymbol{\beta}}'_m(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^\pi).$$

Here $\hat{\boldsymbol{t}}_a^\pi$ and $\hat{\boldsymbol{t}}_m^\pi$ are the Horvitz-Thompson/Narain estimators for $\boldsymbol{t}_a$ and $\tilde{\boldsymbol{t}}_m$ and $\widehat{\boldsymbol{\beta}}_a$ and $\widehat{\boldsymbol{\beta}}_m$ the corresponding estimates for the regression coefficients. With a first order Taylor approximation it follows that:

$$\hat{t}_y^R \doteq \hat{t}_y^\pi + \boldsymbol{\beta}'_a(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^\pi) + \boldsymbol{\beta}'_m(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^\pi) \equiv \hat{t}_y^{R0}.$$

An expression for the variance of $\hat{t}_y^{R0}$ must account for two sources of variation; sampling error of the sample design of the LFS and the measurement error of the time series model. This is achieved by conditioning on the measurement error of the time series models using the following decomposition:

$$Var(\hat{t}_y^{R0}) = E_m Var_s(\hat{t}_y^{R0}|m) + Var_m E_s(\hat{t}_y^{R0}|m), \quad (3)$$

where E_m and Var_m denote the expectation and variance with respect to the time series model and E_s and Var_s the expectation and variance with respect to the sample design. For the first term in (3) it follows that the variance of the regression estimator, conditionally on the time series model is equal to the variance of the regression estimator treating the population totals obtained with the time series model as fixed known values, i.e.:

$$E_m Var_s(\hat{t}_y^{R0}|m) = E_m Var_s(\hat{t}_y^{R0}) = Var_s(\hat{t}_y^{R0}).$$

For the second term in (3) it follows that:

$$Var_m E_s(\hat{t}_y^{R0}|m) = Var_m E_s(\hat{t}_y^\pi + \boldsymbol{\beta}'_a(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^\pi) + \boldsymbol{\beta}'_m(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^\pi)|m).$$

Taking the expectation with respect to the sample design implies that $E_s \hat{t}_q^\pi = t_q$, (for $q = y, a, m$) such that

$$Var_m E_s(\hat{t}_y^{R0}|m) = Var_m(t_y + \boldsymbol{\beta}'_m(\tilde{\boldsymbol{t}}_m - \boldsymbol{t}_m)).$$

In this step it is assumed that $\tilde{\boldsymbol{t}}_m$ and $\hat{\boldsymbol{t}}_m^\pi$ are uncorrelated. This assumption is not necessarily true since the sample data used to construct $\hat{\boldsymbol{t}}_m^\pi$ for a particular quarter, are also used in the time series model to estimate $\tilde{\boldsymbol{t}}_m$. Since $\tilde{\boldsymbol{t}}_m$ is based on a time series model that uses a long series starting in 2000, it is conjectured that this correlation is negligible. Since t_y and $\boldsymbol{t}_m$ are constant with respect to the time series model it follows that

$$Var_m E_s\left(\hat{t}_y^{R0} \middle| m\right) = \boldsymbol{\beta}'_m Var_m(\tilde{\boldsymbol{t}}_m)\boldsymbol{\beta}_m, \qquad (4)$$

with $Var_m(\tilde{\boldsymbol{t}}_m)$ a diagonal matrix containing the variances of the time series model estimates on the diagonal, which are available from the software used to produce the monthly labour force figures. As a result we have the following variance approximation for the GREG estimator

$$Var\left(\hat{t}_y^{R0}\right) = Var_s\left(\hat{t}_y^{R0}\right) + \boldsymbol{\beta}'_m Var_m(\tilde{\boldsymbol{t}}_m)\boldsymbol{\beta}_m, \qquad (5)$$

with $Var_s\left(\hat{t}_y^{R0}\right)$ being the variance of the regression estimator assuming fixed population totals.

5 References

Doornik, J. (2009). An Object-oriented Matrix Programming Language Ox 6. Timberlake Consultants Press.

Durbin, J., and Koopman, S.J. (2012). Time series analysis by state space methods. Oxford: Oxford University Press.

Koopman, S.J., N. Shephard, and J. Doornik (2008). Ssfpack 3.0: Statistical algorithms for models in state-space form. Timberlake Consultants, Press London.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. Journal of Business & Economic Statistics, 9, pp. 163-175.

Särndal, C-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer Verlag.

Van den Brakel, J.A., S. Krieg (2009). Structural time series modelling of the monthly unemployment rate in a rotating panel. Discussion paper, January 2009. Statistics Netherlands, Heerlen.

Van den Brakel, J.A. (2022). Monthly Labour Force Figures during the 2021 redesign of the Dutch Labour Force Survey. Discussion paper, January 2022. Statistics Netherlands, Heerlen.

Van den Brakel, J.A., S. Krieg (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. Survey Methodology. Vol. 41, pp. 267-296.

Van den Brakel, J.A., M. Souren and S. Krieg (2022). Estimating monthly Labour Force Figures during the COVID-19 pandemic in the Netherlands. Journal of the Royal Statistical Society, Series A. Vol 185, pp 1560-1583.