Small area estimation for key LFS variables

**Keywords:** Small area estimation, Labour force survey (LFS), Bayesian hierarchical modelling, post-stratification, survey unemployment rate (ILO definition).

# Introduction

Statistical office of the republic of Slovenia (SURS) currently publishes annual survey unemployment estimates (ILO definition) on NUTS3 level. These annual estimates are often judged as less reliable (CV>10 %). In the last five years there were 31 estimates designated as less reliable, on average for more than six regions out of twelve per year [1]. Quarterly estimates of survey unemployment are only available on NUTS2 and national level.

The goal of our study was to produce a small area estimation for quarterly and yearly survey unemployment rate on regional level (NUTS3).

# Methods

We use hierarchical Bayesian modelling and post-stratification. We selected hierarchical Bayesian regression due to its useful property of partial pooling [2]. Predictors that carry little information are shrunk towards zero. For example, predictor for age-education group of people with doctoral degree who are over 89 years old (n<5) is shrunk close to zero and the probability of unemployment for these two cases is derived from other variables. This avoids many problems of overfitting that would occur in classical regression. Another welcome property of Bayesian modelling is that sampling from a posterior distribution of model coefficients is a useful way to propagate variance of predictions in post-stratification [2,3]. A drawback of Bayesian regression is its computational intensity [2,3]. We tested models in a classical frequentist framework and fit the final selected model with Stan [4,5].

We use person-level data from the LFS survey, supplemented with register rate of unemployment on the municipal level. We model each quarterly survey separately. We divide age into 5-year categories. Education is classified into twelve categories. We include interaction variable between age and education categories. Geographic variable is administrative unit.

For the purposes of this analysis, we use a working age subset of population. This quantity is a denominator for calculating survey activity rate, which in turn is denominator for calculating survey unemployment rate [6]. In order to estimate survey unemployment rate we first need to estimate the rate of survey activity in population. We model the probability that the person in survey is active:

$$Pr⁡(y\_{i}=active)=logit^{-1}(α\_{j }\*age\_{\left[i\right]}+ α\_{k}\*educ\_{\left[i\right]}+ α\_{l }\*age-educ\_{\left[i\right]}+ α\_{m} adm\\_unit\_{\left[i\right]}+β\_{1}\*gender\_{i}+ β\_{2}\*reg\\_unemp\_{i} + ϵ\_{i}) $$

To estimate the probability that the person in survey is unemployed we use the following model:

$$Pr⁡(y\_{i}=unemployed)=logit^{-1}(α\_{j }\*age\_{\left[i\right]}+ α\_{k}\*educ\_{\left[i\right]}+ α\_{l }\*age-educ\_{\left[i\right]}+ α\_{m} adm\\_unit\_{\left[i\right]}+β\_{1}\*gender\_{i}+ β\_{2}\*reg\\_unemp\_{i} + ϵ\_{i}) $$

We create a post-stratification table from demographic registry, where we count the number of people in Slovenia, stratified by all variables used in the model. We use the model to predict the probability of survey activity and unemployment for each row in post-stratification table. We multiply this probability by number of people and sum by statistical region. We get the size of working age population, size of the labour force and number of unemployed people. From these quantities, we can calculate the number of inactive persons and number of people in employment.

We get the confidence intervals for the estimates by sampling from the posterior distributions of the model parameters and post-stratifying. The process is repeated 200 times. We use mean as a point estimate and 2.5 and 97.5 percentile as a 95 % confidence interval for our estimations.

# Results



Graph 1: comparison of direct estimates (red) and model estimates (black) of rate of survey unemployment by Slovenian statistical regions for year 2019 with 95 % confidence intervals.

Graph 1 shows remarkable improvements in precision estimates for Slovenian regions on annual level. There are two effects in play. A difference in point estimate is caused by post-stratification. The survey is weighted by age, gender and statistical region and is representative for all of the combinations of those variables. Post-stratification includes additional variable education and geographical information on administrative unit level. All possible combinations of these variables amount to more than 60.000 rows of data as opposed to weighting which comprises 192 combinations of data. Post-stratification is much more granular.

The reduced width of the confidence intervals is a consequence of using more information in the estimate. Direct estimator uses only weighted survey unemployment variable for estimation [7]. Regression model uses information provided by supplementary variables, which are highly correlated with estimated independent variable. Direct estimation discards available information while regression model includes it in the estimation. Average span of 95 % confidence interval is 3.3 % for direct regional estimators and 0.17 % for Bayesian hierarchical estimators of unemployment. This amounts to more than 15-fold increase in precision of estimates.

# Conclusions

We successfully applied small area estimation techniques to quarterly and yearly regional estimation of survey unemployment and survey activity rates. Contrary to [3] Bayesian hierarchical estimates proved significantly more precise compared to direct estimates. We ascribe this improvement to the inclusion of rich variety of auxiliary data on personal and area level. Bayesian hierarchical modelling allows for inclusion of many predictors. Inclusion of predictors is somewhat limited only by available computing power. Final models took days to estimate, but results proved to be worth the wait.

 **References**

1. Rutar, Katja, Tomažič, M. (2019). Measures of the activity (%), by statistical regions, Slovenia, annually. Statistical Office of the Republic of Slovenia. Retrieved on 1.10.2020 from <https://pxweb.stat.si/SiStatData/pxweb/en/Data/-/0762115S.px/>
2. Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research). Cambridge University Press.
3. Meindl, B. (2008). Estimating Unemployment-Rates for Small Areas – A Simulation-Based Approach. Austrian Journal of Statistics, 37(3&4), 349–360.
4. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.
5. Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 <https://mc-stan.org/rstanarm>.
6. Rutar, Katja, Tomažič, M. (2020). Labour force survey results. Retrieved 1.10.2020 from Statistical office of the Republic of Slovenia webpage: <https://www.stat.si/statweb/File/DocSysFile/8357/07-008-ME.pdf>
7. D. G. Horvitz & D. J. Thompson (1952) A Generalization of Sampling Without Replacement from a Finite Universe, Journal of the American Statistical Association, 47:260, 663-685, DOI: 10.1080/01621459.1952.10483446
8. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
9. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
10. Rutar, Katja. Personal communication. October 2020.