WEIGHTING IN MODELS OF UNEMPLOYMENT SPELL DURATION

Ivana Malá

Abstract

Weighting is a tool to consider the deviations from simple random sampling, nonresponse, and possible non-representativeness of the sample to (known) population characteristics. Positive weights are assigned to all observations by a calibration process and should be included in the analyses. In the text, the impact of the use of weights on results of the probability distribution fitting to incomplete (censored) data. Data from the Labour Force Sample Survey in the Czech Republic (in years 2010-2018) are used to model the duration of unemployment with the lognormal distribution. The weighted and unweighted (weights provided by the Czech Statistical Office are used) survival models are compared concerning the point and interval estimates of parameters and characteristics of the distribution. Maximum likelihood estimates in the accelerated failure time model are used for right (if an unemployed remains unemployed) and interval-censored (if an unemployed found a job) data. The point estimates are very similar in our study; the standard errors are smaller for the weighted model as the weights reflect a number of similar population members.

Key words: weights, censored data, unemployment, maximum likelihood

JEL Code: J64, C83

Introduction

Statistical methods are usually based on an assumption of a random sample from a probability distribution of interest. This assumption is crucial, and its violation can cause problems with the quality and interpretability of results. If the data are obtained by sampling from a finite population, only a simple sample with equal probabilities of appearance with replacement fulfils this assumption. If we sample a small part from a large population (usually up to 5%), sampling without replacement can be applied with assumed properties of the same distribution and independence (Thompson, 2012). In practice, more complicated sample schemes are used for different reasons – expenses (given budget), time, missing list of units in the target population, or various technical limitations. Therefore, post-stratification is usually applied to

assign survey weights to the observations. The individual data from sophisticated and complex surveys are usually supplemented with calibrated weights to be used by users and analysts. The weights are frequently applied not only in the estimation of population characteristics as a mean or variance but in regression models and other procedures (Gelman, 2007, Lohr, 2007), especially for maximum likelihood estimates.

The Labor force sample survey (LFSS, 2022) is performed quarterly by the Czech Statistical Office (CZSO) and weights are provided in the datasets at the level of households (sampling units) and individual respondents. The impact of weighting in the official statistics survey EU-SILC (European Union – Statistics on Income and Living Conditions, EU SILC, 2022) is studied in (Bartošová, Bína, 2010). In the more general SHARE data (Survey of Health, Ageing and Retirement in Europe, SHARE, 2022), a large spectrum of weights is provided and the weights should be selected according to the purpose of the analysis. The survey units are households, for this reason, household weights are given. Only the household members above 50 years of age are eligible for the survey (and their partners). If the analysis concerns individuals, weights on an individual level should be applied. And the survey is organised in waves and for panel data type analyses panel weights are provided.

In addition to the availability of prepared (calibrated) weights, frequently used statistical software allows all users to incorporate weights in the model. For example, the R programming language (R CORE TEAM, 2020, used in this contribution for all computations) is highly flexible in using weights in the modelling from basic to highly sophisticated models and procedures. The weights can be added also if more user-friendly packages for the analysts and researchers (SAS, SPSS, STATA, Excel, …) are used.

Survival analysis provides statistical models for time-to-event data. In the text, the accelerated time model (referred to as AFT) is applied to fit the lognormal distribution to the sample values of the duration of unemployment spell from the LFSS survey for the years 2010–2018. For the model, we need a skewed probability distribution with a hazard function with a maximum. We chose the lognormal distribution as in (Malá, Čabla, 2022), having considered Čabla, Malá (2017), where loglogistic and Weibull distributions for the unemployment spell duration are also discussed in connection with this application. Nonparametric Kaplan-Meier and Weibull parametric models to model the length of unemployment spell are used in Grogan and van den Berg (2001) for the data from Russian Longitudinal Monitoring. Kupets in (Kupets, 2009) uses individual data from the Ukrainian Longitudinal Monitoring Survey and estimates a discrete time-independent competing risks model with gamma distribution to describe random

errors. For Czech data, the loglogistic and lognormal distributions give similar results (according to the AIC criterion) and are superior to a gamma distribution.

In the present paper, the 2010–2018 LFSS data are analysed and the models fitted based on both unweighted and weighted models are compared. Individual data on the unemployed in the LFSS samples were used. We take into account not only information on a given quarter but also on the previous quarter and the next quarter. We estimate each quarter separately (not using the rotating panel structure of the survey), but we use at least information from previous and consecutive quarters. This approach allows us to obtain the most detailed information and find more respondents who found a job (event occurrence).

1 Data and methods

The LFSS survey is organised as a continuous, quarterly rotating panel. All selected households are included for five consecutive quarters (covering one year) and every quarter the fifth of the selected households is substituted. We use only the information on whether a respondent above the age of 16 is employed, unemployed or nonactive (for Figure 1 and some comments also gender of the unemployed). The exact duration of an unemployment spell is not recorded, respondents report their unemployment duration intervals only in intervals 0-1, 1-3, 3-6, 6-2, 12-18, 18-24, 24-48 and over 48 months. It follows from the origin of the data that they are all censored (incomplete), and no exact durations are given. For unemployed people who have found a job, the analysed unemployment spell is taken in the model as interval-censored; for those who have not, the duration is considered right-censored.

Suppose T is a positive time-to-event variable time to reemployment in the month. The model fits the two-parametric lognormal distribution $LN(\mu;\sigma^2)$ to the data using the maximum likelihood estimation (MLE) in the model

$$\ln T = \mu + \sigma\varepsilon,$$

where ε is a standard Gaussian distribution. For a given quarter and $i = 1, 2, ..., n$ (where n is a sample size of this quarter) we suppose interval-censored data in (ll_i, ul_i), right-censored data in (t_i, ∞).

Then for an unweighted model, we maximise the likelihood function (loglikelihood function $\ln L$)

$$L(\mu,\sigma) = \prod_{i \text{ interval-censored}} [F(ul_i;\mu,\sigma) - F(ll_i;\mu,\sigma)] \prod_{i \text{ right-censored}} [1 - F(t_i;\mu,\sigma)], \qquad (1)$$

for a weighted model we maximise function

$$L(\mu,\sigma) = \prod_{i \text{ interval-censored}} [F(ul_i;\mu,\sigma) - F(ll_i;\mu,\sigma)]^{w_i} \prod_{i \text{ right-censored}} [1 - F(t_i;\mu,\sigma)]^{w_i}, \quad (2)$$

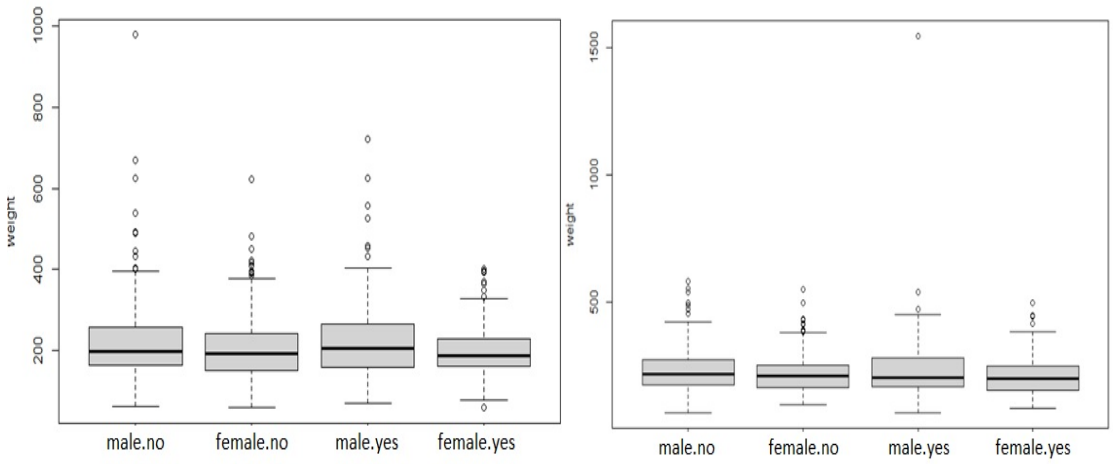
where w_i denotes a weight assigned to an *i*-th respondent and *F* is a cumulative distribution function of a Gaussian distribution. The estimates are asymptotically normally distributed, in figures we use asymptotical Gaussian distribution to construct confidence limits. It follows from (2) that the loglikelihood function equals the weighted mean of logarithms of the expressions based on the distribution of unemployment duration. Based on the lognormal model, the median duration is independent of the parameter σ (is given as e^{μ}), the expected value or variance depending on both parameters.

Taking data from the years 2010–2018, we obtain time series of estimated pairs of parameters and derived characteristics of distributions of the length 36 (9 years, 4 quarters). The individual data were accessed in the SafeCentre of the Czech Statistical Office (SaveCenter, 2022). The program R and the package Survival (Therneau, 2022) is used to fit models and estimate parameters in both – unweighted and weighted approaches for datasets including incomplete data. The time series are smoothed by applying a moving average with a length of four time points covering one year. Our data contains from 654 to 2,496 (sample sizes *n* for particular quarters, *i* in (1) or (2) 1, 2, …, *n*) unemployed respondents in the sample in the analysed quarters with the ratios of new jobs from 0.19 to 0.36.

2 Results

To show the values of weights, two years and the second quarters were selected. The boxplots of weights for the second quarter of 2015 (left) and 2018 (right) are given in Figure 1 depending on gender of respondents (689 male, 293 resp., 781 female, 366 resp.) and whether a new job was found (yes 411, 218 resp., no 1059, 441 resp.). Different scale on the vertical axis is used in both parts, due to one outlier in 2018. The medians of weights are similar for both years and no relationship between gender and the existence of a new job, the values are positively skewed with very rare observations with extremely large values of weights. Similar figures can be obtained for all 36 analysed quarters.

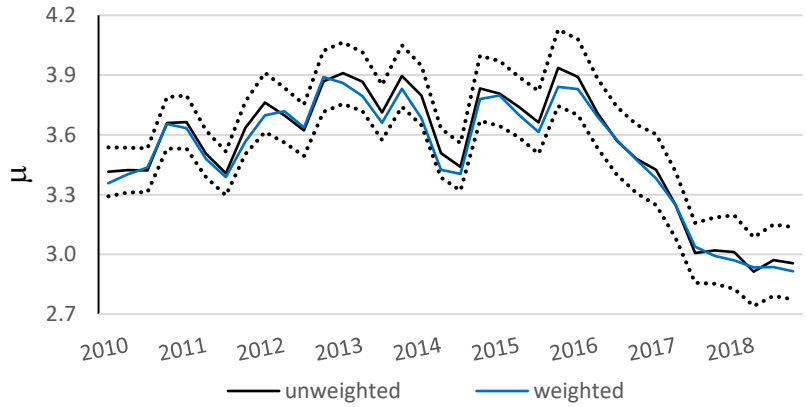
Fig. 1: Boxplots of weights for 2015 (left) and 2018 (right), groups by gender (male/female) and a new job yes/no



Source: own computations

Figures 2 and 3 display the estimated parameters based on formulas (1) and (2). Comparing the weighted (in blue) and unweighted fit (in black, asymptotical confidence limits dotted curves). Point estimates of parameters are very close in values. The estimated parameters differ only in the range of 10 percentage points, the differences are greater for the parameter for the variance of a Gaussian distribution (Figure 3). There is a decline in the estimated parameter μ in 2017 and 2018; it implies decreasing median duration (see Figure 4).

Fig. 2: Asymptotic 95% confidence intervals for parameters μ

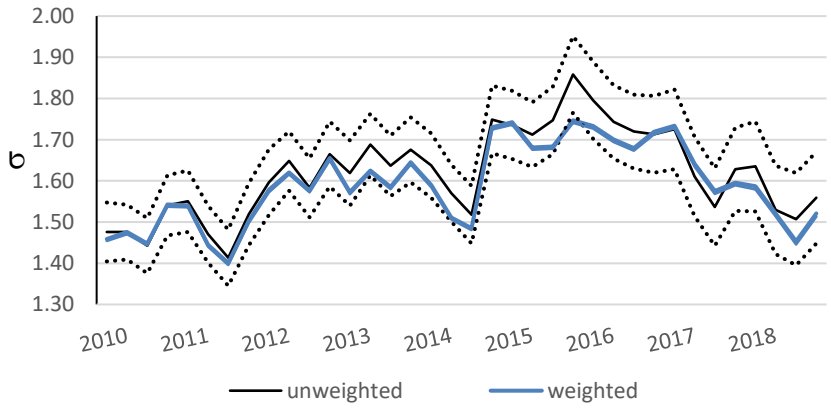


Source: own computations

The point estimates from the weighted model are included in the region given by confidence limits. If weights are used, the accuracy of the estimates (quantified by the standard deviation) is significantly higher, as the weights indicate that the model is fitted (to the whole population of the unemployed in this case) using a much larger number of observations.

Asymptotical limits would be almost unvisible in Figures 2 and 3, for this reason, only point estimates are given.

Fig. 3: Asymptotic 95% confidence intervals for parameters σ

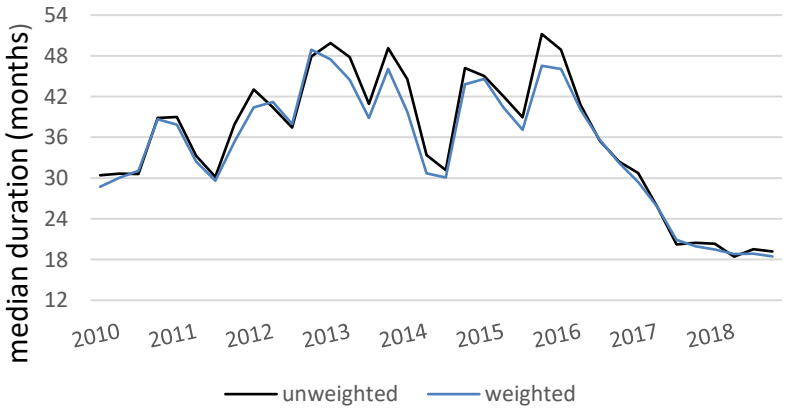


Source: own computations

In Figure 4, estimated medians of the analysed duration are given. The ratio of medians for a particular quarter is given as

$$\frac{e^{median\ unweighted}}{e^{median\ weighted}} = e^{median\ unweighted - median\ weighted}.$$

Fig. 4: Estimated median duration of the unemployment spell



Source: own computations

These values vary from -3% to 12%; the difference in estimated median duration (in months) takes values from -.8 to 4.8 months. Values from the unweighted model tend to be higher than in the weighted model. In survival analysis, usually medians are preferred to expected values

because of heavy tails of distributions and censoring. For this reason, we show only medians in Figure 4. Ratios of expected values (both parameters are included) are from -9% to 34%.

Conclusion

Using weights is a commonly used tool to consider the deviations from simple random sampling and possible non-representativeness of the sample with respect to population characteristics. After completing a survey, calculation and calibration of sampling weights become an integral part of the data preparation for analysis. For comprehensive sample surveys with a large spectrum of users from different backgrounds and approaches to research problems and data analysis, usually calibrated weights are included and provided to be used in analyses.

The maximum likelihood estimates of parameters are similar for both the weighted and unweighted models, the estimate accuracy of the former ones being much higher. Weights reflect a sample drawn from a large population study, the respondents in the presented model representing the Czech Republic's 16+ population. The weight values and distributions for all unemployed respondents – men and women, successful and unsuccessful jobseekers – were explored and presented.

The findings suggest that the use of weights results in a significant reduction in the variability of estimates, the effect on point estimates being small for the models analysed. We do not estimate population characteristics based on plug-in sample counterparts (where the use of weights is crucial), but we use a probability model that included incomplete data. For this reason, the use of weights causes similar point estimates and too small and unrealistic errors.

Accommodation of weights is highly recommended for results from the population but causes changes in variability. They increase variance when using weighted data for direct evaluation of sample characteristics (for example means), but decrease it in maximum likelihood estimation.

Acknowledgment

Access to the individual Labour Force Sample Survey data in the SafeCentre of the Czech Statistical Office is gratefully acknowledged.

References

Bartošová, J., Bína, V. (2010). Influence of the Calibration Weights on Results Obtained from Czech SILC Data. In: *COMPSTAT 2010*. Paris, 22.08.2010 – 27.08.2010. Paris: Physica-Verlag, pp. 753–760.

Čabla, A., Malá, I. (2017). Modelling of Unemployment Duration in the Czech Republic. *Prague economic papers* [online], 26(4), 438–449.

EUROSTAT WEIGHTS (2020). Quality report of the European Union Labour Force Survey 2018. https://ec.europa.eu/eurostat/documents/7870049/10381077/KS-FT-20-001-EN-N.pdf/9945a36a-4166-eae6-47a6-7153346915de.

EU SILC. [online]. [Accessed 05.05.2022] https://www.czso.cz/csu/czso/living-conditions-eu-silc-methodology.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science,* 22(2), pp. 153–164.

Grogan, L. - van den Berg, G. J. (2001). The Duration of Unemployment in Russia. *Journal of Population Economics*, 14, pp. 549-568.

Kupets, O. (2006): Determinants of unemployment duration in Ukraine. *Journal of Comparative Economics*, 34, pp. 228-247.

LFSS. Labour Force Sample Survey (LFSS). [online]. [Accessed 20.05.2022] http://www.czso.cz/eng/redakce.nsf/i/lfs_analyses_in_news_releases.

Lohr, S. L. (2007). Comment: Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, 22(2), pp. 175-178.

Malá, I., Čabla, A. (2022). Modelling of the unemployment Duration in the Czech Republic based on aggregated complete and individual censored data. *Ekonomický časopis* [online], 70(2), pp. 171–187.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review*, 61(2), pp. 317-337.

R CORE TEAM (2020). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. URL: http://www.r-project.org/.

SAFECENTER (2022). Statistical Data for Scientific Research Purposes. https://www.czso.cz/csu/czso/statistical_data_for_scientific_research_purposes

SHARE. SHARE Release Guide 8.0.0. [online]. [Accessed 10.03.2022] http://www.share-project.org/data-documentation/waves-overview/wave-8. html.

Thernau, T. (2022). A Package for Survival Analysis in R. R package version 3.3-1. https://CRAN.R-project.org/package=survival.

Thompson, S. K. (2012). *Sampling*. Wiley, Hoboken, New Jersey.

Contact

Ivana Malá

Prague University of Economics and Business

W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov

Czech Republic

malai@vse.cz