**European Establishment Statistics Workshop**
(Nuremberg, Germany, 9-11 September 2013)

# CONSTRUCTION OF FULL TIME EQUIVALENTS FOR THE REGISTER-BASED SWISS STRUCTURAL BUSINESS STATISTICS

Prepared by Desislava Nedyalkova (Desislava.Nedyalkova@bfs.admin.ch) and Daniel Assoulin (Daniel.Assoulin@bfs.admin.ch), Swiss Federal Statistical Office

## I. INTRODUCTION

In the past, a periodical Business Census (BC) played an important role for producing various statistics on the structure of the Swiss economy. The BC was held for the last time in 2008. For the reference year 2011 it will be replaced by the integrated system SWIS (Swiss Structural Business Statistics). SWIS will be mainly based on the business register (BR), the social security register (SR) and complementary surveys like the Quarterly Survey of Employment (JobStat). Social security data providing information about gender, employment and wages (annual salary) at employee level is linked to the business register at the enterprise level. Therefore, it is the main source of information about employment at the enterprise level while the business register provides information like NACE-code, legal form or enterprise structure (multi-establishments).
Full-time equivalent (FTE) of employment by gender is an important target variable which in the past could be derived from the information in the BC. In the social security register the FTE is not directly available. For this reason the construction of such a variable plays an important role for SWIS.
This document presents how full-time equivalents (by gender) will be constructed using a model based on data coming from the Quarterly Survey of Employment (JobStat) matched with the SR. JobStat is the survey, which is the closest in structure to the BC. It gives total employment and FTE by gender for approximately 31'000 enterprises. Matched data contains also information on the economic activity, geographical location (NUTS 2) and wages.
The chosen explanatory and dependent variables will be discussed in relation to the aim of estimating FTE for the enterprises which do not belong to the JobStat sample and for which the predictions have to be based on employment data coming from the SR only.

## II. Variable of interest

First we consider an enterprise, for which employment data from both sources, JobStat and SR, is available. The JobStat data is coming from Jobstat survey (Renaud, A. 2008; Renaud, A., Panchard, C. and Potterat, J. 2008), quarter 4 with reference month December 2011. The considered SR data contains information concerning employees which received a salary in December. Our focus is on the following variables by gender:

- BETOT_SR: the number of employees according the SR.
- BETOT_JS: number of employees according the JobStat survey in the reference period
- FTE_JS: the number of FTE's per enterprise according the JobStat survey

Then the variable FTE_SR is defined as follows :

$$\text{FTE\_SR} = \frac{\text{FTE\_JS}}{\text{BETOT\_JS}} \times \text{BETOT\_SR}, \tag{1}$$

Hence, FTE_SR is equal to FTE_JS, but adapted for differences between the number of employees according to JobStat (BETOT_JS) and the number of employees according to SR (BETOT_SR). Possible explanations for these differences:

- Measurement errors in either of the sources (JobStat or SR).
- Differences in definitions with regard to retained employees.

Further investigations regarding this point are planned. For the time being, (1) ensures the definition of a FTE that is coherent with the number of employees according to SR. It is important to note, that FTE_SR according to (1) can only be calculated for enterprises, where beside the SR data, information from JobStat is available (i.e. the respondents of JobStat for the considered period). For the other units [1] it is planned to estimate FTE_SR (by gender) using the model that will be presented and discussed in the coming sections.

## III. Linear Regression Model

A separate model was estimated by gender and economic sector (second, third). Hence, four models are used for the estimation of full-time equivalents. As in all the four cases the same approach was used we will not discuss these models separately. The development of the model that was finally chosen is based on the number of employees in four different salary classes. It has the following form:

$$y_i = \sum_{j=1}^{4} \beta_{jkl} V_{ij} + \epsilon_i, \tag{2}$$

where:

- $y_i$ = FTE_SR according to (1),
- $V_{ij}$, number of employees of enterprise $i$ in salary class ($j = 1, ..., 4$),
- $\beta_{jk\ell}$, regression coefficient for the number of employees in salary class $j = 1, ..., 4$, in NUTS2 $k$ ($k = 1, ..., 7$) and NACE (rev.2) section $\ell$,
- $\epsilon_i$, residual with E($\epsilon_i$) $= 0$ and Var($\epsilon_i$) $= \sigma^2$BETOT_SR$_i$.

Discussion about heteroscedasticity follows in Subsection A. The estimation of the regression parameters is performed by Weighted Least Squares. The weight for enterprise $i$ is given by $w_{1,i} = w_i/$BETOT_SR$_i$, where $w_i$ is the extrapolation weight of enterprise $i$ according to JobStat and the factor $1/$BETOT_SR$_i$ takes into account the assumed heteroscedasticity. The inconvenience of this model is that there is a large number of parameters to be estimated. For example for the economic sector 3 our model will have a parameter for each combination of salary class (4) with NUTS 2 (7) and activity section (12) or in total more than 300 parameters. This led to the desire for a more parsimonious model discussed in Section B below.

### A. Some details concerning the explanatory variables of the model

In order to reflect the best the situation of the economy in December 2011, the calculation of an employee's standardized wage in the register was necessary. For example, for an employee that has worked only during the 5 last months of the year (including December), the register reports an annual wage bill of 25 000 CHF. We suppose that his monthly salary in December is 5 000 CHF (the same each working month of the year), which gives an estimated standardized annual wage bill of 60 000 CHF. In this way, we construct the annual wage distribution by gender.

[1] Including units with BETOT_JS $= 0$ and BETOT_SR $> 0$ for which (1) is obviously not defined

The salary classes (by gender) are constructed in the following way: Based on JobStat we estimate for each activity sector OFS50[2] the proportions of part time III (degree of employment less than 15%), part time II (15%-49%), part time I (50%-89%) and fulltime (>90%) employees. Then we calculate quantiles in the distribution of standardized annual salaries according to SR that correspond to the cumulative proportions. Based on these quantiles we define in each OFS50 four salary classes. (In some cases the salary classes had to be adapted to avoid classes with just a few employees). Let $I_1$ denote the variable which for each employee takes the value 1 if an employee is in the first salary class and 0 otherwise. In the same manner, we construct the variables $I_2, I_3$ and $I_4$. Then the variable number of employees per salary class for enterprise $i$ is defined as follows: $V_{ij} = \sum_{k \in E_i} 1(k \in I_j)$, where $E_i$ is the set of employees of an enterprise $i$ and $j = 1, ..., 4$ is the subscript for the class. Note that for an enterprise $i$ , BETOT_SR is equal to the sum of the $V_{ij}$ . This gives also a motivation to assume that the variance of the residuals in (2) is proportional to BETOT_SR.

## B. Model improvement and realization

As mentioned before, model (2) has a large number of parameters. Starting with the model

$$y_i = \sum_{j=1}^{4} \alpha_j V_{ij} + \sum_{j=1}^{4} \beta_{jk\ell} V_{ij} + \epsilon_i, \tag{3}$$

we applied a stepwise selection based on Akaike Information Criterion (AIC) for choosing the best model. This is done through the GLMSELECT procedure, selection criteria=stepwise, in SAS (Cohen 2009). It is important to note that, as a result, for each gender and economic sector the term $\sum_{j=1}^{4} \alpha_j V_{ij}$ is selected for the model. Then for certain NUTS2, $k$, and NACE sections, $\ell$, a specific parameter $\beta_{jk\ell}$ may be added in order to adjust the parameter $\alpha_j$ for the salary class $j$.

B.1. *Weights used in model selection.* The matched data coming from the JobStat survey and the register contains the extrapolation weights, $w_i$, for the single-establishment enterprises (EUNT) which are used in the model. As described above, these weights were adjusted in order to account for the heteroscedasticity present in the model. Another issue with the model was how to reduce the effect of extreme values on model selection and parameter estimation. Thus a robustification procedure for the model was needed. We have chosen not to eliminate observations from the model for the estimation, but rather adjust the weights. We have applied the SAS ROBUSTREG procedure (Chen 2002) on a simple model using only $\sum_{j=1}^{4} \alpha_j V_{ij}$ as predictions. The chosen technique for estimating the regression coefficients is LTS estimation (Rousseeuw 1984). Then we obtain estimates for the scale parameter $\sigma$ and residuals $\epsilon_i^*$. By choosing an appropriate tuning constant $c$, we define a correction factor for the weights $w_{1,i}$ of the model as follows:

$$u_i = \begin{cases} \frac{c\sigma\sqrt{\text{BETOT\_SR}_i}}{|\epsilon_i^*|} & \text{if } |\epsilon_i^*| > c\sigma\sqrt{\text{BETOT\_SR}_i} \\ 1 & \text{otherwise} \end{cases}$$

Taking into account the number of robustified enterprises the tuning constant was set to $c = 3$, respectively to $c = 5$, for the men of sector 3. The new weights $w_i^* = u_i \times w_{1,i}$ are used to select and fit the final model by means of the GLMSELECT procedure.

[2] In Switzerland there is a standard of aggregations of the economic divisions called OFS50

TABLE 1. *reldif* by NACE OFS50 for economic sector 2

| NACE | extrapolated prediction errors | | extrapolated FTE | | *reldif*en % | |
|---|---|---|---|---|---|---|
| OFS 50 | Men | Women | Men | Women | Men | Women |
| total | -788.07 | -1'064.31 | 783'950.57 | 192'753.37 | -0.10 | -0.55 |
| 5.9 | 0.58 | -1.84 | 3'798.63 | 383.05 | 0.02 | -0.48 |
| 10.2 | -206.96 | -158.68 | 41'678.96 | 23'301.49 | -0.50 | -0.68 |
| 13.5 | -11.57 | 100.58 | 6'613.57 | 9'168.60 | -0.17 | 1.10 |
| 16.8 | 67.85 | -219.64 | 53'586.01 | 13'883.82 | 0.13 | -1.58 |
| 19.2 | -9.98 | -13.05 | 23'443.97 | 7'119.84 | -0.04 | -0.18 |
| 21 | 0.06 | 0.06 | 22'065.98 | 14'028.48 | 0.00 | 0.00 |
| 22.3 | -28.66 | -27.23 | 31'705.99 | 8'126.38 | -0.09 | -0.34 |
| 24.5 | -94.28 | -215.34 | 78'052.09 | 15'113.61 | -0.12 | -1.42 |
| 26 | -33.00 | 25.26 | 65'804.96 | 37'872.16 | -0.05 | 0.07 |
| 27 | -24.14 | -37.11 | 26'678.77 | 9'220.19 | -0.09 | -0.40 |
| 28 | -11.09 | -108.49 | 67'072.31 | 11'080.71 | -0.02 | -0.98 |
| 29.3 | 4.31 | -8.93 | 13'608.50 | 1'743.78 | 0.03 | -0.51 |
| 31.3 | -165.12 | -168.63 | 36'601.51 | 12'146.29 | -0.45 | -1.39 |
| 35 | -7.22 | -5.98 | 21'304.91 | 4'116.98 | -0.03 | -0.15 |
| 36.9 | -69.25 | 0.38 | 11'628.31 | 1'429.56 | -0.60 | 0.03 |
| 41.2 | 8.56 | -88.15 | 96'753.14 | 5'454.28 | 0.01 | -1.62 |
| 43 | -208.15 | -137.49 | 183'552.95 | 18'564.15 | -0.11 | -0.74 |

## IV. **Evaluation of the model**

For model (2), the four $R^2$ vary between $0.94$ and $0.99$ (better prediction for Industry than for Services). Beside of $R^2$ different measures have been used in the process of comparing different models and assessing their performance on different aggregation levels. Among these measures were the the weighted mean of absolute prediction errors and a quality measure based on estimated relative prediction errors for totals, *reldif*. More precisely, *reldif* is the ratio between

- the extrapolated prediction errors (residuals) when estimating FTE_SR by our model
- and the extrapolated FTE_SR,

where extrapolation is performed at the desired level of aggregation using appropriate weights. Table 1 gives an example where *reldif* is applied on OFS 50 level in economic sector 2. This measure of quality takes into account the values of the extrapolated FTE not only for the data used in the model, but globally all data from the JobStat survey including the multi-establishment enterprises (MUNT).

## V. **CONCLUSIONS**

With regard to the variable of interest:

- Full-time equivalent is not contained as such in the Social Security Register data which is the main source for employment information in SWIS. Therefore, this variable has to be constructed using information about full-time equivalent on enterprise level from survey data (JobStat),
- The definition of full-time equivalents has to take into account differences between employment information coming from JobStat and SR, respectively. Some further investigations regarding these differences are planned.

With regard to enterprises not contained in JobStat:

- Full-time equivalents must be estimated using an appropriate technique.

- The considered approach based on a linear regression predicting the full-time equivalent of a firm by the number of employees in four different salary classes:
  - is relatively easy to interpret and to understand,
  - has to be based on carefully chosen salary classes which reflect different degrees of employment,
  - allows for separate slopes for NUTS2 and NACE (Rev. 2) sections where this improves the model in terms of AIC,
  - should be relatively stable over time (parsimonious model by model selection procedure, robustification by reducing the weights of potential leverage points),
  - leads to satisfactory prediction accuracy assessed by considered quality measures.
- Quality measures proved to be a very important tool when assessing and comparing the performance of different models for predicting full-time equivalents of employment.

## References

Chen, C. (2002). Robust regression and outlier detection with the robustreg procedure. In *Proceedings of the SUGI conference*.

Cohen, R. A. (2009). Introducing the glmselct procedure for model selection. In *Proceedings of the SUGI conference*.

Renaud, A. (2008). Statistique de l'emploi. révision 2007 : cadre de sondage et échantillonage. Technical report, Swiss Federal Statistical Office.

Renaud, A., Panchard, C. and Potterat, J. (2008). Statistique de l'emploi. révision 2007 : méthodes d'estimation. Technical report, Swiss Federal Statistical Office.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*, 871–880.