Working Papers

03/2014

Additional questions to better measure the self-declared professional status and how to link the mismatches produced in previous series through an econometric model.

Javier Orche Galindo

Miguel Ángel García Martínez

First draft: February 2014

This draft: February 2014

Additional questions to better measure the self-declared professional status and how to link the mismatches produced in previous series through an econometric model.

Abstract

From 2009 onwards, it was decided to include in the Spanish LFS questionnaire some additional questions for workers who self-declared being members of cooperatives, unpaid family workers or self-employed so that the professional status was better measured. Since then, the previously observed mismatch upward in the level on the total number of self-employed workers was almost completely adjusted.

In the new data on professional status, it was also distinguished which of them had changed from self-employment to wage employment due to the supplementary questions. Therefore, after several quarters, it was possible to fit the change in professional status through an econometric model and a set of significant explanatory variables obtained from the rest of the questionnaire. Finally, we managed to get a good enough model and could be able to set down in the self-employed 2005-2008 series and the corresponding rise (by the same amount) in the wage employment series.

Keywords

Labour Force Survey, professional status, self-employment, backcasting, logistic model, imputation, goodness of logistic models.

Authors and Affiliations

Javier Orche Galindo

Miguel Ángel García Martínez

National Statistics Institute of Spain

Additional questions to better measure the self-declared *professional status* and how to link the mismatches produced in previous series through an econometric model.

Javier Orche Galindo
Miguel Ángel García Martínez
National Statistics Institute of Spain

Abstract

From 2009 onwards, it was decided to include in the Spanish LFS questionnaire some additional questions for workers who self-declared being members of cooperatives, unpaid family workers or self-employed so that the *professional status* was better measured. Since then, the previously observed mismatch upward in the level on the total number of self-employed workers was almost completely adjusted.

In the new data on *professional status*, it was also distinguished which of them had changed from self-employment to wage employment due to the supplementary questions. Therefore, after several quarters, it was possible to fit the change in *professional status* through an econometric model and a set of significant explanatory variables obtained from the rest of the questionnaire. Finally, we managed to get a good enough model and could be able to set down in the self-employed 2005-2008 series and the corresponding rise (by the same amount) in the wage employment series.

1 Introduction

Within the Spanish Labour Force Survey (LFS), the ***professional status*** variable provides information to determine the status in employment for the main job, according to the resolution concerning the International Classification of Status in Employment (ICSE), adopted by the Fifteenth International Conference of Labour Statisticians (January 1993).

The mentioned variable has, in the case of Spain, the following classification codes for the **self-employed workers** :

Code 01: Employer

Code 03: Business person without wage earners, or independent worker

Code 05: Member of a cooperative

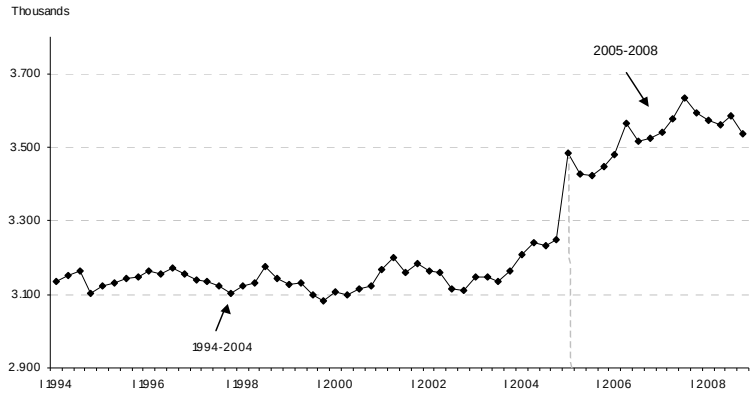
Code 06: Family worker

The following section explains the mismatch in the Spanish LFS self-employed series and the need for readjustment. Section 3 describes the model of readjustment, the significat explanatory variables and interactions observed. Section 4 evalues the goodness of the final fitted model. Section 5 shows the practical aplication to the readjustment of the 2005-2008 *professional status* LFS series. Finally, section 6 gives guidelines over the final conclusions.

2 Additional questions to better measure the *professional status*

As seen in Figure 1, there is a **mismatch** in the series level for total self-employed persons from the fourth quarter of 2004 to first quarter of 2005. The self-employed group is composed in the 2005-2008 period as follows: employers (code 01) with a 30.7% average, the self-employed (code 03) by 59.9%, to members of cooperatives with a 2.3% (code 05) and family workers by 7.1% (code 06).

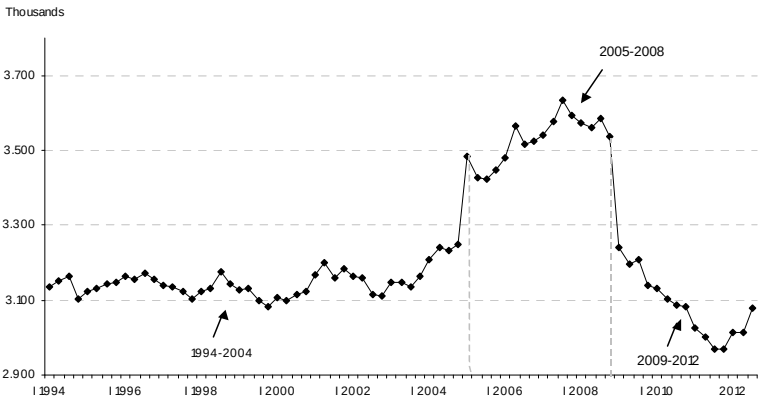
Figure 1. Self-employed workers: total



In 2009 and after analyzing the problems obtaining this variable over 2005-2008 it was decided to include some **additional questions** in the case of workers who self-reported being members of cooperatives, family support or self-employed for clearing the nature of employment in the workplace.

Figure 2 shows **the effect of this measure** on the total number of self-employed persons before and after the inclusion of these questions in 2009. As shown therein, the leap in the level that occurred in the first quarter of 2005 seems completely reversed from the first quarter of 2009.

Figure 2. Self-employed workers: total



Thus, from the above results, it seems appropriate to **adjust downward** the series of independent workers, the members of cooperatives and family workers (codes 03, 05 and 06) for the period 2005-2008. These adjustments produce, as a consequence, a fall in the total number of self-employed and an increase (of the same amount) of the series of private sector employees (code 08). The total employment remains unchanged, since it is a mere redistribution of the same between different professional situations.

3 How to link the mismatches produced in series

The Spanish LFS quarterly series (from 2009 onwards) collected the *professional status* classification in respect to the original series and fitted with new questions incorporated in that year. Therefore, we can forecast using an **econometric model**, the lower self-employed level and the resulting increase in the same amount of employees in the quarterly series from 2005 to 2008.

To fit the model it is used a "**binary logit model**", where the response variable (SITUVAR) takes on only two possible values 0 (holding the same classification of self-employed with new questions incorporated) and 1 (if varies becoming employee and, it is necessary to assess which LFS variables are statistically significant in that classification change in *professional status* and possible interactions between them for that change.

The **input data** are obtained for the series since the 1st quarter of 2009 to the 3rd quarter of 2012 (latest data available when the model was adjusted) of the group of people prior to the inclusion of new questions that they would have been classified in 03, 05 or 06 codes. Of this group, those which does not change its code with the inclusion of new questions will have a value of 0 in the response variable and those which change its code to 08 will have a value of 1.

Operationally, it is used the **logistic SAS procedure** with a logit response function for modelling of the probabilities of change and the stepwise method to select the explanatory LFS variables (and interactions) which are statistically significant at a level of at least 5% (this means that a significance level of 0.05 is required to allow a variable to enter into the model, and also a significance level of 0.05 is required for a variable to stay in the model).

The resulting **significant variables** and **interactions** observed are the following (in this order of importance, acording to the likelihood criterion):

1. Self-declared *professional status*. People without the inclusion of new questions, would be classified in any of the codes 03, 05 or 06.

2. Occupation. Major groups of occupation to one digit of the ISCO-88.

3. Interaction between self-declared *professional status* and occupation.

4. Supervisory responsibilities. Question of the annual subsample, i.e. only applies to one sixth of the total sample but has great explanatory value of real *professional status*.

5. Seniority. Time in months derived from the date in which begins working in the company. It was observed that with increased seniority is more likely to be truly a self-employed person (or less probability of change to employee).

6. Activity. Sections of the NACE Rev 1.1 (to 1 digit, with some regrouping).

7. Interaction between self-declared *professional status* and supervisory responsibilities.

8. Age. This variable is also a quantitative explanation of the professional situation such that the older they are more likely to be truly self-employed person (or less likely to change).

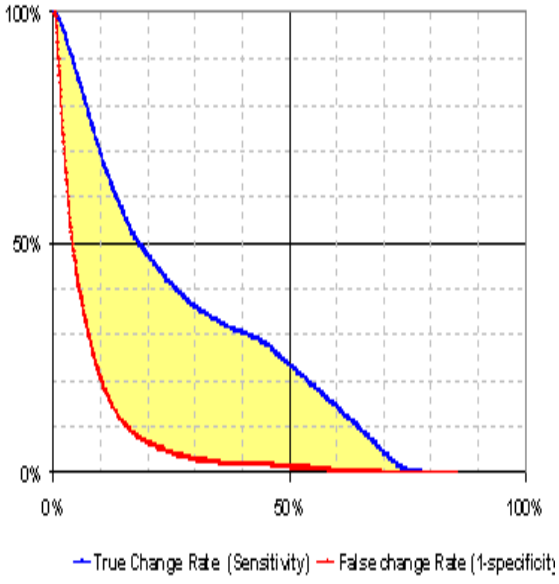
9. Interaction between self-declared *professional status* and seniority.

10. Region of residence. Autonomous communities (NUTS-2).

4 Evaluation of the goodness of fitted model

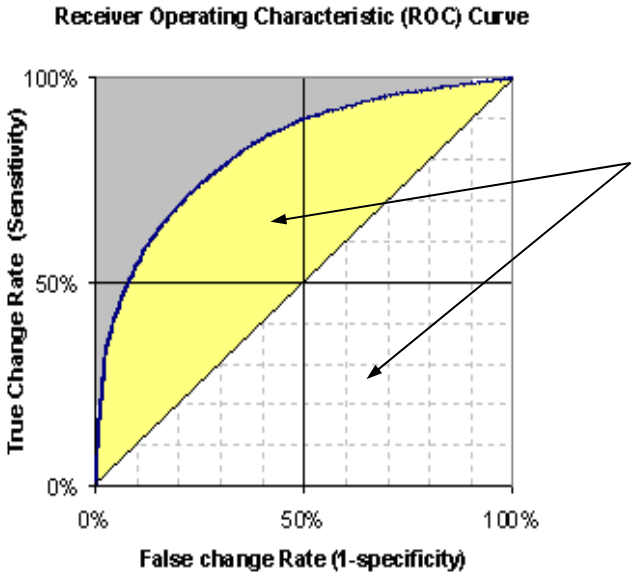
For the evaluation of the goodness of fitted model are calculated measures of association between predicted probabilities and observed responses in the period 2009-2012, yielding a **concordance rate** of 82.2% between predicted and observed values.

The following graph illustrates the **performance of the fitted model** when the discrimination threshold is varied (horizontal axis) of the 'true' and 'false' changes (and 'true' and 'false' no-changes) in the professional status (in vertical axis).



The chart above shows the percentage of 'predicted and actual' changes out of the actual changes (or **sensitivity**, the blue line) and the fraction of 'predicted change but no-actual changes' out of the actual no-changes (or **one minus the specificity**, the red line), varying the cutpoint in the horizontal axis we find in the vertical axis the estimated event probabilities to predict the event.

A **ROC curve** is created below by plotting the fraction of 'predicted an actual changes' out of the actual changes (sensitivity) versus the fraction of 'predicted change but actual no-changes' out of the actual no-changes (one minus the specificity), at various threshold settings.

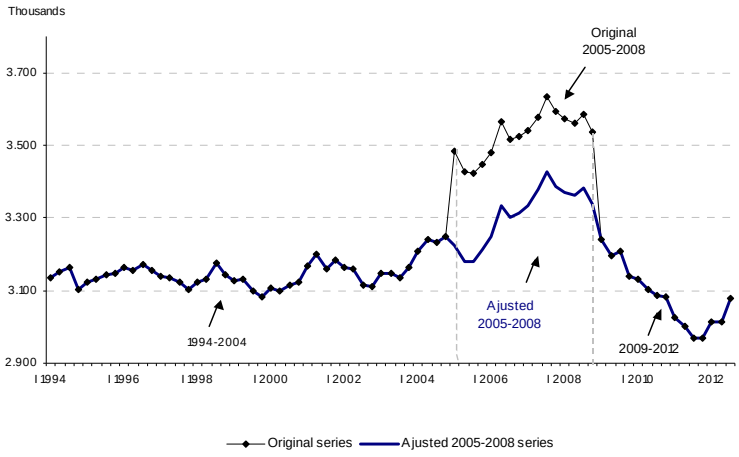


The area under the ROC curve is also a statistic of the goodness of fitted model, which is named "**statistic c**", and in this case has a score of 0.825. It equals 0.5 (white area) plus 0.325 (yellow area).

5 Final adjustment in 2005-2008 *professional status* LFS series

Figure 3 shows the **original and adjusted series** for total employment classified as self-employed. The stability in the adjustment over periods observed fairly constant, assuming an effective lowering of -6.2% on average in the period 2005-2008.

Figure 3. Self-employed workers: total



6 Conclusions

The method of "**logistic regression**" allows to predict the behaviour of a (qualitative or discrete quantitative) response variable based on (qualitative or quantitative) explanatory variables.

This technique is especially useful in **social surveys** like LFS where most variables are qualitative with few quantitative variables and it can be used by statistical offices either for the **imputation** of missing values or for **backcasting** (micro conversion of historical series under a new classification).

In this case, this technique allows the **micro conversion** of historical series in some groups (codes 03, 05 and 06) of self-employed workers through probabilities of reallocation and according to the other variables in the LFS questionnaire.

7 References

Eurostat (2007). Back Casting Handbook.

Hosmer, D. W., Jr. and Lemeshow, S. (2000), Applied Logistic Regression, Second Edition, New York: John Wiley & Sons.

National Statistics Institute of Spain (2008). Economically Active Population Survey. Methodology 2005. Description of the survey, definitions and instructions for completion of the questionnaire.

SAS Institute Inc. (2010), SAS / STAT User's Guide, Second Edition.

SAS Institute Inc. (1995), Logistic Regression Examples Using the SAS System, Cary, NC: SAS Institute Inc.