View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911

Dr. Piero Montebruno

pfm27@cam.ac.uk

Working Paper 11:
Working paper series from ESRC project ES/M010953:
Drivers of Entrepreneurship and Small Businesses
PI Prof. Robert J. Bennett.

University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure, Downing Place, Cambridge, CB2 3EN, UK.

August 2018

Supplementary material for the paper "*The Population of Non-corporate Business Proprietors in England and Wales 1891-1911*", by Bennett, Robert J., Montebruno, Piero, Smith, Harry J. (2018).

Comments are welcomed on this paper: contact the author as above.

© Piero Montebruno, University of Cambridge, member of the Cambridge Group for the History of Population and Social Structure assert his legal and moral rights to be identified as the author of this paper; it may be referenced provided full acknowledgement is made: ***Cite*** (Harvard format):

> Montebruno, Piero (2018) *Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911* Working Paper 11: ESRC project ES/M010953: 'Drivers of Entrepreneurship and Small Businesses', University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure.

Keywords: Entrepreneurship, Employers, Self-employment, Small businesses, Census 1891, Census 1901, Census 1911, non-response bias, misallocation bias

JEL Codes: L26, L25, D13, D22

Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911

Dr. Piero Montebruno

Working Paper 11: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge

1. Introduction.

This paper explains the use of weights to adjust the Censuses 1891-1911 for non-response and misallocation bias. The weights themselves are in a separate file available for download. The weights allow adjustment of observations to 'correct' values of when using data from I-CeM or the Entrepreneurs database at UKDA 1851-1911 developed from the ESRC project ES/M010953 *Drivers of Entrepreneurship and Small Businesses*. The paper provides detailed documentation of how the data base should be adjusted and the weighted data interpreted. More detailed discussion of the difficulties that arise in these three censuses is provided in the paper by Bennett et al. (2018) to which this working paper is linked.

The data referred to are derived from the electronic census data made available through the database deposit of the original CEBs at the UKDA: *The Integrated Census Microdata* (I-CeM). The version used derives from version 2 of these data.[1]

The paper is divided into three sections. The next section discusses how the weighting model was constructed. Then the use of the weights is outlined, and how covariates can be added to add greater control or detail to the resulting weights, tailored by the user.

[1] Higgs, Edward and Schürer, Kevin (University of Essex) (2014) *The Integrated Census Microdata (I-CeM)* UKDA, SN-7481; K. Schürer, E. Higgs, A.M. Reid, E.M Garrett, *Integrated Census Microdata, 1851-1911, version V. 2 (I-CeM.2)*, (2016) [data collection] UK Data Service SN: 7481.

2. The Weighting Model.

The censuses 1891-1911 have non-responses to the employment 'status' question (of whether a respondent was a worker, employer or own-account). In the Censuses 1891, 1901, 1911, this was a separate question, and each economically active person should have crossed one of the three columns (or written in their status). A large percentage did not answer: once the non-economically active are removed (scholars, retired, those living off own means and so on) 16%, 18% and 20% of people in, respectively, 1891, 1901 and 1911. Once the data are cleaned to remove definite non-entrepreneurs such as domestic servants, labourers and those on own means, as described in Bennett et al. (2018), the non-responses are reduced substantially, to 4.6%, 4.8% and 5.3%, respectively, for 1891, 1901 and 1911. However, the remaining non-respondents were not randomly distributed; with position within household (RELA code in I-CeM), gender, and sector being important correlates of non-response. These correlates can be used to correct for non-response bias.

In our case, the weights come from a logit regression (one for each of the I-CeM NewOccode - 797 categories) so more than 2,000 regressions are required for the three years as follows:

```
Logit RESPONSE i.Sex i.RELA_10
```

RESPONSE is a binary variable equal to one or answered if any employment status is given by the individual listed in the enumeration book and zero or blank if no employment status is provided. Hence, the logit model permits to predict the probability of responding for each individual after considering his or her sex and RELA.

The next step is to predict the probability of responding and from there to calculate the inverse of that probability to get the weights or the times each individual should be amplified to account for the non-response bias. Categories, where the probability of response is low, will command larger weights so that they will be greatly amplified. Categories, where the response is high, will receive lower weights.

As an example of the method, suppose a NewOccode has five individuals and three of them are men (M1, M2, M3) and two are women (F1, F2). Suppose also that one of the women doesn't give an employment status as follows:

Sex	Employment Status	RESPONSE
M1	W	1 / Answered
M2	W	1 / Answered
M3	E	1 / Answered
F1	E	1 / Answered
F2	.	0 / Blank

Then the probabilities of response are calculated. In this simple example the values are immediate from simple arithmetic but in the data they are much more complicated and have to be calculated with an in-sample extrapolation of the predicted values of the RESPONSE variable using the coefficients from the logit model and the sex and RELA of each individual. After calculating the probability of response calculation of the weights is straightforward: just take the inverse. The following table summarises this for our worked example:

Sex	probability of response	Weight
M1	1	1
M2	1	1
M3	1	1
F1	0.5	2
F2	0.5	.

That is, the men are not amplified at all because their probability of response is 1, i.e., they provided an answer for all the employment status questions. But the probability of response of the women is just one half since only one of the two answered the employment status question. Hence to re-weight their response gets a weight of 2. So Female 1 (F1) gets a weight of 2 - and her answer is by this means duplicated - while Female 2 (F2)'s answer does not count as she does not answer the employment status question. Thus, to obtain amplified number of workers (Ws), employers (Es) and own accounts (OAs) you sum the weights for each type. From now on we drop the variables RESPONSE and probability of response and we just use the derived variable Weight. This gives the complete responses to the previous example as follows:

```
Sex   Weight  Employment Status
M1      1             W
M2      1             W
M3      1             E
F1      2             E
F2      .             .
```

Then the employment status (ES) is simply the sum of the weights (note that there is just one column before the hash-symbol column: Employment status to which weights are linked):

```
ES #  sum of weights
W  2      (1+1)
E  3      (1+2)
----
T  5
```

Thus, the number of weighted responses is now five because the amplification permits weighting the responses and allocating the values of employment status to account for non-responses. F1 is counted twice because F2 has not answered the employment status. In other words, blank responses have been allocated by the method.

This example uses just Sex for ease of explanation. But the actual weights also uses Relationship to the Head (RELA; which is defined here for RELA_10, a reduced number of categories form those in I-CeM) and implicitly each NewOccode (797 categories).

3. Using the Weights

Once the weights are calculated, you need to add them to find the overall count of each category. If it is needed to count the individuals by sex, both the total and by employment status you don't need to count the number of men and women. Instead of adding the number of individuals by sex and employment status you just need to add the weights by sex and employment status. Always remember, that what you are weighting is only the answer to the employment status question. Thus, you always and only need to add weights. Take for instance farmers in 1901 (NewOccode=173) and display the employment status by Sex:

```
-> Sex = Male
```

EmployCode2	Freq.	Percent	Cum.
Worker	10,526	5.25	5.25
Employer	113,925	56.84	62.09
Own-account	61,252	30.56	92.65
.	14,737	7.35	100.00
Total	200,440	100.00	

-> Sex = Female

EmployCode2	Freq.	Percent	Cum.
Worker	634	3.04	3.04
Employer	10,683	51.25	54.30
Own-account	6,286	30.16	84.46
.	3,240	15.54	100.00
Total	20,843	100.00	

The blanks are 14,737, and 3,240. Just by adding the weights by Sex you get

Sex	Sum of Weights
Male	200,451
Female	20,817.95

This produces a result statistically close to the overall count, i.e. the correct answer. Although we did not add individuals, just weights.

To gain an insight of how weights are calculated look at the following table where each category of RELA_10 and Sex is given a weight again for farmers, NewOccode=173 (for the year 1911):

RELA_10	Sex	Weight
Head	Male	1.0691
Working title	Male	1.0976
Siblings	Male	1.1584
Head	Female	1.1952
Other family	Male	1.2300
Lodgers/boarders	Male	1.2374
Working title	Female	1.2758
CFU member	Male	1.2851

Servants	Male	1.2934
Unknown	Male	1.3084
Siblings	Female	1.4474
Older generation	Male	1.4603
Other family	Female	1.6498
Lodgers/boarders	Female	1.6707
CFU member	Female	1.8053
Servants	Female	1.8287
Unknown	Female	1.8711
Older generation	Female	2.3001
Non-household	Male	2.7069
Non-household	Female	5.8216

From the Table, you can see that men tend to receive lower weights while women larger. This is simply because the non-response by women is generally higher while for men it is lower. Also, Head is a category associated with low non-response, while Non-household has high non-response. Hence, Head commands low weights, and Non-household has large weights because low response categories need to be amplified.

4. Other Covariates

A researcher may wish to use the weights in conjunction with other categories, e.g. the number of females and males by employment status. To calculate this it is necessary to add the weights by Employment Status and Sex. For instance,

Sex	Weight	Employment Status
M1	1	W
M2	1	W
M3	1	E
F1	2	E
F2	.	.

Then the Sex and Employment Status (ES) are simply the sum of the weights:

ES	Sex	#	sum of weights
W	M	2	(1+1)
W	F	0	0
E	M	1	1
E	F	2	(2+0)
---	---	---	---
T		5	

Note that there are now two columns before the hash-symbol column: Employment status and Sex. In the example, females are two; males are three. Females and workers are zero, male and workers are two, females and employers are two, and males and employers are one.

In the same way, other covariates can be calculated by summing the weights either alone or together. If another covariate is known as the relationship to the head of the family, the numbers are calculated similarly:

Rela	Sex	Weight	Employment Status
Head	M1	1	W
CFU	M2	1	W
Head	M3	1	E
Head	F1	2	E
CFU	F2	.	.

.

Then the RELAs are simply the sum of the weights:

Rela	#	sum of weights
Head	4	(1+1+2)
CFU	1	(1+0)
T	5	

And the Sex and RELA are:

Sex	Rela	#	sum of weights
M	Head	2	(1+1)
F	Head	2	2
M	CFU	1	1
F	CFU	0	0
T		5	

The Sex and Rela and Employment Status (ES) are:

Sex	Rela	ES	#	sum of weights
M	Head	W	1	1
F	Head	W	0	0
M	CFU	W	1	1
F	CFU	W	0	0
M	Head	E	1	1
F	Head	E	2	2
M	CFU	E	0	0
F	CFU	E	0	0
T			5	

Note that there are now three columns before the hash-symbol column: RELA, Employment status and Sex.

5. Conclusion

This paper shows how to handle weights to adjust the Censuses 1891-1911 for non-response and misallocation bias. The main scope of the paper is to give worked examples to explain practically how the weights are used. This is not a theoretical discussion but a series of worked examples to show what the weights are supposed to do.

Acknowledgments:

This research has been supported by the ESRC under project grant ES/M010953: **Drivers of Entrepreneurship and Small Businesses.** Piloting of the research for 1881 draws from Leverhulme Trust grant RG66385: **The long-term evolution of Small and Medium-Sized Enterprises (SMEs).**

The database used for 1891 and 1901-11 derives from K. Schürer, E. Higgs, A.M. Reid, E.M Garrett, *Integrated Census Microdata, 1851-1911, version V. 2 (I-CeM.2),* (2016) [data collection]. UK Data Service, SN: 7481, http://dx.doi.org/10.5255/UKDA-SN-7481-1; enhanced; E. Higgs, C. Jones, K. Schürer and A. Wilkinson, *Integrated Census Microdata (I-CeM) Guide, 2nd ed*. (Colchester: Department of History, University of Essex, 2015).

The GIS boundary files for RSDs were constructed by Joe Day for the ESRC fertility project directed by Alice Reid:

http://www.geog.cam.ac.uk/research/projects/victorianfertilitydecline/publications.html

These used as a starting point the GIS parish files of Satchell, A.E.M., Kitson, P.M.K., Newton, G.H., Shaw-Taylor, L., Wrigley E.A. (2006) *1851 England and Wales census parishes, townships and places*, 2006, ESRC RES-000-23-1579, supported by Leverhulme Trust and the British Academy; Satchell, A.E.M. (2015) *England and Wales census parishes, townships and places*; which is an enhanced and corrected version of Burton, N, Westwood J., and Carter P. (2014) *GIS of the ancient parishes of England and Wales, 1500-1850*, UKDA, SN 4828; which is a GIS version of Kain, R.J.P., and Oliver, R.R. (2001) *Historic parishes of England and Wales: An electronic map of boundaries before 1850 with a gazetteer and metadata*, UKDA, SN 4348.

Other Working Papers:

Working paper series: ESRC project ES/M010953: *'Drivers of Entrepreneurship and Small Business',* University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure.

WP 1: Bennett, Robert J., Smith Harry J., van Lieshout, Carry, and Newton, Gill (2017) *Drivers of Entrepreneurship and Small Businesses: Project overview and database design.*

WP 2: Bennett, Robert J., Smith Harry J. and van Lieshout, Carry (2017) *Employers and the self-employed in the censuses 1851-1911: The census as a source for identifying entrepreneurs, business numbers and size distribution.*

WP 3: van Lieshout, Carry, Bennett, Robert J., Smith, Harry J. and Newton, Gill (2017) *Identifying businesses and entrepreneurs in the Censuses 1851-1881*.

WP 4: Smith, Harry J., Bennett, Robert J., and van Lieshout, Carry (2017) *Extracting entrepreneurs from the Censuses, 1891-1911*.

WP 5: Bennett, Robert J., Smith Harry J., van Lieshout, Carry, and Newton, Gill (2017) *Business sectors, occupations and aggregations of census data 1851-1911*.

WP 6: Smith, Harry J. and Bennett, Robert J. (2017) *Urban-Rural Classification using Census data, 1851-1911*.

WP 7: Smith, Harry, Bennett, Robert J., and Radicic, Dragana (2017) *Classification of towns in 1891 using factor analysis.*

WP 8: Bennett, Robert J., Smith, Harry, and Radicic, Dragana (2017) *Classification of occupations for economically active: Factor analysis of Registration Sub-Districts (RSDs) in 1891*.

WP 9: Bennett, Robert, J., Montebruno, Piero, Smith, Harry, and van Lieshout, Carry (2018) *Reconstructing entrepreneurship and business numbers for censuses 1851-81*.

WP10: Bennett, Robert, J., Smith, Harry and Radicic, Dragana (2018) *Classification of environments of entrepreneurship: Factor analysis of Registration Sub-Districts (RSDs) in 1891*.

Full list of all current Working Papers available at:
http://www.geog.cam.ac.uk/research/projects/driversofentrepreneurship/