The Use of Pre-aggregated Administrative Data Sets in Compiling Population Estimates

Part 1: Data Zones Using Multiple Linear Regression

1. Introduction

The topic referred to in the title is likely to be of central important to National Statistical Institutes over the coming years. The research reported here describes how administrative data sets (referred to below as ‘sets’) available in the Spring of 2011 could have been used to compile population estimates at Data Zone (DZ) level and compares these to actual DZ counts from the 2011 Census to assess the extent to which the two sets are consistent. A variety of procedures could be adopted to derive these estimates and potentially unlimited effort could be expended on researching the most effective. The one reported here, based on orthodox multiple linear regression, seems to represent a good combination of simplicity and efficiency.

2. Procedure

The procedure adopted rests on the following assumptions:

a. DZ counts from sets will be related to DZ counts from the census but the nature of the relationship will ‘drift’ over time and so periodic recalibration will be required.
b. There will always be a minority of DZs where the predictability of the census count from administrative data will be so poor that accurate figures can only be obtained by direct enumeration, adjusted by methods currently used for the census. These will be referred to as ‘red’ DZs and will often be associated with the existence of large communal establishments such as prisons, armed forces bases and university halls of residence.
c. Since several sets exist but only one set of estimates is required, some method of integrating the sets is required. The method used here is multiple linear regression. Although this has drawbacks (which are detailed below), it has the important advantage of automatically weighting the contributions which the different sets make to the estimates by taking account of their correlations with the census and with each other. Whether this advantage is enough to outweigh its disadvantages is one of the many matters which could be explored in subsequent stages of this research project.

These three assumptions lead to the following stages in the procedure (the red, white and blue model):

a. Identifying the red DZs and removing them from the analysis, reintegrating them only at the final stage.
b. Identifying from amongst the remaining DZs a calibration sample (the ‘blue’ DZs) which will be used to calculate regression parameters.
c. Using these parameters to calculate estimates for the remaining ‘white’ DZs.
d. Integrating the red, white and blue DZs to provide estimates for all DZs.

In the following account, it was assumed that census counts were not initially available for any DZs but that they became available for the red and blue DZs after they had been identified as the two parts of the direct enumeration sample and surveyed.

Five sets were used for this work. These were: NHS Central Register (covering all ages), Electoral Register or ER (16+ years), Child Benefit or CB (0 to 16 years), School Census or SC (5 to 18 years) and Super Older Person's Database or SOPD (65+ years).

2A. The red DZs

The two questions here concern the number of red DZs and the way in which they will be identified in the absence of a census. The former question will probably be determined on economic rather than statistical grounds. The unpredictability which defines red DZs is a continuum rather than a dichotomy and the cut-off point will inevitably be arbitrary to some extent. Here it will be assumed that about 5% of DZs will be deemed to be red. This could be implemented either by defining cut-offs which identify about 2½% of DZs at each tail or by defining cut-offs which are equidistant from zero and which identify a total of around 5%. If the distribution of error scores is symmetric, the two will effectively give the same answer. If it is not, the 2½% method will mean that relatively accurately predicted DZs in the shorter tail will be deemed red while less accurately predicted DZs in the longer tail will not. Therefore equidistant cut-off scores were used to define red DZs so that the same criterion is used at each tail.

The second question concerns the basis on which red DZs will be identified. The available figures which are closest to the census are the Mid-Year Estimates (MYEs) derived from applying the cohort component method to intercensal years since the most recent census. Therefore the unre-based 2011 MYEs were used in the present research. The method used was to take from each of the five sets the raw count of people in each DZ. These differed greatly over sets in their absolute size since most of the sets covered only a subset of ages. From the point of view of comparability, it would have been better if each set had predicted the census count of the number of people in the age range covered by that set. However this would have meant five dependent variables and hence five separate regression analyses. This would in turn have lost the automatic optimal weighting which is only available for an integrated multiple regression with a single dependent variable.

The raw counts from the five sets were used as predictor variables with MYE as the dependent variable, the regression being taken over all 6,505 DZs. Table 1 gives the results of the analysis. The squared multiple correlation coefficient was 0.91 and the root mean square error was 78.7.

It can be seen that two of the five sets have negative coefficients which means that optimisation involves partially subtracting them from, rather than adding them to, the prediction equation. SC has the weakest relationship with census counts of that age range (probably this is in large part due to the exclusion of private schools) and younger age ranges are already covered by CB. Subtracting SOPD means that excluding the over-65s makes the set-based counts more consistent with the census than including them. This again is something which could be subjected to further research attention but it is probably related to the large degree of intercorrelation between the predictor variables.

Table 1: Regression analysis for 6,505 DZs (target variable = unre-based MYE)

Variable	DF	Coefficient	Standard Error	t Value	Pr > |t|	Variance Inflation[1]
Intercept	1	55.25050	3.51	15.72	<.0001	0.00
NHSCR	1	0.57227	0.01	59.00	<.0001	7.85
ER	1	0.41348	0.01	27.63	<.0001	8.70
CB	1	0.55700	0.05	10.42	<.0001	13.92
SC	1	-0.43215	0.08	-5.70	<.0001	11.71
SOPD	1	-0.12175	0.02	-5.65	<.0001	1.63

From the regression equation we can calculate the difference between the observed MYE and its predicted value, expressed as a percentage of the observed value. These were used as error scores. From the distribution of these scores it was found that cut-off scores of ±15% identified 191 DZs at the bottom and 150 at the top or 5.24% of all DZs. This was the nearest fit to the 5% criterion. To assess the extent of the agreement between this and the 2011 census, the regression analysis was run again, this time using the observed census DZ count as dependent variable, the error score definition being analogous to that used above. From the resulting error scores distribution, the cut-off scores which isolated 5% of the DZs were ±10%. The respective cut-offs were applied to the two distributions of error scores to classify each DZ as 'upper red', 'lower red' or 'neither' for each method. These were then cross-tabulated against each other, resulting in table 2.

Table 2: Red cells from MYEs and the Census

		Census			Total
		lower	neither	upper	
MYE	lower	54	125	12	191
	neither	128	5,951	85	6,164
	upper	3	115	32	150
	Total	185	6,191	129	6,505

As would be hoped, there is a noticeable measure of correlation between red status using MYEs and using the Census. Note in particular that there are 15 DZs which are red on both criteria but in opposite directions. In practice, this would not be important as they would be directly enumerated anyway but they

Footnote

[1] Variance inflation is related to the squared multiple correlation for each predictor variable regressed against all the other predictor variables. High values indicate a high degree of inter-correlation between the predictors, as would be expected in this case, and in extreme cases it can make the regression algorithm unstable. In this case it does not seem to have prevented satisfactory convergence but it is a matter which should be kept in mind when using multiple regression in this way.

are a measure of the infrequency of extreme disagreement between the two criteria. There are 213 DZs which would be red using the census but not using MYEs and 240 for which the reverse is true. In this respect, table 2 may be a little misleading. The cut-offs have been applied for ease of presentation but the result disguises the fact that a bivariate normal distribution underlies table 2 and that the majority of the 453 DZs which would be red on one criterion but not the other are in fact quite close to the borderline. Since DZ counts both from the census and from MYE themselves contain an error component, each could not be expected to be wholly consistent with itself if the counts had been compiled in a different way. The fact that about 93% of the DZs were classified in the same way on the two criteria will be taken as justification for using the red DZs as defined by the MYEs. In practice, the method could be refined not by identifying the red DZs solely on the basis of MYE but by using other information which might be relevant to the decision to enumerate directly such as the presence of communal establishments.

2B. The blue and white DZs

The non-red DZs were sorted by DZ code and every 20th DZ was deemed to be blue. There were therefore 308 of these. It was assumed that the census count could be used as a proxy for the enumeration and a multiple regression was run on the blue DZs only. The census count was used as the dependent variable and the raw set counts as the predictors. Table 3 gives the results of the analysis. The squared multiple correlation coefficient was 0.96 and the root mean square error was 40.3. The absence of the red DZs has led to these goodness-of-fit statistics being better than they were in table 1.

Table 3: Regression analysis for 308 blue DZs (dependent variable = Census)

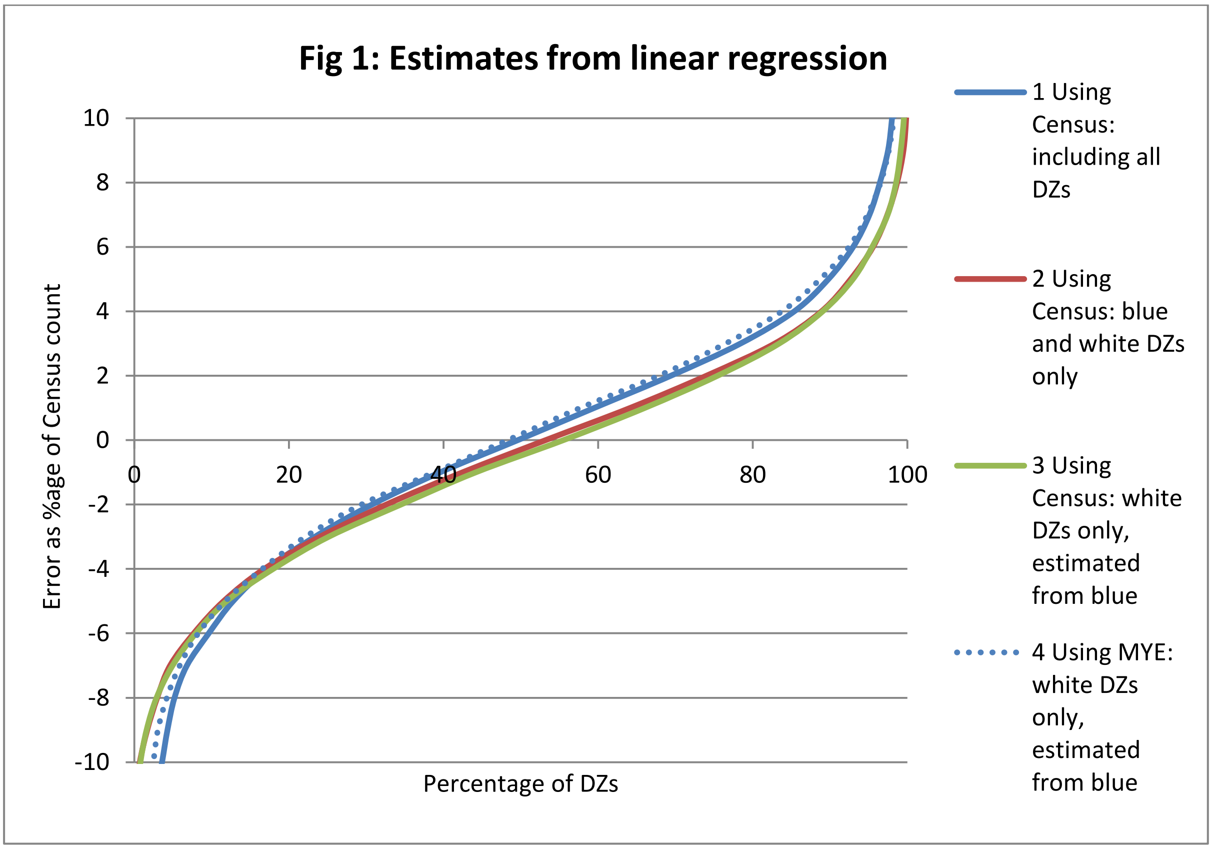
Variable	DF	Coefficient	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	14.00437	10.29468	1.36	0.1747	0
NHSCR	1	0.47953	0.03492	13.73	<.0001	9.20105
ER	1	0.57701	0.04097	14.08	<.0001	8.53623
CB	1	0.99128	0.14251	6.96	<.0001	11.03098
SC	1	-0.88660	0.18044	-4.91	<.0001	8.41959
SOPD	1	-0.16844	0.05103	-3.30	0.0011	1.55449

Again, the same two sets have negative coefficients and the variance inflation is large but did not prevent convergence. The values in table 3 were used, along with raw set counts for the remaining 5,856 white DZs (90% of all DZs), to generate a predicted DZ count. This was then compared, using the error score defined above, with the actual census DZ counts for the white DZs. The actual DZ counts from the census were used for the red and blue DZs. The results are reported in the next section.

For the white DZs, a further comparison was made by calculating the root mean square error between the actual DZ count from the census and the predicted value using the regression coefficients from the blue DZs. This had the value 46.8 which is higher than the 40.3 for the blue DZs because there, the total of the observed scores and the predictions were guaranteed to be the same. For the white DZs however the root mean square error contained a component due to the observations and predictions having different totals.

3. Consistency with Census counts

Fig 1 shows a consistency chart in which DZ counts are estimated using four different procedures and compared to the observed 2011 census DZ counts. In the first procedure, the regression was run over all 6,505 DZs using observed census count as the target variable. This is the simplest regression. The second and third procedures differed from this in that the red DZs have been excluded which is why the consistency curve is closer to the horizontal axis (denoting higher consistency). The difference between them is that for the second curve, the regression was run over all blue and white DZs while for the third, it was run only over the blue DZs and the results generalised to the white DZs. That the second and third lines are so close to each other (almost overlying each other for most of their length) is evidence that the parameter estimates from only the blue DZs give predictions which are almost as accurate as those using both the blue and the white DZs.



These first three curves are all shown with solid lines and none would be possible in the absence of a census. The fourth line (shown dotted) however would be possible. Here, the red DZs were identified using MYEs rather than census data and the curve covers only the white DZs. The closest comparison between what would be possible with and without a census is between curves 3 and 4.

Perhaps the most striking aspect of fig 1 is how close all four lines are. Removing the red DZs in curves 2 and 3 has the expected improvement in consistency while moving from the census to MYEs as the method of identifying red DZs causes a reduction in consistency. These appear to cancel each other out with the effect that curves 1 and 4 are very comparable.

4. Conclusion

Multiple linear regression is a familiar technique in statistical modelling and its strengths and weaknesses are well known. In this case, its weaknesses do not seem to have been unduly problematic. The degree of correlation between the predictor variables does not seem to have led to such instability as to prevent convergence, though the variance inflation factors were large enough to justify continuing vigilance. As the estimates are not bounded, there is the possibility of negative population estimates being produced. This did not occur in the above calculations since the degree of aggregation (over both gender and age) was enough to prevent this. For a more detailed cell structure however this is also a matter which should be monitored.

These cautions aside, the results were encouraging. The squared multiple correlation coefficient was 0.96 (though it should be remembered that this is a second order statistic and these often give rather flattering impressions of goodness of fit). The root mean square error of 46.8 however is a good guide to the mean DZ level residual for the non-enumerated DZs. It seems that, despite the existence of a minority of DZs where predictive accuracy is poor despite not being flagged as red on the basis of the un-rebased MYEs, overall accuracy is reasonably encouraging.

Alternative Sources
June 2014