LOGISTIC REGRESSION ANALYSIS WITH AGGREGATE DATA: TACKLING THE ECOLOGICAL FALLACY

David G. Steel, University of Wollongong, Australia
Mark Tranmer, University of Southampton, UK
D. Holt, Office for National Statistics, London, UK
David G. Steel, School of Mathematics and Applied Statistics,
University of Wollongong, NSW 2522, Australia (david_steel@uow.edu.au)

Key Words: Ecological analysis, ecological fallacy, multilevel models, random effects, grouping variables.

1. Introduction

An ecological analysis uses aggregate group level data to estimate individual level relationships. The ecological fallacy arises when ecological analysis provides biased estimates of individual level relationships. The groups involved are often small geographic areas such as census Enumeration Districts (EDs). Suppose that we are interested in investigating the relationship between two variables Y and X. The aggregate data available for area g usually consists of the group level means $\bar{Y}_g$ and $\bar{X}_g$.

The ecological fallacy arises because the individuals within the groups are not equivalent to randomly formed groups. Progress in understanding aggregation effects requires allowance for the population structure in the model underpinning the analysis. Steel and Holt (1996) proposed a model for cases where the relationship between the two variables of interest, Y and X, is linear. This model incorporated auxiliary variables, Z, which explain much of the within area homogeniety of the variables of interest. Random group level coefficients were also included to reflect group level effects additional to those due to the auxiliary variables. Based on this model Steel and Holt (1996) developed a method to adjust group level covariance matrices using limited individual level data available for the auxiliary variables. Steel, Holt and Tranmer (1996) evaluated this approach for estimating correlation coefficients using ED level data from the 1991 UK population census. Using a combination of basic demographic and housing variables as auxiliary variables they were able to reduce the aggregation effects in a set of ecological correlations by up to 70 per cent.

In social research the variables are usually categorical at the individual level and the area level means are the corresponding proportions. Various methods of ecological inference in this situation were evaluated by Cleave, Brown and Payne (1995). These included linear ecological regression as originally suggested by Goodman (1959) and a method based on an Aggregated Compound Multinomial (ACM) model. The ACM model assumed that the frequencies in each group have independent multinomial distributions and used a Dirichlet compounding distribution. Their evaluation favoured the ACM based method but they recognized that it is not easy to implement. Recently King (1997) proposed a new method of ecological analysis for categorical data. This method models the conditional proportions of Y given X as random effects with a joint truncated Normal distribution and exploits the constraints implied by the group means.

In this paper we develop and evaluate some simple adjusted ecological analysis procedures based on the idea of incorporating auxiliary variables and using individual level data available for these variables.

2. Adjusted Ecological Analysis for Dichotomous Variables

Let Y_i and X_i be the values of the variables of interest for the i th individual in the population. Suppose there is a sample, s, of groups and within the sampled group, g, a sample of n_g individuals

This research was supported by the UK Economic Social Research Council, grant R000 23 6135

is used to calculate sample group means:

$$\bar{Y}_g = \frac{1}{n_g} \sum_{i \in s,g} Y_i \quad \text{and} \quad \bar{X}_g = \frac{1}{n_g} \sum_{i \in s,g} X_i$$

The variables are dichotomous and so these means are proportions. An important special case occurs when the means are available for all areas and each mean is based on all individuals within the relevant area.

For the sample in the group g let n_{abg} be the number of individuals for which $Y_i = a$ and $X_i = b$. The corresponding population counts are N_{abg}. We use " + " to indicate summation over a subscript. Define $p_{ab|g} = n_{abg}/n_{++g}$ and $P_{ab|g} = N_{abg}/N_{++g}$ as the sample and finite population proportions for group g. The conditional proportions are $p_{a|bg} = n_{abg}/n_{+bg}$ and $P_{a|bg} = N_{abg}/N_{+bg}$. Define $p_{ab+} = n_{ab+}/n_{+++}$ and $P_{ab+} = N_{ab+}/N_{+++}$ as the overall sample and finite population proportions. The corresponding conditional proportions are $p_{a|b+} = n_{ab+}/n_{+b+}$ and $P_{a|b+} = N_{ab+}/N_{+b+}$.

The basis of the approach proposed by Steel and Holt (1996) is that, for continuous variables, the conditional probability density function of Y given X can be expressed as

$$f(y|x) = \frac{\int f(y|x,z) f(x|z) f(z) dz}{\int f(x|z) f(z) dz} \qquad (1)$$

When the parameters of $f(y|x,z)$, $f(x|z)$, and $f(z)$ are distinct, analysis can proceed by using individual level data to estimate the parameters of $f(z)$ and aggregate data are used to estimate the parameters of $f(y|x,z)$ and $f(x|z)$.

Assume that there is a single auxiliary variable Z and that Y, X and Z are all categorical. The target of inference is the conditional probability distribution of Y given X, $P(Y|X)$. The approach we develop here is based on

$$P(Y|X) = \frac{\sum_Z P(Y|X,Z) P(X|Z) P(Z)}{\sum_Z P(X|Z) P(Z)} \qquad (2)$$

Estimation of $P(Y|X)$ can be attempted by using aggregate data to estimate $P(Y|X,Z)$ and $P(X|Z)$ and then using the individual level data to estimate $P(Z)$. The analysis using aggregate data will be based on linear or logistic regression.

The potential benefit of using group level information about covariates in ecological analysis is discussed by Cleave, Brown and Payne (1995) and King (1997). In both cases the covariates are used at the area level to explain some of the variation in the random coefficients characterizing the relationship between the variables across the areas. Our approach is motivated by the idea that a large part of the variation in the relationships across areas is due to compositional effects and can be removed by inclusion of the auxiliary variables. If this is the case then handling the remaining variation between areas should be easier. We propose attempting to average over the auxiliary variables to estimate the marginal relationship between Y and X. This requires information about the relevant parameters of the individual level distribution of the covariates.

For a single categorical auxiliary variable the information needed to calculate the adjusted marginal probabilities consists only of the proportions in each category. These can be calculated from the weighted group level data or could come from some other source, such as a survey. If two or more categorical auxiliary variables are used then the marginal cross tabulations of these variables are required, unless further assumptions can be made concerning the relationship between the auxiliary variables. No individual level data about the variables of direct interest are used.

3. Empirical Evaluation

Individual and group level data from the 1991 UK population census were used to evaluate several methods. The Small Area Statistics (SAS) database provided data in the form of totals for a range of categorical variables for EDs. This was the source of the group means $\bar{Y}_g$, $\bar{X}_g$ and $\bar{Z}_g$. For the variables analysed in this paper these means are all based on 100 per cent of the census records for the relevant ED. Individual level data are available from a 2 percent Sample of Anonymized Records (SAR) for Local Authority Districts (LADs). The evaluation used data for the LAD of Manchester, which contained 897 EDs and 7613 individuals in the SAR, of which 5802 were aged 16 or more.

Using these data it is possible to calculate adjusted ecological estimates of the marginal probability distribution $P(Y|X)$ based on equation (2) using $\bar{Y}_g$, $\bar{X}_g$ and $\bar{Z}_g$ obtained from the SAS database and using individual level data obtained from the SAR concerning variables chosen as auxiliary variables. In this evaluation we considered the following estimators of the conditional probability $P(Y=1|X=b)=\pi_{1|b}$, for $b=0,1$.

(a) SAR relative frequencies. The relative frequency obtained from the SAR, n_{1b+}/n_{+b+}.

(b) SAS relative frequencies. For pairs of variables for which the SAS contains the relevant cross tabulation we can calculate N_{1b+}/N_{+b+}.

(c) Ecological linear regression. This is built around the relationship

$$\bar{Y}_g = P_{1|0g}(1-\bar{X}_g) + P_{1|1g}\bar{X}_g \tag{3}$$

If $P_{1|0g}$ and $P_{1|1g}$ are random variables with $E[P_{1|0g}|\bar{X}_g] = \pi_{1|0}$ and $E[P_{1|1g}|\bar{X}_g] = \pi_{1|1}$ then a linear regression of $\bar{Y}_g$ on $\bar{X}_g$ gives unbiased estimates of $\pi_{1|0}$ and $\pi_{1|1}$. This is the classical Goodman regression approach (Goodman, 1959). This approach can produce estimates outside [0,1] and simple direct use of this model is not usually recommended.

(d) Ecological logistic regression. A common model used in analysing a dichotomous response variable is logistic regression which assumes Y_i is a Binomial variable based on one trial, $B(1, E[Y_i|X_i])$, where

$$log\left(\frac{E[Y_i|X_i]}{1-E[Y_i|X_i]}\right) = \alpha + \beta X_i$$

If groups were completely homogeneous with respect to X then each group total Y_g would be a Binomial variable based on N_g trials with probability such that

$$log\left(\frac{E[\bar{Y}_g|\bar{X}_g]}{1-E[\bar{Y}_g|\bar{X}_g]}\right) = \alpha + \beta\bar{X}_g$$

Groups are not homogeneous with respect to X but this model has the advantage of not giving predicted probabilities outside [0,1].

(e) Adjusted ecological linear regression. Here we use linear regression based on aggregate data to estimate $P(Y|X,Z)$ and $P(X|Z)$. Estimation of $P(Y|X)$ is then based on equation (2) with the individual level data being used to estimate $P(Z)$.

(f) Adjusted ecological logistic regression. Here we use logistic regression based on aggregate data to estimate $P(Y|X,Z)$ and $P(X|Z)$. Estimation of $P(Y|X)$ is then based on equation (2) with the individual level data being used to estimate $P(Z)$.

(g) Adjusted correlation approach. For two dichotomous variables the correlation combined with the marginal totals determines the proportions in the cross classification i.e.

$$P_{11+} = P_{1++}P_{+1+} + R_{YX}\sqrt{P_{1++}(1-P_{1++})P_{+1+}(1-P_{+1+})} \tag{4}$$

where R_{YX} is the correlation coefficient based on the table of proportions P_{ab+}. The method proposed by Steel and Holt (1996) is used to produce adjusted estimates of the correlation $R_{YX}(Z)$ which can then be substituted into equation (4). This method enables use of information about several auxiliary variables even if only the two way cross tabulations are available.

(h) King's ecological inference method. This method is also built around equation (3). The group level conditional proportions $P_{1|0g}$ and $P_{1|1g}$ are assumed to have a joint Normal distribution which is truncated so that they each lie in the [0,1] interval. The method incorporates the fact that given $\bar{Y}_g$ and $\bar{X}_g$, equation (3) implies bounds and constraints for $P_{1|1g}$ and $P_{1|0g}$. This approach produces estimates of $P_{1|1g}$ and $P_{1|0g}$ which can then be combined to produce estimates of $P_{1|1}$ and $P_{1|0}$. This method can incorporate group level covariates to help model the variation between groups.

The variables used in the analysis were as follows:

Y: employed, unemployed,

X: marital status,

Z: age 45-59, age 60+, living in owner occupied dwelling, renting from local authority.

These auxiliary variables were chosen because of their success in removing aggregation effects in correlations in the evaluation by Steel, Holt and Tranmer (1996). The analysis was confined to those aged 16 or more.

Table 1 gives the estimated probabilities of being employed ($Y=1$) given marital status (X). The estimates are obtained using the methods listed above. The ecological methods and adjusted correlation methods were implemented using the

SAS package and King's method was implemented using EzI software, developed by King and colleagues (KIng, 1997). The first row of the table give the "true" values of the probabilities as estimated from census cross tabulations, available for these particular variables from the SAS data base, which were the same as the corresponding estimates obtained from the SAR.

When adjustment variables are not included, the ecological linear and ecological logistic estimates are considerably different from the true values. The inclusion of "owner occupied housing" as an adjustment variable leads to ecological linear and logistic estimates that are much closer to the true values. Used as a single adjustment variable "aged 60 and over" does not improve the estimates. The estimates obtained by the adjusted ecological linear method are fairly close to the true values when several adjustment variables are used. In particular, those combinations of adjustment variables that include housing tenure. The adjusted linear regression method generally works better than the adjusted correlation method suggested by Steel and Holt (1996). In this example, King's method without covariates gives results similar to those for the ecological linear and logistic methods without covariates. When "owner occupied" is used as a covariates, the estimated probabilities are further from the true values than those obtained using the adjusted linear regression method. However, when the covariate "renting from a local authority" is added the estimates from King's method are slightly better than those based on the adjusted linear regression method. Use of King's method with more than two covariates proved difficult in practice and no results for such cases are included.

For the estimates of the proportion unemployed by marital status given in Table 2, the unadjusted ecological linear and logistic methods provide poor estimates. The linear method leads to an out of range estimate. Including "owner occupied" as an adjustment variable in these methods leads to some improvement. The adjusted ecological linear regression method works reasonably well for combinations of adjustment variables that include housing tenure and age. King's method without covariates works somewhat better in this case than in the previous example, because differences in the proportions unemployed and married across the EDs lead to more informative bounds. However, the estimates are still worse than those obtained from the adusted logistic regression method. When the variable "owner occupied" is used as a covariate, the estimates obtained from King's method are effectively equal to the true values, while those based on the adjusted linear and logistic regression methods are not.

While these results are limited to two relationships, they highlight the importance of using auxiliary variables in ecological analysis to obtain reasonable estimates. The choice of auxiliary variables is important and methods of identifying effective adjustment variables need to be used, as suggested by Steel and Holt (1996). If appropriate auxiliary variables are used, quite simple methods can perform almost as well as more sophisticated, computer intensive ones. We expect that the simple adjusted ecological methods can be extended to also include random effects within the sort of multilevel framework developed, for example, by Goldstein (1995). Methods that solely use random effects to account for the variation in the relationship between groups will, in general, not be very successful. More information needs to be incorporated in order to get useful estimates. This information may be in the form of individual level data on auxiliary variables, group level covariates or the constraints exploited in King's method.

4. References

Cleave, N., Brown, P.J. and C. D. Payne (1995) Evaluation of Methods for Ecological Inference. *Journal of the Royal Statistical Society, A,* **158**, pp 55 - 72

Goldstein, H. (1995). *Multilevel Statistical Models, 2nd Edition*, Edward Arnold, London.

Goodman, L.A. (1959). Some alternatives to Ecological regression. *American Journal of Sociological Review*, **18**, 663-664.

King, G. (1997). *A Solution to the Ecological Inference Problem :Reconstructing Individual Behavior from Aggregate Data.* Princeton Univ Press

Steel, D. and Holt, D. (1996). Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisted. *International Statistical Review*, **64**, pp 39-60

Steel D., Holt D. and Tranmer M. (1996). Making unit level inferences from aggregated data, *Survey Methodology* **22**, 3-15

Table 1: Estimated probabilities: Y = employed; X = married

1(a): Ecological linear

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.41	.50
covariate(s):		
none	.26	.67
age2	.23	.71
age1, age2	.28	.65
oo	.48	.42
oo, rla	.47	.43
oo, rla, age1,2	.45	.45

1(b): Ecological logistic

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.41	.50
covariate(s):		
none	.26	.67
age60+	.29	.65
age1, age2	.33	.59
oo	.49	.41
oo, rla	.46	.44
oo, rla, age1,2	.45	.46

1(c): Correlation method

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.41	.50
covariate(s):		
none	.27	.66
age2	.22	.70
age1, age2 +	.20	.73
oo	.55	.35
oo, rla	.51	.39
oo, rla, age1,2	.47	.43

1(d): King's method

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.41	.50
covariate(s):		
none	.25	.69
age60+		
age4559, age60+		
oo	.50	.39
oo, rla	.44	.46
oo, rla, age1,2		

Source: 1991 UK census data.
Population: Residents aged 16 or more in households, Manchester LAD.
Y takes the value 1 for 'employed' 0 for 'not employed'.
X takes the value 1 for 'married' and 0 for 'not married'.
$P_{1|0}$ means $P(Y = 1 \mid X = 0)$; $P_{1|1}$ means $P(Y = 1 \mid X = 1)$
Adjustment variables: oo = owner occupied; rla = rented from local authority;
age1 = persons aged 45 - 59; age2 = persons aged 60 and over.

Table 2: Estimated probabilities: Y = unemployed; X = married

1(a): Ecological linear

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.14	.07
covariate(s):		
none	.24	-.06
age2	.24	-.05
age1, age2	.22	-.03
oo	.18	.01
oo, rla	.19	.01
oo, rla, age1,2	.16	.04

1(b): Ecological logistic

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.14	.07
covariate(s):		
none	.17	.02
age60+	.17	.02
age1, age2	.11	.09
oo	.16	.04
oo, rla	.15	.04
oo, rla, age1,2	.11	.09

2(c): Correlation method

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.14	.07
covariate(s):		
none	.27	-.09
age2	.27	-.09
age1, age2	.28	-.09
oo	.19	-.00
oo, rla	.18	.01
oo, rla, age1,2	.16	.03

2(d): King's method

	$P_{1\|0}$	$P_{1\|1}$
SAS 'truth'	.14	.07
covariate(s):		
none	.19	.00
age2		
age1, age2		
oo	.14	.07
oo, rla		
oo, rla, age1,2		

Source: 1991 UK census data.

Population: Residents aged 16 or more in households, Manchester LAD.

Y takes the value 1 for 'unemployed' 0 for 'not unemployed'.

X takes the value 1 for 'married' and 0 for 'not married'.

$P_{1|0}$ means $P(Y = 1 \mid X = 0)$; $P_{1|1}$ means $P(Y = 1 \mid X = 1)$

Adjustment variables: oo = owner occupied; rla = rented from local authority; age1 = persons aged 45 - 59; age2 = persons aged 60 and over.