Modelling the effect of covariates for unbiased estimates in ecological inference methods

Analisi degli effetti delle covariate per stime non distorte nei metodi di inferenza ecologica

Venera Tomaselli, Antonio Forcina and Michela Gnaldi

Abstract After showing that the estimates provided by three main ecological inference methods are heavily biased when compared to multilevel logistic models applied to a set of real individual data, the paper argues that ecological bias can be corrected only by accounting for relevant covariates. In addition, a data generating mechanism where bias cannot even be corrected by using covariates is described.

Abstract *Dopo aver dimostrato che le stime ottenute mediante i tre principali metodi di inferenza ecologica sono fortemente distorte rispetto a quelle ottenute applicando modelli logistici multilivello, lo studio conclude che le distorsioni possono essere corrette soltanto tenendo conto di covariate pertinenti. Inoltre, descritto un sistema di generazione di dati dove le distorsioni non possono essere corrette neppure usando covariate.*

Key words: ecological fallacy, biased estimates, covariates effects, multilevel logistic regression.

1 Ecological Fallacy and ecological inference models

Since Robinson's seminal paper [**?**], it is well known that the association between two variables estimated from data aggregated within geographical units, like polling stations, may be substantially biased compared to the association that would emerge

Venera Tomaselli (*corresponding author*)
Department of Political and Social Sciences, University of Catania, e-mail: tomavene@unict.it
Antonio Forcina
Department of Economics, University of Perugia, e-mail: forcinarosara@gmail.com
Michela Gnaldi
Department of Political Sciences, University of Perugia, e-mail: michela.gnaldi@unipg.it

if data recorded at the individual level were available. This phenomenon became known as the *ecological fallacy*.

Subramanian et al. [?] pointed out that, in certain contexts, the degree of association at the individual level may depend on modelling assumptions and thus may not be such an objective quantity as Robinson seemed to believe. An important implication of this result is that, when the estimates from ecological and individual level studies do not agree, additional investigation may be necessary before concluding that the ecological estimates are inappropriate.

Since the introduction of ecological regression by Goodman [?], several methods of ecological inference have been developed by King [?], [?] and coworkers [?]; their merits are debated [?], [?], [?]. Though less popular, the methods proposed by Brown and Payne [?] and Greiner and Quinn [?] may be considered as valid alternatives.

In a recent paper [?] the authors, by elaborating on the work of Wakefield [?] and Firebaugh [?], argue that, when certain assumptions are violated, the estimates produced by any method of ecological inference are going to be biased. Though bias might be corrected by modelling the effect of relevant covariates, even this may fail under certain data generating mechanisms.

2 Ecological Estimates by Logistic Multilevel Models

This paper is based on the analysis of an extensive set of individual data on voting behaviour from the Democratic Party primary election for the candidate mayor in the city of Palermo, Italy, in 2012. For each polling station, the data provide the joint distribution of voters classified by their decision (vote or not at the primary election) on one side and their age and sex on the other.

The estimates of ecological inference methods rely on the marginal distribution of voting decisions and that of sex, age; when individual data are available, these estimates can be compared with the actual proportions of voters within each age by sex group. In addition, by applying logistic multilevel models one can check if the propensity to vote depends on the relative size of the sex by age groups together with other covariates: when this happen it can be shown that ecological estimates are going to be biased,

The estimates of voting probabilities provided by the Goodman [?] regression model, the King [?] multinomial-Dirichlet and the modified Brown and Payne model [?] without covariates (Table 1) are substantially different from those based on individual data: for certain age groups, estimated probabilities are close to 0 while, for other age groups, they are much higher than the observed proportions.

Next we apply multilevel models [?],[?], [?] to the individual dataset with three objectives: (i) to verify whether the estimates provided by the raw proportions (used by Robinson) and those obtained from multilevel models are substantially different as in [?], (ii) to obtain an estimate of the different variance components and (iii) to detect the appropriate covariates to be used in the ecological inference models. For each polling station, voters are cross classified by sex (M, F) and 6 age groups and, for each of these 12 categories, according to their decision to vote or not. These data may be seen as 12 binomial variables nested within polling station, with polling stations grouped into 31 seats. For these data there are three sources of random variation:

(i) binomial within polling stations;
(ii) among polling stations within seats with an estimated standard deviation of 0.2311;
(iii) among seats with an estimated standard deviation of 0.2547.

To investigate how the propensity to vote depends on available covariates, several different models were explored and the following highly significant covariates were selected:

- *pd*, the proportion of voters for the Democratic Party at the municipal election held a month later (3.8% on average for all the polling stations within total eligible voters);
- *idv*, the proportion of voters for the *Italia dei Valori* Party at the same municipal election (5.1% on average for all the polling stations within total eligible voters);
- *mol*, the proportion of males aged between 45 and 74 (45.0% within male eligible voters);
- *fol*, the proportion of females aged between 45 and 74 (45.4% within female eligible voters).

By the logistic multilevel models fitted separately to each age group, where the observations at the lowest level are the number of voters (classified by sex and age group) nested within polling stations which, in turn, are nested within seats, as potentially relevant covariates, we considered the proportions of eligible voters belonging to each age group separately for males and females, in addition to the *pd* and *idv* covariates described above. The parameter estimates for the effect of the relevant covariates are displayed in Table 2.

Though the proportions of voters aged 45-65 and 65-75 were significant most of the times, when *pd* and *idv* were also used, some of the previous covariates appeared to no longer have a significant effect. This could be due to the fact that *pd* and *idv* are closely related to the age distribution within each polling station: when either *pd*

Table 1 Ecological inference estimates of the probability of voting by sex and age groups, without covariates.

	Sex	Age groups					
Method		18-25	25-30	30-45	45-65	65-75	over 75
Goodman	M	0.000	0.000	0.000	0.123	0.147	0.000
	F	0.000	0.000	0.000	0.124	0.000	0.028
Brown-Payne	M	0.000	0.000	0.000	0.000	0.278	0.000
(revised)	F	0.000	0.000	0.000	0.148	0.000	0.152
King OLS	M	0.001	0.001	0.000	0.002	0.264	0.007
	F	0.000	0.001	0.000	0.144	0.004	0.161

Table 2 Estimated parameters for the multilevel logistic models for the propensity to vote; F is the intercept within females, $M-F$ is the difference in intercept between males and females; $\circ$ = non significant, $\star$ = 5% significant, $*$ = 1% significant, $\bullet$ = p-value smaller than 0.001.

Parameters	Age groups					
	18-25	25-30	30-45	45-65	65-75	over 75
F	$-4.9207^{\bullet}$	$-4.5096^{\bullet}$	$-4.4335^{\bullet}$	$-3.6369^{\bullet}$	$-4.7301^{\bullet}$	$-5.0218^{\bullet}$
pd	$13.1413^{\bullet}$	$14.8040^{\bullet}$	$8.7711^{\bullet}$	$12.4372^{\bullet}$	$7.9891^{\bullet}$	$7.1119^{\bullet}$
idv	$4.5733^{\star}$	0.0000°	4.2559^{*}	$5.2272^{\bullet}$	$4.0019^{\star}$	$10.3799^{\bullet}$
$P(45-65)$	2.3815^{*}	2.6580^{*}	1.6830^{*}	0.0000°	$2.0258^{\star}$	0.0000°
$P(65-75)$	2.3888^{*}	0.0000°	1.9564^{*}	1.3247^{*}	$2.9426^{\bullet}$	0.0000°
$M-F$	0.0256°	-0.0570°	$0.0853^{\bullet}$	$0.0898^{\bullet}$	$0.4785^{\bullet}$	$0.8171^{\bullet}$

or idv increases, the proportion of eligible voters in the 18-45 age group decreases while the proportion in the age range from 45 to 75 and over increases.

The fact that in the logistic multilevel models applied to the individual data the propensity to vote depends significantly on covariates measured at the level of polling stations provides an explanation for the bias present in ecological inference estimates.

However, both the King and the Brown-Payne methods allow modelling the effect of covariates on the logit of the propensity to vote. A rather disappointing (but at the same time rather intriguing) feature of the Palermo data is that both the King and the Brown and Payne method continue to provide biased estimates even if we allow the logit of the propensity to vote to depend on the same covariates which were detected as significant in the logistic multilevel models applied to the individual data.

An important result of this paper is that the failure of covariates to correct the ecological bias is not a feature specific to the Palermo data set. To support this claim we describe a plausible data generating mechanism which may have been working in the Palermo Primary election and explain why, under these conditions, modelling the effect of covariates may not correct the bias.

Let q_{as} be the probability of voting at the primary election for an eligible voter with sex s and affiliation to center-left parties $a = 0,\ 1$. Let also v_{us} denote the proportion of eligible voters of sex s who are affiliated to the same parties in polling station u. Then, it is easily shown that:

$$\pi_{us1} = q_{0s}(1 - v_{us}) + q_{1s} v_{us}. \tag{1}$$

There are two important feature in this equation: (i) it depends on the proportion of voters affiliated to center-left parties (which cannot be observed), rather than on the proportion of females; (ii) the dependence is linear rather than logistic. Artificial electoral data based on equation (1) were generated at random as follows:

1. in each polling station split eligible voters at random between affiliated and not and then among females and males in such a way that sex and affiliation are correlated;
2. assign plausible values to the q_{as} probabilities;

3. generate random electoral data as in a revised Brown and Payne model.

Table 3 shows the estimated proportions of females and males voters using individual data and the Brown and Payne and King OLS models. These estimates are computed on an artificial data set generated according to the procedure described above. Because it contains 16,000 polling stations with 1,000 voters each, standard errors of the raw proportions in the individual data are very small (less than 0.0005). As a consequence, any new replication should produce, essentially, the same estimates. Since the differences between estimates from ecological and individual data are substantially large, relative to the very large sample size, they are certainly due to bias and not to random variation.

Table 3 Estimated proportions of voters in the artificial data; Ind=Individual data, BP=Brown and Payne and King=King OLS.

Females			Males		
Ind	BP	King	Ind	BP	King
0.0464	0.1012	0.1015	0.0548	0.0000	0.000

3 Conclusions

The findings in this paper indicate that the only possibility to correct ecological bias is to model the effect of covariates; also Liu [**?**] noted that the estimates from the King's model improved substantially by including certain covariates. However, while Liu was searching among all possible covariates, the results in this paper show that only covariates strongly correlated with the marginal proportions in the explanatory variables (sex and age in our context) are relevant.

An interesting result here is that, while the extended version of an ecologic inference model with covariates does provide a very accurate fit of the total number of voters in the primary election in each polling station, the estimated number of the same voters by sex and age groups are not much better than those obtained by the same model without covariates.

When voter's choices depend on covariates measured at the level of polling stations, any method of ecological inference that does not account for this is going to provide biased estimates. But even this may fail, as in the Palermo data and on a artificial data set generated at random according to a model which assumes that voting decisions depend on party affiliation rather than sex and age.

References

1. Brown, P. J. & Payne, C. D.: Aggregate data, ecological regression, and voting transitions. J. Am. Statist. Assoc. **81**, 452–460 (1986)
2. Cho, W. K. T.: If the assumption fits: A comment on the King ecological inference solution. Polit. Anal. **7**, 143–163 (1998)
3. Firebaugh, G.: A rule for inferring individual-level relationships from aggregate data. Am. Sociol. Rev. **43**, 557–572 (1978)
4. Forcina, A., Gnaldi, M., & Bracalente, B: A revised Brown and Payne model of voting behaviour applied to the 2009 elections in Italy. Stat. Methods Appl. **21**, 109–119 (2012)
5. Freedman, D.A., Klein, S.P., Ostland, M., & Roberts, M.R.: On solutions to the ecological inference problem. J. Amer. Statist. Assoc. **93**, 1518–1522 (1998)
6. Gnaldi, M., Tomaselli, V., & Forcina, A.: Ecological Fallacy and Covariates: New Insights based on Multilevel Modelling of Individual Data. I. Statist. Rev. **86**, (2018) doi:10.1111/insr.12244
7. Goldstein, H.: Multilevel Statistical Models. 4th ed. John Wiley & Sons, Chichester, UK (2011)
8. Goodman, L. A.: Ecological regressions and behavior of individuals. Am. Sociol. Rev., **18**, 351–367 (1953)
9. Greiner, J. D. & Quinn, K. M.: R×C ecological inference: Bounds, correlations, flexibility and transparency of assumptions. J. Roy. Statist. Soc. Ser. A **172**, 67–81 (2009)
10. Hox, J. J., Moerbeek, M., & van de Schoot, R.: Multilevel Analysis: Techniques and Applications. 2nd ed. Routledge, New York, NJ (2010)
11. King, G.: A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton NJ (1997)
12. King, G.: The future of ecological inference research: A comment on Freedman et al.. J. Amer. Statist. Assoc. **94**, 352–355 (1999)
13. King, G., Rosen, O., & Tanner, M. A.: Binomial-Beta hierarchical models for ecological inference. Sociol. Methods Res. **28**, 61–90 (1999)
14. Liu, B.: EI extended model and the fear of ecological fallacy. Sociol. Methods Res. **20**, 1–23 (2007)
15. Ng, K. W., Tian, G. L., & Tang, M. L.: Dirichlet and Related Distributions: Theory, Methods and Applications. John Wiley & Sons, Chichester UK (2011)
16. Robinson, W. S.: Ecological correlations and the behavior of individuals. . Am. Sociol. Rev. **15**, 351–357 (1950)
17. Rosen, O., Jiang, W., King, G., & Tanner, M. A.: Bayesian and frequentist inference for ecological inference: The R×C case. Stat. Neerl. **55**, 134–156 (2001)
18. Snijders, T. A. B. & Bosker, R. J.: Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage, London UK (2012)
19. Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N.:Revisiting Robinson: the perils of individualistic and ecologic fallacy. Int. J. Epidem. **38**, 342–360 (2009)
20. Wakefield, J.: Ecological inference for 2×2 tables (with discussion).J. Roy. Statist. Soc. Ser. A **167**, 1–42 (2004)