CPB Netherlands Bureau for Economic Policy Analysis

CPB Discussion Paper | 284

Maximum likelihood estimation of the Markov chain model with macro data and the ecological inference model

Arie ten Cate

Maximum likelihood estimation of the Markov chain model with macro data and the ecological inference model

Arie ten Cate *

September 15, 2014

Abstract

This paper merges two isolated bodies of literature: the Markov chain model with macro data (MacRae, 1977) and the ecological inference model (Robinson, 1950). Both are choice models. They have the same likelihood function and the same regression equation.

Decades ago, this likelihood function was computationally demanding. This has led to the use of several approximate methods. Due to the improvement in computer hardware and software since 1977, exact maximum likelihood should now be the preferred estimation method.

Key words: Aggregated data, Markov chain, Ecological inference, Computational statistics; EM algorithm

*Email: a.ten.cate@cpb.nl (ending November 2014) and arietencate@gmail.com. The author thanks Sander Muns and Marno Verbeek for their comments on earlier versions and Gary King for his willingness to reply to my questions about his book and his data.

Contents

1 Introduction

This paper merges two isolated bodies of literature: respectively about the first order Markov chain model with macro data and about the ecological inference model. They contain very few references to each other, while in fact they are equivalent in the sense that they have the same likelihood function and the same regression equation.

The second purpose of this paper is to disprove that the computational burden of the likelihood function (which is indeed heavy) is prohibitive. I have found only one study which uses this function (Steel et al. (2004)), although the computing speed has increased very much since 1977. This is due to improved hardware and also to improved software such as programming languages with built-in array operations.

In sections 2 and 3, the two models are described with their respective typical examples. followed by a common notation in section 4, linking the symbols to each of the two models. In sections 5 and 6 the likelihood function and the regression equation are given in the common notation, with references to the two bodies of literature.

In section 8 some numerical examples are given, followed by the conclusions in section 9. The appendices show some technical details.

All discussions below are limited to binary classifications. This is sufficient for the purpose. Also, most of the ecological literature is (double) binary.

2 The first order Markov chain model with macro data

A first order Markov chain model is a time series model for panels with discrete data.

In the binary case, at each time period the individuals in the panel are in one of two states. For instance, employed or unemployed: the probability for an individual to be employed in a given time period depends on being employed or not in the previous time period. (In a higher order model, the dependence is also on earlier time periods.)

With the original micro panel data, these probabilities can be estimated easily, based on (for each time period except the first) the cross tabulation of the state in that time period against the state in the previous time period.

With aggregated panel data this tabulation is not available; only the marginal frequencies are available and estimation of the probabilities is harder.

A short review of the subject's history over the past four decades is given. Lee et al. (1970) and Dent and Ballintine (1971) are forerunners, discussing mainly regression analysis. They also discuss an approximate maximum likelihood where the data are multinomial. In the seminal MacRae (1977), regression analysis and exact maximum likelihood are compared. She concludes:

> While it is possible to develop computational algorithms to search for a maximum [likelihood], the iterative generalized least squares estimator may represent a better combination of numerical and statistical efficiency.

Rosenqvist (1986) discusses the combined use of micro and macro data and has a large list of references. Crowder and Stephens (2011) review the history of theoretical and applied work on the subject (though not including MacRae (1977)) and conclude that the computation of the exact likelihood function is "unfeasible" (p.3202).

3 The ecological inference model

Here, the word "ecological" has little to do with subjects like pollution, extinction of species, growing crops without chemicals, etcetera. Rather, as the title of King (1997) indicates, it is about "reconstructing individual behavior from aggregate data".

Hence, by definition in an ecological inference model the data are aggregated. Time usually does not play a role and instead of time periods we usually have regions. In the 2×2 case we have two dichotomies, with data for each region. One of the two dichotomies plays the role of the lagged information in the Markov chain model; this dichotomy might refer to a property of individuals which is constant over their life.

The standard example of this dichotomy is race, where the other dichotomy is political preference: the probability of having a given political preference depends on one's race. The data consist of the two marginal frequency distributions, by race and by political preference, for multiple regions. Naturally, where political preferences are expressed by some secret

ballot, they are only available in the form of frequencies by election district, and not in the form of individual records with personal characteristics.

The seminal paper is Robinson (1950). Wakefield (2004a) discusses the background and the state of the research at the time. He notes that the exact likelihood "has rarely been explicitly considered in the ecological inference literature." (top of p.391). See also the introductory chapter in King et al. (2004). To the best of my knowledge, Steel et al. (2004) are the first to apply the exact likelihood.

4 The notation

The Markov chain model

In the Markov chain model, we have $I+1$ observations over time. This gives a chain of I transitions from one time period to the next. Referring to the employment example above, the x_i are the first I frequencies of employed individuals and the y_i are the last I frequencies, giving

$$x_i = y_{i-1} \tag{1}$$

Note that we have no initial conditions problem here.

The symbol p_{1i} is the probability that an individual included in the number x_i is also included in the number y_i. Likewise, p_{2i} is the probability that an individual *not* included in the number x_i is included in the number y_i. Below I will use "unit" as the general term for the i index.

Ignoring panel attrition, the n_i series in the Markov chain model is constant over i, say $n_i = n$. With also constant probabilities p_1 and p_2, the fraction y_i/n moves, with random ups and downs, in the direction of $P = p_2/(1 - p_1 + p_2)$, the solution of $p_1P + p_2(1 - P) = P$.

The ecological inference model

Here, typically the units are regions, without inherent ordering, indexed by $i = 1, \ldots, I$. Unit i has n_i individuals. Using the above standard example in the ecological model, y_i is the number of individuals in unit i who have the reference political preference and x_i is the number of individuals with the reference race.

Table 1: The frequencies in unit i

lagged (Markov chain) or race (typical Ecological)	unlagged (Markov chain) or political preference (typical Ecological)		
	in reference state	else	total
in reference state	k_i	$x_i - k_i$	x_i
else	$y_i - k_i$	$n_i - x_i - y_i + k_i$	$n_i - x_i$
total	y_i	$n_i - y_i$	n_i

Notes. Only totals are observed. The unobserved frequencies are indexed by k_i. The word "state" does not refer to the states of the United States.

Hence p_{1i} is the probability that an individual in unit i with the reference race has the reference political preference and p_{2i} is the probability that an individual in i, not with the reference race, has the reference political preference.

The relations between the frequencies

Table 1 shows the frequencies. Since the data are aggregated, only the totals are observed; the remaining four numbers are not observed. However, if any one of these four numbers would be known then the other three would also be known. Without loss of generality I choose k_i as the index of all possible sets of four unobserved frequencies, given the observed totals.

Exogenous variables

The probabilities may depend on macro exogenous variables as follows. For all i:

$$p_{1i} = F(\boldsymbol{z}_{i1}\boldsymbol{\beta}_1) \qquad \text{and} \qquad p_{2i} = F(\boldsymbol{z}_{i2}\boldsymbol{\beta}_2) \tag{2}$$

where the $\boldsymbol{z}_{i1}$ and $\boldsymbol{z}_{i2}$ are rows from exogenous data matrices $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, respectively. (These two matrices may have columns in common.) The $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown column vectors of parameters. (They may have elements in common.) The function F is a strictly monotonous transformation from the real line to the zero-one range, such as the logit and probit functions. See MacRae (1977), p.185.

5 The likelihood function

The unconditional distribution[1] of k_i and of $y_i - k_i$; i.e., not considering y_i as given, is:

$$\begin{array}{ll} \Pr\left(k_i|x_i, \boldsymbol{\beta}_1\right) & = \mathrm{B}(k_i, \quad x_i, \quad p_{1i}) \\ \Pr\left(y_i - k_i|x_i, \boldsymbol{\beta}_2\right) & = \mathrm{B}(y_i - k_i, \; n_i - x_i, \; p_{2i}) \end{array} \tag{3}$$

taking into account formula's (2). The B indicates the binomial probability with $\mathrm{B}(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$.

The y_i are distributed as follows:

$$\Pr\left(y_i|x_i, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) = \sum_{k_i} \Pr\left(k_i|x_i, \boldsymbol{\beta}_1\right) \Pr\left(y_i - k_i|x_i, \boldsymbol{\beta}_2\right) \tag{4}$$

where the range of the summation is

$$\max(0, x_i + y_i - n_i) \;\leq\; k_i \;\leq\; \min(x_i, y_i) \tag{5}$$

The likelihood function is:

$$L\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) = \prod_{i=1}^{I} \Pr\left(y_i|x_i, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) \tag{6}$$

For the first order Markov chain model, see MacRae (1977), with the general case, not limited to binary choice. For the ecological model, see for instance Wakefield (2004a) equation (4), discussed at the top of page 391 (as noted above) and Wakefield (2004b), equation (1.6). Steel et al. (2004) give the first order derivatives of the loglikelihood and the Fisher information matrix, with a numerical application.

This likelihood is computationally more demanding than ordinary binary choice models. The computing time of (6) is roughly equal to the computing time of an ordinary binary choice model, times the sum over the units i of the size of the range of k_i in (5). Of course, this sum might be anything from a few hundred to, say, a few hundred thousand.

In order to distinguish between this likelihood and its approximations[2], in the binary case this likelihood is often called the convolution likelihood,

[1] For brevity I write for example $\Pr(x) = \phi(x)$, instead of the more precise $\Pr(x{=}X) = \phi(X)$.

[2] A quite different distribution of k_i and $y_i - k_i$ (given x_i) is presented in King (1997), pp.93/94: the fractions k_i/x_i and $(y_i - k_i)/(n_i - x_i)$ are drawn from a truncated bivariate normal distribution. Compare the first line of the first formula on his page 308 with our (4). This model has been adapted and refined; see King et al. (2004), section 0.1.4.

named after the convolution sum in discrete form in the right-hand side of (4).

6 Least squares regression

With (3) we have:

$$\begin{aligned} \mathrm{E}\left[y_i|x_i,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right] &= \mathrm{E}\left[k_i|x_i,\boldsymbol{\beta}_1\right]+\mathrm{E}\left[y_i-k_i|x_i,\boldsymbol{\beta}_2\right] \\ &= x_i p_{1i}+\left(n_i-x_i\right)p_{2i} \end{aligned} \tag{7}$$

With constant p_1 and p_2 we have for all i:

$$\frac{\mathrm{dE}\left[y_i|x_i,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right]}{\mathrm{d}x_i}=p_1-p_2 \tag{8}$$

In words: the difference between the two probabilities is a reflection of the correlation over the regions between x_i and y_i.

Equation (7) is a regression equation with error variances

$$\mathrm{V}\left[y_i|x_i,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right]=x_i p_{1i}\left(1-p_{1i}\right)+\left(n_i-x_i\right)p_{2i}\left(1-p_{2i}\right) \tag{9}$$

Hence equation (7) can be estimated with nonlinear Feasible Generalized Least Squares (FGLS) by minimizing

$$\sum_i \frac{\left(y_i-\mathrm{E}[y_i|x_i,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2]\right)^2}{\mathrm{V}\left[y_i|x_i,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right]} \tag{10}$$

See MacRae (1977), p.189/190.

In section 8.1 below, this regression estimate will be compared with the maximum likelihood estimate. As we shall see, the regression estimate and the maximum likelihood estimate converge to each other with increasing sample size[3].

[3] Loosely speaking this follows from: (a) with increasing n, a binomial mass distribution converges to a normal density distribution and (b) the convolution integral of two normal density distributions is again a normal density distribution and (c) with normally distributed y, least squares gives the same result as maximum likelihood. See also for instance Wakefield (2004b), p.20/21.

7 The Ecological Fallacy

One of the examples below is about the following identification problem. A given correlation between x_i and y_i can result from two quite different models, with the same number of parameters. The first model is the basic ecological inference model, with constant parameters p_1 and p_2; the direction of the correlation depends on which parameter is the largest, as expressed by equation (8).

The second model is a special case of (2), with only one $\boldsymbol{\beta}$ parameter vector:

$$p_{1i} = p_{2i} = F(\beta_1 + \beta_2 x_i) \tag{11}$$

The direction of the correlation depends on the sign of β_2. A person's individual property does not influence that person's behaviour; only the unit is relevant. This can be important for the choice of a policy measure[4].

Unfortunately it is hard to distinguish empirically between (8) and (11). If F is the identity function (with $F(a) = a$) then in both models the expectation of y_i is a linear function of x_i. On the other hand, the variance of y_i differs between the two. Hence, with least squares regression, identification depends on the feeble basis of technicalities as the F function and the variance function. In section 8.2 the two models are estimated with maximum likelihood, allowing the comparison of their likelihoods.

Estimating the basic ecological inference model of (8) while in fact model (11) is true, is an example of the Ecological Fallacy, widely discussed in the ecological inference literature. Slightly more general, we might have in (11) a z_i variable which is correlated with x_i, instead of x_i itself.

8 Examples

As noted above, Steel et al. (2004) apply the exact likelihood function. In their Example section, they select $I = 50$ out of the 1541 Census Districts

[4] With the Markov example of unemployment, the relation with (economic) policy is obvious. For the ecological inference example of the racial aspect of voting, see for instance King (1997), p.8: "Under present law, legally significant discrimination only exists when plaintiffs (or the Justice Department) can first demonstrate that members of a minority group (usually African American or Hispanic) vote both cohesively and differently from other voters. Sometimes they must also prove that majority voters consistently prevent minorities from electing a candidate of their choice."

Table 2: Percentage employment for the Markov chain model

year	%	year	%
1986	40.0	1991	47.6
1987	40.6	1992	51.5
1988	42.5	1993	52.5
1989	43.2	1994	52.6
1990	44.8	1995	55.7

Source: Pelzer et al. (2001)

of Brisbane (Australia), with on average $22323/50 = 446$ individuals. They don't report on the computing time, but this suggests that for this problem size, the likelihood function is feasible.

In order to report computing times and to illustrate some issues, I computed two other examples: a Markov chain model with time series data and some ecological inference experiments with data from King (1997).

8.1 Markov chain: employment of women

I apply the Markov chain model to the data of Pelzer et al. (2001), with emphasis on the relation between maximum likelihood and least squares regression.

Although their paper is not about panels, their data come from a panel. The panel contains women who are either employed or unemployed. The employment percentages are in table 2. The first nine percentages are the x_i/n_i series and the last nine are the y_i/n_i series. The sample size n_i (not shown here) changes somewhat over time, with an average of 2200 persons.

The estimates are in table 3, with simulated sample sizes. A contour map of the loglikelihood shows one maximum, surrounded by convex contour lines.

As discussed above: unlike the maximum likelihood estimate, the least squares estimate hardly changes with the sample size. The difference between the two decreases with increasing sample size.

Last but not least: the computing times are short. For the largest sample size (22000), the computing time was well under ten seconds on a PC from the year 2010. I used the built-in array facilities of the R language, with the `nlm` function. I did not program analytical derivatives and started with all β

Table 3: Estimates of probabilities for the Pelzer data (percent)

	$n_i = n$	$1 - p_1$	p_2	long run $y_i/n = p_2/(1 - p_1 + p_2)$
maximum likelihood	220	1.6	4.6	75
	2200	8.5	10.5	55
	22000	9.8	11.7	54
FGLS regression	any	10	12	54

Note: when employed: p_1 = Pr(keeping employment);
when unemployed: p_2 = Pr(finding employment)

at zero; see appendix D. Assuming that the computing time is proportional with $\sum n_i$, the example of Steel et al. (2004) would take me less than $10 \times 22323/(22000 \times 9) \approx 1$ second. With all of Brisbane, it would take less than half a minute.

8.2 Ecological inference: to vote or not to vote

Figure 4.1 in King (1997) shows a scatter diagram of the fraction Black (x_i) against the fraction Voter Turnout (y_i) in a US Senate election in the precincts of Marion County (Indiana). A file with these data is included among the data files supplied by King. We have here $\sum n_i \approx 446{,}000$. The fraction $\sum x_i / \sum n_i$ of Black people is 16%; the fraction $\sum y_i / \sum n_i$ of voters is 32%. The (unweighted) correlation between x_i and y_i is -0.20. For more details about the data file, see appendix E.

I estimated both models discussed in section 7 above. See table 4. The computer program was quite similar to appendix D. As it should be, I found both $p_1 - p_2$ and β_2 negative, the same sign as the correlation between x_i and y_i. The standard errors in the table are computed from the "observed Fisher information matrix". The model (11) has the largest loglikelihood. The likelihood is a probability here, unlike with models with continuous stochastic variables. For both likelihoods in table 4, exp(loglikelihood/$\sum n_i$) = 94.8% (rounded).

Not shown in the table: for the model (11), the weighted average of the estimated probability to vote is 32%.

Again, last but not least: the computing time was well under 10 seconds.

Table 4: Maximum likelihood estimates for the King data

	model (8)		model (11)	
	p_1	p_2	β_1	β_2
	%			
estimate	26	33	−0.69	−0.39
estimated standard error	0.2	0.1	0.004	0.01
loglikelihood	−23794		−23741	

Note: when Black: $p_1 = \Pr(\text{voting})$;
when not Black: $p_2 = \Pr(\text{voting})$

As noted above, this holds also for the maximum likelihood estimate of table 3 with the largest n_i. Hence, if $\sum n_i$ is a few hundred thousand, up to half a million, it takes not more than 10 seconds on my PC from 2010.

9 Conclusions

The first order Markov chain model with macro data can be considered as a special case of the ecological inference model, with the two classifications of the ecological model being the same, recorded at two subsequent time periods. Students of this Markov model might consult Steel et al. (2004) for details of the exact likelihood function, which can be translated to the Markov model using the current paper.

Students of ecological inference might have benefited from reading MacRae (1977) at that time.

Both might do well no longer to dismiss out of hand the exact likelihood as unfeasible.

The maximum likelihood estimate differs most from the least squares regression in small samples, where the computation of the likelihood function is less of a problem

Remaining work: find conditions for a stationary point of the loglikelihood being unique.

A The $\mathcal{E}$ and $\mathcal{V}$ operators

In the next two appendices, we shall make use of an operator indicated by $\mathcal{E}$, operating on a function of k_i, as follows. For a function $f(.)$, $\mathcal{E}f(k_i)$ is a weighted average over k_i:

$$\begin{aligned}
\mathcal{E}\, f(k_i) &\equiv \frac{\sum_{k_i} f(k_i)\Pr\left(k_i|x_i,p_{1i}\right)\Pr\left(y_i-k_i|x_i,p_{2i}\right)}{\sum_{k_i}\ \Pr\left(k_i|x_i,p_{1i}\right)\Pr\left(y_i-k_i|x_i,p_{2i}\right)} \\
&= \frac{\sum_{k_i} f(k_i)\Pr\left(k_i|x_i,p_{1i}\right)\Pr\left(y_i-k_i|x_i,p_{2i}\right)}{\Pr\left(y_i|x_i,p_{1i},p_{2i}\right)} \\
&= \sum_{k_i} f(k_i)\,\mathrm{FNH}(k_i;n_i,y_i,x_i,\omega_i) \qquad (12)
\end{aligned}$$

where FNH is Fisher's noncentral hypergeometric distribution (or the "extended hypergeometric distribution") with noncentrality parameter

$$\omega_i = \frac{p_{1i}(1-p_{2i})}{(1-p_{1i})p_{2i}} \qquad (13)$$

The y_i and x_i in the last line of (12) can be swapped. If evaluated with the parameters at their true value, this is the conditional expectation of k_i (conditional on the observed value y_i).

With

$$\mathcal{V}\, k_i \equiv \mathcal{E}\left[(k_i-\mathcal{E}k_i)^2\right] = \mathcal{E}\left[k_i^2\right] - (\mathcal{E}k_i)^2 \qquad (14)$$

we have:

$$\frac{\partial \mathcal{E}k_i}{\partial \omega} = \frac{\mathcal{V}k_i}{\omega} \qquad (15)$$

as a property of Fisher's noncentral hypergeometric distribution.

The R package BiasedUrn provides functions which compute $\mathcal{E}k_i$ and $\mathcal{V}k_i$ respectively.

B The EM algorithm

Below it is shown how to apply the EM algorithm to the likelihood function discussed in this paper. In each iteration, the EM algorithm temporarily reduces this likelihood to the likelihood of an ordinary binary choice model.

The EM algorithm was introduced in Dempster et al. (1977) and has been widely used ever since. It is a slow but robust method of maximizing

a likelihood such as ours, with missing, or "latent", data. Whether or not its robustness is needed depends on whether or not there might be multiple stationary points of the loglikelihood; this question remains to be answered. The EM algorithm is not discussed in King et al. (2004) or its second edition (2012).

Without loss of generality I choose the k_i $(i = 1, \ldots, I)$ as the latent class variables. Let $\boldsymbol{\beta}_1^{(\nu)}$ and $\boldsymbol{\beta}_2^{(\nu)}$ be the parameter vectors obtained from the last M step. Then the E step of iteration ν of the EM algorithm is the sum over i of (12) with ω_i computed from $\boldsymbol{\beta}_1^{(\nu)}$ and $\boldsymbol{\beta}_2^{(\nu)}$ and with

$$\begin{aligned} f(k_i) &= \log \Pr\left(k_i | x_i, \boldsymbol{\beta}_1\right) \Pr\left(y_i - k_i | x_i, \boldsymbol{\beta}_2\right) \\ &= \log \Pr\left(k_i | x_i, \boldsymbol{\beta}_1\right) + \log \Pr\left(y_i - k_i | x_i, \boldsymbol{\beta}_2\right) \\ &= \log \binom{x_i}{k_i} \binom{n_i - x_i}{y_i - k_i} \\ &\quad + k_i \log p_{1i} + (x_i - k_i) \log\left(1 - p_{1i}\right) \\ &\quad + (y_i - k_i) \log p_{2i} + (n_i - x_i - (y_i - k_i)) \log\left(1 - p_{2i}\right) \end{aligned} \tag{16}$$

Hence, the E step is:

$$\begin{aligned} \sum_i \mathcal{E} f(k_i) &= \sum_i \mathcal{E}\left[\log \binom{x_i}{k_i} \binom{n_i - x_i}{y_i - k_i}\right] \\ &+ \sum_i \left\{k_i^{(\nu)} \log p_{1i} + \left(x_i - k_i^{(\nu)}\right) \log\left(1 - p_{1i}\right)\right\} \qquad (17) \\ &+ \sum_i \left\{\left(y_i - k_i^{(\nu)}\right) \log p_{2i} + \left(n_i - x_i - (y_i - k_i^{(\nu)})\right) \log\left(1 - p_{2i}\right)\right\} \qquad (18) \end{aligned}$$

with

$$k_i^{(\nu)} = \mathcal{E}\left[k_i \left| \boldsymbol{\beta}_1^{(\nu)}, \boldsymbol{\beta}_2^{(\nu)}, y_i, x_i \right.\right] \tag{19}$$

Only lines (17) and (18) depend on the parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, through p_{1i} and p_{2i}.

In the M step, this is maximized over the parameters. This can be done by maximizing lines (17) and (18) separately: with given $k_i^{(\nu)}$ and hence also given $y_i - k_i^{(\nu)}$, we have now two ordinary binary choice models.

Without $\boldsymbol{z}$ variables (with all $p_{1i} = p_1$ and all $p_{2i} = p_2$) the result of the M step is:

$$\boldsymbol{\beta}_1^{(\nu+1)} = p_1^{(\nu+1)} = \sum_i k_i^{(\nu)} \Big/ \sum_i x_i \tag{20}$$

$$\boldsymbol{\beta}_2^{(\nu+1)} = p_2^{(\nu+1)} = \sum_i \left(y_i - k_i^{(\nu)}\right) \Big/ \sum_i (n_i - x_i) \tag{21}$$

and we iterate between (19) and (20) + (21).

C The derivatives of the loglikelihood

With positive $L_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ we have (with conditionals after a vertical bar omitted for brevity):

$$\begin{aligned}
\frac{\partial \log L_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_1} &= \frac{1}{\Pr(y_i)} \frac{\partial \Pr(y_i)}{\partial \boldsymbol{\beta}_1} \\
&= \frac{1}{\Pr(y_i)} \sum_{k_i} \frac{\partial \Pr(k_i)}{\partial \boldsymbol{\beta}_1} \Pr(y_i - k_i) \\
&= \frac{1}{\Pr(y_i)} \sum_{k_i} \frac{\partial \log \Pr(k_i)}{\partial \boldsymbol{\beta}_1} \Pr(k_i) \Pr(y_i - k_i) \\
&= \mathcal{E} \frac{\partial \log \Pr(k_i)}{\partial \boldsymbol{\beta}_1} = \mathcal{E}\left[\frac{\mathrm{d} \log \Pr(k_i)}{\mathrm{d}p_{1i}} \frac{\partial p_{1i}}{\partial \boldsymbol{\beta}_1}\right] \\
&= \mathcal{E}\left[\frac{\mathrm{d} \log \Pr(k_i)}{\mathrm{d}p_{1i}}\right] \frac{\partial p_{1i}}{\partial \boldsymbol{\beta}_1} \qquad (22)
\end{aligned}$$

with

$$\mathcal{E} \frac{\mathrm{d} \log \Pr(k_i)}{\mathrm{d}p_{1i}} = \mathcal{E}\left[\frac{k_i}{p_{1i}} - \frac{x_i - k_i}{1 - p_{1i}}\right] = \frac{\mathcal{E}k_i}{p_{1i}} - \frac{x_i - \mathcal{E}k_i}{1 - p_{1i}} = \frac{\mathcal{E}k_i - p_{1i}x_i}{p_{1i}(1 - p_{1i})} \qquad (23)$$

Similarly we have for $\boldsymbol{\beta}_2$:

$$\frac{\partial \log L_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2} = \mathcal{E}\left[\frac{\mathrm{d} \log \Pr(y_i - k_i)}{\mathrm{d}p_{2i}}\right] \frac{\partial p_{2i}}{\partial \boldsymbol{\beta}_2} \qquad (24)$$

with

$$\mathcal{E} \frac{\mathrm{d} \log \Pr(y_i - k_i)}{\mathrm{d}p_{2i}} = \frac{y_i - \mathcal{E}[k_i] - p_{2i}(n_i - x_i)}{p_{2i}(1 - p_{2i})} \qquad (25)$$

The last member of (23) and of (25) are given by Steel et al. (2004), p.58, summed over i. (Of course, here, with varying $\partial p_{1i}/\partial \boldsymbol{\beta}_1$ and $\partial p_{2i}/\partial \boldsymbol{\beta}_2$, this summation is not meaningful.)

For a single unit i, there is a stationary point which is a saddle point, with $p_{1i} = p_{2i}$; see for instance the various 3D graphs in King et al. (2004) and figure 6(a) at page 403 of Wakefield (2004a).

For the second order derivatives, use might be made of (15).

D The R program of section 8.1

```
print(paste("n=",n <- 2200))
minimand <- function(beta) {
  value <- 0
  for (i in 1:NROW(x)) {
    # A = reference (employed)
    xA <- x[i]
    xB <- n-x[i]
    yA <- y[i]
    pAA <- toZeroOneRange(beta[1])
    pBA <- toZeroOneRange(beta[2])
    AA  <- max(0,yA-xB) : min(xA,yA)  # paper: k_i
    BA  <- yA - AA
    if (maxlik) {
      liki <- sum(dbinom(AA,xA,pAA) * dbinom(BA,xB,pBA))
      value <- value - log(liki)
    } else {
      residual <- yA - (xA*pAA + xB*pBA)
      residVar <- xA*pAA*(1-pAA) + xB*pBA*(1-pBA)
      value <- value + (residual^2)/residVar
    }
  }
  return(value)
}
toZeroOneRange <- function(x) {1/(1+exp(-x))}
series <- c(0.406,0.425,0.432,0.448,0.476,0.515,0.525,0.526)
y <- round(n * c(series,0.557))
x <- round(n * c(0.400,series))
for (maxlik in c(TRUE,FALSE)) {
  print(paste("maxlik=",maxlik))
  start <- c(0,0)
  result <- nlm(minimand, start, hessian=TRUE)
  print(paste("return code=",result$code))
  print(paste("beta=",beta <- result$estimate))
  print(paste("p1=",  p1   <- toZeroOneRange(beta[1])))
  print(paste("p2=",  p2   <- toZeroOneRange(beta[2])))
  if (maxlik) {
    print(paste("stderr",sqrt(-diag(solve(-result$hessian)))))
  }
  print(paste("long run=",p2/(1-p1+p2)))
}
```

E The data of section 8.2

I used data file in90.asc from `http://doi.org/10.3886/ICPSR01132.v1`. In this file, four records have a turnout fraction y_i/n_i of more than 100%, for reasons such as the census being collected at a different point in time than electoral data (private communication with the book's author). I removed these records, with $I = 657$ records remaining.

These data were chosen because this file is one of the three ascii files in the book's set of data files. (I could not read the binary Gauss files with Gauss 13 for Windows, or the unix files or the Windows NT files.) The two other ascii data files are pa90.asc (with more and much larger errors in the turnout fraction) and hisp.asc (without the n_i).

References

Crowder, M. and Stephens, D. (2011). On inference from Markov chain macro-data using transforms. *Journal of Statistical Planning and Inference*, 141:3201–3216.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Dent, W. and Ballintine, R. (1971). A review of the estimation of transition probabilities in Markov chains. *The Australian journal of agricultural economics*, 15:69–81.

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

King, G., Rosen, O., and Tanner, M. A. (2004). *Ecological Inference: New Methodological Strategies*. Cambridge University Press, 1st edition.

Lee, T., Judge, T., and Zellner, A. (1970). *Estimating the Parameters of the Markov Probability Model From Aggregate Time Series Data*. North Holland.

MacRae, E. C. (1977). Estimation of time-varying Markov processes with aggregate data. *Econometrica*, 45:183–198.

Pelzer, B., Eisinga, R., and Franses, P. H. (2001). Estimating transition probabilities from a time series of independent cross sections. *Statistica Neerlandica*, 55:249–262.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357.

Rosenqvist, G. (1986). *Micro and Macro Data in Statistical Inference on Markov Chains*. PhD thesis, Publications of the Swedish School of Economics and Business Administration, nr 36.

Steel, D. G., Beh, E. J., and Chambers, R. L. (2004). The information in aggregate data. In King, G., Rosen, O., and Tanner, M. A., editors, *Ecological Inference: New Methodological Strategies*, chapter 2, pages 51–68. Cambridge University Press, 1st edition.

Wakefield, J. (2004a). Ecological inference for 2×2 tables. *Journal of the Royal Statistical Society, Series A*, 167, Part 3:385–445.

Wakefield, J. (2004b). Prior and likelihood choices in the analysis of ecological data. In King, G., Rosen, O., and Tanner, M. A., editors, *Ecological Inference: New Methodological Strategies*, chapter 1, pages 13–50. Cambridge University Press, 1st edition.

Publisher:

CPB Netherlands Bureau for Economic Policy Analysis
P.O. Box 80510 | 2508 GM The Hague
T (070) 3383 380

September 2014 | ISBN 978-90-5833-654-5