Southampton Statistical Sciences Research Institute
University of Southampton

The Information in Aggregate Data

David G. Steel, Eric J. Beh, Ray L. Chambers

Abstract

Ecological analysis involves using aggregate data for a set of groups to make inferences concerning individual level relationships. Typically the data available for analysis consists of the means or totals of variables of interest for geographical areas, although the groups can be organisations such as schools or hospitals. Attention has focused on developing methods of estimating the parameters characterising the individual level relationships across the whole population, but also in some cases the relationships for each of the groups.

Applying standard methods used to analyse individual level data, such as linear or logistic regression or contingency table analysis, to aggregate data will usually produce biased estimates of individual level relationships. Thus much of the effort in ecological analysis has concentrated on developing methods of analysing aggregate data that can produce unbiased, or less biased, parameter estimates. There has been less work done on inference procedures, such as constructing confidence intervals and hypothesis testing. Fundamental to these inferential issues is the question of how much information is contained in aggregate data and what evidence such data can provide concerning important assumptions and hypotheses.

S^{3}RI Methodology Working Paper M03/14

CHAPTER 1

The Information in Aggregate Data

David G. Steel[†], Eric J. Beh[†] and Ray L. Chambers[‡]
† School of Mathematics and Applied Statistics,
University of Wollongong,
Wollongong, NSW, 2522,
Australia.
‡ Department of Social Statistics,
University of Southampton,
Southampton, S017 1BJ,
United Kingdom.

1.1 Introduction

Ecological analysis involves using aggregate data for a set of groups to make inferences concerning individual level relationships. Typically the data available for analysis consists of the means or totals of variables of interest for geographical areas, although the groups can be organisations such as schools or hospitals. Attention has focused on developing methods of estimating the parameters characterising the individual level relationships across the whole population, but also in some cases the relationships for each of the groups.

Applying standard methods used to analyse individual level data, such as linear or logistic regression or contingency table analysis, to aggregate data will usually produce biased estimates of individual level relationships. Thus much of the effort in ecological analysis has concentrated on developing methods of analysing aggregate data that can produce unbiased, or less biased, parameter estimates. There has been less work done on inference procedures, such as constructing confidence intervals and hypothesis testing. Fundamental to these inferential issues is the question of how much information is contained in aggregate data and what evidence such data can provide concerning important assumptions and hypotheses.

In Section 2 we describe a general approach to determining the information in aggregate data and how it compares with the information in individual level data

for likelihood based inference, including hypothesis testing. In Section 3 we illustrate how the approach applies in the case of data from several 2×2 tables. We also consider the information contributed by aggregate and individual information when both are available in Section 4. Section 5 gives empirical results based on some real data, illustrating the loss of information due to aggregation and how hypothesis testing and analysis of residuals can be done using aggregate data. Section 6 provides a brief discussion.

1.2 Information Lost by Aggregation

Suppose that we have individual level data $d^{(1)}$, which has associated probability function $f^{(1)}\left(d^{(1)};\psi\right)$. The vector ψ contains the parameters of the distribution of the individual level data. Likelihood inference about the parameter vector ψ would be based on the likelihood $L^{(1)}\left(\psi;d^{(1)}\right)=f^{(1)}\left(d^{(1)};\psi\right)$ or the associated log-likelihood

$$l^{(1)}\left(\psi;d^{(1)}\right)=\log L^{(1)}\left(\psi;d^{(1)}\right)$$

The score function for ψ based on $d^{(1)}$ is

$$\mathrm{sc}^{(1)}\left(\psi;d^{(1)}\right)=\frac{\partial}{\partial\psi}l^{(1)}\left(\psi;d^{(1)}\right) \tag{1.1}$$

Maximum likelihood estimates (MLEs) would usually be obtained by solving

$$\mathrm{sc}^{(1)}\left(\psi;d^{(1)}\right)=0 \tag{1.2}$$

resulting in the MLE $\hat{\psi}$.

For inference based on the MLEs we would also be interested in the (observed) information matrix

$$\begin{aligned}\mathrm{info}^{(1)}\left(\psi;d^{(1)}\right) &= -\frac{\partial}{\partial\psi}\mathrm{sc}^{(1)}\left(\psi;d^{(1)}\right)\\ &= -\frac{\partial^2}{\partial\psi\partial\psi^T}l^{(1)}\left(\psi;d^{(1)}\right)\end{aligned} \tag{1.3}$$

The expected information is

$$\mathrm{Info}^{(1)}\left(\psi;d^{(1)}\right)=E\left[\mathrm{info}^{(1)}\left(\psi;d^{(1)}\right)\right] \tag{1.4}$$

The expectation is over the distribution of $d^{(1)}$. Under several regularity conditions the variance matrix of the asymptotic distribution of $\hat{\psi}$ is $\left[\text{Info}^{(1)}\right]^{-1}$ (see for example Cox and Hinkley, 1974, Chapter 9).

Suppose we are interested in testing the hypothesis H_0. Let $\hat{\psi}_0$ be the MLE of ψ under H_0. There are three common approaches to testing H_0.

1. Likelihood Ratio Test (LRT) is based on the likelihood ratio

$$R^{(1)} = \frac{L^{(1)}\left(\hat{\psi}_0; d^{(1)}\right)}{L^{(1)}\left(\hat{\psi}; d^{(1)}\right)}$$

   and

$$-2\log R^{(1)} = 2\left[l^{(1)}\left(\hat{\psi}; d^{(1)}\right) - l^{(1)}\left(\hat{\psi}_0; d^{(1)}\right)\right]$$

   is tested against the χ^2_q distribution with $q = \dim\{\psi\} - \dim\{\psi_0\}$.
2. Wald Test is based on

$$W^{(1)} = \left(\hat{\psi} - \hat{\psi}_0\right)^T \left[\text{Info}^{(1)}\left(\hat{\psi}; d^{(1)}\right)\right]\left(\hat{\psi} - \hat{\psi}_0\right)$$

3. Score Test is based on

$$ST^{(1)} = \text{sc}^{(1)}\left(\hat{\psi}_0; d^{(1)}\right)^T \left[\text{Info}^{(1)}\left(\hat{\psi}_0; d^{(1)}\right)\right]^{-1} \text{sc}^{(1)}\left(\hat{\psi}_0; d^{(1)}\right)$$

The score test does not require the calculation of $\hat{\psi}$, only $\hat{\psi}_0$, which in some situations will be an advantage over the Wald test. However, the Wald test does not require inversion of the information matrix. All these tests may be used to produce confidence regions for ψ. Efron and Hinkley (1978) argue that it is preferable to use the observed rather than the expected information matrix for inference. We will follow this approach.

Instead of individual level data we have available the aggregate data $d^{(2)}$. Let $f^{(2)}\left(d^{(2)}; \psi\right)$ denote the associated probability function. Likelihood based inference can then be undertaken using $f^{(2)}$. In general, deriving $f^{(2)}$ from $f^{(1)}$ may be difficult. Since $f^{(2)}$ is derived from $f^{(1)}$ it will depend on the same parameters as $f^{(1)}$. However, not all these parameters may be identifiable using aggregate data.

We assume that the individual level data set comprises n individuals divided into m groups. In general, the n individuals are obtained from a sample of individuals, $S^{(1)}$, and the sample of m groups is $S^{(2)}$. The sample of individuals in group g is S_g. An important special case is when the samples are the entire finite population, i.e. $S^{(1)} = U^{(1)}$, $S^{(2)} = U^{(2)}$ and $S_g = U_g$. We will assume that any sampling involved is ignorable, for example simple random sampling.

Breckling, Chambers, Dorfman, Tam and Welsh (1994) described an approach for maximum likelihood inference using sample data. Sampling is a process by

which data are unobserved or reduced and aggregation is also a process that leads to the observed data being reduced. The basic results of Breckling et al. (1994) can then be applied to examine the effect of using aggregate data.

Let $\mathrm{sc}^{(2)}\left(\psi; d^{(2)}\right)$ and $\mathrm{info}^{(2)}\left(\psi; d^{(2)}\right)$ be the score function and observed information matrix based on $d^{(2)}$. The key results of Breckling et al. (1994) are

$$\mathrm{sc}^{(2)}\left(\psi; d^{(2)}\right) = \mathrm{E}\left[\mathrm{sc}^{(1)}\left(\psi; d^{(1)}\right) | d^{(2)}\right] \tag{1.5}$$

$$\mathrm{info}^{(2)}\left(\psi; d^{(2)}\right) = \mathrm{E}\left[\mathrm{info}^{(1)}\left(\psi; d^{(1)}\right) | d^{(2)}\right] - \mathrm{Var}\left[\mathrm{sc}^{(1)}\left(\psi; d^{(1)}\right) | d^{(2)}\right] \tag{1.6}$$

The expectations in (1.5) and (1.6) are over the distribution of $d^{(1)}$ conditional on $d^{(2)}$, that is the individual level data given the aggregate data. Hypothesis testing can also be done using this score function and information matrix as well as the likelihood based on $d^{(2)}$.

In some cases using (1.5) to obtain the score function may be more convenient than direct differentiation of $l^{(2)} = \log f^{(2)}$. Result (1.6) is the key to determining the information loss due to the use of aggregate data. The variance-covariance matrix of the individual level score function conditional on $d^{(2)}$ can be interpreted as the loss of information due to aggregation. In Section 3 we will illustrate this approach for the case of m 2×2 tables, but the result can be applied in general.

1.3 Several 2×2 Tables

1.3.1 Data Available

Suppose that the individual level data consists of m 2×2 tables giving the frequencies associated with two dichotomous variables, Y and X. Table 1.1 illustrates the data for group g.

X/Y	$Y = 1$	$Y = 0$	Total
$X = 1$	n_{11g}	n_{12g}	$n_{1\bullet g}$
$X = 0$	n_{21g}	n_{22g}	$n_{2\bullet g}$
Total	$n_{\bullet 1g}$	$n_{\bullet 2g}$	n_g

Table 1.1 *Individual Level Data for Group g*

It is assumed that the marginal frequencies for X are fixed, or conditioned upon, and that the values of Y are independent given X. Hence, for group g

$$n_{11g} \sim \text{Bin}\left(n_{1\bullet g}, \pi_{1g}\right) \qquad n_{21g} \sim \text{Bin}\left(n_{2\bullet g}, \pi_{2g}\right)$$

where $\pi_{1g} = \text{Prob}\left(Y = 1 | X = 1\right)$ and $\pi_{2g} = \text{Prob}\left(Y = 1 | X = 0\right)$ for group g. The associated odds ratio is

$$\theta_g = \frac{\pi_{1g}}{1 - \pi_{1g}} \frac{1 - \pi_{2g}}{\pi_{2g}}$$

Let $d_g^{(1)} = \left\{n_{11g}, n_{1\bullet g}, n_{\bullet 1g}, n_g\right\}$ be the individual level data for group g and $d^{(1)} = \left\{d_g^{(1)}, g \in S^{(2)}\right\}$ be the entire individual level data set. In ecological inference the individual level data are not available, so the n_{11g} values are not available. However, the marginal frequencies and n_g are available giving the aggregate data $d_g^{(2)} = \left\{n_{1\bullet g}, n_{\bullet 1g}, n_g\right\}$ for group g and $d^{(2)} = \left\{d_g^{(2)}, g \in S^{(2)}\right\}$ for the m groups.

1.3.2 Analysis Using Individual Level Data

Let $\phi_g = \left(\pi_{1g}, \pi_{2g}\right)^T$ and $\psi = \left[\phi_1^T, \ldots, \phi_m^T\right]^T$. If no assumptions are made concerning the parameters ϕ_g, each table could be analysed separately with individual level data. The likelihood for ϕ_g based on $d_g^{(1)}$ is denoted $L_g^{(1)}\left(\phi_g; d_g^{(1)}\right)$ and the log-likelihood is

$$\begin{aligned} l_g^{(1)}\left(\phi_g; d_g^{(1)}\right) &= n_{11g} \log \pi_{1g} + n_{12g} \log\left(1 - \pi_{1g}\right) \\ &\quad + n_{21g} \log \pi_{2g} + n_{22g} \log\left(1 - \pi_{2g}\right) \end{aligned}$$

The individual level score function for ϕ_g is

$$\text{sc}^{(1)}\left(\phi_g; d_g^{(1)}\right) = \begin{bmatrix} \frac{n_{11g} - n_{1\bullet g}\pi_{1g}}{\pi_{1g}\left(1-\pi_{1g}\right)} \\ \frac{n_{\bullet 1g} - n_{11g} - n_{2\bullet g}\pi_{2g}}{\pi_{2g}\left(1-\pi_{2g}\right)} \end{bmatrix} \tag{1.7}$$

The resulting MLEs are $\hat{\phi}_g = \left(\hat{\pi}_{1g}, \hat{\pi}_{2g}\right)^T = \left(\frac{n_{11g}}{n_{1\bullet g}}, \frac{n_{\bullet 1g} - n_{11g}}{n_{2\bullet g}}\right)^T$. The observed information matrix is

$$\text{info}^{(1)}\left(\phi_g; d_g^{(1)}\right) = \begin{bmatrix} \frac{n_{11g}\left(1-2\pi_{1g}\right)+n_{1\bullet g}\pi_{1g}^2}{\pi_{1g}^2\left(1-\pi_{1g}\right)^2} & 0 \\ 0 & \frac{\left(n_{\bullet 1g}-n_{11g}\right)\left(1-2\pi_{2g}\right)+n_{2\bullet g}\pi_{2g}^2}{\pi_{2g}^2\left(1-\pi_{2g}\right)^2} \end{bmatrix} \tag{1.8}$$

and the expected information matrix is

$$\text{Info}^{(1)}\left(\phi_g; d_g^{(1)}\right) = \begin{bmatrix} \frac{n_{1\bullet g}}{\pi_{1g}\left(1-\pi_{1g}\right)} & 0 \\ 0 & \frac{n_{2\bullet g}}{\pi_{2g}\left(1-\pi_{2g}\right)} \end{bmatrix} \tag{1.9}$$

It may be of interest to test whether there is evidence that the tables are homogeneous with respect to the conditional probabilities, i.e. $\pi_{1g} = \pi_1$, $\pi_{2g} = \pi_2$ for $g \in S^{(2)}$, which can be written as $\phi_g = \phi = \left(\pi_1, \pi_2\right)^T$ for all $g \in S^{(2)}$. This hypothesis may be of substantive interest or it may be convenient for further analysis and interpretation. For example, if we have a sample of groups then assuming group specific parameters means that no inferences can be made concerning groups that are not in the sample. Even if all groups in the population of interest are included in $S^{(2)}$, the large number of groups may make interpretation of the analysis difficult if each group is assumed to have different parameter values. One approach to this issue is to allow for variation in ϕ_g by including random effects, but for non-linear models, this introduces considerable complexities in the analysis.

If $\phi_g = \phi$, then the log-likelihood for ϕ based on $d^{(1)}$ is

$$\begin{aligned} l^{(1)}\left(\phi; d^{(1)}\right) &= \sum_{g \in S^{(2)}} l_g^{(1)}\left(\phi; d_g^{(1)}\right) \\ &= n_{11\bullet}\log\pi_1 + n_{12\bullet}\log\left(1-\pi_1\right) + n_{21\bullet}\log\pi_2 + n_{22\bullet}\log\left(1-\pi_2\right) \end{aligned}$$

Hence the tables can be collapsed and the analysis can be based on the 2×2 table for the entire sample, $S^{(1)}$. The MLEs, score and information functions are as in (1.7), (1.8) and (1.9) with the g for the elements of $d^{(1)}$ replaced with the summation subscript $\bullet$. That is

$$\text{sc}^{(1)}\left(\phi; d^{(1)}\right) = \begin{bmatrix} \frac{n_{11\bullet}-n_{1\bullet\bullet}\pi_1}{\pi_1\left(1-\pi_1\right)} \\ \frac{n_{\bullet 1\bullet}-n_{11\bullet}-n_{2\bullet\bullet}\pi_2}{\pi_2\left(1-\pi_2\right)} \end{bmatrix} \tag{1.10}$$

$$\left.\begin{array}{rcl} \text{info}_{11}^{(1)}\left(\phi; d^{(1)}\right) & = & \frac{n_{11\bullet}\left(1-2\pi_1\right)+n_{1\bullet\bullet}\pi_1^2}{\pi_1^2\left(1-\pi_1\right)^2} \\ \text{info}_{21}^{(1)}\left(\phi; d^{(1)}\right) & = & 0 \\ \text{info}_{22}^{(1)}\left(\phi; d^{(1)}\right) & = & \frac{\left(n_{\bullet 1\bullet}-n_{11\bullet}\right)\left(1-2\pi_2\right)+n_{2\bullet\bullet}\pi_2^2}{\pi_2^2\left(1-\pi_2\right)^2} \end{array}\right\} \tag{1.11}$$

The resulting MLEs are $\hat{\phi} = \left(\frac{n_{11\bullet}}{n_{1\bullet\bullet}}, \frac{n_{\bullet 1\bullet}-n_{11\bullet}}{n_{2\bullet\bullet}}\right)^T$.

The hypothesis $\phi_g = \phi$ can be tested using the likelihood ratio, Wald or score test. The latter two can be based on the observed or expected information matrix. Also the likelihood can be directly examined to see what evidence it provides (see Royall, 1997). For example, when the tables are homogeneous, $\psi_0 = \left[\phi^T, \ldots, \phi^T\right]^T$ and the score test using the observed information matrix is

$$\begin{aligned} \text{ST}^{(1)} &= \sum_{g\in S^{(2)}} \text{sc}^{(1)}\left(\hat{\phi}; d_g^{(1)}\right)^T \left[\text{info}^{(1)}\left(\hat{\phi}; d_g^{(1)}\right)\right]^{-1} \text{sc}^{(1)}\left(\hat{\phi}; d_g^{(1)}\right) \\ &= \sum_{g\in S^{(2)}} \text{ST}_g^{(1)} \end{aligned}$$

The likelihood ratio is

$$R^{(1)} = \prod_{g\in S^{(2)}} \frac{L_g^{(1)}\left(\hat{\phi}; d_g^{(1)}\right)}{L_g^{(1)}\left(\hat{\phi}_g; d_g^{(1)}\right)} = \prod_{g\in S^{(2)}} R_g^{(1)}$$

1.3.3 Analysis Using Aggregate Data

In ecological inference the data available from each table are $d_g^{(2)}$ so that n_{11g} is not available. We could attempt an analysis without making any assumptions concerning ϕ_g. This amounts to analysing each group separately. Applying (1.5) to (1.7) immediately gives

$$\text{sc}^{(2)}\left(\phi_g; d_g^{(2)}\right) = \begin{bmatrix} \frac{\text{E}\left(n_{11g}|d_g^{(2)}\right)-n_{1\bullet g}\pi_{1g}}{\pi_{1g}\left(1-\pi_{1g}\right)} \\ \frac{n_{\bullet 1g}-\text{E}\left(n_{11g}|d_g^{(2)}\right)-n_{2\bullet g}\pi_{2g}}{\pi_{2g}\left(1-\pi_{2g}\right)} \end{bmatrix}$$

Conditional on $d_g^{(2)}$, n_{11g} has a non-central hypergeometric distribution (see for example McCullagh and Nelder, 1989, pg 257-259) and

$$\mathrm{E}\left(n_{11g}|\,d_g^{(2)}\right)=\frac{P_1\left(\theta_g;d_g^{(2)}\right)}{P_0\left(\theta_g;d_g^{(2)}\right)}$$

where

$$P_r(\theta_g;d_g^{(2)})=\sum_{j=a_g}^{b_g}\binom{n_{1\bullet g}}{j}\binom{n_{2\bullet g}}{n_{\bullet 1g}-j}j^r\theta_g^j$$

The limits of the sum are the lower and upper bounds on n_{11g} given $d_g^{(2)}$ and are $a_g=\max\left(0,n_{\bullet 1g}-n_{2\bullet g}\right)$ and $b_g=\min\left(n_{1\bullet g},n_{\bullet 1g}\right)$. Denote $\mathrm{E}\left(n_{11g}|\,d_g^{(2)}\right)$ by $\kappa_1\left(\theta_g;d_g^{(2)}\right)$. Also

$$\mathrm{Var}\left(n_{11g}|\,d_g^{(2)}\right)=\frac{P_2\left(\theta_g;d_g^{(2)}\right)}{P_0\left(\theta_g;d_g^{(2)}\right)}-\kappa_1\left(\theta_g;d_g^{(2)}\right)^2$$

which will be denoted by $\kappa_2\left(\theta_g;d_g^{(2)}\right)$.

From (1.7)

$$\mathrm{Var}\left(\mathrm{sc}^{(1)}\left(\phi_g;d_g^{(1)}\right)|\,d_g^{(2)}\right)=\kappa_2\left(\theta_g;d_g^{(2)}\right)\begin{bmatrix}\frac{1}{\pi_{1g}^2\left(1-\pi_{1g}\right)^2} & \frac{-1}{\pi_{1g}\pi_{2g}\left(1-\pi_{1g}\right)\left(1-\pi_{2g}\right)}\\ \frac{-1}{\pi_{1g}\pi_{2g}\left(1-\pi_{1g}\right)\left(1-\pi_{2g}\right)} & \frac{1}{\pi_{2g}^2\left(1-\pi_{2g}\right)^2}\end{bmatrix}$$

Applying (1.6) with (1.7) and (1.8) gives

$$\begin{aligned}\mathrm{info}_{11}^{(2)}\left(\phi_g;d_g^{(2)}\right) &= \frac{\kappa_1\left(\theta_g;d_g^{(2)}\right)\left(1-2\pi_{1g}\right)+n_{1\bullet g}\pi_{1g}^2-\kappa_2\left(\theta_g;d_g^{(2)}\right)}{\pi_{1g}^2\left(1-\pi_{1g}\right)^2}\\ \mathrm{info}_{21}^{(2)}\left(\phi_g;d_g^{(2)}\right) &= \frac{\kappa_2\left(\theta_g;d_g^{(2)}\right)}{\pi_{1g}\pi_{2g}\left(1-\pi_{1g}\right)\left(1-\pi_{2g}\right)}\\ \mathrm{info}_{22}^{(2)}\left(\phi_g;d_g^{(2)}\right) &= \frac{\left(n_{\bullet 1g}-\kappa_1\left(\theta_g;d_g^{(2)}\right)\right)\left(1-2\pi_{2g}\right)+n_{2\bullet g}\pi_{2g}^2-\kappa_2\left(\theta_g;d_g^{(2)}\right)}{\pi_{2g}^2\left(1-\pi_{2g}\right)^2}\end{aligned}$$

Setting $\mathrm{sc}^{(2)}\left(\phi_g; d_g^{(2)}\right) = 0$ yields the relationship

$$\pi_{1g} n_{1\bullet g} + \pi_{2g} n_{2\bullet g} = n_{\bullet 1g}$$

or

$$\pi_{2g} = \frac{n_{\bullet 1g}}{n_{2\bullet g}} - \frac{n_{1\bullet g}}{n_{2\bullet g}} \pi_{1g} \tag{1.12}$$

which corresponds to the tomography line for group g discussed in King (1997, pg 80) .

The aggregation of the data has resulted in each element of the information matrix being modified by a term proportional to $\kappa_2\left(\theta_g; d_g^{(2)}\right)$ arising from the conditional variance of the individual level score function. Also n_{11g} is replaced by its expectation conditional on $d_g^{(2)}$.

For each group there is only one observed random variable, $n_{\bullet 1g}$ and two parameters unless some further assumptions are made. For m groups there are m observations, $n_{\bullet 1g}$, $g \in S^{(2)}$, but $2m$ parameters. Hence standard asymptotic properties of likelihood based methods cannot be relied upon. Beh, Steel and Booth (2002) consider the likelihood associated with aggregate data for a single group. This is given by McCullagh and Nelder (1989, pg 353) :

$$L_g^{(2)}\left(\phi_g; d_g^{(2)}\right) = \left(1 - \pi_{1g}\right)^{n_{1\bullet g}} \pi_{2g}^{n_{\bullet 1g}} \left(1 - \pi_{2g}\right)^{n_{2\bullet g} - n_{\bullet 1g}} P_0\left(\theta_g; d_g^{(2)}\right) \tag{1.13}$$

Wakefield (2001) uses the same likelihood, but presents it in the form of a convolution likelihood of two binomials.

Beh, Steel and Booth (2002) show that the likelihood surface has a ridge along the tomography line (1.12). Along the tomography line the likelihood is minimised when $\pi_{1g} = \pi_{2g}$, i.e. at independence, and the maximum occurs at one of the ends of the tomography line. They also show that except for cases when $n_{\bullet 1g}$ is very close to $n_{1\bullet g}$ or $n_{2\bullet g}$ the likelihood surface is not able to provide useful evidence concerning the values of π_{1g} and π_{2g} other than they should be on the tomography line. Notice that the score and information function in this case can also be obtained directly from the likelihood $L_g^{(2)}$ given by (1.13).

Beh, Steel and Booth (2002) obtain the exact values of $\hat{\phi}_g = \left(\hat{\pi}_{1g}, \hat{\pi}_{2g}\right)^T$ that maximise the likelihood. Wakefield (2001) also obtains these values using an approximation. The resulting maximum of the likelihood $L_g^{(2)}\left(\hat{\phi}_g; d_g^{(2)}\right)$ can also be obtained. Notice $\hat{\phi}_g$ is unique, except when $n_{1\bullet g} = n_{2\bullet g}$, in which case the likelihood is maximised at $(0, 1)^T$ and $(1, 0)^T$.

The inferential problem that arises from wishing to estimate $2m$ parameters from

m observations can be tackled if we assume $\phi_g = \phi$ for all $g \in S^{(2)}$. Of course this is a very strong assumption and it is more realistic to assume that ϕ_g varies in some way across the m groups. The variation may be related to group level covariates z_g and random effects. However, analysis is relatively straight forward if this homogeneity assumption holds. More importantly the question arises of whether, in practice, it is possible from aggregate data alone to assess whether the homogeneity assumption is reasonable before attempting to use methods that allow for variation in ϕ_g.

When $\phi_g = \phi$, we can obtain the score and information functions based on the aggregate data for the m groups in the sample by applying (1.5) and (1.6) to (1.10) and (1.11) or summing the score and information functions arising from each group with $\phi_g = \phi$. This gives

$$\mathrm{sc}^{(2)}\left(\phi; d^{(2)}\right) = \begin{bmatrix} \frac{\sum_g \kappa_1\left(\theta; d_g^{(2)}\right) - n_{1\bullet\bullet}\pi_1}{\pi_1\left(1-\pi_1\right)} \\ \frac{n_{\bullet 1\bullet} - \sum_g \kappa_1\left(\theta; d_g^{(2)}\right) - n_{2\bullet\bullet}\pi_2}{\pi_2\left(1-\pi_2\right)} \end{bmatrix}$$

$$\begin{aligned}
\mathrm{info}_{11}^{(2)}\left(\phi; d^{(2)}\right) &= \frac{\sum_g \kappa_1\left(\theta; d_g^{(2)}\right)\left(1-2\pi_1\right) + n_{1\bullet\bullet}\pi_1^2 - \sum_g \kappa_2\left(\theta; d_g^{(2)}\right)}{\pi_1^2\left(1-\pi_1\right)^2} \\
\mathrm{info}_{12}^{(2)}\left(\phi; d^{(2)}\right) &= \frac{\sum_g \kappa_2\left(\theta; d_g^{(2)}\right)}{\pi_1\pi_2\left(1-\pi_1\right)\left(1-\pi_2\right)} \\
\mathrm{info}_{22}^{(2)}\left(\phi; d^{(2)}\right) &= \frac{\left(n_{\bullet 1\bullet} - \sum_g \kappa_1\left(\theta; d_g^{(2)}\right)\right)\left(1-2\pi_2\right) + n_{2\bullet\bullet}\pi_2^2 - \sum_g \kappa_2\left(\theta; d_g^{(2)}\right)}{\pi_2^2\left(1-\pi_2\right)^2}
\end{aligned}$$

Setting $\mathrm{sc}^{(2)}\left(\phi; d^{(2)}\right) = 0$ gives the overall sample level tomography line

$$\pi_1 n_{1\bullet\bullet} + \pi_2 n_{2\bullet\bullet} = n_{\bullet 1\bullet}$$

The correlation between the two elements of the individual level score function conditional on $d^{(2)}$, obtained from $\mathrm{Var}\left[\mathrm{sc}^{(1)}\left(\phi; d^{(1)}\right) \mid d^{(2)}\right]$, is -1 and corresponds to the constraint arising from the tomography line.

Comparing $\mathrm{info}^{(2)}$ with $\mathrm{info}^{(1)}$ we see that in addition to the reduction in the diagonal elements, a positive term appears in the off-diagonal elements. This suggests that inferences concerning $\pi_1 - \pi_2$ will be particularly badly affected.

The same score function can be obtained directly from the likelihood of the aggregate data

$$L^{(2)}\left(\phi; d^{(2)}\right) = \prod_{g \in S^{(2)}} L_g^{(2)}\left(\phi; d_g^{(2)}\right)$$

McCullagh and Nelder (1989, pg 353) obtain an equivalent score function for a different parameterisation.

The equations $\mathrm{sc}^{(2)}\left(\phi; d^{(2)}\right) = 0$ can be solved to obtain the estimates $\hat{\phi} = \left(\hat{\pi}_1, \hat{\pi}_2\right)^T$ under the hypothesis of homogeneity. This can be done in several ways as reviewed by Beh and Steel (2002). Here we obtain the estimate of ϕ using the Newton-Raphson iterative procedure

$$\phi^{(j+1)} = \phi^{(j)} - \alpha A^{-1}\left(\frac{\partial l}{\partial \phi}\right)\Bigg|_{\phi=\phi^{(j)}}$$

with the secant approximation of hessian matrix A to accelerate convergence. Reddien (1986) comments that the use of this approximation is often preferred to the standard Newton-Raphson procedure and that its rate of convergence is both satisfactory and stable. The value of α is chosen such that $0 \leq \alpha \leq 1$ and dictates the step length taken at iteration of the procedure (see McCulloch and Searle, 2001, pg 269).

Once an estimate of the common probabilities, $\hat{\phi}$, is obtained we can produce estimates of the group specific proportions $P_{1g} = n_{11g}/n_{1\bullet g}$ and $P_{2g} = n_{21g}/n_{2\bullet g}$ by evaluating the expectation $\mathrm{E}\left[n_{11g} | d^{(2)}\right] = \kappa_1\left(\hat{\theta}; d_g^{(2)}\right)$ where $\hat{\theta}$ is the odds ratio calculated from $\hat{\phi}$. This gives the estimates $\hat{P}_{1g} = \kappa_1\left(\hat{\theta}; d_g^{(2)}\right)/n_{1\bullet g}$ and $\hat{P}_{2g} = \left(n_{\bullet 1g} - \kappa_1\left(\hat{\theta}; d_g^{(2)}\right)\right)/n_{2\bullet g}$. For each group these estimates of the proportions are obtained by projecting the estimates of the common probabilities, $\hat{\phi}$, onto the tomography line (1.12) for that group, using the expectation of the non-central hypergeometric distribution $\kappa_1\left(\hat{\theta}; d_g^{(2)}\right)$.

The likelihood ratio for testing the hypothesis $\phi_g = \phi$ is

$$R^{(2)} = \prod_{g \in S^{(2)}} \frac{L_g^{(2)}\left(\hat{\phi}; d_g^{(2)}\right)}{L_g^{(2)}\left(\hat{\phi}_g; d_g^{(2)}\right)} = \prod_{g \in S^{(2)}} R_g^{(2)}$$

We will not use the Wald test as $\mathrm{info}^{(2)}$ is not defined at the $\hat{\phi}_g$ values. The score test based on the observed information matrix is

$$ST^{(2)} = \sum_{g \in S^{(2)}} \mathrm{sc}^{(2)}\left(\hat{\phi}; d_g^{(2)}\right)^T \left[\mathrm{info}^{(2)}\left(\hat{\phi}; d_g^{(2)}\right)\right]^{-1} \mathrm{sc}^{(2)}\left(\hat{\phi}; d_g^{(2)}\right) = \sum_{g \in S^{(2)}} ST_g^{(2)}$$

1.4 Using Aggregate and Unit Level Data

In some situations it may be feasible to obtain both individual level and aggregate data. For example, we may have a reasonably large number of groups and could consider conducting a small sample of individuals to supplement the aggregate data. Alternatively, we could have a reasonable sized sample of individuals and consider supplementing it by some aggregate data. The latter case could be useful in producing estimates of group specific quantities. This leads to the general issue of what is the relative value of the two types of data. This can help us decide at what sample size the information in the aggregate data has little additional value.

Suppose that we have a simple random sample, $S^{(0)}$ of n_0 individuals selected from the population of interest. We assume that the sampling fraction is small so that we can treat the data in $S^{(0)}$ as independent of that in $S^{(2)}$. The sample $S^{(0)}$ produces the data $d^{(0)}$. The aggregate and individual level data can be combined, giving $d^{(c)} = \left\{ d^{(2)}, d^{(0)} \right\}$. Because of the independence of the data sets the score function and information matrices can be added giving

$$\begin{aligned} \mathrm{sc}^{(c)}\left(\psi; d^{(c)}\right) &= \mathrm{sc}^{(2)}\left(\psi; d^{(2)}\right) + \mathrm{sc}^{(0)}\left(\psi; d^{(0)}\right) \\ \mathrm{info}^{(c)}\left(\psi; d^{(c)}\right) &= \mathrm{info}^{(2)}\left(\psi; d^{(2)}\right) + \mathrm{info}^{(0)}\left(\psi; d^{(0)}\right) \end{aligned}$$

Consider the case of m 2×2 tables. Suppose that the group that each individual comes from is not known. This could be for reasons of confidentiality or because the sample was selected in a way that did not make recording the groups convenient. Then $d^{(0)} = \left\{ n_{11}^{(0)}, n_{1\bullet}^{(0)}, n_{\bullet 1}^{(0)}, n^{(0)} \right\}$.

Assuming $\phi_g = \phi$ the information associated with $d^{(0)}$ is

$$\left.\begin{aligned} \mathrm{info}_{11}^{(0)}\left(\phi; d^{(0)}\right) &= \frac{n_{11}^{(0)}\left(1-2\pi_1\right)+n_{1\bullet}^{(0)}\pi_1^2}{\pi_1^2\left(1-\pi_1\right)^2} \\ \mathrm{info}_{21}^{(0)}\left(\phi; d^{(0)}\right) &= 0 \\ \mathrm{info}_{22}^{(0)}\left(\phi; d^{(0)}\right) &= \frac{\left(n_{\bullet 1}^{(0)}-n_{11}^{(0)}\right)\left(1-2\pi_2\right)+n_{2\bullet}^{(0)}\pi_2^2}{\pi_2^2\left(1-\pi_2\right)^2} \end{aligned}\right\}$$

The addition of the unit level data increases the diagonal elements of the information matrix and leaves the off-diagonal elements unchanged. Besides reducing the asymptotic variance of the estimates of π_1 and π_2 this will also dampen the correlation of the estimates resulting in additional benefits for the estimation of $\pi_1 - \pi_2$.

1.5 Example

To illustrate the application of these results we will consider a simple example using data from the 1996 Australian census. The data corresponds to the census district (CD) level data for the city of Brisbane in Australia where the individual level classifications are known. There are a total of 1541 CDs, but for simplicity we will focus our discussion on a random sample of 50 CDs.

For comparison, King's method is also applied to these data using the $\mathrm{E_ZI}$ package (Benoit and King, 1998) with its default global parameters.

Consider the data with variables "Income" and "Age" so that for CD g the classification of individuals is

$X = 1$; if a person is aged between 15 and 24 years,

$X = 0$; if a person is aged at least 25 years,

$Y = 1$; if a person's weekly income is between \$AU0 and \$AU159,

$Y = 0$; if a person's weekly income is at least \$AU160.

For the 50 CDs considered there are 22323 individuals classified, with 4238 individuals aged between 15 and 24 years and 5674 with a weekly income of between \$AU0 and \$AU159. These values correspond to the marginal frequencies $n_{\bullet}$, $n_{1\bullet\bullet}$ and $n_{\bullet 1\bullet}$ respectively. The proportion of people aged between 15 and 24 was 0.1898 and varied from 0.1053 to 0.2861 with a coefficient of variation 0.1970. A plot of the values of the group specific proportions P_{1g} and P_{2g} is given in Figure 1.1 and shows a considerable amount of variation.

Based on the individual level data we obtain $\hat{\pi}_1^{(1)} = 0.5054$ and $\hat{\pi}_2^{(1)} = 0.1953$ which have estimated standard errors of 0.0077 and 0.0029 respectively.

Based only on the aggregate level data the maximum likelihood estimates assuming homogeneous parameters using the accelerated Newton-Raphson iterative procedure gives $\hat{\pi}_1^{(2)} = 0.5184$ and $\hat{\pi}_2^{(2)} = 0.1922$ and estimated standard errors of 0.0353 and 0.0085 respectively. The initial values of π_1 and π_2 were set at 0.6 and 0.1966 so that the overall tomography line is satisfied. Instability of the convergence was experienced with $\alpha = 1$ so smaller steps throughout the iterative procedure were carried out with $\alpha = 0.4$. Using King's (1997) method via $\mathrm{E_ZI}$ produced estimates $\tilde{\pi}_1^{(2)} = 0.4769$ and $\tilde{\pi}_2^{(2)} = 0.2020$ with estimated standard errors of 0.1606 and 0.0376 respectively. The point estimates obtained from the two methods are quite similar although there is a large difference between the estimated standard errors. This may be due to the random effects incorporated into the King method while our approach does not include any random variation in the group specific parameters.

The estimates of the group specific proportions P_{1g} and P_{2g} using King's approach

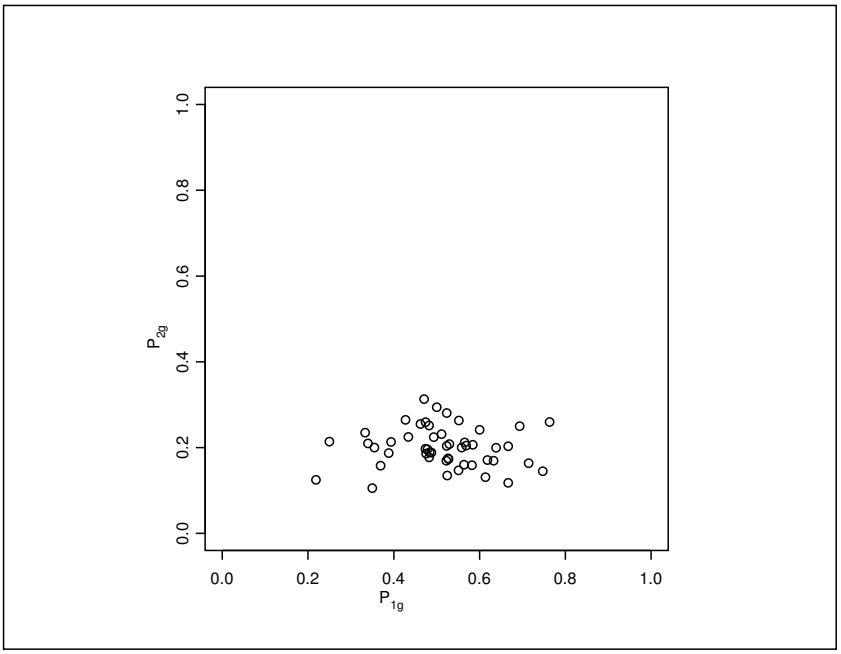

Figure 1.1 *Plot of P_{2g} versus P_{1g}*

and assuming homogeneity of the associated probabilities are very similar. In the latter approach, even though the probabilities π_{1g} and π_{2g} are assumed to be constant across the groups, the associated proportions, P_{1g} and P_{2g} are not assumed to be constant across the groups. Figure 1.2 compares the individual level proportions P_{1g} and P_{2g} with the estimates $\hat{P}_{1g}$ and $\hat{P}_{2g}$ obtained by considering the expectation $\mathrm{E}\left[n_{11g}|\, d_g^{(2)}\right]$ using the parameter values $\hat{\pi}_1^{(2)}$ and $\hat{\pi}_2^{(2)}$, that is $\kappa_1\left(\hat{\theta}^{(2)}; d_g^{(2)}\right)$. These values are very similar to those produced when estimating P_{1g} and P_{2g} using King's approach and these are produced in Figure 1.3. Chambers and Steel (2001) considered using the relative root-mean-squared errors

$$V_1 = \frac{1}{\hat{\pi}_1^{(1)}}\sqrt{m^{-1}\sum_g\left(\hat{P}_{1g} - P_{1g}\right)^2} \qquad V_2 = \frac{1}{\hat{\pi}_2^{(1)}}\sqrt{m^{-1}\sum_g\left(\hat{P}_{2g} - P_{2g}\right)^2}$$

to assess how well these estimates reproduce the true values. For the method assuming homogeneity between the groups $V_1 = 0.1993$ and $V_2 = 0.1204$, while King's method produces the similar values $V_1 = 0.2066$ and $V_2 = 0.1317$. This indicates that for these CDs there is no advantage in allowing for group heterogeneity in the conditional probabilities.

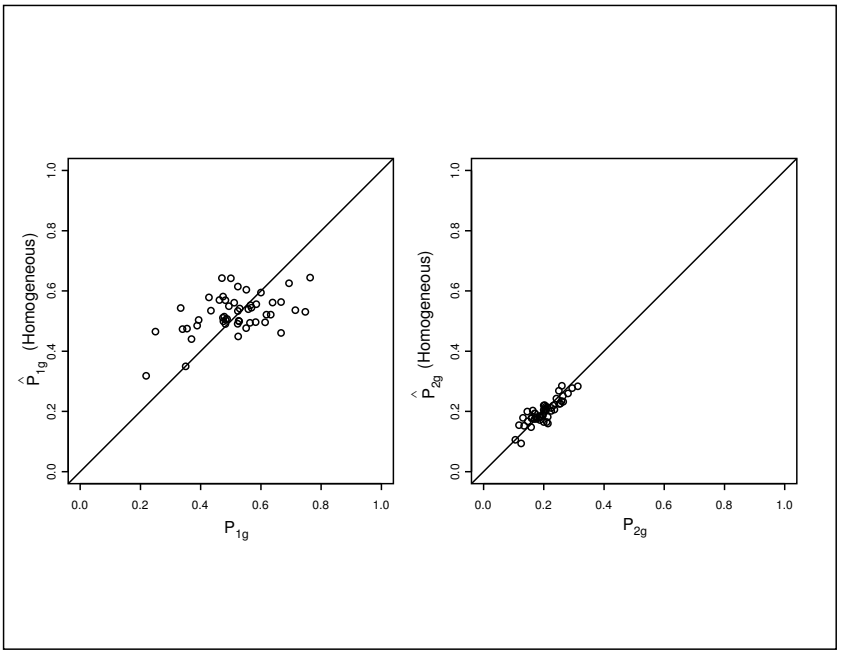

Figure 1.2 *Plot of* $\hat{P}_{1g}$ *versus* P_{1g} *and* $\hat{P}_{2g}$ *versus* P_{2g} *using* $\kappa_1\left(\hat{\theta}^{(2)}; d_g^{(2)}\right)$

Based on the individual level parameter estimates $\hat{\pi}_1^{(1)}$ and $\hat{\pi}_2^{(1)}$ the information matrix and its inverse are

$$\text{info}^{(1)} = \begin{pmatrix} 16953.96 & 0 \\ 0 & 115075.3 \end{pmatrix}$$

$$\left[\text{info}^{(1)}\right]^{-1} = \begin{pmatrix} 0.00005898323 & 0 \\ 0 & 0.00000868996 \end{pmatrix}$$

This gives the estimated standard errors $\widehat{\text{SE}}^{(1)}\left(\hat{\pi}_1^{(1)} | \, d^{(1)}\right) = 0.0077$ and $\widehat{\text{SE}}^{(1)}\left(\hat{\pi}_2^{(1)} | \, d^{(1)}\right) = 0.0029$.

The conditional expectation of this information matrix can be evaluated by replacing n_{11g} by its conditional expectation evaluated at $\hat{\theta}^{(2)}$. Doing so yields

$$E\left[\text{info}^{(1)} | \, d^{(2)}\right] = \begin{pmatrix} 16991.07 & 0 \\ 0 & 117205.8 \end{pmatrix}$$

which is very close to $\text{info}^{(1)}$.

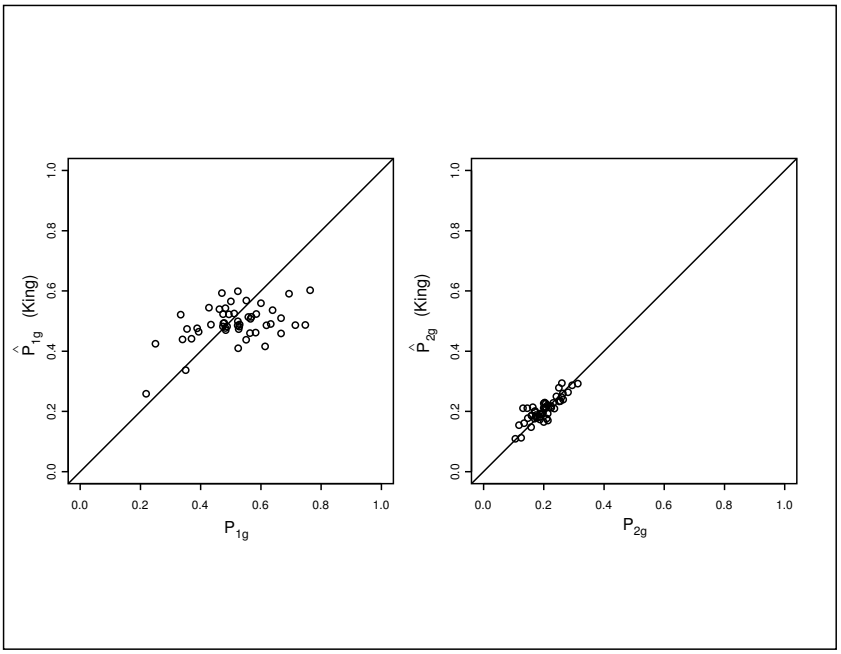

Figure 1.3 *Plot of $\hat{P}_{1g}$ versus P_{1g} and $\hat{P}_{2g}$ versus P_{2g} using King's Methodology*

Using $\hat{\pi}_1^{(2)}$, $\hat{\pi}_2^{(2)}$ and $\hat{\theta}^{(2)}$ from the Newton-Raphson procedure

$$Var\left[\mathrm{sc}^{(1)} \,|\, d^{(2)}\right] = \begin{pmatrix} 11927.53 & -19179.83 \\ -19179.83 & 30841.74 \end{pmatrix}$$

which has an associated correlation of -1. Applying (1.6) the resulting information matrix based only on the aggregate level data is

$$\mathrm{info}^{(2)} = \begin{pmatrix} 5063.538 & 19179.83 \\ 19179.83 & 86364.03 \end{pmatrix}$$

and

$$\left[\mathrm{info}^{(2)}\right]^{-1} = \begin{pmatrix} 0.0012436897 & -0.00027620009 \\ -0.00027620009 & 0.00007291774 \end{pmatrix}$$

and so the estimated standard errors are $\widehat{\mathrm{SE}}^{(2)}\left(\hat{\pi}_1^{(2)} \,|\, d^{(2)}\right) = 0.0353$ and $\widehat{\mathrm{SE}}^{(2)}\left(\hat{\pi}_2^{(2)} \,|\, d^{(2)}\right) = 0.0085$.

The difference in the probabilities, $\pi_1 - \pi_2$ will often be of particular interest. From info$^{(1)}$ we obtain

$$\begin{aligned}\widehat{\text{Var}}^{(1)}\left[\hat{\pi}_1^{(1)} - \hat{\pi}_2^{(1)} | \, d^{(1)}\right] &= \widehat{\text{Var}}^{(1)}\left(\hat{\pi}_1^{(1)} | \, d^{(1)}\right) + \widehat{\text{Var}}^{(1)}\left(\hat{\pi}_2^{(1)} | \, d^{(1)}\right) \\ &\qquad -2\widehat{\text{Cov}}^{(1)}\left(\hat{\pi}_1^{(1)}, \hat{\pi}_2^{(1)} | \, d^{(1)}\right) \\ &= 0.00005879863 + 0.000008480757 - 2 \times 0 \\ &= 0.00006727939\end{aligned}$$

Hence

$$\widehat{\text{SE}}^{(1)}\left[\hat{\pi}_1^{(1)} - \hat{\pi}_2^{(1)} | \, d^{(1)}\right] = 0.008202401$$

From info$^{(2)}$

$$\begin{aligned}\widehat{\text{Var}}^{(2)}\left[\hat{\pi}_1^{(2)} - \hat{\pi}_2^{(2)} | \, d^{(2)}\right] &= \widehat{\text{Var}}^{(2)}\left(\hat{\pi}_1^{(2)} | \, d^{(2)}\right) + \widehat{\text{Var}}^{(2)}\left(\hat{\pi}_2^{(2)} | \, d^{(2)}\right) \\ &\qquad -2\widehat{\text{Cov}}^{(2)}\left(\hat{\pi}_1^{(2)}, \hat{\pi}_2^{(2)} | \, d^{(2)}\right) \\ &= 0.0012436897 + 0.00007291774 + 2 \times 0.00027620009 \\ &= 0.001869008\end{aligned}$$

giving

$$\widehat{\text{SE}}^{(2)}\left[\hat{\pi}_1^{(2)} - \hat{\pi}_2^{(2)} | \, d^{(2)}\right] = 0.04323203$$

The estimated correlation between $\hat{\pi}_1^{(2)}$ and $\hat{\pi}_2^{(2)}$ obtained from info$^{(2)}$ is -0.917.

Parameter	$\widehat{\text{Var}}^{(2)}/\widehat{\text{Var}}^{(1)}$	Ind. Sample Equiv. to 50 CD's	Ind. Sample Equiv. Per CD
π_1	21.2	1053	21
π_2	8.6	2596	52
$\pi_1 - \pi_2$	27.8	803	16

Table 1.2 *Effect of Aggregation on Variance Estimates: Income by Age*

The effect of aggregation can be examined by looking at the ratio of the estimated variances obtained from info$^{(1)}$ and info$^{(2)}$. These are given in Table 1.2. Here

the estimation of π_1 is affected by aggregation more than π_2, possibly because π_1 is larger and P_{1g} varies more across the CDs. The increase in the asymptotic variance of the parameters π_1 and π_2 is more than the increase in the diagonal elements of the information matrix, i.e. more than 3.3 and 1.3 respectively. This is due to the large covariance term introduced by the aggregation. The estimation of $\pi_1 - \pi_2$ is affected even more than that of π_1 due to the affect of aggregation on the correlation of the estimates. In looking at these ratios, it must be remembered that the individual level data consists of 22323 people whereas the aggregate data relates to 50 CD's, a ratio of 446. There are 4238 people who are 15-24 years old which contribute to the estimation of π_1, an average of 84.8 people per CD. While there is clearly a loss of information through the use of aggregate data, it does not correspond to each CD being equivalent to an individual. In Table 1.2 we show the individual level sample size required to obtain the same variance, and therefore standard error, as using these aggregate data for 50 CD's. For example, the sample of 50 CD's gives the same variance for the estimation of $\pi_1 - \pi_2$ as 803 individuals. Dividing by 50 gives an indication of the information per CD compared with the information per individual. For this example, on average, each CD is as useful as 16 individuals in terms of estimating $\pi_1 - \pi_2$. These results depend on the variation in the proportion of 15-24 year olds across the CD's.

Using the results in Section 1.3.4 we can also examine the likely impact of supplementing aggregate data with individual level survey data. This is shown in Table 1.3 which gives the variance $\mathrm{Var}^{(c)}$ of the estimate of $\pi_1 - \pi_2$ based on aggregate data for 50 CD's plus an independent sample of n_0 individuals for $n_0 = 0, 1, 10, 50, 100, 500, 1000$. For comparison, we also give the variance for these sample sizes when there is no aggregate data, $\mathrm{Var}^{(0)}$.

n_0	$\mathrm{Var}^{(c)}\left(\pi_1 - \pi_2\right)$	$\mathrm{Var}^{(0)}\left(\pi_1 - \pi_2\right)$
0	0.001869	–
1	0.001866	1.501876
10	0.001845	0.150188
50	0.001756	0.030037
100	0.001656	0.015019
500	0.001138	0.003004
1000	0.000818	0.001502
5000	0.000253	0.000300

Table 1.3 *Comparison on Var* $\left(\pi_1 - \pi_2\right)$ *for the analysis of aggregate data and a sample of individual level data of various sizes*

The results in Table 1.3 are consistent with the aggregate data being equivalent to 803 individuals.

We can also compare the use of individual and aggregate data in testing for homgeneity, using the likelihood ratio and score test as described in Section 3, page 11. Both tests should be compared with χ^2_{98}, for which the critical value for a 5% test is 122.

For the likelihood ratio test the results are

$$-2\log R^{(1)} = 502.7287 \qquad -2\log R^{(2)} = 339.2903$$

Both these values suggest that the null hypothesis of $\phi_g = \phi$ be rejected. The test statistic calculated from the individual level data is larger, which is consistent with it having more power. Each of these test statistics can be decomposed into a term for each group, i.e.

$$-2\log R^{(1)} = \sum_g \left(-2\log R_g^{(1)}\right) \qquad -2\log R^{(2)} = \sum_g \left(-2\log R_g^{(2)}\right)$$

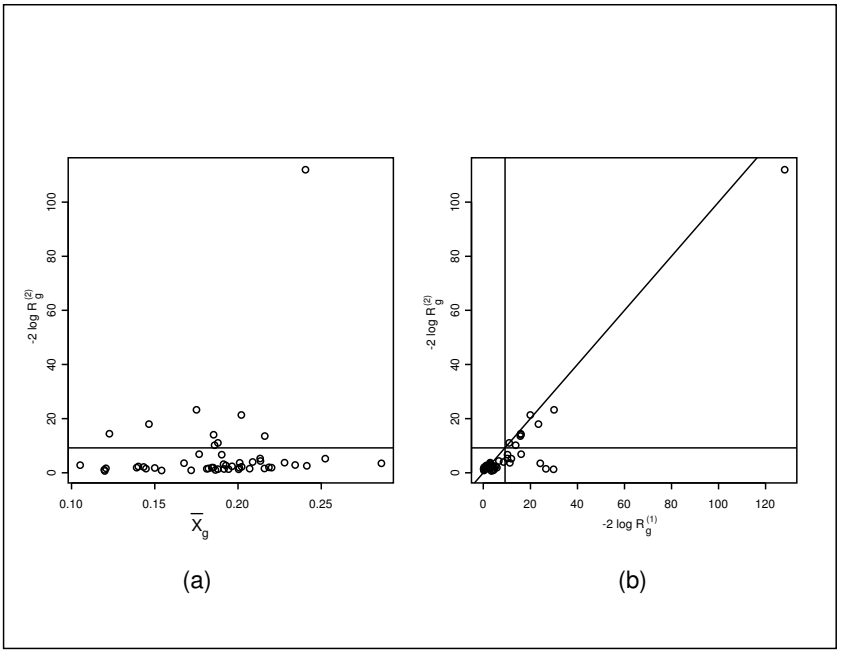

Figure 1.4 *Plot of* $-2\log R_g^{(2)}$ *versus* $\overline{X}_g$*, and* $-2\log R_g^{(1)}$

Figure 1.4(a) gives a plot of $-2\log R_g^{(2)}$ versus $\overline{X}_g = n_{1\bullet g}/n_g$, the proportion of people aged 15-24 years for each CD. This plot may be useful as a diagnostic in

terms of identifying groups with large values which indicates that they are particularly affecting the statistical significance of the test. This will suggest those groups having parameters π_{1g} and π_{2g} which are statistically significantly different from the overall parameters values. It may also be useful in suggesting any trends in departures from homogeniety that may be related to $\overline{X}_g$.

In examining these values we suggest comparing them with the 1% critical value of χ^2_2, i.e. 9.210. The horizontal and vertical lines on the figures correspond to this value.

Figure 1.4(b) gives a plot of $-2\log R_g^{(2)}$ versus $-2\log R_g^{(1)}$. Of the 17 cases that would be identified as statistically significant using individual level data, 9 are also identified using the group level data. Also no cases that are statistically non-significant using $-2\log R_g^{(1)}$ are identified as statistically significant using $-2\log R_g^{(2)}$. Hence, while there is, as expected, a loss of power in using the aggregate data, it is still possible to undertake a useful analysis of residuals.

Both the analyses of $-2\log R_g^{(1)}$ and $-2\log R_g^{(2)}$ identify one particular CD as having a large influence on the hypothesis test. This CD was investigated and found to have more than twice the usual population size, low values of P_{1g} and P_{2g}, and a reasonably high value of $\overline{X}_g$. This is probably a CD in a newly developed area of the city.

A similar approach can be used with the score test, giving

$$\mathrm{ST}^{(1)} = 496.8291 \qquad \mathrm{ST}^{(2)} = 359.9741$$

Figure 1.5 gives a plot of $ST_g^{(2)}$ versus $\overline{X}_g$ and $ST_g^{(1)}$.

Again these results both lead to the rejection of the null hypothesis. However, we encountered a problem with the score test. For 24 of the 50 CDs, $\mathrm{info}^{(2)}\left(\hat{\phi}; d_g^{(2)}\right)$ was not positive-definite, leading to a negative $\mathrm{ST}_g^{(2)}$ value. In our analysis we set such cases to zero. Numerically this situation arises because the subtraction of the estimate of the conditional variance of the score function for the CD reduces the diagonal elements and increases the off-diagonal elements too much. We are investigating modifications to the score test to overcome this issue. Notwithstanding this issue using $ST_g^{(2)}$ identifies 10 of the 15 cases that $ST_g^{(1)}$ would identify as having parameters statistically significantly different from the overall values. However, it also identified one case as statistically significant that was not so identified using $ST_g^{(1)}$.

Signed residuals can also be determined and examined.

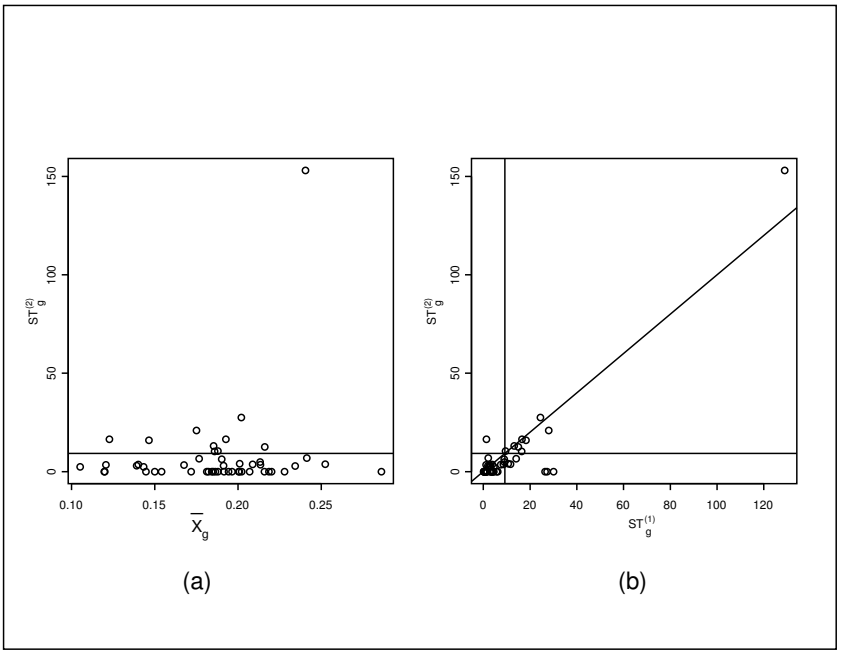

Figure 1.5 *Plot of* $ST_g^{(2)}$ *versus* $\overline{X}_g$ *and* $ST_g^{(1)}$

1.6 Discussion

We have described a general approach to clearly identify the loss of information in using aggregate rather than individual level data. Let Y_i denote the value of the response variable for individual i. In many situations determining the score function and information loss through aggregation will involve determining $\mathrm{E}\left(Y_i | d_g^{(2)}\right)$, $\mathrm{Var}\left(Y_i | d_g^{(2)}\right)$ and $\mathrm{Cov}\left(Y_i, Y_j | d_g^{(2)}\right)$ for $i, j \in g$.

In the example of homogeneous 2×2 tables, this approach is not much simpler than direct use of the likelihood based on the aggregate data $d^{(2)}$. However, equation (1.6) clearly shows the information loss. Much of the effect of aggregation in this case arises from the change to the off-diagonal elements of the information matrix.

The example considered in this chapter shows how we can test the hypothesis of the parameters of interest being constant across groups from aggregate data alone. Decomposing the resulting test statistics into contributions from each group enables an analysis of the impact that each group has on the hypothesis test. This can be useful in identifying groups with parameter values very different from the overall parameters.

The example suggests that residuals obtained from the Likelihood Ratio Test using aggregate data are preferable to those obtained from the Score Test.

We are currently considering how the general approach applies in the more complex models, especially those including random effects to allow for the variation in group specific parameters.

1.7 References

Beh, E. J., and Steel, D. G., "Maximum likelihood estimation and homogeneous 2×2 tables", Preprint 3/02, School of Mathematics and Applied Statistics, University of Wollongong, Australia, 2002.

Beh, E. J., Steel, D. G. and Booth, J. G., "What useful information is in the marginal frequencies of a 2×2 table?", Preprint 4/02, School of Mathematics and Applied Statistics, University of Wollongong, Australia, 2002.

Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M. and Welsh, A. H., "Maximum likelihood inference from sample survey data", International Statistical Review, 62, 349–363, 1994.

Chambers, R. L. and Steel, D. G., "Simple methods for ecological inference in 2×2 tables, Journal of the Royal Statistical Society A, 164, 175-192, 2001.

Cox, D. R. and Hinkley, D. V., "Theoretical Statistics", Chapman and Hall, London, UK, 1974.

Efron, B. and Hinkley, D. V., "Assessing the accuracy of the maximum likelihood estimator : Observed versus expected Fisher information (with discussion)", Biometrika, 65, 457–487, 1978.

King, G., "A Solution to the Ecological Inference Problem", Princeton University Press, Princeton, USA, 1997.

McCullagh, P. and Nelder, J. A., "Generalized Linear Models", Chapman and Hall, London, UK, 1989.

McCulloch, C. E. and Searle, S. R., "Generalized, Linear, and Mixed Models", Wiley, New York, 2001.

Reddien, G. W., "Newton-Raphson methods", Encyclopedia of Statistical Sciences, 6, 210-212, 1986.

Royall, R. M., "Statistical Evidence : A Likelihood Paradigm", Chapman and Hall, London, UK, 1997.

Wakefield, J., "Ecological inference for 2×2 tables", Techincal Report, Department of Statistics and Biostatistics, University of Washington, USA, 2001.

Acknowledgment

This research was supported by grants from the Australian Research Council. We would also like to thank John Rayner for some useful discussions.