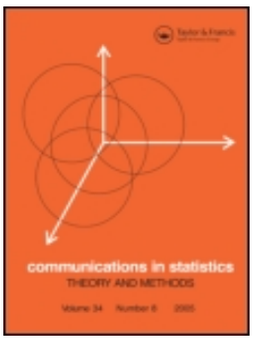
Hierarchical Bayesian models for the estimation of correlated effects in multilevel data: A simulation study to assess model performance



Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

Hierarchical Bayesian Models for the Estimation of Correlated Effects in Multilevel Data: A Simulation Study to Assess Model Performance

Giulia Roli & Paola Monari

To cite this article: Giulia Roli & Paola Monari (2015) Hierarchical Bayesian Models for the Estimation of Correlated Effects in Multilevel Data: A Simulation Study to Assess Model Performance, Communications in Statistics - Theory and Methods, 44:12, 2644-2653, DOI: 10.1080/03610926.2013.806662

To link to this article: http://dx.doi.org/10.1080/03610926.2013.806662

Communications in Statistics—Theory and Methods, 44: 2644–2653, 2015

ISSN: 0361-0926 print / 1532-415X online
DOI: 10.1080/03610926.2013.806662

Hierarchical Bayesian Models for the Estimation of Correlated Effects in Multilevel Data: A Simulation Study to Assess Model Performance

GIULIA ROLI AND PAOLA MONARI

Department of Statistical Sciences, University of Bologna, Bologna, Italy

In this article, we aim at assessing hierarchical Bayesian modeling for the analysis of multiple exposures and highly correlated effects in a multilevel setting. We exploit an artificial data set to apply our method and show the gains in the final estimates of the crucial parameters. As a motivating example to simulate data, we consider a real prospective cohort study designed to investigate the association of dietary exposures with the occurrence of colon-rectum cancer in a multilevel framework, where, e.g., individuals have been enrolled from different countries or cities. We rely on the presence of some additional information suitable to mediate the final effects of the exposures and to be arranged in a level-2 regression to model similarities among the parameters of interest (e.g., data on the nutrient compositions for each dietary item).

Keywords Hierarchical Bayesian modeling; Correlated effects; Multilevel data.

Mathematics Subject Classification 62J12; 62H20; 62C12; 62P10.

1. Introduction and Background

The estimation of multiple effects often faces problems of some sort of complications that need to be somehow controlled during the analysis. Typical examples lie in case-control studies aiming at investigating the exposures which can cause the occurrence of a disease. In such cases, the use of the models conventionally employed becomes improper, yielding unstable and biased estimates.

In this article, we consider two kinds of such complications. The first one concerns the structure of the data and occurs whenever units are nested into higher level units involving their own variability and a dependence among the related observations. The nested or multilevel structure of data is a common phenomenon, especially in behavioral and social research, where the evaluation of the relationship between individuals and society is a focal interest of the research. In other cases, the hierarchy of data can be viewed as a nuisance. For instance, in the multi-stage sampling, which is frequently employed in the traditional surveys to reduce the costs of data collection, the nested structure of the data is directly

Received November 14, 2012; Accepted May 14, 2013.
Address correspondence to Giulia Roli, Dipartimento di Scienze Statistiche "P. Fortunati", via delle Belle Arti, 41, 40126, Bologna; E-mail: g.roli@unibo.it

generated by the sampling design and thus requires some adjustments during the inferential process. Whatever the dependence arises from, it is "neither accidental nor ignorable" (Goldstein, 1995) and the risks of drawing wrong conclusions are high if the clustering of the data is disregarded.

The joint analysis of multiple exposures gives rise to the second complication. Indeed, many studies involve a set of potential effects to be compared, especially in epidemiologic field, yielding problems of multiple inference. When a conventional analysis is carried out, these are revealed by failures in the convergence of the estimation process or by implausible large and unstable estimates, especially when the samples are small and sparse (Witte et al., 1994). The main reason is that these effects are often correlated. Therefore, we need to take into account for a covariance structure among them to reduce the random errors in the estimates.

Both these complications have been tackled separately in various applications and simulations by using hierarchical modeling (Diex-Roux, 2000, 2004; Witte et al., 1994; Greenland, 1992). When the structure of the data is nested, hierarchical modeling allows to handle simultaneously multiple levels of information and dependencies (Raudenbush and Bryk, 2002; Leyland and Goldstein, 2001; Snijders and Bosker, 1999; Hox, 1995). In this setting, we also refer to *multilevel* regression models. These can appropriately address different research aims: (i) improved estimation of the individual effects under investigation (i.e., all the available information at both levels are efficiently used in order to exploit both the group features and the relations existing in the overall sample); (ii) evaluation of the cross-level effects (e.g., how variables measured at one level affect relations occurring at another); and (iii) decomposition of the variance-covariance components at each level.

As far as the multiple exposure issue is concerned, numerous authors have shown that empirical and semi-Bayes estimates from hierarchical models can improve standard regression estimation, allowing for correlated associations and showing to be less sensitive to sampling error and model misspecification (Morris, 1983; Greenland, 1992, 1993, 1997). Indeed, relying on the presence of some additional information suitable to mediate the final effects of the exposures, these can be arranged in a second-stage regression to model similarities among the parameters of interest (Witte et al., 1994; Rothman et al., 2008).

Although developed separately and for different purposes, hierarchical modeling for correlated effects and for nested data have important communalities, which can be strengthened when a Bayesian perspective is adopted. The use of Bayesian methods to analyze real data is a relevant topic discussed by several authors (Greenland, 2006, 2007; MacLehose et al., 2007; Graham, 2008). They all support the employment of prior assumptions as they are more reasonable than those implicitly made by frequentist models and able to address the problems of sparse data, multiple comparisons, subgroup analysis and study bias. In this framework, the assignment of prior judgements becomes of primary importance and different strategies can be adopted. A fully-Bayesian (FB) approach forces all the parameters in the model to be random and corresponding probability distributions to be assigned. When these prior distributions are in the form of prior data, we refer to *empirical priors*, arising from frequentist shrinkage-estimation or empirical-Bayes (EB) methods (Maritz and Lwin, 1989). Instead of assigning a full prior distribution, another strategy consists in fixing in advance a specific value for one or more parameters using background information. This method, called semi-Bayes (SB) approach, is commonly employed to avoid the drawback of absurd estimates of some (hyper-) parameters (Greenland, 1992, 2000). Finally, in a Bayes empirical-Bayes (BEB) setting both FB and EB criteria are jointly adopted by exploiting the available empirical data for some (hyper-) parameters and some kinds of proper distributions for the others (Deeley and Lindley, 1981).

In this article, we aim at extending the hierarchical approach for the analysis of multiple exposures and highly correlated effects to a multilevel setting. We attempt to improve the ordinary estimates of such effects by using some descriptive information to develop a second-stage regression model mediating the effects of the exposure variables, separately by group membership but into a single analysis. We adopt a BEB perspective and exploit the previous knowledge on the other (hyper-) parameters to specify prior distributions, which are suitable regarding the problem at hand. The method we propose has been already applied to an empirical study in a previous work (Roli and Monari, 2011). Here, we simulate data by resembling a real prospective cohort study designed to investigate the association of dietary exposures with the occurrence of colon-rectum cancer in multilevel data (e.g., individuals enrolled from different countries or cities). We rely on the presence of some additional information suitable to mediate the final effects of the exposures and to be arranged in a level-2 regression to model similarities among the parameters of interest (e.g., data on the nutrient compositions for each dietary item). Using these artificial data, we assess our method to show and measure the gains in the final estimates of the crucial parameters with respect to the conventional analysis results and to different prior specifications.

2. Modeling Framework

We consider J groups, commonly defined by geographical areas. For each group j, we have the total number of individuals, N_j (with $j = 1, \ldots, J$), and the presence/absence of a disease, denoted by the individual indicator y_{ij} ($y_{ij} = 1$ for cases, $y_{ij} = 0$ for control units). We wish to model the number of disease cases in terms of K explanatory variables, or exposures, denoted by X_k for each exposure k, further controlling for the effects of P potential confounders (such as age, sex and smoking status of individuals), denoted by W_p for each confounder p. These can either be continuous, binary or categorical variables.

We assume that the data are generated through the following underlying model. Individual i in group j, with exposures in $(x_{ij1}, \ldots, x_{ijK})$ and confounders in $(w_{ij1}, \ldots, w_{ijP})$, experiences the binary outcome y_{ij} with probability p_{ij}, where:

$$logit(p_{ij}) = \alpha_j + \sum_{k=1}^{K} \beta_{jk} x_{ijk} + \sum_{p=1}^{P} \gamma_p w_{ijp}. \tag{1}$$

The paramter α_j represents the group-specific logit-baseline risk, which, according to a conventional hierarchical approach, could be structured in terms of group-level covariates, or exchangeable or spatially correlated random effects.

The effects of confounders are reasonably assumed to be the same in all the groups and, thus, simply denoted by γ_p.

As far as the exposures effects β_{jk} are concerned, they represent the key objective of the investigation and we assume they vary across groups. In order to tackle the problem of interactions among the multiple exposures and the correlation among corresponding effects, we assume that some kinds of Q group-specific prior or level-2 data are available (denoted by z_{jkq} for each group j exposure k and level-2 covariate q). These are arranged to form the exposures' coefficients through a level-2 regression model:

$$\beta_{jk} = \pi_{0k} + \sum_{q=1}^{Q} \pi_q z_{jkq} + \delta_{jk}, \tag{2}$$

where π_q are the effects of such prior information on the exposures (and on the disease), which are assumed to be common to all the exposures; π_{0k} is the intercept reflecting our knowledge about any residual effects of the exposure k due to prior information not included in the second-stage model; δ_{jk} are the residuals, which are assumed to hold the simple hypothesis of independence and normal distribution with null means and constant variances, denoted by σ^2_β.

2.1. Prior Distributions

Previous works on hierarchical modeling for correlated effects have shown that the frequentist estimation methods (such as, maximum-likelihood, penalized quasi-likelihood, or marginal quasi-likelihood) often yields null values for the level-2 variance σ^2_β leading to an extreme shrinkage estimation of the target parameters β_{jk} toward the empirical prior means (EB estimators). This seems more likely to reflect a marginal likelihood for σ^2_β with peak at zero, rather than true under dispersion (Greenland, 1992). Moreover, a credible result would achieve a more reasonable non zero value for σ^2_β, as it represents the uncertainty about the residuals δ_{jk} and therefore also about the estimation of β_{jk} after incorporating the level-2 information. In particular, if σ^2_β tends to ∞, the hierarchical model and the conventional logistic regression come to the same results according to the estimates of β_{jk}. On the contrary, if $\sigma^2_\beta = 0$, then the residuals δ_{jk} result to be null, meaning that we implicitly assume the absence of any effects of exposures beyond those of level-2 covariates.

In 1992, Greenland suggested the SB approach as a good and easy strategy to tackle the problem of null estimation of the level-2 variance parameters by setting specific suitable values for σ^2_β. In particular, SB estimates appear to be better than EB estimates when the sample sizes and the ratio of subjects to parameters are small. Moreover, they are proved to be robust to misspecification (Greenland, 1993).

To borrow estimation power, here we propose to adopt a Bayesian perspective (more precisely, a BEB approach) as a more appropriate framework. Indeed, it requires all the parameters to be random by offering the opportunity of assigning reasonable priors by letting the data contribute to the final estimation. Thus, in our example the specification of the prior distributions for (hyper-) parameters γ_p, α_j, π_q, π_{0k}, as well as for σ^2_β, is needed. We model the intercepts α_j as $N(\alpha_0; \sigma^2_\alpha)$, representing exchangeable random baseline risks of disease for each group. The choice of other prior distributions, together with those for α_0 and σ^2_α, can be problematic. As a first attempt, we assign flat and conjugate prior distributions. Then, we specify weakly informative priors to assess model performances. In particular, we focus our attention on the crucial parameter σ^2_β by reasonably supposing that its value will be small, on the grounds that most important level-2 covariates have been included in the analysis. We believe a 2-fold variation between the Odds Ratios (OR) for the upper and lower 5% of units is reasonable, that is $\beta_{95\%} - \beta_{5\%} = log2$. Hence, our prior guess at the precision term σ^{-2}_β is $\approx 3.29^2/(log2)^2 \approx 22.53$. To reflect our uncertainty in this prior guess, we believe that 4-fold variation between the upper and the lower 5% of units is very unlikely (say, less than a 1% chance). Thus, lower 1% quantile of our prior distribution for the precision σ^{-2}_β can be supposed to be $\approx 3.29^2/(log4)^2 \approx 5.63$. These two assumptions are sufficient to fully specify a proper and informative hyperprior for the precision term that is $\sigma^{-2}_\beta \sim Gamma(4.9; 0.22)$.

The Bayesian hierarchical model we propose will be assessed, under various circumstances, with a simulation study described in the next section and fitted by Markov chain

Monte Carlo simulation. Samples from full-conditional distributions are generated using the WinBUGS software (Spiegelhalter et al., 2003).

3. Simulation Study

We generate an artificial dataset basing on a real prospective cohort study designed to investigate the association of dietary exposures with the occurrence of colon-rectum cancer with individuals enrolled from different centers (Riboli and Kaaks, 1997). We suppose for a sample of healthy subjects information on dietary intakes is assessed by specific instruments and collected at the enrollment, together with anthropometric measurements and lifestyle habits. Then, after a scheduled period of observation (e.g., 10 years) a follow-up for disease incidence is carried out, focusing on the identification of a certain number of study subjects who developed colon-rectum cancer. Further suppose that, for each center of enrollment, additional information on the dietary exposures are available in the form of nutrient compositions (e.g., the amounts of protein, fat, fiber, beta-carotene, etc., in one gram of pasta or rice or milk). This descriptive information on food constituents is usually collected in tables which vary across countries and, more generally, across centers.

Referring to our motivating example, we consider J=30 centers of N_j=400 individuals, $j = 1, \ldots, J$, which are typical numbers of small-area epidemiological studies on disease risks for a subset of the population. We suppose to have K8 dietary exposures, P=2 potential confounders, one binary (such as smoking status) and one continuous (such as age or pollution exposure), and Q=4 nutrients. We fix parameters and generate data according to available information on the study at hand (Roli and Monari, 2011; Witte et al., 1994) and to previous works on data simulation of grouped data (Jackson et al., 2006; Moineddin et al., 2007; Witte and Greenland, 1996).

Under this perspective, we fix the intraclass correlation (ICC) to be equal to 0.2, which, together with sample sizes, ensures accuracy of the estimates (Goldstein, 1995). Accordingly, the logit-baseline risks α_j of disease for the J groups are generated from a normal distribution, with mean $logit(0.2)$ and standard deviation 0.9, giving a 95% sampling interval for the baseline risk of (0.04, 0.59). We take γ_1=$log(1.5)$=0.41 as the effect of the binary confounder and γ_2=$log(1.07)$=0.07 as that of the continuous one, which both correspond to reasonable values of odds ratio associated with a lifestyle factor such as smoking and with factors such as age or pollution exposure, respectively. We fix the nutrients effects included in Eq. (2) to be $log(1.06)$, $log(1.05)$, $log(0.9)$, and $log(1.015)$, basing on a selection of previous results. Level-2 intercepts are specified under two possible scenarios:

Scenario 1. we rely on the presence of data for all the crucial nutrients whose amounts are included in the analysis. Under this assumption, model 2 can be simplified with a small common level-2 intercept, $\pi_{01} = \ldots = \pi_{0k} = \ldots = \pi_0$ assumed to be equal to 0.001.

Scenario 2. some relevant levels of nutrients are supposed to be not available and, thus, we assume k-specific level-2 intercepts: negative (if we expect the residual effects for item k to be preventive), positive (if we expect to be causative), or small/null (if we expect little or no residual effects). In particular, we fix $\pi_{01} = \pi_{02} = -0.05$, $\pi_{03} = \pi_{04} = +0.05$ and $\pi_{0k} = 0$, otherwise (Witte et al., 1994).

The residuals δ_{jk} are generated from a normal distribution with a null mean and a common standard deviation equal to 0.01.

Table 1
Values of the parameters to generate exposure data

Parameters	Values
$M_{(k)}$	63, 224, 5, 32, 2, 13, 7, 11
$S_{(k)}$	8, 5, 1.5, 5, 0.15, 4, 1.5, 1.5
$a_{(k)}$	−5.13, −4.16, −2.08, −4.16, 2.77, −3.58, −1.83, −1.39
$b_{(k)}$	0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6

As far as data generation is concerned, individuals exposed to the binary confounder in the groups are randomly chosen from a binomial distribution by fixing the proportion of cases to be equal to 0.2. The continuous confounder variable is generated from a Normal distribution with mean 58.4 and standard deviation 6.3. For each exposure k, we generate values within group j from Normal distributions with group-specific parameters $N(m_{j(k)}, s^2_{j(k)})$ where $m_{j(k)}$ and $log(1/s^2_{j(k)})$ are simulated as $N(M_{(k)}, S^2_{(k)})$ and $N(a_{(k)}, b^2_{(k)})$, respectively. Parameters $M_{(k)}, S^2_{(k)}, a_{(k)}, b^2_{(k)}$ are chosen basing on real data and fixing the ratios of the between-center standard deviation (the standard deviation of $m_{j(k)}$) to the within-center standard deviation (the mean of $s_{j(k)}$) to be always ≤ 0.75 (Table 1). This quantity describes the amount of information at the ecological or center level concerning the true individual-level variability of exposures (Jackson et al., 2006).

The amounts of each nutrient q in the composition of the dietary items are separately generated across center units from normal distributions defined by k-specific parameters, $N(m_{k(p)}, s^2_{k(p)})$, following the same rationale described above for dietary exposures. The means $m_{k(p)}$ are simulated as K equally spaced quantiles of independent P normal distributions, $N(M_{(p)}, S^2_{(p)})$; $log(1/s^2_{k(p)})$ as $N(a_{(p)}, b^2_{(p)})$; and we fix the ratios of the between-dietary standard deviation to the within-dietary standard deviation to be always ≥ 60, i.e., reasonably assuming (within the same nutrient) much greater variability across dietary items than across centers (Table 2).

The $J \times K$=240 coefficients β_{jk} are then formed by the artificial quantities involved in Eq. (2). These result to vary from a minimum of −0.095 (OR=0.910) to a maximum of 0.180 (OR=1.197). Finally, the disease status Y of individual i in group j is generated taking the value 1 with probability p_{ij}, given by Eq. (1). An overall prevalence of the disease which is close to 15% is attained.

Table 2
Values of the parameters to generate level-2 covariates

Parameters	Values
$M_{(p)}$	5.4, 1, 3.7, 7.3
$S_{(p)}$	4, 0.5, 2.5, 6
$a_{(p)}$	6.44, 13.82, 7.82, 4.61
$b_{(p)}$	0.7, 0.7, 0.7, 0.7

4. Results

Results on the target parameters, i.e., OR of dietary exposures , are considered. Several circumstances are assessed under the two different scenarios of data generation described above, according to the following scheme:

Scenario 1. The parameters are estimated under three model specifications:

(a) a conventional approach where several logistic regression analysis are carried out (one for each center separately);
(b) a standard multilevel framework with random-slopes and intercepts, where at level 1 the same model as in Eq. (1) is considered and at level 2 we have

$$\begin{aligned} \alpha_j &= \alpha_0 + u_{0j} \\ \beta_{jk} &= \beta_{0k} + u_{jk} \ \ \forall\, k \end{aligned} \tag{3}$$

$$\text{with} \begin{bmatrix} u_{0j} \\ \vdots \\ u_{jk} \\ \vdots \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix}, \begin{bmatrix} \tau_{00} & \dots & \tau_{0k} & \dots \\ \vdots & \dots & \vdots & \dots \\ \tau_{0k} & \dots & \tau_{kk} & \dots \\ \vdots & \dots & \vdots & \dots \end{bmatrix} \right); \text{ and}$$

(c) under our modeling framework, namely a Hierarchical BEB approach, with the same conditions of data generation, and by considering vague priors.

Scenario 2. Parameter estimation from Hierarchical BEB model:

(a) under flat non-informative priors;
(b) when informative or credible assumptions are fixed on hyperparameter (see Sec. 2.1); and
(c) under hypothesis of scenario 1 (i.e., misspecification with respect to the level-2 intercepts).

In order to compare the results, two measures are considered. The relative bias is used to quantify the accuracy of the parameter estimates and the observed coverage of the nominal 95% credibility interval (CI) to assess the accuracy of the standard error of the parameter estimates (Maas and Hox, 2005). Given the large number of parameters, we present summarized results in terms of percentage of the coverage and minimum, maximum, mean and standard deviation (SD) measures for the absolute value of the relative bias, separately by scenario (Tables 3 and 4).

Under Scenario 1, the method we propose works rather well, yielding good coverages of credibility intervals (98% of cases including the true parameter) and reduced measurements for bias. As a benchmark, estimates from the center-specific conventional analysis are shown to be very biased (190.7 on average) and some problems during the estimation process occur (for 9 centers the algorithm does not converge). In order to quantify the gains in our estimates, we further consider results from a standard multilevel analysis, where the shrinkage of the estimates is made only by the same food across centers, but part of the correlation (e.g., foods that share similar nutrient compositions) remains uncontrolled. In such case, even if an improvement is achieved against conventional approach, results remain quite biased.

The estimates from the Hierarchical BEB approach seem to be robust when some level-2 information are missing

Table 3
Results - Scenario 1

Parameters: OR(β_{jk})	Conventional*	Multilevel	Hierarchical BEB
Coverage of 95% CI	94.5%	97.9%	98.0%
Relative bias			
Mean	190.7	17.641	1.719
SD	1608.2	162.7	6.6
Min	0.008	0.000	0.000
Max	23976.5	2495.5	84.7

*For 9 centers algorithm did not converge.

(Scenario 2), when the priors are not informative (Scenario 2(a) and 2(b)) and to model misspecification (Scenario 2(c)). In particular, specifying informative and reasonable priors seems to mainly improve the accuracy of the standard error of the estimates (coverage from 92.9–96.3%), while the bias measures hold steady. Similarly, when we wrongly assume a common level-2 intercept the performance in the parameter estimates is good, but the accuracy of standard error estimates consistently decreases (only 88.3% of intervals include the true parameter). Indeed, forcing the level-2 intercepts to be the same yields a too large shrinkage effect of the estimates of dietary effects. As a result, corresponding posterior distributions result to be too much concentrated around an estimate (taken as the mean of the distribution) which is less precise. In such cases, a good practice could be to adopt a direct likelihood approach to firstly select a suitable model specification. In Bayesian analysis the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) can be used to assess model complexity and goodness-of-fit. In our example, DIC properly identifies the model with k-specific level-2 intercepts as the best specification (DIC=4991.450) rather than one with a common intercept (DIC=5054.760).

5. Concluding Remarks

We considered two complications that can affect the estimation of multiple exposures and that need to be somehow controlled during the analysis: the nested structure of data and an high correlation across effects. We proposed a Hierarchical Bayes empirical Bayes

Table 4
Results - Scenario 2

Parameters: OR(β_{jk})	Vague priors	Informative priors	Under Scenario 1
Coverage of 95% CI	92.9%	96.3%	88.3%
Relative bias			
Mean	0.218	0.218	0.237
SD	0.318	0.187	0.176
Min	0.0003	0.0001	0.002
Max	1.644	1.067	0.816

approach, which exploit additional information suitable to mediate the final effects of the exposures, as well as providing a reasonable framework to assign credible priors.

We implemented a simulation study based on a real study carried out to investigate the association of dietary exposures with the occurrence of colon-rectum cancer. We showed that the method we propose provides results on the effects of exposures which are more precise and less biased than those provided by the standard approaches. Moreover, it seems quite robust in terms of bias of the estimates when a misspecifications in the parameter structure occurs. Conversely, standard errors of the estimates result to be more sensitive, even when specifications of vague rather than informative priors are compared.

Other aspects need to be explored as further developments of this work, such as evaluating model performance when the numbers of exposures and level-2 covariates increase (or decrease) and when more complex residual covariance matrix and related misspecification are considered.

The method we propose can be easily applied in several fields, other than in the epidemiologic application we considered. Some examples lie in occupational studies, where more levels of information can be merged; or to perform polytomous logistic regressions of different causes of death on a set of exposures; or in disease mapping and spatial analysis, where the variations due to random occurrences need to be controlled by exploiting the spatial proximity and the consequent interaction of the geographical areas; or, finally, in genetics, where the adjacency of genes can generate some sorts of interactions in their expression.

References

Deeley, J. J., Lindley, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* 76:833–841.

Diez-Roux, A. V. (2000). Multilevel anlaysis in public health research. *Ann. Rev. Public Health* 21:171–92.

Diez-Roux, A. V. (2004). The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. *Epidemiol. Rev.* 26:104–111.

Goldstein, H. (1995). *Multilevel Statistical Models*. 2nd ed. London: Edwards Arnold; New York: Halstead.

Graham, P. (2008). Intelligent smoothing using hierarchical Bayesian models. *Epidemiology* 19:493–495.

Greenland, S. (1992). A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occipational cancer-mortality study. *Statist. Med.* 11:219–230.

Greenland, S. (1993). Methods for epidemiologic analysis of multiple exposures: a review and a comparative study of maximum-likelihood, preliminary testing and empirical Bayes regression. *Statist. Med.* 12:717–736.

Greenland, S. (1997). Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analysis. *Statist. Med.* 16:515–526.

Greenland, S. (2000). Principles of multilevel modelling. *Int. J. Epidemiol.* 29:158–167.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int. J. Epidemiol.* 35:765–775.

Greenland, S. (2007). Bayesian perspectives for epidemiological research: II. Regression analysis. *Int. J. Epidemiol.* 36:195–202.

Hox, J. J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Pubblikaties.

Jackson, C., Best, N., Richardson, S. (2006). Improving ecological inference using individual-level data. *Statist. Med.* 25:2136–2159.

Leyland, A., Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*. Chichester, UK: John Wiley.

Maas, C. J. M., Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1:86–92.

MacLehose, R. F., Dunson, D. B., Herring, A. H., Hoppin, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology* 18:199–207.

Maritz, J., Lwin, T. (1989). *Empirical Bayes Methods*. London: Chapman and Hall/CRC.

Moineddin, R., Matheson, F. I., Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Med. Res. Methodol.* 7: 34.

Morris, C. (1983). Parametric empirical Bayes; theory and applications (with discussion). *J. Amer. Statist. Assoc.* 178:47–65.

Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical Linear Models - Application and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage.

Riboli, E., Kaaks, R. (1997). The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* 26(Suppl 1):6–14.

Roli, G., Monari, P. (2011). Improving the estimation of multiple correlated dietary effects on colon-rectum cancer in multicentric studies: a hierarchical Bayesian approach. *Statistica* 71:437–452.

Rothman, K. J., Greenland, S., Lash, T. L. (2008). *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott-Williams-Wilkins.

Snijders, T., Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Spiegelhalter, D., Best, N., Carlin, B., Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. B.* 64:583–640.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2003). *WinBUGS User Manual*. Version 1.4.

Witte, J., Greenland, S., Haile, R., Bird, C. (1994). Hierarchical regression analysis applied to a study of multiple dietray exposures and breast cancer. *Epidemiology* 5(6):612–621.

Witte, J., Greenland, S. (1996). Simulation study of hierarchical regression. *Statist. Med.* 15:1161–1170.