Multilevel Modeling in Epidemiology with GLIMMIX

John S. Witte,1 Sander Greenland,2 Lee-Lian Kim,1 and Lenore Arab3

Previous work has shown that multilevel modeling can be a valuable technique for epidemiologic analysis. The complexity of using this approach, however, continues to restrict its general application. A critical factor is the lack of flexible and appropriate software for multilevel modeling. SAS provides a macro, GLIMMIX, that can be used for multilevel modeling, but that is not sufficient for a complete epidemiologic analysis.

We here provide additional code to obtain epidemiologic output from GLIMMIX, illustrated with new data on diet and breast cancer from the European Community Multicenter Study on Antioxidants, Myocardial Infarction, and Breast Cancer (EURAMIC). Our results give epidemiologists an easily used tool for fitting multilevel models. (Epidemiology 2000; 11:684–688)

Keywords: breast neoplasms, diet, epidemiologic methods, logistic regression, multilevel models, odds ratio, relative risk, software.

Multilevel modeling (also known as hierarchical regression) is an important technique for epidemiologic analysis for three key reasons. First, it allows one to incorporate multiple levels of information into a single epidemiologic analysis. Doing this can be critical for adequately modeling exposure-disease relations driven by risk factors arising at both the individual and ecologic levels. This advantage helped spur the use of multilevel modeling in social science research and is currently helping prompt its application in epidemiology. For example, a recent study used a multilevel model to examine concurrently the impact of individual- and neighborhood-level factors on poor health. This analysis found that both individual educational status and neighborhood socioeconomic environment were associated with self-reported poor health.1

Second, multilevel modeling can provide estimates of effect that are more accurate and more plausible than those from conventional models.2 Much research indicates that in the analysis of epidemiologic data on multiple exposures, multilevel models can statistically outperform conventional approaches (for example, one-stage logistic regression).3–11 This outperformance occurs in part because the higher levels of a multilevel model incorporate additional information for estimation. For example, in an application to nutritional epidemiology, a hierarchical model improved conventional estimates of food effects by shrinking them toward each other when their corresponding foods had similar levels of nutrients.7

Third, multilevel modeling provides a solution to problems of multiple comparisons. Proposed solutions of this issue have ranged from ignoring the multiple comparisons to “adjusting” for them by altering the alpha level used in statistical procedures.12–17 Instead of ignoring or adjusting for multiple comparisons, one can solve this problem by using a multilevel model to simultaneously evaluate multiple exposures or outcomes.3–6 For example, hierarchical modeling has been used in genetic epidemiology for evaluating associations between large numbers of candidate human leukocyte antigen genes and insulin-dependent diabetes mellitus18 and for evaluating gene and diet interactions in the etiology of colon polyps.11 In a similar fashion, hierarchical regression (in the form of random coefficient modeling) provides an alternative to variable selection in problems involving many potential confounders.19

Despite the benefits of multilevel modeling, lack of software tailored to epidemiology remains a major obstacle to its use. In a previous report, we provided SAS code for two-stage multilevel modeling.20 This code uses a basic two-step approach whereby regression coefficients from a first-stage model are analyzed with a second-stage linear weighted-least-squares regression.5,6 This approach requires that the data be adequate for fitting the first-stage model without any constraints. One can instead undertake multilevel modeling with the SAS macro GLIMMIX (http://ftp.sas.com/techsup/download/stat/glmm612.sas),21 which allows for more than two stages and does not require fitting distinct models in the different stages. The GLIMMIX fitting
method (penalized quasi-likelihood) has properties superior to the basic two-step approach outlined above and allows one to evaluate the fit of a multilevel model with conventional likelihood ratio tests. The value of GLIMMIX is limited, however, because it does not provide statistics [for example, relative risk estimates and 95% confidence intervals (CIs)] for quantities of primary epidemiologic interest. The additional work required to obtain this information from GLIMMIX may deter its use—and the application of multilevel models—in epidemiology. Therefore, in the present paper, we show how to use GLIMMIX for epidemiologic analysis with a new application of multilevel modeling.

Multilevel Modeling

In a previous report,7 we applied multilevel modeling to a study of diet and breast cancer. In particular, we developed a multilevel application to produce improved relative risk estimates of breast cancer for foods by using a second-stage nutrient-composition model to pull ordinary estimates toward each other when they have similar levels of nutrients (such as fat or fiber). We here use this approach for the Berlin European Community Multicenter Study on Antioxidants, Myocardial Infarction, and Breast Cancer (EURAMIC) data on diet and breast cancer.22 We have complete information from 40 cases and 90 controls and are interested in estimating the effects of the following 12 foods on breast cancer: cauliflower, broccoli, Brussels sprouts, red cabbage, white cabbage, sauerkraut, allium vegetables (that is, onions and garlic), tomato, pizza, tomato salad, and strawberries. These foods were selected for investigation a priori on the basis of their known levels of anticarcinogenic constituents (for example, glucosinolates).

Our conventional analysis of the data used logistic regression to model the risk of breast cancer as

\[ \logit(p) = \alpha + X\beta + W\gamma \]  

where \( p \) = risk of breast cancer, \( X \) is the matrix of food intake information, \( W \) is a matrix of covariate data (on age, calories, body mass index, socioeconomic status, alcohol intake, and supplemental... hormonal use), \( \alpha \) is the intercept term, and \( \beta = (\beta_1, \ldots, \beta_9)' \) and \( \gamma \) are the vectors of logistic regression coefficients corresponding to the 12 foods and six covariates. Fitting model 1 to the data by maximum likelihood yields the odds ratio (OR) estimates and 95% CIs given in the first set of columns in Table 1. (All results in Table 1 correspond to a 100-gm-per-week increase, vs none, in the foods). Some of the ORs show an inverse association with breast cancer, whereas Brussels sprouts and white cabbage show positive associations. The latter results are implausible (a priori we expected all foods to be either inversely associated or not associated with breast cancer owing to their known constituents’ potential anticarcinogenic properties) and may only reflect small-sample bias or instability of the maximum-likelihood estimates from the conventional logistic regression: with only 40/19 = 2.1 cases per parameter, the dataset fails the criterion of at least five cases and five controls per parameter for valid maximum-likelihood estimates of the coefficients.

We can improve our estimates by using a multilevel model. This approach uses external (prior) information to develop a regression model for a second level or stage in which the effects of exposure variables are partially determined by other factors.2,5-7 The second level (stage) of our multilevel model is a linear regression model for the logistic coefficients \( \hat{\beta} \) of the dietary items,

\[ \beta = Z\pi + \delta \]  

where the ith row of \( Z \) contains second-stage covariates for the ith dietary item, \( \pi \) is a vector of coefficients corresponding to the effects of second-stage covariates on breast cancer, and the elements of \( \delta \) are independent normal random variables with zero means and variances \( \tau \). The second-stage covariates are defined as food constituents that may contribute to dietary effects on breast

<table>
<thead>
<tr>
<th>Food*</th>
<th>Maximum-Likelihood Logistic Regression</th>
<th>GLIMMIX Two-Step Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>95% CI</td>
</tr>
<tr>
<td>Cauliflower</td>
<td>1</td>
<td>0.83–1.3</td>
</tr>
<tr>
<td>Broccoli</td>
<td>0.83</td>
<td>0.29–2.4</td>
</tr>
<tr>
<td>Brussels sprouts</td>
<td>6.1</td>
<td>1.5–26</td>
</tr>
<tr>
<td>Red cabbage</td>
<td>1.3</td>
<td>0.53–3.3</td>
</tr>
<tr>
<td>White cabbage</td>
<td>2.5</td>
<td>0.73–8.4</td>
</tr>
<tr>
<td>Savoy cabbage</td>
<td>0.58</td>
<td>0.16–2.1</td>
</tr>
<tr>
<td>Sauerkraut</td>
<td>0.26</td>
<td>0.08–0.88</td>
</tr>
<tr>
<td>Allium vegetables</td>
<td>0.69</td>
<td>0.44–1.1</td>
</tr>
<tr>
<td>Tomato</td>
<td>1.1</td>
<td>0.75–1.6</td>
</tr>
<tr>
<td>Pizza</td>
<td>0.48</td>
<td>0.05–4.5</td>
</tr>
<tr>
<td>Tomato salad</td>
<td>0.54</td>
<td>0.13–2.2</td>
</tr>
<tr>
<td>Strawberries</td>
<td>0.89</td>
<td>0.30–2.6</td>
</tr>
</tbody>
</table>

* Results correspond to a 100-gm-per-week increase (vs none) in the foods.
† Semi-Bayes, \( \tau_i = 0.35 \) for each food.
cancer. Thus, $Z$ is the diet-nutrient table that gives the amount of nutrients in each food (that is, element $z_{ij}$ of $Z$ is the amount of constituent $j$ reported in food $i$). Here we include the following eight second-stage covariates that may contribute to the effects of food items on breast cancer: glucosinolates, alpha- and beta-carotene, lutein, lycopene, vitamin E, organosulfate, and fiber. Table 2 gives the 12-by-eight second-stage design matrix $Z$ used here.

### Table 2: Second-Stage Design Matrix for Multilevel Model

<table>
<thead>
<tr>
<th>Food</th>
<th>Glucosinolates (mg)*</th>
<th>Carotenoids (µg)†</th>
<th>Vitamin E (µg)‡</th>
<th>Total Tocopherol (mg)§</th>
<th>Organosulfate (mg)¶</th>
<th>Fiber (gm)††</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauliflower</td>
<td>48</td>
<td>0</td>
<td>8</td>
<td>33</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>Broccoli</td>
<td>74</td>
<td>1</td>
<td>1,300</td>
<td>1,800</td>
<td>0</td>
<td>660</td>
</tr>
<tr>
<td>Brussels sprouts</td>
<td>192</td>
<td>6</td>
<td>480</td>
<td>1,300</td>
<td>0</td>
<td>557</td>
</tr>
<tr>
<td>Red cabbage</td>
<td>66</td>
<td>0.7</td>
<td>32</td>
<td>58</td>
<td>0</td>
<td>1,740</td>
</tr>
<tr>
<td>White cabbage</td>
<td>63</td>
<td>0</td>
<td>80</td>
<td>150</td>
<td>0</td>
<td>1,720</td>
</tr>
<tr>
<td>Savoy cabbage</td>
<td>87</td>
<td>0</td>
<td>80</td>
<td>150</td>
<td>0</td>
<td>2,560</td>
</tr>
<tr>
<td>Sauerkraut</td>
<td>73</td>
<td>0</td>
<td>80</td>
<td>150</td>
<td>0</td>
<td>153</td>
</tr>
<tr>
<td>Allium vegetable</td>
<td>0</td>
<td>0</td>
<td>481</td>
<td>736</td>
<td>0</td>
<td>300</td>
</tr>
<tr>
<td>Tomato</td>
<td>0</td>
<td>0</td>
<td>567</td>
<td>109</td>
<td>3,780</td>
<td>869</td>
</tr>
<tr>
<td>Pizza</td>
<td>0</td>
<td>0</td>
<td>170</td>
<td>33</td>
<td>650</td>
<td>1,280</td>
</tr>
<tr>
<td>Tomato salad</td>
<td>0</td>
<td>0</td>
<td>263</td>
<td>170</td>
<td>1,550</td>
<td>2,740</td>
</tr>
<tr>
<td>Strawberries</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>31</td>
<td>0</td>
<td>120</td>
</tr>
</tbody>
</table>

Values correspond to the amount of relevant constituents in 100 gm of given foods (if foods are combined, then a weighted average is used here, where the weight reflects the relative consumption of each food).

* From database of German foods developed in Berlin. 33
† From database created by the National Cancer Institute. 32
‡ From database of German foods developed in Berlin. 33

Fitting the Multilevel Model with GLIMMIX

To use GLIMMIX, we combine models 1 and 2 into a “mixed-effects” logistic model,

$$
\logit(p | X, Z, W) = \alpha + XZ\hat{\pi} + X\hat{\delta} + W\gamma
$$

where the parameters are as defined above; $\pi$ and $\gamma$ are treated as vectors of fixed coefficients and, as above, $\hat{\delta}$ is treated as a vector of random coefficients with mean 0 and variance $\tau_0^2$. Note that the $\tau_0^2$ values reflect any residual associations for the 12 foods on breast cancer, after incorporating the eight constituents in $Z$. At first glance, one might think that incorporating $XZ$ in model 3 would overfit the data relative to the conventional one-stage model 1. Just the opposite is the case, however; the conventional approach (1) is equivalent to using model 3 with no constraint on the residual effects $\hat{\delta}$ (that is, with $\tau_0^2 = \infty$) and thus results in more overfitting than the constrained (that is, $\tau_0^2 < \infty$) multilevel approach.

The GLIMMIX default restricts one to empirical-Bayes, $\tau_0^2$, is estimated from the data. Several applications and simulation studies indicate, however, that when the study size is small, as in the present example, prespecifying the $\tau_0^2$ to define a reasonable range of residual values (that is, a semi-Bayes approach) can work as well as or better than empirical-Bayes,5-8 and there are also philosophical reasons for preferring prespecification.19 One can modify GLIMMIX for semi-Bayes multilevel modeling by incorporating in the random statement the option gdata = "filename," where the file in “filename” contains prespecified $\tau^2$ values. In the present example we use a semi-Bayes approach; because Z includes most of the known relevant food constituents, we set the elements of $\hat{\tau}$ to a relatively moderate value (0.35) for all foods. Assuming a normal distribution for the $\hat{\delta}, \tau_0 = 0.35$ corresponds to a 95% prior certainty that the residual relative risk for the effect of a 100-gm increase in food item $i$ lies in a fourfold range (that is, 0.5–2.0), because exp $(3.92) = 4.0$ if $\tau_0 = 0.35$.

Fitting model 3 to the diet and breast cancer data with GLIMMIX gives estimates of the fixed and random coefficients and their corresponding covariances (Table 3). We can use this output from GLIMMIX to calculate the food-effect coefficients $\hat{\beta} = Z\hat{\pi} + \hat{\delta}$, the natural antilogarithms of which give OR estimates. To obtain 95% CIs, we need the covariance matrix of $\hat{\beta}$ given by

$$
cov(\hat{\beta}) = Z \text{cov}(\hat{\pi})Z' + \text{cov}(\hat{\delta}) + Z \text{cov}(\hat{\delta})Z' + \text{cov}(\hat{\delta}) \text{cov}(\hat{\delta})'
$$

where $\text{cov}(\hat{\pi}), \text{cov}(\hat{\delta}),$ and $\text{cov}(\hat{\pi}, \hat{\delta}) = \text{cov}(\hat{\delta}, \hat{\pi})'$ are covariance-matrix estimates from GLIMMIX, and an

### Table 3: Mixed-Model Equations Solution Matrix

<table>
<thead>
<tr>
<th>\text{cov}(\hat{\pi})</th>
<th>\text{cov}(\hat{\pi}, \hat{\delta})</th>
<th>\text{cov}(\hat{\pi})</th>
<th>\text{cov}(\hat{\delta})'</th>
</tr>
</thead>
<tbody>
<tr>
<td>\hat{\pi}</td>
<td>\hat{\delta}</td>
<td>\hat{\pi}</td>
<td>\hat{\delta}</td>
</tr>
<tr>
<td>\hat{\delta}</td>
<td>\hat{\delta}</td>
<td>\hat{\delta}</td>
<td>\hat{\delta}</td>
</tr>
<tr>
<td>\hat{\pi}</td>
<td>\hat{\delta}</td>
<td>\hat{\pi}</td>
<td>\hat{\delta}</td>
</tr>
<tr>
<td>\hat{\delta}</td>
<td>\hat{\delta}</td>
<td>\hat{\delta}</td>
<td>\hat{\delta}</td>
</tr>
</tbody>
</table>

*df = number of observations minus the number of independent random effects.
apostrophe denotes matrix transposition (Table 3). The square root of the diagonal of $\text{cov}(\hat{\beta})$ gives standard error estimates for the $\hat{\beta}$, which we can use to calculate corresponding 95% CIs.

Upon fitting the mixed model 3 to the example data and calculating the semi-Bayes estimates from model 4, estimates that were unstable and nonsensical became more precise and reasonable, whereas those that were relatively precise did not change much. For example, the maximum-likelihood estimate of the OR for eating 100 additional grams of white cabbage was 2.5 (95% CI = 0.75–8.4); in contrast, the semi-Bayes OR estimate from GLIMMIX was 1.2 (95% CI = 0.59–2.6) (Table 1). In contrast, the OR for eating 100 additional grams of allium vegetables changed little upon applying the multilevel model (Table 1). Setting the elements of $\tau$ to larger values (for example, 0.53, which corresponds to a 95% prior certainty that the residual relative risk lies in an eightfold range) gave less stable results; here, the semi-Bayes OR for eating 100 additional grams of white cabbage was 1.7 (95% CI = 0.64–4.4). Using the GLIMMIX default (that is, empirical-Bayes) led to the common $\tau$ being estimated as 0, which ignores any residual second-stage effects and results in overly precise estimates that are pulled too close together.5

For comparison, we also fit the two-level model (model 1 plus model 2) to the diet and breast cancer data using our earlier weight-least-squares method.3–8,20 This method (third set of columns in Table 1) generally gave results almost identical to the penalized likelihood results from GLIMMIX. The similarity of these results, and their limited improvement over the conventional approach (Table 1), occurred because the mixed model 3 with $\hat{\beta} = 0$ fit the data reasonably well.

Whereas the multilevel estimates in Table 1 are certainly more credible than the maximum-likelihood estimates, previous applications1–7,11,19 have shown more improvement from multilevel modeling than the current analysis. For example, in our earlier diet and breast cancer application,2 we obtained a maximum-likelihood estimate of the OR for eating four 4-inch celery sticks per week vs none equal to 5.1 (95% CI = 0.89–29); in contrast, the multilevel modeling estimate was 0.90 (95% CI = 0.28–2.9). Part of the difference in impact may be due to the fact that our earlier example involved even fewer cases per first-stage parameter (140/92 = 1.5).

The Appendix provides the SAS IML code used to calculate ORs and 95% CIs from our GLIMMIX output. The complete program, a corresponding SAS macro, and instructions on how to obtain results using a multilevel model with GLIMMIX, are available at URL http://darwin.cwru.edu/~witte/hm.html.

Conclusion

With the explication and code given here, epidemiologists can use GLIMMIX to analyze their own data with a multilevel model. Strengths of GLIMMIX include allowing for more than two stages and providing diagnostic statistics. A recent report28 provides further evaluation of GLIMMIX and comparison with variance component software packages specifically written for multilevel modeling.7–20 As with SAS GLIMMIX, however, use of these packages for epidemiologic analysis will require either writing special code or altering output to give the relative risk estimates and corresponding confidence intervals.20 Multilevel modeling can also be undertaken with procedures available in the SAS IML and GAUSS languages.5,6,20,30 These procedures provide standard epidemiologic output and are available from URL http://darwin.cwru.edu/~witte/hm.html.

Acknowledgment

We thank the EURAMIC Study for the example diet and breast cancer data used here.

References

This SAS code was used in conjunction with the SAS macro GLIMMIX to obtain standard epidemiologic output from a multilevel analysis of the diet and breast cancer data. We assume here that the SAS dataset “design” already exists and includes observed values for the outcome variable (“disease”) plus fixed and random effects. In our example, fixed effects include the covariates (“calories,” “BMI” (body mass index), “SES” (socioeconomic status), “alcohol,” “hormones,” and “age”) and the product of the first- and second-stage design matrices (XZ), given as “col1,” “col2,” “col3,” “col4,” “col5,” “col6,” “col7,” and “col8.” Random effects are the exposures of interest (“caulif,” “brocco,” “brusse,” “cabbbrd,” “cabbwh,” “savoy,” “skrout,” “allium,” “tomato,” “pizza,” “tomsalad,” and “strawb”). Second-stage variances are included in the dataset “tau2.” Results are output in “mmeqsol” as shown in Table 3 and then used in SAS proc IML to calculate odds ratios and 95% confidence limits. Note that because we are using a logistic model, a binomial error term is specified in GLIMMIX. Users of this code will need to make minor modifications to reflect their own data structure. An electronic version of this code, a macro that automates some of these steps, and other multilevel modeling software, is available at URL http://darwin.cwru.edu/~witte/hm.html

/* Run the GLIMMIX Macro */
%inc '/glmm612.sas' / nosource;
%glimmix (data=design, procopt=mmeqsol, stmts=%str(
    model DISEASE = calories BMI SES alcohol hormones age
    col1 col2 col3 col4 col5 col6 col7 col8 / solution;
    random
caulif brocco brusse cabbrd cabbwh savoy skrout allium
tomato pizza tomsalad strawb
    /gdata=tau2 solution;
    make 'mmeqsol' out=mmeqsol;
    error=binomial)
run;
/* Use SAS proc IML to read estimates from GLIMMIX output into matrix form */
proc iml;
use mmeqsol;
read all var
   {_col1 _col2 _col3 _col4 _col5 _col6 _col7 _col8 _col9 _col10
into mmeqsol;
p = mmeqsol [8:15, 28]; /* Fixed coefficient estimates (i.e., of XZ) */
covpi = mmeqsol [8:15, 8:15]; /* Covariance of fixed coefficient estimates */
delta = mmeqsol [16:27, 28]; /* Random coefficient estimates */
covd = mmeqsol [16:27, 16:27]; /* Covariance of random coefficient estimates */
cov_pd = mmeqsol [8:15, 16:27]; /* Covariance of fixed and random coeff. */
/* Calculate estimates, odds ratios and CIs. Z needs to be entered here. */
b = Z*pi + delta; /* Food-effect coefficients, equation (4) */
orb = exp (b); /* Odds ratios */
varb = Z*covpi*t (Z) + covd + Z* cov_pd + t (cov_pd) *t (Z); /* Covariance */
stderr = sqrt (vecdiag (varb)); /* matrix for the betas, equation (5) */
lower = exp (b-1.96*stderr); /* Standard errors of betas */
upper = exp (b+1.96*stderr); /* Upper and lower 95% confidence intervals */
print orb lower upper; /* Print results */