MODELS OF VARYING PARAMETRIC FORM IN CASE-REFERENT STUDIES

ALEXANDER M. WALKER1,2 AND KENNETH J. ROTHMAN³


The analysis of case-referent data can be based on a consideration of case and referent counts in various exposure categories as the realizations of a set of binomial processes. After appropriate modeling of the binomial parameters, a joint likelihood function can be formed and maximized to obtain estimates of the parameters constituting the model elements. The procedure has been applied to the problem of additive and multiplicative models of disease incidence rates, as encountered in case-referent studies. Likelihood ratios can be used to compare models with equal numbers of parameters. These ratios do not lead to significance tests, but to estimates of the relative degree of corroboration of different hypotheses by the data at hand.

The issue of "modeling" arises in epidemiologic studies when data are available on multiple determinants of disease. Whether the investigator is searching for an unbiased estimate of the association between disease and a single determinant or whether he wishes to estimate the magnitude of several associations, the question of how to deal with combined effects must be answered as preliminary to quantitative analysis. The answer generally takes the form of a parametric model of combined effect.

In the analysis of epidemiologic data, the implicit model of combined effect has traditionally been a multiplicative one. That is, if the incidence rate of disease in persons exposed to A is \( R(A) \) times the incidence rate in comparable persons not exposed to A, and if the similarly defined relative incidence rate corresponding to B is \( R(B) \), then under the multiplicative model the incidence rate of disease in persons exposed to both A and B, compared to the incidence rate in persons exposed to neither, is \( R(A)R(B) \). A corollary of the multiplicative model is that \( R(A) \) is not dependent on \( B \), and can be estimated in the presence or absence of B.

Multiplicative models have been directly incorporated into many analytic procedures for case-referent studies, and are tacitly assumed in several others. The multiplicative model of effect appears in the likelihood functions of multiple logistic models when disease status is the dependent variable (1, 2). Vitaliano (3) has given an extended example of analysis of interaction based on the multiplicative aspects of logistic modeling. Log-linear functions (4) in the two-sample case provide models equivalent to the logistic formulation. Proportional hazards models...
(5) when applied to the analysis of case-referent data (6) are, like logistic models, explicit in their inclusion of multiplicative relations in the likelihood function. The multiplicative model is moreover implicit in the likelihood function of logistic models in which exposure status is the dependent variable (7), and is a formal justification for estimating a common odds ratio in all forms of stratified analyses (8).

Despite their utility, multiplicative models of disease incidence are not always implied by scientific hypotheses about the joint effect of determinants of disease. In this regard, it is worth recalling that Mantel and Haenszel (9) were guarded about the interpretation of their estimate of a common odds ratio "since the assumption of a constant relative risk can usually be discarded as untenable." Greenland (10) has noted a number of problems that arise when logistic models are applied to the analysis of data generated by processes with additive components. Nonetheless, the application of alternative models, particularly in the context of case-referent studies, has been limited. The restricted use of the additive models which have been developed may stem from difficulty in interpreting their parameters meaningfully: familiar additive parameters, such as attributable risk (11–13), are not directly estimable in case-referent studies. In an effort to provide a measurable, meaningful additive parameter, Rothman (14) proposed an index of synergy, which quantifies deviations from additivity of incidence rates, as observed in case-referent studies. Walker (15) has recast the problem as one of estimating the etiologic fraction (12) attributable to the combined effect of two exposures. Berry (16) has pointed out that it is possible to consider linear functions of the rate ratio or relative risk directly, and that these can be modeled in case-referent studies, with polychotomous and continuous classifications of one or more exposures. Berry (16) and Thompson and Baker (17) have made an important practical contribution by indicating how the parameters of such functions can be estimated using Generalized Linear Interactive Modelling (GLIM) (18), a widely available modeling package.

Our primary purpose in this communication is to illustrate the construction models of joint effect in case-referent studies, in which the impetus for construction of a particular model is an hypothesis concerning the functional form of the incidence rate relations in the source population. Sheehe (19) pioneered this practice of deducing "retrospective" functions from "prospective" models; its advantage lies in the relative ease of interpretation of the case-referent parameters, which follow naturally from the parameterization applied to the source population. We provide the estimating equations for the examples given; the use of GLIM could be substituted for our solutions and would provide numerically equivalent parameter estimates.

Gardner and Munford (20) have recently developed an additive model based on the distribution of two dichotomous exposures in case and referent groups, but chose to sidestep the question of interpretation by using the model as a vehicle for the testing of additive and multiplicative hypotheses. Dayal (21) independently developed a very similar additive model, which he proposed as the basis for tests of interaction on an additive scale. The problem of comparing models in epidemiologic applications has been reviewed in detail recently by Greenland (22), who also addresses the very closely related issue of testing for interaction. In common with Berry (16), Gardner and Munford (20), and Thomas (23) we will suggest not a test, but a likelihood ratio criterion for the comparison of models of different parametric structure. Edwards (24, chap-
MODELS OF VARYING PARAMETRIC FORM

131

ters 2 and 3) provides a justification for likelihood ratio measures which does not depend on the repeated sampling arguments of classical statistical theory.

RELATIVE INCIDENCE RATES AND CASE-REFERENT RATIOS

If individuals are removed by illness with an incidence rate $I$ from a nondiseased equilibrium population of size $P$, then in a period of time $t$, a total of $IPt$ individuals will become ill. Consider an unexposed subpopulation of size $P(0)$ with incidence rate $I(0)$, and a subpopulation exposed to factor $A$, of size $P(A)$ with incidence rate $I(A)$. Over a period of time $t$, the number of individuals developing disease in the two subpopulations will be $I(0)P(0)t$ and $I(A)P(A)t$. Case-referent studies are carried out by sampling diseased individuals (in principle as they become diseased) and nondiseased individuals in the source population from which the diseased persons came. After the individuals have been sampled, their exposure status is ascertained, and they are sorted into a four-fold table, whose expected cell sizes are

<table>
<thead>
<tr>
<th>Level of $C$</th>
<th>Exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>$R(A)M$</td>
<td>$M$</td>
</tr>
<tr>
<td>C1</td>
<td>$R(A,C1)M$</td>
<td>$R(C1)M$</td>
</tr>
<tr>
<td>C2</td>
<td>$R(A,C2)M$</td>
<td>$R(C2)M$</td>
</tr>
<tr>
<td>Etc</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where $F(c)$ is the sampling fraction among cases and $F(r)$ is the sampling fraction applied to the nondiseased population in order to obtain the reference series. The odds ratio for this table is $I(A)/I(0)$, that is $R(A)$ as defined above (25). Using $R(A)$, the number of exposed cases could be re-expressed as $F(c)R(A)I(0)P(A)t$. The case-referent ratio among the nonexposed is $F(c)I(0)tF(r)$. Define this quantity as $M$; the case-referent ratio among the exposed is $R(A)M$.

By extension, consider the classification of cases and referents according to the presence or absence of $A$ and according to the level of some other exposure $C$, which may be absent or which may take on one of several values. As long as the sampling fractions used for cases and referents are the same over strata defined by the levels of $C$, then the case-referent ratios would be as follows:

The goal of the next sections will be to derive expressions for $R(A,C)$ from additive and multiplicative models of incidence rates and to use those expressions for the estimation of effect. The derivations which follow are presented as illustrations of a class of estimation procedures which need not be limited a priori to any particular parameterization, but which can be adapted to whatever incidence rate models might be suggested by prior, scientific evidence.

AN ADDITIVE MODEL

If the effect of exposure $A$ were to change disease incidence rate by a fixed amount, the term $I(A)$ would be decomposable into two parts. Let $E(A)$ be the effect of $A$, expressed as the difference in incidence rates observed in the presence and absence of $A$, then

$$I(A) = I(0) + E(A).$$

Similarly, for any level $i$ of the confounder, $C$, then

$$I(Ci) = I(0) + E(Ci).$$

The essence of an additive model for disease rates is the estimation of combined effect of $A$ and $C$ as the sum of the separate effects. That is

$$E(,Ci) = E(A) + E(Ci)$$

and

$$I(A, Ci) = I(0) + E(A) + E(Ci).$$
The case-referent ratios for a combination of \((A,C_i)\) can be expressed as

\[ R(A,C_i)M = (1 + W + K_i)M \]

where \(W = E(A)/I(0)\) is the relative effect of \(A\) over the baseline, \(I(0)\), and \(K_i\) is the relative effect of the confounder, \(C_i\), at level \(i\) over the same baseline \(I(0)\). \(W\) and \(K\) of the additive model measure the excess rate ratio attributable to the exposures or traits which are under study.

**A multiplicative model**

If, on the other hand, the effect of exposure \(A\) were to increase disease incidence by a constant multiple, the term \(I(A)\) could be decomposed in a different fashion. Let \(W\) in this case be the multiplier by which incidence is changed in the presence of \(A\), then

\[ I(A) = WI(0). \]

Similarly, for any level \(i\) of \(C\), let \(K_i\) be in this case a multiplication of the incidence rate which is effected by the presence of \(C_i\); then

\[ I(C_i) = K_iI(0). \]

The combined incidence rate in the presence of both \(A\) and \(C_i\), under the multiplicative model is

\[ I(A,C_i) = W K_i I(0). \]

The case-referent ratios observed for a combination of \((A,C_i)\) can be expressed as

\[ R(A,C_i)M = W K_i M. \]

\(W\) and \(K\) of the multiplicative model are estimates of the familiar rate ratios attributable to the exposures or traits under study.

**Estimation of parameters**

Under either the additive or the multiplicative models presented above, the parameters \(W, K_i,\) and \(M\) can be estimated using a conditional maximum likelihood procedure. At each level of \((A,C_i)\) consider the observed numbers of cases and referents to be the realization of a binomial process, conditional on the case-plus-referent total count for each level. Thus, if there are \(n\) exposure levels of \(C\) (including the no-exposure category, if there is one) and two levels of \(A\), as outlined here, then there will be \(2n\) binomial processes to consider. Estimation proceeds by forming a joint likelihood function, which is maximized over \(W, K_i,\) and \(M\). Here it is convenient to maximize the function by finding the parameter values for which the vector of first derivatives of the natural log likelihood is zero, using a Newton-Raphson technique for iterative solution of the problem. The matrix of second derivatives, which is obtained in the iterating process, can, at the end of the final iteration, be inverted and multiplied by \(-1\) to produce an asymptotically valid estimate of variance-covariance matrix.

Likelihood functions for the additive and multiplicative models presented here are given in the Appendix, together with formulae for the vectors of first derivatives and the matrices of second derivatives. While the models and solutions presented might be considered prototypes for effect estimation, other formulations, including mixed models, models with more than two factors, models with several levels of each factor, and models incorporating dose-response functions, can be built up in a relatively straightforward manner and their parameters estimated.

**Comparisons of models**

If meaningful hypotheses about the relations between determinant and disease entail specific statistical models (in this case for the binomial processes realized in the case-referent ratios), then it should be possible to compare hypotheses by comparing the accuracy with which their implied models describe the data. For this purpose it is worthwhile to distinguish a hierarchy of models (and corresponding classes of hypotheses).
Models may be arranged in this hierarchy according to the extent to which they specify predicted outcomes, or numerical values in a body of data (26, Sections 31–39). The most general models specify only the form (e.g., additive, multiplicative) of value relations predicted for a set of data. The least general models specify actual predicted values of various data points. We will call the most general models "formal models" and the least general ones "particular models." A particular model is a formal model for which parameter values are specified. The theory of choosing an appropriate particular model, given a formal model and a set of observations, is precisely the theory of parameter estimation and hypothesis testing.

For the purposes of comparing formal models such as the two developed in previous sections in the light of a body of data, we propose to use the following method, which has been suggested for the comparison of models in very similar circumstances (16, 20, 23). Choose in advance the number of parameters to be estimated. Using the same number of parameters in each formal model, find the particular models which maximize the probability of the observed data under each formal model. The ratio of the maximum attainable likelihoods under the two formal models is taken as the relative degree of corroboration of the models by the data at hand. A likelihood ratio obtained by this method provides a quantitative measure of the relative degree of corroboration of the two alternative hypotheses. Given two competing hypotheses accounting for the same observations, the one which is more specific is taken to be more highly corroborated. Greater specificity here means greater probability density in the neighborhood of the expected value of the vector of observations. Pairs of hypotheses with identical predictions will, with this likelihood ratio procedure, always be equally corroborated, no matter how different their formal structures.

**Example**

Postmenopausal women who develop endometrial cancer are on the whole heavier than women who do not develop the disease. One possible explanation of the phenomenon is that heavy women are more exposed to endogenous estrogens, which are produced in postmenopausal women by conversion of steroid precursors to active estrogens in peripheral fat. In the face of varying levels of endogenous estrogen production, one might ask whether the carcinogenic potential of exogenous estrogens would be the same in all women.

Kelsey (27) and coworkers have examined the relation between weight, replacement estrogen therapy, and endometrial cancer in a case-referent study. Table 1 shows the counts of cases with endometrial cancer and reference group women classified by history of replacement estrogen use and weight. The observed relative risk associated with estrogen use declines from 5.0 in women weighing less than 57 kg to 1.3 in women weighing more than 75 kg. Moreover, among nonusers of estrogens, risk increases with weight: the relative risk comparing women over 75 kg to those under 57 kg is 4.6.

Table 1 displays the odds ratio estimates for each of the strata, as derived from the additive and multiplicative models. As would be expected, odds ratios based on the additive model are lower in the higher risk strata. This accords with a diminished relative risk observed with a constant attributable risk and a higher background risk. The odds-ratio estimate from the multiplicative model is constant across strata. The parameter estimates for the models are given in table 2; the fitted estimates of the odds ratios were derived from these by observing that in the case of the jth stratum the cross-
### Table 1

Replacement estrogen use, weight, and endometrial cancer: results of a case-referent study of Connecticut women, 1980

<table>
<thead>
<tr>
<th>Weight (kg)</th>
<th>Estrogen replacement</th>
<th>Odds ratio</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>Referents</td>
<td></td>
</tr>
<tr>
<td>&lt;57</td>
<td>20</td>
<td>12</td>
<td>5.0</td>
</tr>
<tr>
<td>57-75</td>
<td>37</td>
<td>45</td>
<td>2.8</td>
</tr>
<tr>
<td>&gt;75</td>
<td>9</td>
<td>42</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Mantel Haenszel chi = 5.565
Additive chi = (1 + W)ln(1 + W)/SE (W) = 5.479
Multiplicative chi = Wln(W)/SE(W) = 5.427

* Data from Kelsey (27)

### Table 2

Parameters fitted to the data of table 1, a case-referent study of replacement estrogen use, weight, and endometrial cancer in Connecticut women, 1980

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Additive</th>
<th>Multiplicative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Variance</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>Variance</td>
</tr>
<tr>
<td>W</td>
<td>1.87</td>
<td>0.305</td>
</tr>
<tr>
<td>M</td>
<td>0.118</td>
<td>0.000334</td>
</tr>
<tr>
<td>K1</td>
<td>-0.423</td>
<td>0.0353</td>
</tr>
<tr>
<td>K2</td>
<td>0*</td>
<td>-*</td>
</tr>
<tr>
<td>K3</td>
<td>1.46</td>
<td>0.317</td>
</tr>
<tr>
<td>Log likelihood -435.228 + constant</td>
<td>-437.602 + constant</td>
<td></td>
</tr>
</tbody>
</table>

Likelihood ratio (A/M) = 10.7

* Reference stratum.

A verbal interpretation of the additive model the transformed ratio measure is ln(W), with standard error SE(W)/W. Interval estimation can be performed using the same transformation. Had the model-based tests been performed using untransformed variances and point estimates, they would both have yielded normal deviates smaller than that obtained in the Mantel-Haenszel test.

The ratio of likelihoods of the additive to the multiplicative models is 10.7, indicating that the additive model is much more highly corroborated by the data than is the multiplicative model.
K3 = 1.46 would be: "Women weighing 57–75 kg and taking replacement estrogens developed endometrial cancer 187 per cent more commonly than did women of similar weights who had not taken replacement estrogens. Among nonusers of estrogens women weighing less than 57 kg had 42 per cent less endometrial cancer risk and women weighing more than 75 kg had 146 per cent more risk than women weighing 57–75 kg. Total risk appeared to be the sum of risks associated with weight and estrogen use.

REFERENCES
1. Farewell VT. Some results on the estimation of logistic models based on retrospective data Biometrika 1974;66:27-32

APPENDIX

Log likelihood functions and derivative arrays for the additive and multiplicative models

A1. Notation

For each stratum, i, the counts of cases and referents, exposed and nonexposed are referred to as follows

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a_i</td>
<td>b_i</td>
</tr>
<tr>
<td>Referents</td>
<td>c_i</td>
<td>d_i</td>
</tr>
</tbody>
</table>
Let \( n_{1i} = a_i + c_i \),
\( n_{0i} = b_i + d_i \).

Let \( K_i \) be the stratum effect parameter, and \( W \) be the exposure effect parameter. Let there be a reference stratum \( r \) for which \( K_r = 0 \), and for which the parameter \( M \) describes the case-referent ratio in the nonexposed.

**A2. Additive model**

The likelihood equation is the product of a series of binomial probabilities, two to each stratum. The likelihood function is

\[
L = G \prod_i \left[ \frac{[M(1 + K_i)]^{b_i}}{1 + M(1 + K_i)} \right]^{n_{0i}} \left[ \frac{[M(1 + K_i + W)]^{n_{1i}}}{1 + M(1 + K_i + W)} \right]^{n_{1i}}
\]

where

\[
G = \prod_i \frac{n_{0i}! n_{1i}!}{b_i! d_i! a_i! c_i!}.
\]

For the convenience of notation, define the following terms

\[
P_i = 1 + K_i
\]
\[
Q_i = 1 + MP_i
\]
\[
R_i = 1 + K_i + W
\]
\[
S_i = 1 + MR_i.
\]

The log likelihood is

\[
\ln L = \ln G + \sum_i \left[ (a_i + b_i) \ln M + b_i \ln P_i + a_i \ln R_i - n_{0i} \ln Q_i - n_{1i} \ln S_i \right].
\]

The first derivatives of the log likelihood with respect to \( W, M, \) and \( K_i \) \( i \neq r \) are

\[
\frac{\partial \ln L}{\partial W} = \sum_i \left( a_i/R_i - M n_{1i}/S_i \right)
\]

\[
\frac{\partial \ln L}{\partial M} = \sum_i \left[ (a_i + b_i)/M - P_i n_{0i}/Q_i - R_i n_{1i}/S_i \right]
\]

\[
\frac{\partial \ln L}{\partial K_i} = b_i/P_i + a_i/R_i - M n_{0i}/Q_i - M n_{1i}/S_i.
\]

Second derivatives are

\[
\frac{\partial^2 \ln L}{\partial W^2} = \sum_i \frac{\partial \ln L}{\partial W \partial K_i}
\]

\[
\frac{\partial^2 \ln L}{\partial W \partial M} = - \sum_i n_{1i}/S_i^2
\]

\[
\frac{\partial^2 \ln L}{\partial W \partial K_i} = M^2 n_{1i}/S_i^2 - a_i/R_i^2
\]

\[
\frac{\partial^2 \ln L}{\partial M^2} = \sum_i \left[ P_i n_{0i}/Q_i^2 + R_i n_{1i}/S_i^2 - (a_i + b_i)/M^2 \right]
\]
MODELS OF VARYING PARAMETRIC FORM 137

A3. Multiplicative model

The likelihood function is

\[ L = G \prod_i \frac{(MK_i)^{b_i}}{(1 + MK_i)^{a_i}} \frac{(MK_iW)^{n_i}}{(1 + MK_iW)^{n_i}} \]

where \( G \) is defined as in A2. Again, for convenience, define the following terms

- \( P_i = n_{ui}/(1 + MK_i) + Wn_{ui}/(1 + MK_i) \)
- \( Q_i = n_{ui}/(1 + MK_i)^2 + W^2n_{ui}/(1 + MK_i)^2 \)
- \( R_i = K_i n_{ui}/(1 + MK_iW) \)
- \( S_i = R_i/(1 + MK_iW) \)
- \( T_i = K_i S_i. \)

The log likelihood is

\[ \ln L = \ln G + \sum_i \left[ (a_i + b_i) \ln M + (a_i + b_i) \ln K_i + a_i \ln W - n_{ui} \ln (1 + MK_i) - n_{ui} \ln (1 + MK_iW) \right]. \]

The first derivatives of the log likelihood with respect to \( W, M, \) and \( K \) are

\[ \frac{\partial \ln L}{\partial W} = \sum_i \left[ (a_i + b_i)M - MR_i \right] \]

The second derivatives are

\[ \frac{\partial^2 \ln L}{\partial W^2} = \sum_i \left( M^2T_i - a_i/W^2 \right) \]

\[ \frac{\partial^2 \ln L}{\partial W \partial M} = \sum_i \left( MT_i - R_i \right) \]

\[ \frac{\partial^2 \ln L}{\partial W \partial K_i} = -MS_i/K_i \]

\[ \frac{\partial^2 \ln L}{\partial M^2} = \sum_i \left( K_i Q_i - (a_i + b_i)/M^2 \right) \]

\[ \frac{\partial^2 \ln L}{\partial M \partial K_i} = MK_i Q_i - P_i \]

\[ \frac{\partial^2 \ln L}{\partial K_i^2} = M^2 Q_i - (a_i + b_i)/K_i^2 \]

\[ \frac{\partial^2 \ln L}{\partial K_i \partial K_j} = 0, i \neq j. \]