BIOMETRICS 34, 100–105
March, 1978

Are There Two Logistic Regressions for Retrospective Studies?

N. BRESLOW and W. POWERS

Department of Biostatistics SC-32, University of Washington, Seattle, Washington 98195, U.S.A.

Summary

A comparison is made between two different approaches to the linear logistic regression analysis of retrospective study data: the prospective model wherein the dependent variable is a dichotomous indicator of case/control status; and the retrospective model wherein the dependent variable is a binary or polychotomous classification of exposure. The two models yield increasingly similar estimates of the relative risk with increasing degrees of covariate adjustment. When the covariate effects are saturated with parameters, the relative risk estimates are identical. The prospective model is generally to be preferred for studies involving multiple quantitative risk factors.

1. Introduction

Analysis of retrospective study data is oriented toward estimating the relative risk (RR), i.e. the ratio of disease incidence among persons exposed versus those not exposed to the risk factor(s) under investigation (Mantel and Haenszel 1959). Adjustments are usually made for confounding factors, related to both disease and exposure, whose effects might otherwise bias the results. An important feature of the analysis is to identify and quantify the effects of factors which modify the relative risk (Miettinen 1974).

Two distinct versions of the logistic regression model (Cox 1970) have been proposed for such analyses. In one, termed here the *prospective* model, the dependent or outcome variable is formally considered to be the binary indicator of case/control status. Regression variables represent both risk and confounding factors, with interaction terms used to quantify effect modification. Although this model would seem most appropriate for prospective sampling, wherein healthy individuals are classified by exposure and followed up to determine disease occurrence, Anderson (1972, 1973) has shown that it may be applied also to retrospective studies involving separate samples of cases and controls. See Mantel (1973), Siegel and Greenhouse (1973) and Prentice and Breslow (1978) for further accounts of the relationship between prospective and retrospective sampling.

In the *retrospective* logistic model, proposed by Prentice (1976), the outcome variable is a binary indicator of exposure to one particular risk factor. Here the regression variables represent disease (case/control) status, other risk factors and confounding factors, with effect modification expressed through interaction terms involving disease status. This approach accords better with the sampling scheme actually used in retrospective studies.

In this note we compare these two logistic models by relating both to a log-odds ratio regression model for a series of 2×2 tables (Zelen 1971, Breslow 1976). Conditions are derived under which they yield identical estimates of the RR and its modifications. Data

Key Words: Retrospective studies; Odds ratio; Logistic regression.

from the Oxford Childhood Cancer Survey (Kneale 1971) are analyzed to illustrate the principle involved, and the numerical results serve as a basis for further discussion.

2. *Conditions for Equivalence of the Prospective and Retrospective Estimates*

Retrospective studies typically involve three types of factors: (i) the disease d which occurs at two levels $d = 1$ (case) and $d = 0$ (control); (ii) risk factor(s) $\mathbf{f} = (f_1, f_2, \cdots)'$; and (iii) covariates $\mathbf{g} = (g_1, g_2, \cdots)'$. Suppose for the moment there is but a single risk factor coded $f = 1$ for exposed and $f = 0$ for non-exposed. The RR is well approximated by the odds ratio which, for a given set of covariates $\mathbf{g}$, is defined by

$$\psi(\mathbf{g}) = \frac{P(d = 1 | f = 1, \mathbf{g}) P(d = 0 | f = 0, \mathbf{g})}{P(d = 1 | f = 0, \mathbf{g}) P(d = 0 | f = 1, \mathbf{g})} \tag{1}$$

or, equivalently (Cornfield 1951, Prentice 1976),

$$\psi(\mathbf{g}) = \frac{P(f = 1 | d = 1, \mathbf{g}) P(f = 0 | d = 0, \mathbf{g})}{P(f = 1 | d = 0, \mathbf{g}) P(f = 0 | d = 1, \mathbf{g})}\,. \tag{2}$$

Covariates are included in the analysis for either, or both, of two reasons: (i) they are themselves risk factors for the disease and hence may need to be controlled in the analysis; or (ii) they may interact with disease status and risk variables so as to influence the RR.

The prospective logistic regression model is specified by the equation

$$\operatorname{logit} P(d = 1 | f, \mathbf{g}) = \alpha + \beta f + \boldsymbol{\gamma}'\mathbf{z} + f\boldsymbol{\delta}'\mathbf{y} \tag{3}$$

where $\mathbf{z} = \mathbf{z}(\mathbf{g})$ is a p-vector representing the effects of the covariates on d, while $\mathbf{y} = \mathbf{y}(\mathbf{g})$ is a q-vector of covariate terms which interact with f. Similarly, the retrospective model is specified by the equation

$$\operatorname{logit} P(f = 1 | d, \mathbf{g}) = \alpha^* + \beta d + \boldsymbol{\gamma}^{*\prime}\mathbf{z}^* + d\boldsymbol{\delta}'\mathbf{y} \tag{4}$$

where $\mathbf{z}^* = \mathbf{z}^*(\mathbf{g})$ is a p-vector representing the main effects of the covariates on f, while $\mathbf{y}$ again represents the interactions. This notation is justified by the fact that the models (3) and (4) both imply the same linear structure for the ln odds ratio (Zelen 1971). From (1) and (3) or (2) and (4), it follows that

$$\ln \psi(\mathbf{g}) = \beta + \boldsymbol{\delta}'\mathbf{y}. \tag{5}$$

Although β and $\boldsymbol{\delta}$ have the same meaning for both models, the estimates of these parameters derived from them will usually be different. Varying the covariate term $\mathbf{z}$ used in model (3), or the term $\mathbf{z}^*$ used in (4), will change the estimates of β and $\boldsymbol{\delta}$ even though the structure (5) remains fixed. There is, however, the following situation in which (3) and (4) yield identical estimates for β and $\boldsymbol{\delta}$.

When the covariates $\mathbf{g}$ are discrete, the data may be arranged as a series of 2 × 2 tables, one for each of a finite number of covariance outcomes corresponding to the possible values of $\mathbf{g}$:

$$\begin{array}{lccc} & \text{Case} & \text{Control} & \\ \text{Exposed} & n_{11}(\mathbf{g}) & n_{10}(\mathbf{g}) & n_{1+}(\mathbf{g}) \\ \text{Non-exposed} & n_{01}(\mathbf{g}) & n_{00}(\mathbf{g}) & n_{0+}(\mathbf{g}). \\ & n_{+1}(\mathbf{g}) & n_{+0}(\mathbf{g}) & n_{++}(\mathbf{g}) \end{array} \tag{6}$$

For such data Breslow (1976) discusses two methods of estimating β and $\mathbf{\delta}$ directly from (5): (i) "exact" maximum likelihood (ML) based on the non-central hypergeometric distribution; and (ii) "asymptotic" ML based on the normal approximation to this distribution. In cases where the covariate effects on disease are saturated, that is fully modeled by the vector **z**, so that the number of independent parameters α, γ equals the number of covariance outcomes, ML fitting of the prospective model (3) and asymptotic ML fitting of (5) yield identical results. This may be argued as follows.

First, the observed $n_{ij}(\mathbf{g})$ and ML fitted $\hat{n}_{ij}(\mathbf{g})$ frequencies for any prospective model satisfy $n_{+j}(\mathbf{g}) = \hat{n}_{+j}(\mathbf{g})$ for $j = 0$, 1 and all **g**. Second, when the covariate effects in (3) are saturated by **z**, the likelihood equations specify that $n_{1+}(\mathbf{g}) = \hat{n}_{1+}(\mathbf{g})$, which implies that all the marginal totals of the 2 × 2 tables (6) are fit exactly. Third, the likelihood equations from (3) always specify that $\sum_{\mathbf{g}} n_{11}(\mathbf{g})\mathbf{y}(\mathbf{g}) = \sum_{\mathbf{g}} \hat{n}_{11}(\mathbf{g})\mathbf{y}(\mathbf{g})$ whether or not **z** saturates the model.

An identical argument shows that the same three sets of equations are satisfied by the ML fitting of the retrospective model (4) when **z*** saturates the covariate effects on exposure. Since they are precisely the equations used to fit (5) directly by asymptotic ML, this proves that all three approaches yield identical estimates of β and $\mathbf{\delta}$. An extension of the results of Breslow (1976) shows that the usual expressions for the variances and covariances of these estimates, obtained by inversion of the matrices of second partials of the ln-likelihood, are likewise identical under the stipulated conditions.

In practice it is unlikely that one would attempt to include so many parameters as to obtain complete saturation of the covariate effects. Hence the main interest in the previous result is its implication that, as one adds terms to describe more fully the effects of the covariates, the estimates of β and $\mathbf{\delta}$ obtained from the two models will tend to converge towards common values. This phenomenon is illustrated below.

3. *Further Study of the Oxford Survey Data*

For an example having nearly "continuous" covariates, we use data from the Oxford Childhood Cancer Survey reported by Kneale (1971). Cases were 6,347 children under 10 years of age who died of malignant disease during 1944–1964. Controls were neighbor children of the same age, sex and year of birth. *In-utero* radiation, as reported during an interview of the mother, was the risk factor under investigation. The two covariates were age (g_1) coded 0 years = −9, 1 year = −7, · · ·, 9 years = 9; and year of birth (g_2) coded 1944 = −10, 1945 = −9, · · ·, 1954 = 10. Due to the limited period of case ascertainment, not all combinations of the 10 ages and 21 years occurred; in fact only 120 tables of the form (6) were available for analysis. Results are presented in Table 1. Two odds ratio parameters were estimated in each run: β represents the overall ln RR, and δ represents the linear change in ln RR with year of birth ($y = g_2$). Estimates of these two parameters are shown with their standard errors according to the degree of polynomial adjustment for the two covariates. Also shown are the goodness-of-fit χ^2 for each model.

The unadjusted analyses, obtained from (3) and (4) without **z** or **z*** terms, give quite different impressions of the degree of decline in the RR with year of birth. This happens to be underestimated by the retrospective model. The decline itself is probably related to improvement in radiologic technology, which led to lower doses of obstetric radiation during these years (Bithell and Stewart 1975). However, after adjustment for both linear and quadratic effects of age and year, the concordance is quite respectable. By the time a fifth degree polynomial is fitted, requiring the estimation of 20 independent parameters γ or γ^*, the prospective and retrospective estimates agree to the fourth decimal place. In normal practice,

TABLE 1

Comparison of Estimates of the Log-Odds Ratio Relating Childhood Cancer and In-Utero X-Ray Exposure Under "Retrospective" and "Prospective" Regression Models, Depending on the Degree of Adjustment for Age at Death and Year of Birth

Parameter Estimated	Prospective Coef. ± S.E.	Retrospective Coef. ± S.E.	Covariates
$\ln\psi(\beta)$	0.5113 ± 0.0562	0.5036 ± 0.0559	None
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0343 ± 0.0136	−0.0143 ± 0.0083	
χ^2_{237}	114.62	326.86	
$\ln\psi(\beta)$	0.5124 ± 0.0562	0.5105 ± 0.0561	Linear terms in
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0382 ± 0.0143	−0.0317 ± 0.0130	age and year
χ^2_{235}	113.87	325.71	
$\ln\psi(\beta)$	0.5149 ± 0.0563	0.5148 ± 0.0563	Linear and quadratic
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0384 ± 0.0143	−0.0386 ± 0.0144	terms in age
χ^2_{232}	113.57	280.72	and year
$\ln\psi(\beta)$	0.5150 ± 0.0563	0.5151 ± 0.0563	Linear, quadratic
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0385 ± 0.0143	−0.0387 ± 0.0144	and cubic terms
χ^2_{228}	113.55	279.18	in age and year
$\ln\psi(\beta)$	0.5157 ± 0.0564	0.5158 ± 0.0564	Linear, quadratic, cubic
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0385 ± 0.0143	−0.0386 ± 0.0144	and quartic terms
χ^2_{223}	113.43	260.63	in age and year
$\ln\psi(\beta)$	0.5163 ± 0.0564	0.5163 ± 0.0564	Linear, quadratic, cubic,
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0386 ± 0.0144	−0.0386 ± 0.0144	quartic and quintic
χ^2_{217}	113.33	244.60	terms in age and year
$\ln\psi(\beta)$	0.5218 ± 0.0567	0.5218 ± 0.0567	Complete saturation
$\ln\psi \times \mathrm{Yr}$ (δ)	−0.0390 ± 0.0145	−0.0390 ± 0.0145	of age and
χ^2_{118}	112.52	112.52	year marginals

of course, one would probably not consider for adjustment purposes more than a second or at most third degree polynomial in the covariates.

There is very little change in the goodness-of-fit of the prospective model as additional covariate terms are added: the coefficients of these terms are all near zero and the estimates of β and δ remain quite constant. This behavior is explained by the fact that subjects were matched on the two covariates, so that each of the 120 2 × 2 tables had equal numbers of cases and controls. Other numerical examples considered by the authors, in which there was no matching, do not evidence such behavior.

There is a particularly large drop in the goodness-of-fit χ^2 for the retrospective model after the addition of quadratic covariate effects. Subsequent examination of the data revealed that the proportion of children irradiated increased from 1944 to the mid-1950's and declined thereafter, again reflecting changes in radiologic practice.

It was not feasible in this particular example to obtain complete saturation of the covariate effects using a computer program for linear logistic regression analysis, since to do so would have required estimation of 122 separate parameters. However the equivalent ln odds ratio regression analysis had already been carried out, and the last set of estimates in Table 1 was taken from the third line of Table 3 of Breslow (1976). The result is reassuring since it shows that estimation of the additional 99 covariate terms beyond those in the fifth degree polynomial equation leads to negligible changes in the assessment of the RR. If the data were less extensive, one would expect the estimation of these additional parameters to have a greater ffect on the RR estimate (see discussion).

4. *Multiple Levels of a Quantitative Risk Factor*

Consider now the situation where the risk factor f is quantitative, being recorded at J levels $x_1 < \cdots < x_J$ of some variable x, with x_0 denoting zero or baseline exposures. Let $\psi_j(\mathbf{g})$ represent the odds ratio of disease occurrence at exposure level $x_j(f = j)$, relative to exposure $x_0(f = 0)$, for individuals have covariates $\mathbf{g}$. A possible model here is

$$\ln \psi_j(\mathbf{g}) = \beta(x_j - x_0) + (x_j - x_0)\boldsymbol{\delta}'\mathbf{y}(\mathbf{g}) \tag{7}$$

which implies a linear change in the ln RR with increasing exposure, the slope being itself a linear function of $\mathbf{y}$.

This situation is easily handled in the context of prospective logistic regression, since any model of the form

$$\text{logit } P(d = 1 | f = j, \mathbf{g}) = \alpha + \beta x_j + \boldsymbol{\gamma}'\mathbf{z} + x_j\boldsymbol{\delta}'\mathbf{y} \tag{8}$$

implies (7) for the ln odds ratio. It is more difficult to construct retrospective models which satisfy this relationship, due to the necessity of considering more than two levels of exposure. However, if $Q(j | d, \mathbf{g})$ denotes the conditional probability of an exposure level $f = j$, given disease status d and covariates $\mathbf{g}$, then such a model is defined by

$$\ln \frac{Q(j | d, \mathbf{g})}{Q(0 | d, \mathbf{g})} = \alpha_j^* + \beta(x_j - x_0)\, d + \boldsymbol{\gamma}_j^*\mathbf{z}^* + d(x_j - x_0)\boldsymbol{\delta}'\mathbf{y} \tag{9}$$

for $d = 0$, 1 and $j = 1, 2, \cdots, J$. This is a polychotomous logistic regression model as introduced by Mantel (1966). An extension of arguments used for the previous case of a dichotomous risk factor leads to the same condition for agreement of the estimates of common parameters in (8) and (9); that is, complete saturation of the covariate effects by the parameters α, $\boldsymbol{\gamma}$ and α_j^*, $\boldsymbol{\gamma}_j^*$. This of course requires that J times as many parameters of this type be estimated for the retrospective as for the prospective model.

5. *Discussion*

The general principle illustrated by the preceding is that when prospective and retrospective models imply the same structure for the RR, one expects them to yield increasingly similar estimates for the common parameters as further covariate adjustments are made. When the covariates are sufficiently continuous so as to preclude the formation of 2 × 2 tables as in (6), complete saturation of their effects is neither feasible nor desirable. Indeed, one of the strengths of the regression approach is that covariates may be accounted for without close matching of cases and controls. A strategy for analysis which is suggested by the example is to fit a series of models, whether prospective or retrospective, which contain an increasing number of covariate terms. Provided there are no high order interactions in the data, the estimates of the relative risk parameters should eventually stabilize, and the adjustment process may be stopped.

One situation where complete saturation of covariate effects is clearly inappropriate is when cases and controls are collected and analyzed in a matched pairs design, so that each covariate class contains but one case and one control. When there is a single binary risk factor, the proper analysis is the conditional analysis which restricts attention to the pairs which are discordant for exposure (Cornfield and Haenszel 1960). The RR is estimated by the ratio of the number of pairs where the case is exposed and the control not, to the number where the case is not exposed and the control is. Use of the unconditional logistic regression models considered here, wherein a separate covariate parameter is actually estimated for each pair, leads to a RR estimate which is the square of the above. See Prentice and Breslow (1978) for a generalization of the conditional analysis to include multiple, continuous exposure variables.

Both prospective and retrospective models are easily employed with a single binary risk factor. Hence the choice between them could well be made in terms of which model requires less covariate adjustment to achieve an unconfounded estimate of the RR. When cases and controls have similar distributions of covariates, as in the above example, this would be the prospective model. In situations where the covariates were strongly associated with disease status and less so with the risk factor, the retrospective model would be preferred.

The prospective model would seem to offer a clear advantage when dealing with a continuous risk factor or when several risk factors are being considered on an equal footing, since they are modelled quite simply in the binary logistic regression equation. With the retrospective approach one must discretize the exposures into a relatively small number of outcome categories, representing different levels of exposure to a single factor or combinations of exposures to several factors. This is an awkward procedure which may entail the estimation of a large number of parameters. Moreover, if the results presented here are any guide, it is probably unnecessary.

Résumé

On compare deux approches de l'analyse de régression logistique linéaire sur données d'étude rétrospective: le modèle prospectif où la variable dépendante est un indicateur dichotomique de l'état d'un témoin; et le modèle rétrospectif dans lequel la variable dépendante est une classification de l'exposition en deux ou plusieurs niveaux.

Les deux modèles fournissent des estimateurs du risque relatif d'autant plus semblables qu'il y a plus de covariables intervenant dans l'ajustement. Lorsque les effets des convariables sont saturés de paramètres, les estimateurs du risque relatif sont identiques. On préférera en général le modèle prospectif pour étudier des facteurs de risque multiples et quantitatifs.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika 59,* 19–35.

Anderson, J. A. (1973). Logistic discrimination with medical applications. In *Discriminant Analyses and Applications,* T. Cacoullous (ed.), Academic Press, New York, 1–15.

Bithell, J. and Stewart, A. (1975). Pre-natal irradiation and childhood malignancy: a review of British data from the Oxford Survey. *British Journal of Cancer 31,* 271–87.

Breslow, N. (1976). Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics 32,* 409–16.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute 11,* 1269–75.

Cornfield, J. and Haenszel, W. (1960). Some aspects of retrospective studies. *Journal of Chronical Disease 11,* 523–34.

Cox, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

Kneale, G. W. (1971). Problems arising in estimating from retrospective study data the latent period of juvenile cancers initiated by obstetric radiography. *Biometrics 27,* 563–90.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute 22,* 719–48.

Mantel, N. (1966). Models for complex contingency tables and polychotomous dosage response curves. *Biometrics 22,* 83–95.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics 29,* 479–86.

Miettinen, O. (1974). Confounding and effect modification. *American Journal of Epidemiology 100,* 350–353.

Prentice, R. L. (1976). Use of the logistic model in retrospective studies. *Biometrics 32,* 599–606.

Prentice, R. L. and Breslow, N. (1978). Retrospective studies and failure time models. *Biometrika* (to appear).

Siegel, D. and Greenhouse, S. (1973). Multiple relative risk functions in case-control studies. *American Journal of Epidemiology 97,* 324–331.

Zelen, M. (1971). The analysis of several 2 × 2 contingency tables. *Biometrika 58,* 129–37.

Received July 1977; *Revised October* 1977