Use of Two-segmented Logistic Regression to Estimate Change-points in Epidemiologic Studies

Roberto Pastor and Eliseo Guallar

In many epidemiologic data, the dose-response relation between a continuous exposure and the risk of disease abruptly changes when the exposure variable reaches an unknown threshold level, the so-called change-point. Although several methods are available for dose-response assessment with dichotomous outcomes, none of them provide inferential procedures to estimate change-points. In this paper, we describe a two-segmented logistic regression model, in which the linear term associated with a continuous exposure in standard logistic regression is replaced by a two-segmented polynomial function with unknown change-point, which is also estimated. A modified, iteratively reweighted least squares algorithm is presented to obtain parameter estimates and confidence intervals, and the performance of this model is explored through simulation. Finally, a two-segmented logistic regression model is applied to a case-control study of the association of alcohol intake with the risk of myocardial infarction and compared with alternative analyses. The ability of two-segmented logistic regression to estimate and provide inferences for the location of change-points and for the magnitude of other parameters of effect will make this model a useful complement to other methods of dose-response analysis in epidemiologic studies. Am J Epidemiol 1998;148:631-42.

case-control studies; epidemiologic methods; logistic models; risk assessment

Epidemiologic studies are often designed to explore the relation between a continuous exposure variable and disease risk. Frequently, the dose-response relation abruptly changes when the exposure variable reaches an unknown threshold level, the so-called change-point (1-5), but none of the usual methods of dose-response analysis provides inference procedures for estimating the location of the change-point or its confidence interval.

In categorical analysis (6), the exposure range is divided into a few categories (such as tertiles, quartiles, or quintiles), and then a constant disease risk is estimated within each category. Since the disease risk is forced to remain invariant within each category, the cutpoints between adjacent categories will determine in advance the points of risk change, and hence, categorical analysis will be inefficient for detecting the location of the change-point. To avoid the limitations of categorical analysis, spline regression can be used to fit linear or quadratic models within each category (7-9). Spline regression allows the risk to change within categories, and it is more flexible in estimating the shape of the dose-response relation than is categorical analysis, but the location of the knots has to be fixed arbitrarily by the investigator. An alternative approach to exploring risk trend is to use nonparametric logistic regression to display the dose-response relation without parametric assumptions about the risk trend and without the need to categorize the continuous exposure variable (9, 10), but the identification of change-points with this method is subjective.

The aim of this paper is to introduce statistical methods for estimating change-points in epidemiologic studies by using logistic regression. Several methods for change-point estimation have already been applied to explore the relation between two continuous variables, including two-segmented polynomial regression models (1). The approach presented here is an adaptation of these models to data with a dichotomous response variable. A related method has been developed for detecting J-shaped risk curves in proportional hazards models (11), but this method did not provide an algorithm for simultaneous estimation of all model parameters. In this paper, we describe the main theoretic properties of the proposed model and develop inference procedures to obtain simultaneous estimates of the change-point and associated parame-
tters. We also present the results of a simulation study to explore the appropriateness of using large-sample approximations for model inferences. Finally, we show the epidemiologic applicability of this model in a case-control study of the association of alcohol intake with the risk of myocardial infarction.

For concreteness, we shall consider a case-control study in which the dichotomous response variable \( y \) denotes each patient’s disease status (\( y = 1 \) for cases and \( y = 0 \) for controls). For each subject, we have a set of \( p + 1 \) risk factors or covariates \( x, z_1, \ldots, z_p \). The interest is focused on a single continuous exposure variable \( x \), and the remaining risk factors are possible confounders for the association between exposure and disease.

**TWO-SEGMENTED LOGISTIC REGRESSION**

The principal objective of the logistic regression model presented here is to estimate a potential change-point in the exposure variable \( x \) at which the shape of the dose-response relation markedly changes. To address this problem, we propose replacing the linear term associated with the exposure \( x \) in standard logistic regression by a two-segmented polynomial function with unknown change-point. That is, the parametric relation between the logit (i.e., the natural logarithm of disease odds) and the continuous exposure variable \( x \) is assumed to follow two different polynomial forms on either side of an unknown change-point. The model, here termed two-segmented logistic regression, can be expressed in terms of the logit

\[
g(x, z_1, \ldots, z_p) = \log \left( \frac{\pi(x, z_1, \ldots, z_p)}{1 - \pi(x, z_1, \ldots, z_p)} \right) = f(x, \lambda) + \alpha_1 z_1 + \ldots + \alpha_p z_p,
\]

where \( \pi(x, z_1, \ldots, z_p) = P(y = 1|x, z_1, \ldots, z_p) \) denotes the probability of disease for a given set of covariates, \( \lambda \) is the unknown change-point, the two-segmented polynomial function is \( f(x, \lambda) = f_1(x) \) if \( x < \lambda \) and \( f_2(x) \) otherwise, with \( f_1(x) \) and \( f_2(x) \) being different polynomials, and \( \alpha_1, \ldots, \alpha_p \) are the regression coefficients of the remaining covariates. For practical applications, one must specify the order of the polynomials (such as linear, quadratic, or cubic). The desirable property of smooth transition at the change-point requires one of the polynomials to be at least quadratic. On the other hand, the coefficients of polynomials of order greater than quadratic will be difficult to interpret in epidemiologic or biologic terms. The main alternatives for two-segmented logistic regression are thus linear-quadratic, quadratic-linear, or quadratic-quadratic models. In the next section, we describe the properties and applications of the linear-quadratic-quadratic logistic regression model (linear below and quadratic above the change-point), but inferences with respect to other two-segmented logistic models are similar, and the algorithms described in appendix 1 also apply to all model alternatives.

**Linear-quadratic logistic regression**

The two-segmented logistic regression with linear-quadratic form in the logit (figure 1) can be expressed as

\[
g(x, z_1, \ldots, z_p) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 I(x > \lambda) + \alpha_1 z_1 + \ldots + \alpha_p z_p,
\]

where \( \lambda \) represents the unknown change-point; \( \beta_0, \beta_1, \) and \( \beta_2 \) are the unknown regression coefficients of the exposure variable \( x \); and \( I(x > \lambda) = 1 \) if \( x > \lambda \) and 0 otherwise. In the two-segmented model with linear-quadratic logit, the effect is linear in the logit below the change-point \( \lambda \). In this segment, the logit is \( \beta_0 + \beta_1 x \), where \( \beta_0 \) is the intercept and \( \beta_1 \) is the constant slope of the dose-response relation. Above the change-point \( \lambda \), the effect is quadratic in the logit and is given by \( \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 \), where \( \beta_2 \) represents half the change in the slope of the logit per unit of increment in exposure above the change-point. Since the derivative of the logit is continuous in this model, there is a smooth transition of the risk trend across the two segments: The slope is \( \beta_1 \) below the change-point, and it begins to change in a continuous way above the change-point according to the linear expression \( \beta_1 + 2\beta_2 (x - \lambda) \).

To fit the linear-quadratic logistic regression model in equation 2, we must estimate the change-point \( \lambda \), and the regression coefficients of the exposure \( \beta_0, \beta_1, \) and \( \beta_2 \) and of the other covariates \( \alpha_1, \ldots, \alpha_p \). The maximum likelihood (ML) estimates of these unknown parameters are those values, \( \hat{\lambda}, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \) and \( \hat{\alpha}_1, \ldots, \hat{\alpha}_p \), which maximize the log-likelihood function

\[
\ell = \log(L)
= \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)]
= \sum_{i=1}^{n} [y_i g_i - \log(1 + e^{g_i})],
\]

where $y_i$ denotes the disease status for the $i^{th}$ subject, and $\pi_i = \pi(x_i, z_{i1}, \ldots, z_{ip})$ and $g_i = g(x_i, z_{i1}, \ldots, z_{ip})$ are the probability of disease and its logit for the $i^{th}$ subject, respectively. Since the logit (equation 2) is not linear in the parameters and the change-point is not fixed in advance, the standard iteratively reweighted least squares (IRLS) procedure (12) is inappropriate to fit this model. Instead, we propose a modified version of the IRLS procedure to estimate simultaneously all model parameters (a detailed presentation of this algorithm is provided in appendix 1).

Inferences in standard regression are usually conducted using Wald-type approaches, based on the asymptotic normal distribution of the ML estimates, or likelihood-based methods, directly obtained from the asymptotic $\chi^2$ distribution of the likelihood ratio statistic. In two-segmented regression, these asymptotic properties depend on the existence of an underlying change-point (the so-called identified case) (2, 13). A formal test for the existence of a change-point would be represented by the null hypothesis $H_0$: $\beta_2 = 0$, which implies an homogeneous linear pattern of dose-response. This situation, the nonidentified case, represents a degeneracy of the parameter space: if $\beta_2 = 0$, ...
$\lambda$ is not well defined. In this case, the ML estimates are not asymptotically normal, and the asymptotic distributions of the Wald and likelihood ratio statistics do not converge to standard normal and $\chi^2$ distributions, respectively (3). All inferences described in this paper, such as confidence intervals or tests of hypothesis, are based on the assumption of existence of $\lambda$.

The Wald-type 100(1 - $\alpha$) percent confidence interval for $\lambda$ can be estimated as $\hat{\lambda} \pm z_{1 - \alpha/2} s(\lambda)$, where $z_{1 - \alpha/2}$ is the 100(1 - $\alpha/2$) percentile of the standard normal distribution and $s(\lambda)$ is an estimate of the standard error obtained from the inverse of the Fisher information matrix at the last iteration of the modified IRLS algorithm (see appendix 1). However, this method yields a symmetric interval that is often inappropriate due to the asymmetric distribution of the change-point estimator (see Simulation Study). More appropriate confidence intervals for $\lambda$ can be obtained by using the likelihood ratio statistic $L_\lambda = 2 \ell(\lambda, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) - 2 \ell(\hat{\lambda}, \hat{\beta}_0(\lambda), \hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda))$, where $\hat{\beta}_j(\lambda)$ is the ML estimate of $\beta_j$ given $\lambda$ ($j = 0, 1, 2$) (4). The likelihood-based 100(1 - $\alpha$) percent confidence interval consists of those values $\lambda$ verifying $L_\lambda \leq \chi^2_{1, 1-\alpha}$, where $\chi^2_{1, 1-\alpha}$ represents the 100(1 - $\alpha$) percentile of the $\chi^2$ distribution with 1 degree of freedom. This interval is usually more accurate in terms of coverage than is that based on the Wald approach.

**SIMULATION STUDY**

A simulation study was conducted to examine the properties of the parameter estimates in the linear-quadratic logistic regression model. We considered a single exposure variable that followed an uniform distribution in the range 0 ≤ $x$ ≤ 1. The exposure level of each subject was randomly generated from this distribution, and the disease status was randomly assigned according to the conditional probability of disease given by

$$p(x) = P(y = 1|x) = \frac{e^{-5 + 30(x - 0.75)^2}(\lambda, \beta_0, \beta_1)}{1 + e^{-5 + 30(x - 0.75)^2}(\lambda, \beta_0, \beta_1)}.$$ (4)

corresponding to a linear-quadratic logistic model with underlying parameters $\lambda$, $\beta_0$, $\beta_1$, and $\beta_2$, equal to 0.75, -5, 0, and 30, respectively. In this model, the actual odds ratio for an exposure level $x$ compared with the reference exposure level ($x = 0$) was constant and equal to 1 (no effect) below the threshold 0.75, and it increased above this threshold with odds ratios of 1.08, 1.35, 1.96, and 3.32 for exposure levels 0.80, 0.85, 0.90, and 0.95, respectively. The data of the hypothetical case-control study consisted of the first 500 cases generated by this procedure and a random sample of 500 controls from all of the available disease-free subjects. A total of 1,000 replicates of the case-control study were performed in the simulation.

For each of the 1,000 replicates, a linear-quadratic logistic regression model was fitted to obtain the ML estimates $\hat{\lambda}$, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ using the GAUSS matrix language (14). Several measurements of bias and accuracy of the sampling distributions of $\lambda$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are given in table 1. The distribution of the change-point estimates was skewed to the left with a mean of 0.735 and a median of 0.750 (mean and median bias of -0.015 and 0.0002, respectively). The median bias, that is, the difference between the median of the sampling distribution of $\hat{\lambda}$ and the true parameter $\lambda$, is a more general measure of bias for asymmetric distributions (15). Thus, although $\hat{\lambda}$ was slightly mean biased due to its asymmetric distribution, it had the desirable property of being median unbiased in our simulation study. The distribution of the estimates $\hat{\beta}_1$ was close to a normal distribution, with a mean bias of -0.074 and a median bias of -0.027. The distribution of $\hat{\beta}_2$, however, was markedly right-skewed, with a median bias (1.01) that was considerably smaller than its mean bias (18.69), inflated by the asymmetric distribution of the estimates.

Even though the sample size of this hypothetical case-control study was moderate (500 cases and 500 controls), the sampling distributions of the parameter estimates were not well approximated by the normal, and there were substantial differences between the percentile-based confidence intervals (the 2.5th and 97.5th empirical percentiles of the sampling distrib-

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Underlying value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>95% normal interval*</th>
<th>Median</th>
<th>Interquartile range</th>
<th>95% percentile range†</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>0.75</td>
<td>0.735</td>
<td>0.110</td>
<td>0.619 to 0.950</td>
<td>0.750</td>
<td>0.096</td>
<td>0.467 to 0.887</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0</td>
<td>-0.074</td>
<td>0.487</td>
<td>-1.028 to 0.881</td>
<td>-0.027</td>
<td>0.488</td>
<td>-0.978 to 0.624</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>30</td>
<td>48.692</td>
<td>115.090</td>
<td>-176.885 to 274.269</td>
<td>31.013</td>
<td>28.563</td>
<td>6.630 to 167.827</td>
</tr>
</tbody>
</table>

* Mean ± 1.96 standard error.
† Defined by the 2.5th and 97.5th percentiles of the sampling distribution, respectively.
tions (15)) and the standard normal confidence intervals (table 1).

To evaluate further the large sample approximation of Wald and likelihood-based inferences for testing $\lambda = 0.75$, $\beta_1 = 0$, and $\beta_2 = 30$, the Wald statistics $W_\lambda = (\lambda - 0.75)/s(\lambda)$, $W_1 = \beta_1/s(\beta_1)$, and $W_2 = $...
(\hat{\beta}_2 - 30)/s(\hat{\beta}_2), and the signed square root of the likelihood ratio statistics \( LR_\lambda = \text{sgn}(\hat{\lambda} - 0.75) \sqrt{\Lambda_{\lambda=0.75}} \), \( LR_1 = \text{sgn}(\hat{\beta}_1) \sqrt{\Lambda_{\beta_1=0}} \), and \( LR_2 = \text{sgn}(\hat{\beta}_2 - 30) \sqrt{\Lambda_{\beta_2=30}} \) were computed for each replicate of the simulation study (4). The quantile-quantile plots for the statistics show that the sampling distribution of \( W_1 \) was close to the standard normal (figure 2c), while the sampling distributions of \( W_\lambda \) and \( W_2 \) were highly skewed to the right and to the left, respectively (figures 2a and 2e). The sampling distributions of \( LR_\lambda, LR_1, \) and \( LR_2 \) were well approximated by the standard normal (figures 2b, 2d, and 2f), showing that inferences based on the likelihood ratio statistics were more reliable than those based on Wald statistics in our simulation. Similar results were obtained in terms of coverage of the Wald and likelihood-based confidence intervals (not shown).

An important feature of the linear-quadratic logistic regression model was the high dependence of \( \beta_1 \) and \( \beta_2 \) on the change-point estimate \( \hat{\lambda} \). When \( \hat{\lambda} \) was close to the true value 0.75, \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) were also close to their true values of 0 and 30, respectively. As expected, this occurred in most replicates of the simulation study. There was also a pattern of increase in \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) with increasing \( \hat{\lambda} \). As a consequence of the linear-quadratic parameterization of the logit, the in-

**FIGURE 3.** Scatterplots of the estimates \( \hat{\beta}_1 \) (a) and \( \hat{\beta}_2 \) (b) against the change-point estimate \( \hat{\lambda} \) from the 1,000 replicates of the simulation study. The horizontal dashed lines represent the true parameters \( \beta_1 = 0 \) (a) and \( \beta_2 = 30 \) (b).
creasing pattern of $\beta_1$ was additive, while the increasing pattern of $\beta_2$ was multiplicative (figure 3).

Additional simulation studies were performed with modifications of the location of the change-point, the shape of the distribution of the exposure variable, and the sample size, but the conclusions with respect to the performance of model estimators were essentially unchanged.

EXAMPLE: ALCOHOL INTAKE AND RISK OF MYOCARDIAL INFARCTION IN THE EURAMIC STUDY

A two-segmented logistic regression model was used to study the association of alcohol intake and risk of myocardial infarction using data from the EURAMIC (EUROpean study on Antioxidants, Myocardial Infarction, and breast Cancer) Study, an international case-control study conducted in eight European countries and Israel, designed primarily to evaluate the association of antioxidants with the risk of developing a first myocardial infarction in men. The methods of the EURAMIC Study have been described in detail elsewhere (16, 17). Cases were men aged less than 70 years with a confirmed diagnosis of first acute myocardial infarction who had been admitted to the coronary care units of participating hospitals within 24 hours of the onset of symptoms. Controls were obtained through random samples of the population from which the cases originated (five countries), through random samples of the patients of the cases' general practitioners (one country), by inviting friends or relatives of the cases (one country), or through mixed methods (two countries).

To focus on the dose-response of alcohol intake and risk of myocardial infarction, we restricted our analyses to the 330 cases and 441 controls who reported some alcohol intake during the previous year, but the analyses of all study subjects with additional indicator variables for ex-drinkers and never drinkers (18, 19) gave virtually the same results (not shown).

Among controls who were current drinkers, alcohol intake ranged from 0.1 to 142.6 g/day, and the 25th, 50th, and 75th percentiles of intake were 5.7, 13.7, and 30.0 g/day, respectively. We fitted the following quadratic-linear logistic regression model to the EURAMIC Study alcohol data:

$$g(x, z_{1}, \ldots, z_{16}) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 I_{(x>\lambda)}$$

$$+ \alpha_1 z_1 + \ldots + \alpha_{16} z_{16},$$

where $x$ represented alcohol intake in g/day, $\lambda$ was the change-point, and the adjustment covariates $z_1, \ldots, z_{16}$ were age, study center (eight indicator variables), smoking (current number of cigarettes smoked and indicator variables for past smoking and never smoking), waist-hip ratio, history of diabetes, history of hypertension, and family history of coronary heart disease.

Additional simulation studies were performed with modifications of the location of the change-point, the shape of the distribution of the exposure variable, and the sample size, but the conclusions with respect to the performance of model estimators were essentially unchanged.

$\lambda, \beta_1, \beta_2$ were age, study center (eight indicator variables), smoking (current number of cigarettes smoked and indicator variables for past smoking and never smoking), waist-hip ratio, history of diabetes, history of hypertension, and family history of coronary heart disease. The parameter estimates summarizing the association of alcohol intake and risk of myocardial infarction are shown in table 2, and a graphic display of the model is shown in figure 4, where the reference value for the effect of alcohol (odds ratio = 1) has been set at 2.7 g/day, the median of the first quartile of alcohol intake among controls. The risk decreased sharply at low intakes of alcohol, and then it increased, showing a detrimental effect at very high intakes. The estimated change-point was 13.1 g/day (likelihood-based 95 percent confidence interval 4.7-50.1 g/day), and the minimum risk in terms of alcohol intake was reached at 12.6 g/day. The joint 95 percent and 90 percent confidence regions for $\lambda$ and $\beta_2$ are shown in figure 5, where the strong dependence of the model parameters can be appreciated. Although the marginal confidence intervals for $\lambda$ and $\beta_2$ were wide, the joint confidence region showed that the area covered by the most likely values of $\lambda$ and $\beta_2$ was relatively narrow.

For comparison, we fitted four additional models to our data (figure 4): a quadratic model (in which alcohol was introduced as alcohol and alcohol$^2$), a categorical model (in which alcohol was introduced as indicator variables for the quartiles of the control distribution), a restricted spline model (7) with knots at the quartiles of the control distribution (in which quadratic forms were used in the lowest three quartiles and a linear form in the highest), and a nonparametric logistic regression model (9) (in which a locally weighted smoother with 30 percent bandwidth was used to fit alcohol intake using the GAM function in S-PLUS (20)). All models were adjusted for the above-mentioned covariates and used the same reference value for the effect of alcohol intake.

The quadratic model fitted a J-shaped curve with a nadir at 50.2 g/day of alcohol intake. Compared with the other analyses, this model underestimated the decrease in risk at low doses of alcohol and exaggerated

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Point estimate</th>
<th>Likelihood-based 95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>13.118</td>
<td>4.679 to 50.062</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.008</td>
<td>0.000 to 0.021</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.009</td>
<td>0.001 to 0.064</td>
</tr>
</tbody>
</table>

* Adjusted for age, center, smoking, waist-hip ratio, history of diabetes, history of hypertension, and family history of coronary heart disease.
the increase at higher doses. The categorical analysis showed an L-shaped dose-response, with reduced and very similar risk in the three highest quartiles. Finally, the quadratic-linear, spline, and nonparametric models resulted in a similar shape of the dose-response relation, but only the quadratic-linear model provided inference procedures for the change-point.

**DISCUSSION**

In this paper, we have presented a two-segmented logistic regression model to estimate change-points with a dichotomous dependent variable, and we have shown how to obtain point estimates and confidence intervals for model parameters without the need to determine in advance the approximate location of the change-point.

Several regression models have been proposed to explore the shape of the risk trend in dose-response analysis. Fractional polynomial regression (containing fractional and inverse powers of $x$, such as $x^{-2}$, $x^{-1}$, $x^{-1/2}$, $x^{1/2}$, $x$, and $x^2$) and spline regression are flexible enough to reproduce a wide variety of dose-response curves and can be easily programmed with standard statistical software (7). In addition to these parametric models, nonparametric regression provides smooth curves that summarize the exposure-disease relation without imposing any parametric assumption (10). If there is a change-point in the effect of the exposure, it will be reflected in the plots of the dose-response curves estimated from the above regression models. A visual examination of these curves is commonly used to approximately locate the change-point, but this practice is largely subjective and does not provide a systematic inferential procedure for change-point estimation. On the other hand, the two-segmented logistic regression model estimates simultaneously the location of the change-point and other parameters of effect of the exposure variable, avoiding the need for arbitrary cutpoints.

For practical applications, one of the most important consequences of two-segmented logistic regression is derived from the high dependence of the coefficients of effect of the exposure, $\beta_1$ and $\beta_2$, with the change-point $\lambda$ (figures 3 and 5). The estimates of $\beta_1$ and,
especially, of $\beta_2$ will be strongly dependent on the value of $\lambda$. In other epidemiologic analyses involving categorization of the exposure range, the effect estimates will depend on the arbitrary position of the cutpoints, particularly in extreme categories of exposure. Thus, caution must be used in the interpretation of relative risks obtained after post hoc categorization of the range of the exposure variable.

An additional consequence of categorizing the exposure is that the standard errors of the coefficients of effect will be artificially reduced in categorical or spline analyses. In our alcohol example, fitting the quadratic-linear spline model with a single knot at 13.118 g/day of alcohol (the estimated change-point in two-segmented regression),

$$g(x, z_1, \ldots, z_{16}) = \beta_0 + \beta_1 x + \beta_2(x - 13.118)I_{(x \leq 13.118)} + \alpha_1 z_1 + \ldots + \alpha_{16} z_{16},$$

resulted in the same parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, but estimated with higher precision. For instance, the likelihood ratio-based 95 percent confidence interval for $\beta_2$ in the spline model was 0.005–0.013, substantially smaller than the confidence interval in our quadratic-linear model, 0.001–0.084. This increased precision, however, was artificial in the sense that the spline model assumed a fixed, preestablished knot. The impact of the position of the change-point on parameter estimates and the need to obtain realistic estimates of error highlight the need to develop models for change-point estimation.

The two-segmented logistic regression model provides valuable methods to estimate effects in certain dose-response analyses, but this approach is not without limitations. First, model inferences are based on the assumption of existence of a change-point. Although a graphic examination of this assumption can be performed by nonparametric logistic regression, further work is needed to adapt tests of hypothesis useful when there is a degeneracy of the parameter space under the null (21, 22) to segmented logistic regression models.

Second, as in other parametric models, the potential for model misspecification is always a problem (23). Nonparametric models are useful to check the appropriate parameterization of the segmented model. Although we have restricted our presentation to a two-segmented polynomial function, these methods (including the algorithms described in appendix 1) also apply to other parametric forms of the logit and to more than one change-point as long as the transitions among the segments are smooth. For continuous outcomes, segmented models have been extended to avoid the restriction of continuity at the change-points (24) or to impose continuity, but not derivability, across the transitions (13, 25–27). These methods included the particular case of two intersecting straight
lines with a sharp corner at the change-point, which was later generalized to accommodate a smooth transition, as well as an abrupt change using transition functions in the neighborhood of the change-point (28). These generalizations, however, are not available for discrete outcomes. A segmented model has also been developed for the detection of J-shaped risk curves in proportional hazards models by obtaining the profile log-likelihood from sequential quadratic spline models at different cutpoints (11), but this method did not provide an algorithm for simultaneous estimation of the change-point and associated parameters.

Third, we have explored the performance of model estimators through simulation, but the results of our simulation study may not be representative of what might occur in practice. Further simulation scenarios have to be evaluated to assess adequately the appropriateness of large-sample approximations for model inferences (29).

Finally, the lack of availability of two-segmented models in standard statistical packages restricts their widespread use at the moment. GAUSS macros for linear-quadratic, quadratic-linear, or quadratic-quadratic logistic regression are available upon request by mail from the corresponding author or by electronic mail (egualhar@isciii.es).

Our analysis of the alcohol data in the EURAMIC Study further illustrates the limitations of traditional dose-response analysis in epidemiologic studies. A dose-response meta-analysis of alcohol intake and risk of myocardial infarction found an L-shaped association, described as a drop in risk at low doses, followed by a plateau in risk above one drink per day (30). The case-control studies combined in this meta-analysis based their assessment of the effect of alcohol on categorizations of alcohol intake, and their dose-response curves were very similar to our categorical analyses of the EURAMIC data. It is likely that the use of broad categories limited the ability of these studies to detect an increased risk of myocardial infarction at high intakes.

As previous studies (7, 11, 19) have also shown, the usual epidemiologic approach to dose-response assessment based on quadratic models or on categorical analysis will often be inadequate. As a minimum check, spline or nonparametric regression should be used to confirm more traditional methods. If these analyses provide evidence of abrupt changes in risk or if there are a priori reasons to suspect the existence of change-points, two-segmented logistic regression could then be used to estimate the change-point and to provide more realistic measures of statistical error.

ACKNOWLEDGMENTS

Dr. Roberto Pastor was supported by a grant from the "Instituto de Salud Carlos III" (FIS 97/4148). Dr. Eliseo Guallar's work was supported in part by grant FIS 95/0457.

The authors thank the EURAMIC Study group (Professor Frans Kok, Principal Investigator) for providing the original data for the example of alcohol intake and risk of myocardial infarction presented in this paper, and Drs. Leon Gordis, Victor Moreno, Javier Nieto, and Santiago Pérez-Hoyos for their helpful comments on earlier versions of this paper.

REFERENCES

21. Davie RB. Hypothesis testing when a nuisance parameter is
Two-segmented Logistic Regression

APPENDIX

In this appendix, we develop an algorithm to fit two-segmented logistic regression models with logit \( g(x, z_1, \ldots, z_p) = f(x, \lambda) + \alpha_1 z_1 + \ldots + \alpha_p z_p \), where \( f(x, \lambda) \) is a two-segmented polynomial function with continuous first partial derivatives over the entire parameter space. The linear-quadratic, \( f(x, \lambda) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 \), quadratic-linear, \( f(x, \lambda) = \beta_0 + \beta_1 x + \beta_2 (x - \lambda)^2 I(x \geq \lambda) \), and quadratic-quadratic, \( f(x, \lambda) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \lambda)^2 I(x \geq \lambda) \), models are special parameterizations of \( f(x, \lambda) \). The primary objective is to obtain the ML estimate \( \hat{\theta} = (\lambda, \beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_p) \) that maximizes the log-likelihood of the two-segmented logistic regression model

\[
\ell = \sum_{i=1}^{n} [y_i g_i - \log(1 + e^{g_i})].
\]  

(A1)

The score vector \( u(\theta) = \frac{\partial \ell}{\partial \theta} \), with elements

\[
\frac{\partial \ell}{\partial \theta_r} = \sum_{i=1}^{n} (y_i - \pi_i) \frac{\partial g_i}{\partial \theta_r}, \quad r = 1, \ldots, p + q + 2,
\]

(A2)

can be calculated because \( g_i \) has continuous first partial derivatives with respect to \( \theta \) over the entire parameter space. Since \( E(y_i) = \pi_i \) and \( \text{var}(y_i) = \pi_i (1 - \pi_i) \), \( u(\theta) \) has expectation \( \theta \) and covariance matrix equal to the Fisher information matrix \( I(\theta) \), with elements

\[
I_{rs}(\theta) = \{E(u(\theta)u^T(\theta))\}_{rs} = \sum_{i=1}^{n} \pi_i (1 - \pi_i) \frac{\partial g_i}{\partial \theta_r} \frac{\partial g_i}{\partial \theta_s},
\]

(A3)

where \( I_{rs}(\theta) \) is the \((r, s)\)th element of \( I(\theta) \) \((r, s = 1, \ldots, p + q + 2)\). The ML estimates \( \hat{\theta} \) are obtained as the solutions of the \( p + q + 2 \) likelihood equations \( u(\theta) = 0 \). Since these equations are not linear in the parameters \( \theta \), iterative methods must be used to obtain \( \hat{\theta} \). The two-segmented logistic regression model can be fitted by using the Fisher scoring method (12). Let \( \theta^{(k)} \) be the current estimate of \( \theta \), with corresponding score vector \( u(\theta^{(k)}) \) and Fisher information matrix \( I(\theta^{(k)}) \). The new estimate \( \theta^{(k+1)} \) is derived from

\[
I_{rs}(\theta^{(k)}) \frac{\partial g_i}{\partial \theta_s} = \sum_{i=1}^{n} \pi_i (1 - \pi_i) \frac{\partial g_i}{\partial \theta_r} \frac{\partial g_i}{\partial \theta_s}.
\]

After evaluating equations A2 and A3 at \( \theta^{(k)} \), results in the following system of equations \((r = 1, \ldots, p + q + 2)\)

\[
\sum_{i=1}^{n} \pi_i^{(k)} (1 - \pi_i^{(k)}) \frac{\partial g_i}{\partial \theta_r} \left[ \sum_{s=1}^{p+q+2} \frac{\partial g_i}{\partial \theta_s} (\theta^{(k+1)} - \theta^{(k)}) - \frac{y_i - \pi_i^{(k)}}{\pi_i^{(k)} (1 - \pi_i^{(k)})} \right] = 0.
\]

(A4)
Using the first-order Taylor series of $g_t$ around $\theta^{(k)}$, we have $[\partial g_t/\partial \theta^{(k)}]^T(\theta^{(k+1)} - \theta^{(k)}) = g_t^{(k+1)} - g_t^{(k)}$, and then equation A4 can be rewritten as $(r = 1, \ldots, p + q + 2)\sum_{i=1}^{n} \tau_i^{(k)}(1 - \pi_i^{(k)}) \partial g_i/\partial \theta_i^{(k)} \left[ g_i^{(k+1)} - \left( g_i^{(k)} + \frac{y_i - \pi_i^{(k)}}{\pi_i^{(k)}(1 - \pi_i^{(k)})} \right) \right] = 0$,

(A5)

which are the weighted least squares equations of a two-segmented regression with weights $w_i^{(k)} = \pi_i^{(k)}(1 - \pi_i^{(k)})$ and response variables $v_i^{(k)} = g_i^{(k)} + \left( y_i - \pi_i^{(k)} \right) / w_i^{(k)}$. Thus, $\theta^{(k+1)}$ can be calculated by fitting a weighted two-segmented regression model, in which the weight $w_i^{(k)}$ and the response variable $v_i^{(k)}$ depend on $\theta^{(k)}$ and, hence, must be recomputed at each iteration (iteratively reweighted least squares procedure (12)).

Since the two-segmented regression function $g(\theta)$ has continuous first derivatives throughout the parameter space, Hartley’s modification of Gauss-Newton algorithm can be used to solve the weighted two-segmented regression models (1, 4). Let $\theta^{(k,l)}$ be an estimate of $\theta$ at the $l$th iteration of the procedure nested in the $k$th stage of the primary iterative process. For each subject, the first-order Taylor series of $g_i(\theta)$ around $\theta^{(k,l)}$ gives $g_i(\theta) \equiv g_i(\theta^{(k,l)}) + [\partial g_i/\partial \theta(\theta^{(k,l)})]^T(\theta - \theta^{(k,l)})$, and the residual $r_i(k, l) = v_i^{(k,l)} - g_i(\theta^{(k,l)})$ can be expressed as

or in matrix form for all subjects $r(k, l) = G(k, l)(\theta - \theta^{(k,l)}) + e_i$, where $r(k, l)$ is the residual vector and $G(k, l)$ is a matrix with $(i, r)$th element $\partial g_i/\partial \theta_{(r)}$. The solution of this linearized regression model with diagonal matrix of weights $W(k) = \text{diag}(\pi_i^{(k)}(1 - \pi_i^{(k)}))$ can be obtained by using the ordinary weighted least squares method, so that the solution is $d(k, l) = (G^T(k, l)W(k)G(k, l))^{-1}G^T(k, l)W(k)r(k, l)$. Following Hartley’s modified Gauss-Newton algorithm for segmented polynomial regression models (1), the next estimate $\theta^{(k,l+1)}$ is given by $\theta^{(k,l+1)} = \theta^{(k,l)} + \eta(k, l)d(k, l)$, where $\eta(k, l)$ is the scalar between 0 and 1 at which $\theta^{(k,l+1)} = \theta^{(k,l)} + \eta d(k, l)$ minimizes the weighted sum of the errors of the two-segmented regression model.

Each of these iterative procedures is stopped when the relative change between $\theta^{(k,l+1)}$ and $\theta^{(k,l)}$ is negligible, and the last estimate is used as the starting value of the next iterative process $\theta^{(k+1)}$. One must then recompute the weight $w_i^{(k+1)}$ and the response variable $v_i^{(k+1)}$ at $\theta^{(k+1)}$, and fit again the weighted, two-segmented regression model using the above iterative procedure. This process is repeated until convergence.

The choice of the initial estimate $\theta^{(0)}$ is crucial to obtain a proper convergence of the algorithm presented in this appendix because the log-likelihood function (equation A1) may possess local maxima. A good alternative is to use the grid-search approach to obtain the initial estimate (13). The method determines in advance several equally spaced points over the range of the exposure variable $x$. For each point, a spline logistic regression with two-segmented logit (7, 9) is fitted, and the log-likelihood is evaluated. The point at which the log-likelihood reaches the highest value is used along with the remaining parameter estimates from the spline logistic regression as initial values $\theta^{(0)}$. 


Downloaded from https://academic.oup.com/aje/article-abstract/148/7/631/148349 by guest on 10 January 2019