Küchenhoff:

An exact algorithm for estimating breakpoints in segmented generalized linear models

Sonderforschungsbereich 386, Paper 27 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

An exact algorithm for estimating breakpoints in segmented generalized linear models

Helmut Küchenhoff

University of Munich, Institute of Statistics, Akademiestrasse 1, D-80799 München

Summary

We consider the problem of estimating the unknown breakpoints in segmented generalized linear models. Exact algorithms for calculating maximum likelihood estimators are derived for different types of models. After discussing the case of a GLM with a single covariate having one breakpoint a new algorithm is presented when further covariates are included in the model. The essential idea of this approach is then used for the case of more than one breakpoint. As further extension an algorithm for the situation of two regressors each having a breakpoint is proposed. These techniques are applied for analysing the data of the Munich rental table. It can be seen that these algorithms are easy to handle without too much computational effort. The algorithms are available as GAUSS–programs.

Keywords: Breakpoint, generalized linear model, segmented regression

1 Introduction

In many practical regression–type problems we cannot fit one uniform regression function to the data, since the functional relationship between the response Y and the regressor X changes at certain points of the domain of X. These points are usually called breakpoints or changepoints. One important example is the threshold model used in epidemiology (see Ulm, 1991, Küchenhoff and Carroll, 1996), where the covariate X, typically an exposure, has no influence on Y, e.g. the occurrence of a certain disease, up to a certain level. Thus the relationship between X and Y is described by a constant up to this level and for values of X greater than this level it is given by an increasing function.

In such situations, we apply segmented or multiphase regression models which are obtained by a piecewise definition of the regression function $E(Y|X = x)$ on intervals of the domain of X. An overview concerning this topic can be found in Chapter 9 of Seber and Wild (1989). Assuming a generalized linear model and a known number of segments we have

$$E(Y|X=x) = \begin{cases} G(\alpha_1 + \beta_1 x) & \text{if} \quad x \le \tau_1 \\ G(\alpha_2 + \beta_2 x) & \text{if} \quad \tau_1 < x \le \tau_2 \\ \quad \vdots & \\ G(\alpha_K + \beta_K x) & \text{if} \quad \tau_{K-1} < x \end{cases} \tag{1}$$

and

$$f(y|\vartheta, \xi) = \exp\left\{ \frac{y\vartheta - b(\vartheta)}{\xi} + c(y, \xi) \right\}. \tag{2}$$

Here ξ is the nuisance–parameter and $b'(\vartheta) = E(Y|X=x)$, see Fahrmeir and Tutz (1994), Seber and Wild (1989), G is the link–function, e.g. logistic, identity etc, and f denotes the density function of Y given $X = x$. For the threshold model mentioned above, for instance, there are two segments, where G is the logistic link and $\beta_1 = 0$. The endpoints τ_i of the intervals denote the breakpoints. Since they are typically unknown, they have to be estimated. For theoretical, but also practical reasons the breakpoints are assumed to ly between the smallest and the largest sample value $x_i, i = 1, \ldots, n$.

We further assume that the regression function is continuous, i.e.

$$\alpha_i + \beta_i \tau_i = \alpha_{i+1} + \beta_{i+1} \tau_i \ , \ 1 \le i \le K-1.$$

Thus the model can be stated in another parameterisation:

$$E(Y|X=x) = G(\alpha + \beta_1 x + \sum_{i=2}^{K} \beta_i (x - \tau_{i-1})_+), \tag{3}$$

where $$t_+ = \begin{cases} t & \text{if} \quad t \ge 0 \\ 0 & \text{if} \quad t < 0. \end{cases}$$

From this representation it can be seen, that (3) is a usual generalized linear model, if the breakpoints τ_i are known. Therefore the ML–estimation can be performed by a grid–search–type algorithm in case of two segments, see Stasinopoulos and Rigby (1992).

For the linear model an exact algorithm for the least squares estimator was given by Hudson (1966) (see also Schulze, 1987 or Hawkins, 1976). In Section 2 it is shown, that this algorithm also works for the GLM with one breakpoint. In Section 3 the algorithm is extended to models with further covariates. In Section 4 the ideas of Section 3 are used to derive the algorithm for fairly general models with more than one breakpoint and more than one covariate with breakpoints. Giving an algorithm for such general models we fill a gap existing so far in the literature. In Section 5 an example is considered. We investigate the relationship between the net rent of flats in

Munich and the flat size as well as the age of the flats based on data of the Munich rental table. Finally, problems concerning the computing time are discussed and some interesting additional aspects are pointed out.

2 Exact ML–estimation for models with one breakpoint

We consider a GLM with one breakpoint and density (2). The regression function can then be written as

$$E(Y|X=x) = G(\alpha + \beta_1(x-\tau)_- + \beta_2(x-\tau)_+) \quad \text{with} \quad t_- = -(-t)_+. \quad (4)$$

Here β_1 is the slope parameter of the first segment and β_2 is the slope in segment 2.

The log–likelihood function of one observation, conditioned on X, is given by

$$\mathcal{G}(y, \alpha+\beta_1(x-\tau)_-+\beta_2(x-\tau)_+, \xi) = \frac{y\vartheta - b(\vartheta)}{\xi} + c(y,\xi),$$

where the nuisance parameter ξ is assumed to be constant over the segments. If G is the natural link function, then

$$\vartheta = \alpha+\beta_1(x-\tau)_-+\beta_2(x-\tau)_+.$$

Having i.i.d. observations $(x_i, y_i)_{i=1,\ldots,n}$, the log–likelihood function to be maximized in $(\alpha, \beta_1, \beta_2, \tau, \xi)'$ is

$$\sum_{i=1}^{n} \mathcal{G}(y_i, \alpha+\beta_1(x_i-\tau)_-+\beta_2(x_i-\tau)_+, \xi). \quad (5)$$

Since this function is not differentiable in τ at x_i, we first calculate the profile likelihood. That is, we maximize (5) with respect to all other parameters and get

$$\mathcal{P}(\tau) = \max_{\alpha,\beta_1,\beta_2,\xi} \sum_{i=1}^{n} \mathcal{G}(y_i, \alpha+\beta_1(x_i-\tau)_-+\beta_2(x_i-\tau)_+, \xi). \quad (6)$$

Obviously model (4) is a GLM for fixed τ. Thus, the calculation of the profile likelihood corresponds to the ML–estimation of a GLM.

Since (6) is continuous in τ, it can be maximized by a grid search, see Ulm (1991). For a GLIM–macro see Stasinopoulos and Rigby (1992). Though these algorithms give reliable results, if the grid is appropriately chosen, it would be desirable to have an exact algorithm at one's disposal.

We derive such an exact algorithm for maximizing (5) following the ideas of Hudson (1966). Since we have assumed, that the nuisance parameter ξ is constant for the two segments, it can be neglected in maximizing (5). Let the observations $(x_i, y_i)_{i=1,\ldots,n}$ be ordered with respect to x_i such that $x_i \leq x_j$ for $i < j$. The log–likelihood is differentiable with respect to $\theta = (\alpha, \beta_1, \beta_2, \tau)'$ everywhere except for those values of θ with $\tau = x_i$ for one $i \epsilon \{1, \ldots, n\}$. Therefore the algorithm has to be divided into roughly two steps according to the differentiability of the log–likelihood.

In the first step the points of differentiability, i.e. the case $\hat{\tau} \epsilon (x_k, x_{k+1})$ for some k, are considered. Denoting the partial derivative of $\mathcal{G}(\cdot, \cdot, 1)$ with respect to the second argument by $\mathcal{G}_2(\cdot, \cdot)$ we therefore get

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \mathcal{G}(y_i, glp_i, 1) = \sum_{i=1}^{n} [\mathcal{G}_2(y_i, glp_i)] \begin{pmatrix} 1 \\ (x_i - \tau)_- \\ (x_i - \tau)_+ \\ -\beta_1 I_{\{x_i \leq \tau\}} - \beta_2 I_{\{x_i > \tau\}} \end{pmatrix}, \quad (7)$$

where glp is the "broken linear predictor"

$$glp_i = \alpha + \beta_1 (x_i - \tau)_- + \beta_2 (x_i - \tau)_+$$

and I denotes the indicator function.

Let $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\tau})'$ be a zero of (7) with $\hat{\tau} \epsilon (x_k, x_{k+1})$ and $\hat{\beta}_1 \neq \hat{\beta}_2$. The system of equations obtained by equating (7) to 0 results after some algebra in

$$\sum_{i=1}^{k} \mathcal{G}_2(y_i, \hat{\alpha}_1 + \hat{\beta}_1 x_i) = 0 \quad (8)$$

$$\sum_{i=1}^{k} \mathcal{G}_2(y_i, \hat{\alpha}_1 + \hat{\beta}_1 x_i) x_i = 0 \quad (9)$$

$$\sum_{i=k+1}^{n} \mathcal{G}_2(y_i, \hat{\alpha}_2 + \hat{\beta}_2 x_i) = 0 \quad (10)$$

$$\sum_{i=k+1}^{n} \mathcal{G}_2(y_i, \hat{\alpha}_2 + \hat{\beta}_2 x_i) x_i = 0 \quad (11)$$

$$\text{with } \hat{\alpha}_j = \hat{\alpha} - \hat{\beta}_j \hat{\tau}, \; j = 1, 2.$$

From equations (8) and (9), (10) and (11) we conclude that $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$, respectively, are ML–solutions of the regressions in the two segments. Since τ is uniquely determined by the continuity condition, possible zeros with $\hat{\tau} \epsilon (x_k, x_{k+1})$ can be determined by estimating the parameters separately in the two segments based on $(x_i, y_i)_{i=1,\ldots,k}$ and $(x_i, y_i)_{i=k+1,\ldots,n}$,

which yields $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$, respectively. The estimator for τ is then obtained from

$$\hat{\tau} = \frac{\hat{\alpha}_2 - \hat{\alpha}_1}{\hat{\beta}_1 - \hat{\beta}_2}.$$

If $\hat{\tau} \epsilon (x_k, x_{k+1})$, it is a zero of (7). In this case, the estimator $\hat{\alpha}$ is given by

$$\hat{\alpha} = \hat{\alpha}_1 + \hat{\beta}_1 \hat{\tau}.$$

If $\hat{\tau} \notin (x_k, x_{k+1})$, we deduce from (8) – (11) that there is no local maximum with $\tau \epsilon (x_k, x_{k+1})$.

The above mentioned procedure is performed for the finite number of intervals (x_k, x_{k+1}), where these intervals have to be chosen such that the ML–estimators exist in the corresponding segments.

In the second step we calculate the profile likelihood $\mathcal{P}(x_i), i = 1, \ldots, n$, obtaining the maximum of the log–likelihood at all points of non–differentiability. Finally the global maximum of the log–likelihood is given by the maximum of this finite number of local maxima.

Conducting this algorithm the estimation of at most $m + 2m$ GLMs is needed, if there are m observations with different values of x.

3 Models with covariates

In many practical situations there will be further covariates in the regression model, which leads to the following extension of model (4):

$$E(Y|X = x, Z = z) = G(\alpha + \beta_1 (x - \tau)_- + \beta_2 (x - \tau)_+ + z'\gamma), \qquad (12)$$

where Z is the vector of covariates with vector of parameters γ. As in Section 2 $(x_i, y_i, z_i)_{i=1,\ldots,n}$ denote the corresponding observations with $x_i \leq x_j$ for $i < j$.

To derive the ML–estimator, we first consider again the case $\hat{\tau} \epsilon (x_k, x_{k+1})$. Then the derivative of the log–likelihood with respect to $\theta = (\alpha, \beta_1, \beta_2, \tau, \gamma)$ is

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \mathcal{G}(y_i, glp_i, 1) = \sum_{i=1}^{n} [\mathcal{G}_2(y_i, glp_i)] \begin{pmatrix} 1 \\ (x_i - \tau)_- \\ (x_i - \tau)_+ \\ -\beta_1 I_{\{x_i \leq \tau\}} - \beta_2 I_{\{x_i > \tau\}} \\ z_i \end{pmatrix} \qquad (13)$$

with

$$glp_i = \alpha + \beta_1 (x_i - \tau)_- + \beta_2 (x_i - \tau)_+ + z_i' \gamma.$$

Analogously to (8) – (11) we get the following system of equations with $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}, \hat{\tau})'$ denoting a zero of (13):

$$\sum_{i=1}^{k} \mathcal{G}_2(y_i, \hat{\alpha}_1 + \hat{\beta}_1 x_i + z_i'\hat{\gamma}) = 0 \quad (14)$$

$$\sum_{i=1}^{k} \mathcal{G}_2(y_i, \hat{\alpha}_1 + \hat{\beta}_1 x_i + z_i'\hat{\gamma}) x_i = 0 \quad (15)$$

$$\sum_{i=k+1}^{n} \mathcal{G}_2(y_i, \hat{\alpha}_2 + \hat{\beta}_2 x_i + z_i'\hat{\gamma}) = 0 \quad (16)$$

$$\sum_{i=k+1}^{n} \mathcal{G}_2(y_i, \hat{\alpha}_2 + \hat{\beta}_2 x_i + z_i'\hat{\gamma}) x_i = 0 \quad (17)$$

$$\sum_{i=1}^{k} \mathcal{G}_2(y_i, \hat{\alpha}_1 + \hat{\beta}_1 x_i + z_i'\hat{\gamma}) z_i + \sum_{i=k+1}^{n} \mathcal{G}_2(y_i, \hat{\alpha}_2 + \hat{\beta}_2 x_i + z_i'\hat{\gamma}) z_i = 0 \quad (18)$$

with $\hat{\alpha}_j = \hat{\alpha} - \hat{\beta}_j \hat{\tau}, j = 1, 2$.

Equations (14) – (18) correspond to a generalized linear model with an analysis of covariance–type design matrix

$$D_k = \begin{pmatrix} 1 & x_1 & 0 & 0 & z_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_k & 0 & 0 & z_k \\ 0 & 0 & 1 & x_{k+1} & z_{k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_n & z_n \end{pmatrix}.$$

Therefore we obtain zeros of (13) by fitting a generalized linear model with design matrix D_k, which again yields because of the continuity condition

$$\hat{\tau} = \frac{\hat{\alpha}_2 - \hat{\alpha}_2}{\hat{\beta}_1 - \hat{\beta}_2}.$$

If $\hat{\tau} \epsilon (x_k, x_{k+1})$, we have found a local maximum, otherwise there is no maximum with $\hat{\tau} \epsilon (x_k, x_{k+1})$.

The remaining part of the algorithm is now completely analogous to that presented in Section 2.

4 Further extensions

4.1 Models with more than two segments

Let us now consider the case of $K > 2$ segments, i.e. Model (3). In practical problems the number of segments will be typically not greater than three. Williams (1970), for instance, restricts his investigations to this special case. But even in the situation of a linear regression with a normally distributed Y no complete algorithm for calculating the exact ML-estimator can be found in the literature. Williams (1970) states explicitly that his algorithm for three segments shows certain gaps. We describe the complete algorithm for $K = 3$, where it will be formulated such that it can be directly extended to the case of $K > 3$.

We start with the generalized linear model with two breakpoints in the following parametrization:

$$E(Y|X = x) = G(\alpha + \beta_1 x + \beta_2(x - \tau_1)_+ + \beta_3(x - \tau_2)_+) \tag{19}$$

with $\tau_1 < \tau_2$ and $\beta_i \neq 0$ for $i = 2, 3$. Let $(x_i, y_i)_{i=1,\ldots,n}$ again denote the ordered observations with $x_i \leq x_j$ for $i < j$. Then, we get the derivative of the log-likelihood with respect to $\theta = (\alpha, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2)'$ as

$$\sum_{i=1}^{n} \mathcal{G}_2(y_i, glp_i)\,(1, x_i, (x_i - \tau_1)_+, (x_i - \tau_2)_+, -\beta_2 I_{\{x_i > \tau_1\}}, -\beta_3 I_{\{x_i > \tau_2\}})' \tag{20}$$

$$\text{with} \quad glp_i = \alpha + \beta_1 x_i + \beta_2(x_i - \tau_1)_+ + \beta_3(x_i - \tau_2)_+ .$$

Since the log-likelihood is not differentiable at points with $\tau_1 = x_i$ or $\tau_2 = x_i$ for some x_i, the domain of $(\tau_1, \tau_2)'$ is divided into rectangles R_{kl} with

$$R_{kl} = [x_k; x_{k+1}] \times [x_l; x_{l+1}],\ k + 1 < l.$$

In the interior of R_{kl} the log-likelihood is differentiable. Equating (20) to 0 we get after some algebra

$$\begin{aligned}
\sum_{i=1}^{n} \mathcal{G}_2(y_i, \alpha + \beta_1 x_i)(1, x_i)' I_{\{x_i \leq \tau_1\}} &= 0 \\
\sum_{i=1}^{n} \mathcal{G}_2(y_i, \alpha - \beta_2\tau_1 + (\beta_1 + \beta_2)x_i)(1, x_i)' I_{\{x_i > \tau_1\}} I_{\{x_i \leq \tau_2\}} &= 0 \\
\sum_{i=1}^{n} \mathcal{G}_2(y_i, \alpha - \beta_2\tau_1 - \beta_3\tau_2 + (\beta_1 + \beta_2 + \beta_3)x_i)(1, x_i)' I_{\{x_i > \tau_2\}} &= 0.
\end{aligned}$$

Obviously all maxima of the log-likelihood correspond to the maxima of the separate regressions in the three segments. As in the preceding sections we

therefore derive the ML-solutions separately for the different segments which are denoted by $\tilde{\alpha}_1, \tilde{\beta}_1, \tilde{\alpha}_2, \tilde{\beta}_2, \tilde{\alpha}_3, \tilde{\beta}_3$. The continuity assumption yields

$$\tilde{\tau}_1 = \frac{\tilde{\alpha}_1 - \tilde{\alpha}_2}{\tilde{\beta}_2 - \tilde{\beta}_1} \quad \text{and} \quad \tilde{\tau}_2 = \frac{\tilde{\alpha}_2 - \tilde{\alpha}_3}{\tilde{\beta}_3 - \tilde{\beta}_2}.$$

If $\tilde{\tau}_1 \in (x_k; x_{k+1})$ and $\tilde{\tau}_2 \in (x_l; x_{l+1})$ then a local maximum has been found.

Otherwise, the log-likelihood function has its maximum value at the boundary of the rectangle, i.e. $\tau_1 \in \{x_k, x_{k+1}\}$ or $\tau_2 \in \{x_l, x_{l+1}\}$. Let for instance $\tau_1 = x_i$ then the model can be rewritten introducing a new covariate $z = (x - \tau_1)_+$ as

$$E(Y|X = x) = G(\alpha + \beta_1 x + \beta_2 z + \beta_3 (x - \tau_2)_+).$$

Thus, we are in the situation of a model with one breakpoint and an additional covariate as discussed in Section 3. To obtain all maxima at the boundary of R_{kl} we check successively all points $\tau_1 = x_{k+1}$ and $\tau_2 = x_l$ as well as $\tau_2 = x_{l+1}$.

This yields the maximizer $\hat{\theta}_{kl}$ for the rectangle R_{kl}. Finally, the global maximum is obtained as

$$\hat{\theta}_{ML} = \arg\max_{k,l} L(\hat{\theta}_{kl}).$$

4.2 Models with two regressors both having a breakpoint

As a further extension we allow for two regressors with each of them having a breakpoint which is modelled as

$$E(Y|X, Z) = G(\alpha + \beta_1 (x - \tau_1)_- + \beta_2 (x - \tau_1)_+ + \delta_1 (z - \tau_2)_- + \delta_2 (z - \tau_2)_+).$$

Even in this case we can essentially proceed as above. Let $(x_i, y_i, z_i)_{i=1,\ldots,n}$ denote the ordered observations with $x_i \leq x_j$ for $i < j$. The observations of the second regressor Z are ordered by a second index z_{i_r} with $z_{i_r} \leq z_{i_s}, r < s$. For deriving the ML-estimator of the parameter vector $\theta = (\alpha, \beta_1, \beta_2, \delta_1, \delta_2, \tau_1, \tau_2)'$ we again divide the domain of $(\tau_1, \tau_2)'$ into rectangles $R_{kr} = [x_k; x_{k+1}] \times [z_{l_r}; z_{l_{r+1}}]$ and consider first the case where $(\tau_1, \tau_2)'$ lies in the interior of R_{kr}. Using ideas of Section 2 we fit a generalized linear model with parameter vector $(\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2, \gamma_1, \gamma_2)'$ and design matrix

$$D = \left((1, x_i)\, I_{\{x_i \leq \tau_1\}},\ (1, x_i)\, I_{\{x_i > \tau_1\}},\ (1, z_i)\, I_{\{z_i \leq \tau_2\}},\ (1, z_i)\, I_{\{z_i > \tau_2\}} \right)_{i=1,\ldots,n}.$$

The corresponding model equation is given by

$$\begin{aligned} E(Y|X = x, Z = z) \quad = \quad & G((\alpha_1 + \beta_1 x) I_{\{x \leq \tau_1\}} + (\alpha_2 + \beta_2 x) I_{\{x > \tau_1\}} \\ & + (\gamma_1 + \delta_1 z) I_{\{z \leq \tau_2\}} + (\gamma_2 + \delta_2 z) I_{\{z > \tau_2\}}). \end{aligned}$$

From the ML-estimator $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\delta}_1, \hat{\delta}_2)'$ we obtain as estimators for $(\tau_1, \tau_2)'$ using the continuity assumption

$$\hat{\tau}_1 = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{\beta}_2 - \hat{\beta}_1} \quad \text{and} \quad \hat{\tau}_2 = \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\hat{\delta}_2 - \hat{\delta}_1}.$$

If $(\hat{\tau}_1, \hat{\tau}_2)' \in R_{kr}$ we have found a local maximum of the log-likelihood since the score equations can be transformed similarly to equations (14) - (18).

At the boundary of the rectangle the above model reduces to one with only a single regressor, one breakpoint and a covariate. Thus, the approach of Section 3 can be applied.

This leads to the maximizer $\hat{\theta}_{kr}$ for each rectangle R_{kr}. Finally, the global maximum results from

$$\hat{\theta}_{ML} = \arg \max_{k,r} L(\hat{\theta}_{kr}).$$

Further extensions to more than two covariates or more than two breakpoints are straightforward, but the corresponding algorithms become considerably more complicated and thus require an increasing computational effort.

5 An example: the Munich rental table

Rental tables are built up based on surveys in larger cities or communities in Germany. They serve as a formal instrument for rating rents depending on year of construction, flat size, and other covariates. For a detailed description of the data material and the statistical methods used we refer to Fahrmeir, Gieger, Mathes, and Schneeweiß (1995) and Fahrmeir, Gieger, and Klinger (1995).

As a first approach we model the relationship between net rent (Y) and flat size (X), where the assumption of a breakpoint is justified because smaller flats are more expensive relative to bigger flats. Thus, we consider the following model equation

$$E(Y|X = x) = \alpha + \beta_1(x - \tau)_- + \beta_2(x - \tau)_+. \tag{21}$$

Besides the presence of a breakpoint an additional problem occurs when analysing this data caused by heteroscedasticity. Following Fahrmeir et al. (1995) we apply a weighted regression using the weights proposed there.

The results are obtained from the algorithm presented in Section 2 and are given in Table 1, where the estimated variances are calculated by the asymptotic theory derived in Küchenhoff (1995).

As it can be seen from these results the breakpoint takes a value of about 44 m^2. For smaller flats the estimated slope of 5.04 is below the one for bigger flats ($\hat{\beta}_2 = 9.05$). Comparing our results with those gained from a linear regression no essential differences can be stated regarding the fit of the

Table 1: Parameter estimates of the weighted broken linear regression model (21) for the Munich rental table. In the column $\hat{\sigma}$ the estimated standard deviations are listed.

parameter	estimate	$\hat{\sigma}$
$\hat{\tau}$	44.0	4.6
$\hat{\alpha}$	601	36
$\hat{\beta}_1$	5.04	1.3
$\hat{\beta}_2$	9.05	0.43

data. For a more detailed analysis and a comparison with the results from Fahrmeir, Gieger, Mathes, and Schneeweiß (1995) see Küchenhoff (1995).

In a second step we take into account the age of the flat (Z) as additional regressor which is defined as 1994 minus year of construction. Possible changes in the way of building flats are reflected in the model equation by allowing for a breakpoint in the second regressor, age:

$$E(Y|X,Z) = \alpha + \beta_1(x-\tau_1)_- + \beta_2(x-\tau_1)_+ + \delta_1(z-\tau_2)_- + \delta_2(z-\tau_2)_+ . \quad (22)$$

Here, the application of the algorithm proposed in Section 4.2 yields the results given in Table 2.

Table 2: Parameter estimates of the weighted broken linear regression model (22) for the Munich rental table. τ_1 and τ_2 denote the breakpoints belonging to flat size and age. In the column $\hat{\sigma}$ the estimated standard deviations are listed.

parameter	estimate	$\hat{\sigma}$
$\hat{\tau}_1$	37.0	3.7
$\hat{\tau}_2$	30.6	0.97
$\hat{\alpha}$	501	29
$\hat{\beta}_1$	3.54	2.2
$\hat{\beta}_2$	8.81	0.35
$\hat{\delta}_1$	-12.9	1.4
$\hat{\delta}_2$	-0.391	0.28

In this case, the application of a segmented model yields a much better fit than the use of a linear model without breakpoints. The estimated breakpoint of 30.6 for the variable age corresponds to the year 1963. Since δ_2 is not significantly different from 0, the age of flats built before 1963 is not essential for their net rent. For flats built after 1963 the net rent increases with the

year of construction ($\hat{\delta}_1 = -12.9$). As already indicated, this effect is possibly due to a substantial change in the way of how houses were built.

6 Discussion

From our experience when analysing data by applying the above algorithms as for instance in the example presented in Section 5 we can state a good practicability of these algorithms. For the Munich rental table, where we have about 2.000 data points, the computing time for the model with one breakpoint was about 5 minutes on a sun-sparc 10 work station. For the model with two regressors each having a breakpoint it took about 10 minutes for calculating the ML-estimators using the algorithm presented in Section 4.2. Thus, the computing times do not seem to constitute a real problem. In addition, algorithms for linear, logistic, and weighted linear models can be obtained on request from the author. The algorithms are written in Gauss and can easily be rewritten for other generalized linear models.

Finally, we should address two important aspects related to the proposed algorithms.

When deriving the above algorithm we assumed that the nuisance parameter is constant over the segments. Note that this assumption fails e.g. in the linear regression when there are different variances in the segments. In case of logistic regression this problem does not occur since there is no nuisance parameter ($\xi = 1$). This case of the logistic regression where ξ can be treated as known can be generalized to other models. That means, alternatively to the assumption of a constant nuisance parameter the derivation of the algorithm remains valid if the nuisance parameter, itself, or the ratio of the nuisance parameters of the different segments is known.

One of the most interesting aspects of our idea for deriving an exact algorithm concerns the models with covariates. The approach proposed in this paper is not only useful for such models, for which this approach was designed, but it is also an essential part of the derivation of the corresponding algorithm for the case of more than two segments. Thus, we were also able to solve the problem pointed out by Williams (1970).

References

Fahrmeir, L., Gieger, C., and Klinger, A. (1995). Additive, dynamic and multiplicative regression, *Discussion Paper 1*, SFB 386, Munich.

Fahrmeir, L., Gieger, C., Mathes, H., and Schneeweiß, H. (1995). *Gutachten zur Erstellung des Mietspiegels für München 1994, Teil B: Statistische Analyse der Nettomieten*, Institute of Statistics, Ludwig-Maximilians-University of Munich.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*, Springer-Verlag, New York.

Hawkins, D. M. (1976). Point estimation of the parameters of piecewise regression models, *Applied Statistics* **25**: 51–57.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated, *Journal of the American Statistical Association* **61**: 1097–1129.

Küchenhoff, H. (1995). *Schätzmethoden in mehrphasigen Regressionsmodellen*, Habilitationsschrift, Institute of Statistics, Ludwig-Maximilians-University of Munich.

Küchenhoff, H. and Carroll, R. J. (1996). Biases in segmented regression with errors in predictors, *Statistics in Medicine* **15**: to appear.

Schulze, U. (1987). *Mehrphasenregression: Stabilitätsprüfung, Schätzung, Hypothesenprüfung*, Akademie-Verlag, Berlin.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear regression*, John Wiley & Sons, New York.

Stasinopoulos, D. M. and Rigby, R. A. (1992). Detecting break points in generalised linear models, *Computational Statistics & Data Analysis* **13**: 461–471.

Ulm, K. (1991). A statistical method for assessing a threshold in epidemiological studies, *Statistics in Medicine* **10**: 341–349.

Williams, D. A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures, *Biometrics* **26**: 23–32.