STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:3109–3121 (DOI: 10.1002/sim.928)

Bayesian analysis of logistic regression with an unknown change point and covariate measurement error

Christoff Gössl[1] and Helmut Küchenhoff[2,*,†]

[1] *Max-Plank-Institute of Psychiatry, Kraepelinstraße 10, D-80804 München, Germany*
[2] *University of Munich, Institute of Statistics, Akademiestraße 1, D-80799 München, Germany*

SUMMARY

We discuss Bayesian estimation of a logistic regression model with an unknown threshold limiting value (TLV). In these models it is assumed that there is no effect of a covariate on the response under a certain unknown TLV. The estimation of these models in a Bayesian context by Markov chain Monte Carlo (MCMC) methods is considered with focus on the TLV. We extend the model by accounting for measurement error in the covariate. The Bayesian solution is compared with the likelihood solution proposed by Küchenhoff and Carroll using a data set concerning the relationship between dust concentration in the working place and the occurrence of chronic bronchitis. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

In toxicology, environmental and occupational epidemiology, the assessment of threshold limiting values (TLVs) is an important task. In a dose–response relationship the TLV is the dose of a toxin or a substance under which there is no influence on the response. In many applications there is a controversy about the existence of such a TLV from a substantive point of view. In empirical studies evidence for the existence of a TLV and its estimation is often difficult, since distinguishing between no effect and a small effect can only be done with huge data sets. There are different models and methods for assessing a TLV, see, for example, Küchenoff and Ulm [1]. In this paper we concentrate on a fully parametric logistic regression model proposed by Ulm [2]. In this model, which is a segmented regression model, the TLV is treated as an unknown parameter, which can be estimated assuming its existence. The interval estimates of the TLV give some evidence about its existence, since a TLV which is smaller than the smallest observed dose is equivalent to a non-existing TLV. While the theoretical and practical problems in maximum likelihood estimation and the frequentist treatment of this model have been discussed by Küchenhoff and Wellisch [3], we use a

[*] Correspondence to: Helmut Küchenhoff, University of Munich, Institute of Statistics, Akademiestraße 1, D-80799 München, Germany
[†] E-mail: helmut@stat.uni-muenchen.de

Contract/grant sponsor: German Research Foundation (DFG); contract/grant number: SFB 386

Received April 1999
Accepted November 2000

Bayesian approach. In this context no differentiability assumptions are necessary and it can be implemented with Markov chain Monte Carlo (MCMC) methods. We apply our methods to a study concerning the relationship between dust concentration in the working place and the occurrence of chronic bronchitis. In this study the exposure can only be measured with substantial measurement error. Therefore we also show how to incorporate this measurement error in our model. Since there are different approaches and possibilities concerning the MCMC algorithm and the assumption of the distribution of the regressor variable, we give a detailed discussion of the bronchitis example. The results are compared with those of a frequentist approach.

The paper is organized as follows. In Section 2 we introduce and formalize the problem using an example from occupational epidemiology regarding the assessment of a TLV for dust concentration in the working place. We present a Bayesian solution to the task of estimating the limiting value of a logistic threshold model and propose a way to calculate the estimates by means of MCMC methods. The modelling and the handling of measurement error in the dose covariate of our model is treated in Section 3. In Section 4 we apply our methods to analyse our example in detail. Further, our methods are compared with the different approaches as investigated by Küchenhoff and Carroll [4].

2. A BAYESIAN APPROACH TO THE LOGISTIC THRESHOLD MODEL

Workplace exposure to unhealthy or even toxic agents is a well known source of severe impairments and diseases in occupational studies, but in many cases such exposure is inevitable. For example, in mining and some other industrial branches a certain dust concentration at the workplace cannot be avoided completely. Thus, it is important to assess reasonable exposure limits for workers, because one will expect that these concentrations have no adverse effects. This is exactly the task of determining a threshold limiting value (TLV). The data set with which we want to illustrate the problem in this paper describes the dependence between such a dust concentration at the workplace and the occurrence of chronic bronchitis. In an occupational study by the German Research Foundation (DFG) the disease was measured based on medical examinations, including questionnaire about symptoms, chest X-rays and lung function analysis. Further covariates were smoking and the duration of the exposure.

To formalize the problem, in the following we focus our analysis on the logistic threshold model proposed by Ulm [2]:

$$P(Y=1|X=x,Z=z) = G(z'\beta_{k-} + \beta_k(x-\tau)_+) \tag{1}$$

where

$$G(t) = (1+\exp(-t))^{-1}$$

$$\beta \in \mathbb{R}^k, \quad \beta_{k-} = (\beta_1,\dots,\beta_{k-1})' \quad \text{and} \quad (x-\tau)_+ = \max(0, x-\tau)$$

Here Y denotes the response variable, which is the occurrence of chronic bronchitis in our example, X is the dose variable and Z refers to further covariates. The unknown model parameters are β and the TLV τ. As can be seen from (1), there is no influence of X on Y if X is smaller than τ, which exactly reflects the concept of a TLV.

In contrast to the classical frequentist inference, the parameters of a Bayesian model are not assumed to be fixed but at random. For each of them there exists a probability function, which reflects the prior knowledge of their value, the so-called priors. Now it is possible, according to the theorem of Bayes, to determine in combination with the likelihood function of the data a so-called posterior of the parameters. This posterior distribution includes all knowledge relating to the parameters first from the prior and secondly from the likelihood.

The theorem of Bayes in its simplest form is

$$p(\theta|\text{data}) = \frac{p(\theta, \text{data})}{p(\text{data})}$$

Here, 'data' denotes the observed and θ the unknown parameters and latent variables. The numerator is the product of the likelihood and the priors. Note that in contrast to likelihood analysis no further assumptions on $p(\theta, \text{data})$, like differentiability in θ, are needed for the analysis.

From this posterior the Bayesian point and interval estimates are derived. In this paper we calculate for each parameter the empirical mean and the 2.5 per cent and 97.5 per cent quantiles as estimates of the parameter value and 95 per cent credible interval.

Thus, to derive the Bayesian posterior for our logistic threshold model we have to determine the conditional likelihood function and the prior distributions.

The conditional likelihood of the i.i.d. sample (y_i, z_i, x_i) $i = 1, \ldots, n$, is according to (1) given by

$$[\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau] = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \tag{2}$$

where

$$\pi_i = G(z_i' \beta_{u-} + \beta_k (x_i - \tau)_+).$$

As usual '[]' refers to the density (or probability) of the corresponding random variables. We assume that the threshold is in the range of our observed data **X** and use a uniform prior on this range for τ. For β a flat prior is assumed.

Thus, the posterior density of the parameters is

$$[\beta, \tau|\mathbf{Y}, \mathbf{Z}, \mathbf{X}] = \frac{[\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau][\beta][\tau]}{\int [\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau][\beta][\tau] \mathrm{d}\mu(\beta, \tau)} \tag{3}$$

Based on this density, the above mentioned estimates are calculated. Although the determination of the numerator is easy, in most practical cases it is not possible to evaluate the denominator in an analytic way. For the threshold model we solve this problem by means of MCMC methods. These methods allow sampling from a density only known up to a normalizing constant, which is in our particular problem the denominator. Then, on the basis of this sample of the posterior, the Bayesian estimates can simply be calculated, see, for example, Gilks *et al.* [5].

For the logistic threshold we use a two-step Metropolis-Hastings (MH) algorithm with multivariate random walk proposals in each step. We sample the parameter β_k and the threshold τ in one step and the parameter vector β_{k-} in a second. The full conditionals are straightforward. Since the densities cannot be determined analytically it is not possible to apply the

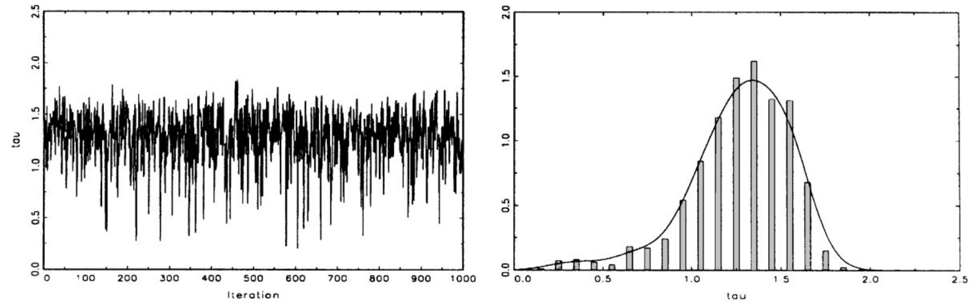


Figure 1. Trajectory and histogram with kernel density estimate of an MH algorithm.

Gibbs sampler. We take two steps because of the strong dependence between the threshold τ and β_k. The covariance matrices of the proposals are tuned according to test runs to acceptance rates from 0.3 to 0.4. The starting values are chosen to be overdispersed at random and the burn-in phase has to be determined by comparing several runs and then discarded. Owing to high autocorrelation and slow convergence in the MH output, it is often necessary to thin out the simulated chain by taking only every kth observation into the sample. We choose k such that the autocorrelation decreases to a sufficiently low level. The total extent of the runs depends on the convergence of the Markov chains and can be determined by comparing the point estimates of several runs. Figure 1 shows a trajectory of such a run and the associated histogram with kernel density estimate for the bronchitis example.

3. ERRORS IN VARIABLES

In many practical regression problems the regressors cannot be observed without any measurement error. In our present example this holds for the dust concentration. In most cases the concentration could not have been observed directly for the whole time period, but measurements were gained by averaging over several single measurements scattered over the period and raw estimates for earlier time points. Since it is well known that measurement error leads to biased estimators, we propose a solution for incorporating measurement error of the variable X in our Bayesian model. A general introduction to the measurement error problem in threshold models is given by Küchenhoff and Carroll [4]; see also Carroll *et al.* [6].

We discuss the two most common measurement error models. For the first model, instead of X only the variable $W = X + U$ can be observed, where U is the measurement error which is independent of X, Y, Z and normally distributed with $E(U) = 0$ and $V(U) = \sigma_u^2$. This case can be observed when measurements are disturbed by a number of uncontrollable factors and influences. Secondly, because the independence assumption between U and X is often too strong, we also investigate the effect of applying the Berkson error model, which assumes $X = W + U, U$ is independent of W, Y, Z and normally distributed. Strongly simplifying measurements are a typical source for this error process. We limit ourselves to these two additive models because most practical problems are covered by these approaches and cases where a multiplicative error process is advisable can be transformed to fit in the additive framework.

Under the assumption of conditional independence of Y and W given X, now we split our model into three parts:

$$\begin{array}{rl} \text{main model} & [Y|X,Z,\beta,\tau] \\ \text{error model} & [W|X,\sigma_u^2] \text{ respectively } [X|W,\sigma_u^2] \\ \text{covariable model} & [X|Z,\lambda] \text{ respectively } [W|Z] \end{array}$$

where β,τ,σ_u^2 and λ are the model parameters.

Further, using the independence assumptions for U, the likelihoods of the whole models can be written down first as

$$[\mathbf{Y},\mathbf{W}\,|\,\mathbf{Z},\theta]=\prod_{i=1}^{n}\int[y_i\,|\,x,z_i,\beta,\tau][w_i\,|\,x,\sigma_u^2][x\,|\,z_i,\lambda]\,\mathrm{d}\mu(x) \qquad (4)$$

and second for the Berkson model

$$[\mathbf{Y},\mathbf{W}\,|\,\mathbf{Z},\theta]=\prod_{i=1}^{n}\int[y_i\,|\,x,z_i,\beta,\tau][x\,|\,w_i,\sigma_u^2][w_i\,|\,z_i]\,\mathrm{d}\mu(x)$$

Since the distribution of W given Z conveys no information about the critical parameters, this can be simplified to

$$[\mathbf{Y}|\mathbf{W},\mathbf{Z},\theta]=\prod_{i=1}^{n}\int[y_i|x,z_i,\beta,\tau][x|w_i,\sigma_u^2]\mathrm{d}\mu(x) \qquad (5)$$

For our Bayesian approach the underlying model is the threshold model (1), as measurement error models we define $W|X{\sim}\mathrm{N}(X,\sigma_u^2)$, respectively, $X|W{\sim}\mathrm{N}(W,\sigma_u^2)$, and finally for the first error process we assume X to be independent from Z and to have a normal distribution with mean μ_x and variance σ_x^2. These parameters are denoted by λ. This approach is extended in Section 4 to a finite mixture of normal distributions which is with more parameters a flexible model for X.

With respect to the priors, we propose as in Section 2 that no additional information is available and therefore define again non-informatives for β and τ.

In the first model the parameters μ_x and σ_x^2 of the covariable model are assumed to have a normal distribution with mean 0 and a very large variance s^2 and a highly dispersed inverse-gamma distribution with parameters 1 and 0.005 so that its expectation equals infinity, similarly to the non-informative distributions above. The use of proper priors is advisable here so as to avoid improper posteriors, see, for example, Besag *et al.* [7]. Owing to severe identification/estimation problems we further assume the error variance σ_u^2 to be known. An evaluation of the practical implications of this measure will be given in the next section for the bronchitis example. Thus, for the posterior of the unknown parameters we get, by suitable conditional independence assumptions

$$[\beta,\tau,\mu_x,\sigma_x^2\,|\,\mathbf{Y},\mathbf{Z},\mathbf{W},\sigma_u^2]\propto[\mathbf{Y},\mathbf{W}\,|\,\mathbf{Z},\beta,\tau,\sigma_u^2,\mu_x,\sigma_x^2][\beta][\tau][\mu_x,\sigma_x^2]$$

As in the previous section, it is not possible to derive the posterior analytically. Thus, we again have to apply the MH algorithm. In addition we have to cope with the fact that the integral of formula (4) is in general not evaluable. In this paper we also want to solve the latter by means of MCMC methods. For another way of obtaining the integral analytically, which works with an approximation of the logit model by the probit model, see Carroll *et al.*

[6]. For the foundations of the following method we refer to Richardson [8]. The idea is to add the unknown variable X to the parameters with a prior according to the covariable model and sample it from its full conditional. The partial densities of the other parameters will not be affected and the integration of formula (4) is implicitly carried out by the algorithm.

Now we describe the problem more formally. As mentioned above we assume X to be independent from Z and for simplicity distributed according to a single normal distribution $\mathrm{N}(\mu_x, \sigma_x^2)$. We regard the latent variables X_i as parameters and decompose the likelihood as follows:

$$[\mathbf{Y}, \mathbf{W} \mid \mathbf{Z}, \mathbf{X}, \beta, \tau, \sigma_u^2] = [\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \beta, \tau][\mathbf{W} \mid \mathbf{X}, \sigma_u^2]$$

With the above priors and $X \sim \mathrm{N}(\mu_x, \sigma_x^2)$, the posterior takes the form

$$\begin{aligned}[\beta, \tau, \mu_x, \sigma_x^2, \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \sigma_u^2] &\propto [\mathbf{Y}, \mathbf{W} \mid \mathbf{Z}, \mathbf{X}, \beta, \tau, \sigma_u^2][\mathbf{X} \mid \mu_x, \sigma_x^2][\beta][\tau][\mu_x, \sigma_x^2] \\ &\propto [\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau][\mathbf{W} \mid \mathbf{X}, \sigma_u^2][\mathbf{X}|\mu_x, \sigma_x^2][\beta][\tau][\mu_x, \sigma_x^2]\end{aligned}$$

For the MH algorithm, we use, as in Section 2, two metropolis steps with random walk proposals for the parameters β and τ. Furthermore, we add an Metropolis step for X. Thus, full conditionals are given by

$$\begin{aligned}&[\beta_{k-} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{X}, \beta_k, \tau, \sigma_u^2, \mu_x, \sigma_x^2] \propto [\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau] \\ &[\beta_k, \tau \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{X}, \beta_{k-}, \sigma_u^2, \mu_x, \sigma_x^2] \propto [\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau][\tau] \\ &\quad [x_i|y_i, z_i, w_i, \beta, \tau, \sigma_u^2, \mu_x, \sigma_x^2] \propto [y_i|z_i, x_i, \beta, \tau][w_i|x_i, \sigma_u^2][x_i|\mu_x, \sigma_x^2], \quad i = 1, \ldots, n\end{aligned}$$

Owing to our choice of normal and inverse-gamma distributions, respectively, for the parameters μ_x and σ_x^2 we can derive their full conditionals as

$$(\mu_x|\mathbf{X}, \sigma_x^2) \sim \mathrm{N}\left(\frac{s^2 \sum x_i}{s^2 n + \sigma_x^2}, \frac{s^2 \sigma_x^2}{s^2 n + \sigma_x^2}\right)$$

$$(\sigma_x^2|\mathbf{X}, \mu) \sim \mathrm{IG}\left(\frac{n}{2} + 1, \frac{1}{2} \sum (x_i - \mu_x)^2 + 0.005\right)$$

In the case of the Berkson error model, the formulation of the posterior and the full conditionals is far less extensive. Because the covariable model vanishes in the calculation of the posterior and $X|W$ is normal with again known variance σ_u^2, the posterior becomes

$$[\beta, \tau \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \sigma_u^2] \propto [\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \beta, \tau, \sigma_u^2][\beta][\tau]$$

Further, we treat the integral problem similarly to the above, which yields

$$[\beta, \tau, \mathbf{X}|\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \sigma_u^2] \propto [\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \beta, \tau][\mathbf{X}|\mathbf{W}, \sigma_u^2][\beta][\tau]$$

In the MCMC algorithm we adopt the above full conditionals for β and τ, for the unobserved variable X they are

$$[x_i \mid y_i, z_i, w_i, \beta, \tau, \sigma_u^2] \propto [y_i \mid z_i, x_i, \beta, \tau]\,[x_i \mid w_i, \sigma_u^2], \quad i = 1, \ldots, n$$

Further steps, for example, for μ_x and σ_x^2, are not necessary

For the details of the implementation of the algorithm, such as fixing the starting values, we refer to Section 2. An application of the algorithm is reported in the following section.

4. BRONCHITIS STUDY

In this section some results are shown for the application of the presented approaches to the dust data set described above. As already mentioned this study investigates the relationship between average dust concentration in the working place and the occurrence of a chronic bronchitis reaction (Y) with covariates smoking (SMK) and duration of exposure (DUR). We use the data of 1256 Munich workers which were also anlaysed by Küchenhoff and Carroll [4]. It should be mentioned that we use the quantity $X = \log(1 + \text{dust concentration})$ in our calculations. Thus, the model runs as follows:

$$P(Y=1) = G(\beta_1 + \beta_2 \text{SMK} + \beta_3 \text{DUR} + \beta_4 (X - \tau)_+) \tag{6}$$

As well as the simple threshold model, the measurement error model of Section 3 will also be applied.

For the simple model without measurement error, the algorithm of Section 2 can be adopted directly. Only the proposals of the Markov chain have to be modified in the described manner. Apart from that we derived our estimates from one chain with 110 000 iterations, where, due to the high autocorrelations, we took every 100th observation in our sample. With respect to the burn-in, we discarded the first 10 000 iterations. Table I shows the estimates and those of Küchenhoff and Carroll [4] gained by the classical methods. While the estimators are nearly identical, the estimators for the variance are higher for the classical methods. For example, the estimated standard deviation for the threshold estimate $\hat{\tau}$ was 0.41 compared to 0.28 in the Bayesian analysis.

In order to take into account the measurement error, more complex modifications of the presented model have to be made. The first is to regard the measured dust concentration as the error-exposed surrogate W. The true concentration X is unknown.

For the first error model, in a preliminary attempt we applied a single normal distribution to model the data, but due to the ill-fitting of this covariable model (see Figure 2) for reasonable choices of the error variance ($\sigma_u^2 < 0.5^2$), we decided to discard this simple approach. Therefore, following Küchenhoff and Carroll [4], we model the distribution of the unknown variable X by a mixture of two normal distributions. Assuming an additive measurement error $W = X + U$ where $U \sim \mathrm{N}(0, \sigma_u^2)$, then W is also a mixture of normals with the same number of mixing distributions as X. Thus, taking two mixing distributions is justified by the empirical distribution of W. Consequently, for $\lambda \in [0;1]$ we assume X to be distributed according to

$$X \sim \mathrm{MixN}(\mu_{x1}, \sigma_{x2}^2, \mu_{x2}, \sigma_{x2}^2, \lambda)$$

Table I. Likelihood and Bayes estimates of the simple model.

Parameter	ML estimate	Bayes estimate	Bayes STD
β_1	−3.00	−3.01	0.24
β_2	0.68	0.69	0.17
β_3	0.039	0.040	0.01
β_4	0.85	0.91	0.35
τ	1.27	1.27	0.28

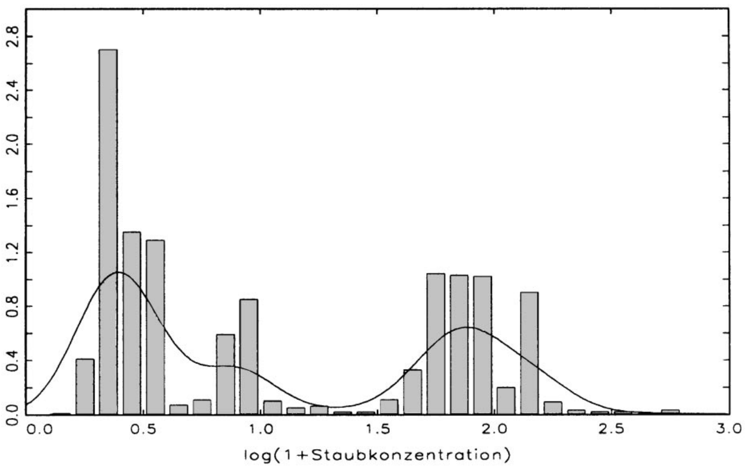


Figure 2. Histogram and kernel density estimate of the dust data set.

with

$$[x \mid \mu_{x1}, \sigma_{x2}^2, \mu_{x2}, \sigma_{x2}^2, \lambda] = \lambda \sigma_{x1}^{-1} \phi\left(\frac{x-\mu_{x1}}{\sigma_{x1}}\right) + (1-\lambda)\sigma_{x2}^{-1}\phi\left(\frac{x-\mu_{x2}}{\sigma_{x2}}\right)$$

The covariable model is given above, and the underlying model the logistic threshold model is straightforward and we suppose an additive measurement error model, where $W \mid X = x$ has a normal distribution, that is $(W \mid X = x) \sim \mathrm{N}(x, \sigma_u^2)$. Thus, the likelihood is complete and is given by (4).

The prior distributions are chosen as in the previous section, we only have to take into account that the covariable model has more parameters than above. Again, we assume the case of no prior information. Accordingly we define the priors for the additional parameters μ_{x2}, σ_{x2}^2 for which we use the same dispersed normal and gamma distribution as for μ_{x1}, σ_{x1}^2, and we define the prior for λ where we use a uniform distribution on $[0;1]$. Because the integral of the likelihood is not evaluable, we also have to consider the different prior of the unknown variable X according to our covariable model.

With respect to a practical solution, we assume the variance of the measurement error is known and take the value proposed in Küchenhoff and Carrol [4], $\sigma_u^2 = 0.187^2$. Because of problems with model identification, other assumptions, like setting another prior distribution on σ_u^2, did not work in our model. Therefore the posterior is of the form

$$\begin{aligned}&[\beta, \tau, \mu_{x1}, \mu_{x2}, \sigma_{x1}^2, \sigma_{x2}^2, \lambda, \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \sigma_u^2]\\ &\quad \propto [\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \beta, \tau][\mathbf{W} \mid \mathbf{X}, \sigma_u^2][\mathbf{X} \mid \mu_{x1}, \mu_{x2}, \sigma_{x1}^2, \sigma_{x2}^2, \lambda][\beta][\tau][\mu_{x1}][\mu_{x2}][\sigma_{x1}^2][\sigma_{x2}^2][\lambda]\end{aligned}$$

Although the derivation of the above formula is quite simple, the adaptation of the MH algorithm is a far more sophisticated task. The main problem results from the determination of the original distribution of the mixture for the variable X, which is necessary for defining the full conditionals of the associated parameters. Here we use a method which is given in detail in Robert [9] and works in principle by defining an indicator variable m_i, which for all observations of X states from which distribution of the mixture it comes.

In general, suppose for X a mixture of m normal distributions is given, with parameters $\mu_1,\dots,\mu_m$, $\sigma_1^2,\dots,\sigma_m^2$, $\lambda_1,\dots,\lambda_m$: $x_i \sim \sum_{j=1}^{m} \lambda_j \mathrm{N}(\mu_j, \sigma_j^2)$. Then weights $g_{ij} = \lambda_j \sigma_j^{-1} \phi(\frac{x_i-\mu_j}{\sigma_j})$ can be calculated that correspond to the contributions of the several mixing distributions to the density of x_i. Here $\phi(x)$ denotes the standard normal density. Thus, defining a discrete random variable M_i on the set of distributions $\{1,\dots,m\}$ with the standardized weights $p_{ij} = \frac{g_{ij}}{\sum_j g_{ij}}$ as probabilities, yields a variable that allocates each element j of the mixture a probability of having generated the observation x_i. Consequently, realizations of this distribution assign every observation one particular underlying distribution.

After having sampled an origin for every observation, the means, variances and mixing parameters of these distributions can be updated in familiar Gibbs or MH steps, according to the priors and the observations that come from this distribution. In our case, we use as above conjugate normal and inverse gamma priors for the means and variances and a non-informative prior for the mixing parameter so that Gibbs steps and one MH step can be applied. Thus, the full conditionals for our particular algorithm with the mixture of two normal distributions are

$$[\beta_{4-} \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{X}, \beta_4, \tau, \sigma_u^2, \mu_{x1}, \sigma_{x1}^2, \mu_{x2}, \sigma_{x2}^2, \lambda] \propto [\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \beta, \tau]$$

$$[\beta_4, \tau \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{X}, \beta_{4-}, \sigma_u^2, \mu_{x1}, \sigma_{x1}^2, \mu_{x2}, \sigma_{x2}^2, \lambda] \propto [\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \beta, \tau][\tau]$$

$$[\lambda \mid \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{X}, \beta, \tau, \sigma_u^2, \mu_{x1}, \sigma_{x1}^2, \mu_{x2}, \sigma_{x2}^2] \propto [\mathbf{X} \mid \mu_{x1}, \sigma_{x1}^2, \mu_{x2}, \sigma_{x2}^2, \lambda][\lambda]$$

$$[x_i \mid y_i, z_i, w_i, \beta, \tau, \sigma_u^2, \mu_{x1}, \sigma_{x1}^2, \mu_{x2}, \sigma_{x2}^2, \lambda] \propto [y_i \mid z_i, x_i, \beta, \tau][w_i \mid x_i, \sigma_u^2][\mathbf{X} \mid \mu_{x1}, \sigma_{x1}^2, \mu_{x2}, \sigma_{x2}^2, \lambda]$$

for the MH steps and for Gibbs sampling we get

$$(\mu_j \mid \mathbf{X}, \sigma_j^2) \sim \mathrm{N}\left(\frac{10\sum x_i^j}{10n_j + \sigma_j^2}, \frac{10\sigma_j^2}{10n_j + \sigma_j^2}\right)$$

$$(\sigma_j^2 \mid \mathbf{X}, \mu_j) \sim \mathrm{IG}\left(\frac{n_j}{2} + 1, -\frac{1}{2}\sum (x_i^j - \mu_j)^2 + 0.005\right), \quad j = 1, 2$$

$$(m_i | \mathbf{X}, \mu_{x1}, \mu_{x2}, \sigma_{x1}^2, \sigma_{x2}^2, \lambda) = \begin{cases} P(M_i = 1) = p_{i1} \\ P(M_i = 2) = 1 - p_{i1} \end{cases}$$

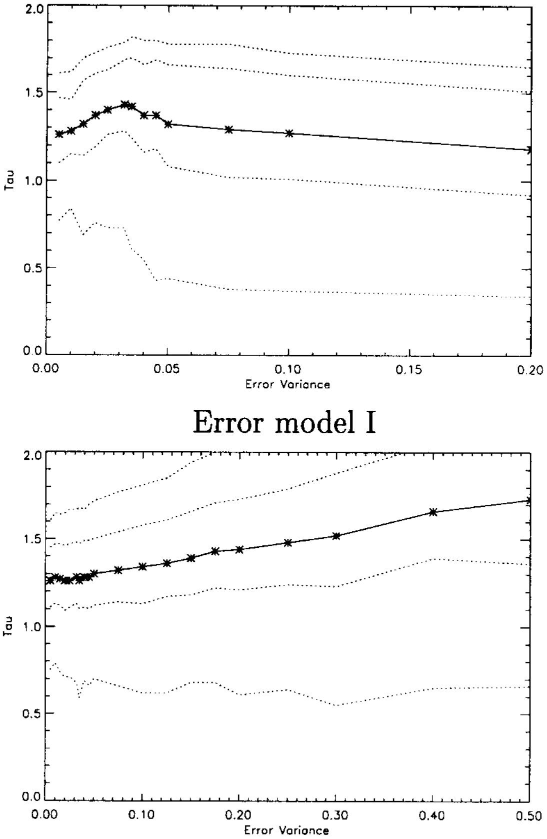
where p_{i1} is defined as above, n_j denotes the number of observations x_i coming from distribution $M_i = j$ and x_i^j are their values themselves.

For the Berkson model the full conditionals and the MCMC algorithm of Section 3 can be applied. No new parameters for the auxiliary variables have to be introduced. We fixed the error model variance again at 0.187^2.

Further, we derived our estimates from one chain the length of 110 000 iterations where we took every 100th observation into our sample. Autocorrelations decreased to values of about 0.1 to 0.25, which seemed to be acceptable. The discarded burn-in extended again to 10 000 iterations. Table II shows the Bayes estimates for both error models. For comparison the estimate for the threshold derived by Küchenhoff and Carroll [4] under the assumption of a mixture of normal distributions for X, where $X \sim \mathrm{MixN}(0.52, 0.144^2, 1.93, 0.106^2, 0.61)$, and a variance for the error model of $\sigma_u^2 = 0.187^2$, takes a value of 1.76.

Table II. Bayes estimates for the dust data set, threshold model with errors in variables.

	β_1	β_2	β_3	β_4	τ	μ_{x1}	μ_{x2}	σ_{x1}^2	σ_{x2}^2	λ
Mean (Model I)	−3.02	0.69	0.04	1.63	1.43	0.519	1.927	0.023	0.013	0.607
STD (Model I)	0.23	0.17	0.01	0.98	0.36	<0.01	<0.01	<0.01	<0.01	<0.01
Mean (Berkson)	−3.01	0.70	0.04	0.92	1.28					
STD (Berkson)	0.25	0.18	0.01	0.37	0.30					



Error model I

Berkson model

Figure 3. Results of sensitivity analysis.

As can be seen, the Bayes estimates for the two error models differ remarkably. The reason for this discrepancy should lie in the different sensitivity of the models with regard to the choice of the error variance. Evaluations of the models for a grid of error variances have confirmed this assumption. The results are shown in Figure 3, where the mean together with 90 per cent and 50 per cent credibility regions are displayed. Whereas the estimation of the TLV for the first error model is most sensitive to changes of the error variance in the

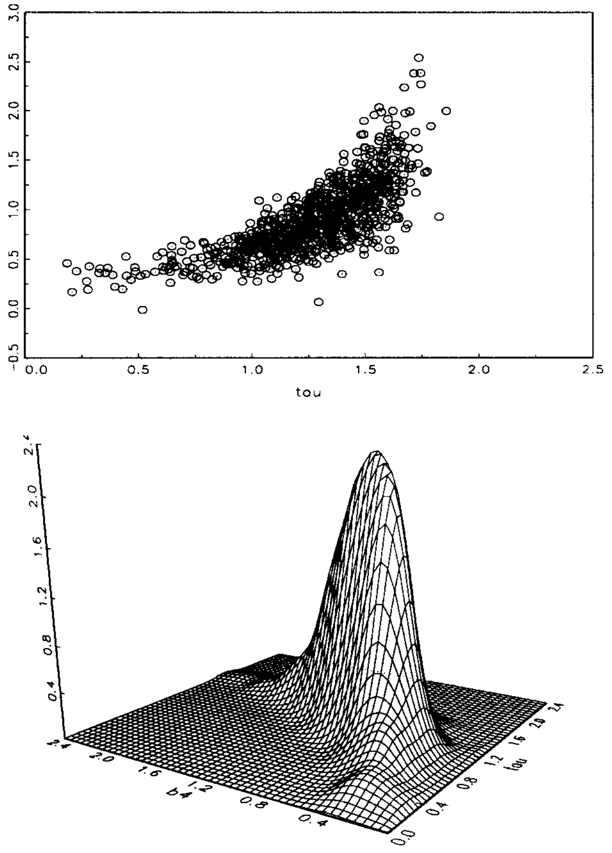


Figure 4. Scatter plot and two-dimensional kernel density estimation of the parameters β_4 and τ of the simple model of the bronchitis study.

range from 0.1^2 to 0.7^2, the Berkson model shows the first remarkable effects for values higher than 0.7^2. Furthermore, while the estimated TLV for the first error model decreases for variances larger than 0.2^2, in the Berkson model the estimates increase steadily with higher error variances. Even though the first result is astonishing, the increase in the Berkson model could be expected. The less reliably an influencing factor can be measured, the less strongly its influence should be weighted. In the logistic threshold model this means that the TLV moves to the upper end of the factor's scale, practically excluding the covariable from the regression. For the first model an explanation of the results could be the noticeable separation of the components of the mixture of densities. Up to a certain limit the choice of the error variance influences the value of the TLV; exceeding this 'reasonable' imprecision of the covariable, the TLV moves back to the robust location between the two densities. In conclusion, for the first model the choice of an error variance of $\sigma_u^2 = 0.187^2$ is advisable, reflecting the most severe possible changes in the estimates when concerning measurement error. This could be interpreted as a kind of an upper bound for the TLV. For the Berkson model the results are very similar to the simple model when using the suggested error variance, but here the question arises as to whether this approach is adequate at all, being sensitive only for unrealistically large error variances.

With respect to the classical methods, with the same assumptions they also give different estimates for the parameter τ depending on the method, see Küchenhoff and Carroll [4]. While the likelihood estimator is 1.76, the simex estimator of Cook and Stefanski [10] which does not use the information about the distribution of X gives a result of 1.40, which is close to our result. An interesting point is that variance estimation is higher for the Bayesian approach than it is for all classical methods where the SE varies from 0.12 to 0.23. A reason could be that in the Bayesian approach all sources of variability are modelled automatically, while in the classical approach this is not the case.

As mentioned above, β_4 and τ are sampled in the same MH step, because of their high correlation. The estimate for this correlation in the dust data set was 0.71. In Figure 4 a scatter plot and a two-dimensional kernel density estimate of the posterior's sample of β_4 and τ for the simple model in the bronchitis study also shows this high correlation.

5. CONCLUSION

We have shown that Bayesian analysis for the complicated model of a segmented regression with errors in the regressors can be done by MCMC methods. We use a mixture of normals for the distribution of the regressor variable, which is flexible parametric model. In this setting a classical analysis is very difficult, both theoretically and from a practical point of view. Another important argument for Bayesian analysis for finding threshold limiting values in epidemiology is the possibility of including knowledge from other studies or from substantive considerations by selecting a suitable prior distribution for the TLV.

A possible extension of our model would be to drop the assumption of an existing threshold, but to find out whether there is a threshold by data analysis. Another point is to treat the number of normals in the mixture distribution as a further parameter, which has been discussed by Carroll *et al.* [11].

ACKNOWLEDGEMENTS

C. Gössl's research was supported by grant SFB 386 from the German Research Foundation (DFG). We would like to thank Leonhard Knorr-Held for useful discussions and valuable comments.

REFERENCES

1. Küchenhoff H, Ulm K. Comparison of statistical methods for assessing threshold limiting values in occupational epidemiology. *Computational Statistics* 1997; **12**:249–264.
2. Ulm K. A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine* 1991; **10**:341–349.
3. Küchenhoff H, Wellisch U. Asymptotics for generalized linear segmented regression models with unknown breakpoint. Discussion Paper Nr. 83, SFB 386, München, 1997.
4. Küchenhoff H, Carroll RJ. Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statistics in Medicine* 1997; **16**:169–188.
5. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London, 1996.
6. Carroll RJ, Ruppert D, Stefanski LA. *Nonlinear Measurement Error Models*. Chapman & Hall: New York, 1995.

7. Besag JE, Green PJ, Higdon D, Mengersen K. Bayesian computation and stochastic systems (with discussion). *Statistical Science* 1995; **10**:3–66.
8. Richardson S. Measurement error. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996; 401–418.
9. Robert CP. Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996; 441–464.
10. Cook JR, Stefanski LA. Simulation extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 1994; **89**:1314–1328.
11. Carroll JR, Roeder K, Wassermann L. Flexible parametric measurement error models. *Biometrics* 1998; **55**:44–54.