STATISTICS IN MEDICINE, VOL. 16, 169–188 (1997)

SEGMENTED REGRESSION WITH ERRORS IN PREDICTORS: SEMI-PARAMETRIC AND PARAMETRIC METHODS

H. KÜCHENHOFF

Seminar für Ökonometrie und Statistik, Akademiestrasse 1, D-80799 München, Germany

AND

R. J. CARROLL

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A

SUMMARY

We consider the estimation of parameters in a particular segmented generalized linear model with additive measurement error in predictors, with a focus on linear and logistic regression. In epidemiologic studies segmented regression models often occur as threshold models, where it is assumed that the exposure has no influence on the response up to a possibly unknown threshold. Furthermore, in occupational and environmental studies the exposure typically cannot be measured exactly. Ignoring this measurement error leads to asymptotically biased estimators of the threshold. It is shown that this asymptotic bias is different from that observed for estimating standard generalized linear model parameters in the presence of measurement error, being both larger and in different directions than expected. In most cases considered the threshold is asymptotically underestimated. Two standard general methods for correcting for this bias are considered; regression calibration and simulation extrapolation (simex). In ordinary logistic and linear regression these procedures behave similarly, but in the threshold segmented regression model they operate quite differently. The regression calibration estimator usually has more bias but less variance than the simex estimator. Regression calibration and simex are typically thought of as functional methods, also known as semi-parametric methods, because they make no assumptions about the distribution of the unobservable covariate X. The contrasting structural, parametric maximum likelihood estimate assumes a parametric distributional form for X. In ordinary linear regression there is typically little difference between structural and functional methods. One of the major, surprising findings of our study is that in threshold regression, the functional and structural methods differ substantially in their performance. In one of our simulations, approximately consistent functional estimates can be as much as 25 times more variable than the maximum likelihood estimate for a properly specified parametric model. Structural (parametric) modelling ought not be a neglected tool in measurement error models. An example involving dust concentration and bronchitis in a mechanical engineering plant in Munich is used to illustrate the results.

1. INTRODUCTION

Regression models are used for modelling the relationship between a response Y and a predictor X. Usually it is assumed that this relationship remains unchanged over the complete range of X. There are, however, practical problems, where this assumption is not suitable. Consider, for example, the occupational study which motivated our research and will be discussed in Section 6.

CCC 0277–6715/97/020169–20

In this study the main issue is the influence of occupational dust concentration on the development of chronic bronchitis. Here it is assumed that there is no influence of the exposure 'dust concentration' on the response under a certain limiting value. Thus the relationship is modelled in two segments. Under the threshold value the regression function is a constant, while it is increasing when the dust concentration exceeds the threshold. This is a special case of segmented regression. There a response Y depends on a predictor X and is modelled by a function $f(\cdot)$, which has a different analytical form and different parameters in different intervals of the domain of X. The endpoints of these intervals are called changepoints. There is wide application of such models, see for example Shaban[1] and Seber and Wild.[2] There are many possible special cases, including that the function f can be continuous or discontinuous at the changepoints, the changepoints can be known or unknown, and the number of changepoints can be known or unknown. An important example is broken line regression with changepoint τ_0, where f is continuous and linear in the intervals $[a, \tau_0]$ and $[\tau_0, b]$. As already mentioned, another special case is the threshold model. This can be modelled by a segmented regression model with $f(\cdot)$ being constant for X smaller than the threshold value τ_0. A formal description is given in Section 2.

In this paper, we discuss methods for estimation in the threshold segmented regression model where X is unobservable but is instead subject to additive measurement error, that is, $W = X + U$ is observed. In our occupational study, for instance, the predictor 'averaged individual dust concentration at the working place' cannot be measured exactly. Instead we have averages of different measurements. While there is a large literature on the problem when X is observable, to the best of our knowledge there have been no attempts to study the measurement error problem with an unknown threshold. Gbur and Dahm[3] and Grimshaw[4] consider the case of a known threshold. Our main issue is the estimation of the unknown threshold.

The outline of the paper is as follows. In Section 2, we describe the asymptotic biases incurred in linear threshold segmented regression when measurement error is ignored, in the particular special case that X and W are jointly normally distributed. Under these circumstances, it is possible to perform exact bias calculations, one of the few such examples in the literature for non-linear models. We find that the bias of the naive estimator of τ_0 does *not* behave according to the 'correction for attenuation' formula of linear regression, with the biases having qualitatively different behaviour depending on whether the changepoint occurs before or after the mean of X. The results are extended to logistic regression and to the case that X has a mixture normal distribution.

In Section 3, we review two general methods for approximate estimation in generalized linear measurement error models, *regression calibration* (Rosner *et al.*,[5] Carroll and Stefanski,[6] Gleser,[7] Pierce *et al.*[8]) and *simulation extrapolation* or *simex* (Cook and Stefanski,[9] Carroll *et al.*[10]). In generalized linear models, and in particular for linear, logistic and log-linear models, these two methods have similar bias and variance performance over a wide variety of conditions, with regression calibration often used because of its ease of computation. In contrast, as shown in Section 4, in the threshold segmented regression model, regression calibration and simex behave fundamentally differently with respect to bias. Our conclusion is that the simex procedure is better at correcting the bias induced by measurement error, especially in the case that the threshold occurs at a value larger than the mean of X. Likelihood methods are also discussed; when the likelihood is specified correctly, the maximum likelihood estimator can be far superior to the other methods.

Section 5 describes the results of a small simulation study. It is well-known that in order to correct the naive estimate for bias in linear regression, one incurs the cost of extra variability. This so-called 'bias-variance trade-off' also occurs in the threshold segmented regression model. The size of this trade-off is somewhat of a surprise, however, to those accustomed to linear regression.

For example, for normal regressors X, using regression calibration to estimate the threshold limiting value (TLV) τ_0 actually results in a decrease in variance, while using simex causes a much larger increase in variance than one might expect.

As already mentioned, we consider finally in Section 6 an occupational study involving the relationship between chronic bronchitis (Y) and dust concentration in the workplace (X). It is of main interest to assess a threshold limiting value (TLV) under which no harm to health can be expected. Therefore, we applied a logistic threshold model estimating the TLV taking into account the non-neglible measurement error in the predictor dust concentration.

2. BIAS ANALYSIS FOR LINEAR AND LOGISTIC REGRESSION

Before discussing the problem of measurement error in the predictor, we give the model equation for the threshold model, where the expectation of the response conditional on the predictor is written as a function of the so-called broken line predictor, that is

$$E(Y|X) = H\{\alpha_0 + \beta_0(X - \tau_0)_+\} \tag{1}$$

where

$$(x-\tau)_+ = \begin{cases} 0 & \text{if} \quad x < \tau; \\ x - \tau & \text{if} \quad x \geqslant \tau. \end{cases}$$

The term $(x - \tau_0)_+$ reflects the fact that the predictor X only has influence on Y above the threshold (changepoint) τ_0. In (1), H is the link-function, for example, linear, logistic, etc. For simplicity of exposition, except in the example we leave out further regressors which would usually be in the model. For the linear link, Feder[11] discussed the asymptotic properties of the maximum likelihood estimate of τ_0. Ulm[12] discussed logistic regression. Stasinopoulous and Rigby[13] gave concrete methods of calculation for generalized linear models.

We now consider the important case that the regressor can only be observed with an additive non-differential measurement error U, that is, instead of X the variable $W = X + U$ is observed where U is independent of X and Y and has mean zero and variance σ_u^2. In our discussion, we will assume that the measurement error variance σ_u^2 is known, although with additional data it can be estimated (Carroll *et al.*,[10] Chapter 3).

We allow for linear, logistic and general regression models by writing the log-likelihood function without measurement error as

$$\mathscr{L}\{Y, \alpha + \beta(X - \tau)_+, \xi\}, \tag{2}$$

where ξ denotes nuisance parameters, for example, the variance of Y given X in linear regression. In logistic regression there is no nuisance parameter and then we get the well-known formulae

$$\mathscr{L}\{Y, \alpha + \beta(X-\tau)_+\} = Y\log[H\{\alpha + \beta(X-\tau)_+\}] + (1-Y)\log[1 - H\{\alpha + \beta(X-\tau)_+\}]$$

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

The observations are independent and identically distributed and denoted by (Y_i, W_i) for subjects $i = 1, \dots, n$.

The naive estimator which ignores the measurement error is obtained by maximizing

$$\sum_{i=1}^{n} \mathscr{L}\{Y_i, \alpha + \beta(W_i - \tau)_+, \xi\} \tag{3}$$

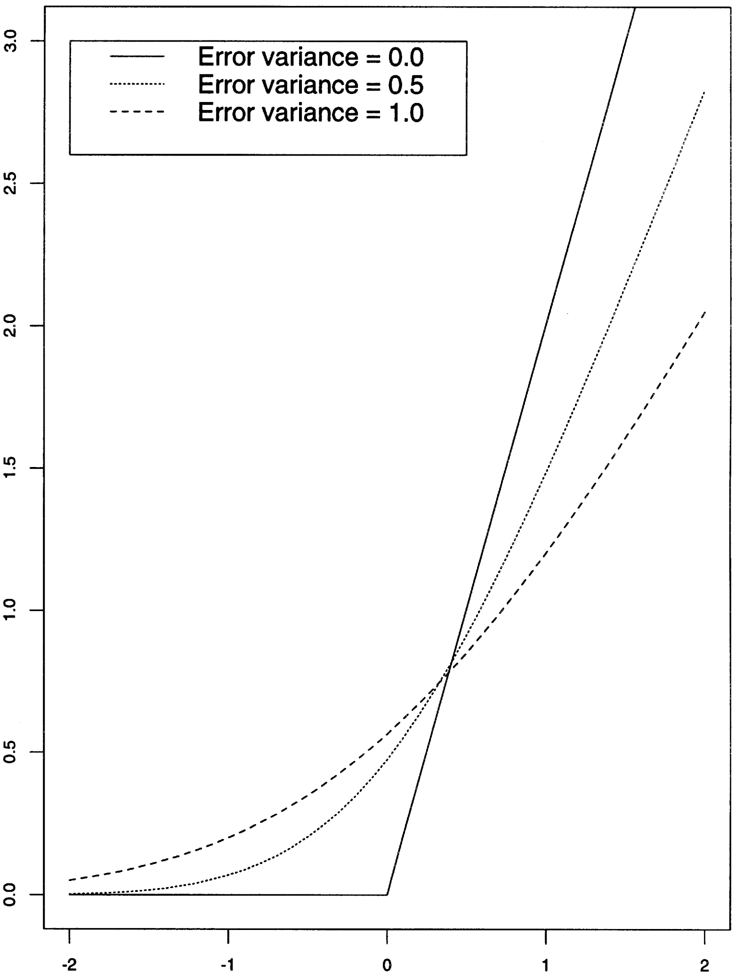


Figure 1. The true segmented regression line and the observed lines with measurement error variance 0·5 and 1·0. Here $\alpha_0 = \tau_0 = 0$, $\beta_0 = 2$ and $\sigma_x^2 = 1$

in the parameters $(\alpha, \beta, \tau, \xi)$. In practice this is done by finding the profile likelihood for a given τ

$$\mathscr{L}_{\mathrm{P}}(\tau) = \max_{\alpha,\beta,\xi} \sum_{i=1}^{n} \mathscr{L}\{Y_i, \alpha + \beta(W_i - \tau)_+, \xi\}$$

and then $\mathscr{L}_{\mathrm{P}}(\tau)$ is maximized by a grid search. The naive estimator is generally inconsistent in the presence of measurement error.

Part of the reason for this inconsistency is that the observed data do not follow the threshold segmented regression model. To see this, consider the case of linear regression with normally distributed measurement error, and let X also be normally distributed with mean μ and variance σ_x^2. Because of $X = W + U$ and $E(U) = 0$, X and W have the same mean denoted by μ. Defining $\gamma_1 = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$, X given W is also normally distributed with $E(X|W) = m(W) =$

$\mu(1-\gamma_1)+\gamma_1 W$ and $\mathrm{var}(X|W)=\gamma_1\sigma_u^2$. The observed regression can be calculated by integration:

$$E(Y|W)=E\{\alpha_0+\beta_0(X-\tau_0)_+|W\}=\alpha_0+\beta_0\gamma_1^{1/2}\sigma_u\{a\Phi(a)+\phi(a)\} \tag{4}$$

where $a=\{m(W)-\tau_0)\}/(\gamma_1\sigma_u^2)^{1/2}$ and ϕ and Φ are the normal density and distribution functions, respectively. In Figure 1 we plot (4) as a function of W, when $\mu=\alpha_0=0$, $\sigma_x=1$ and $\sigma_u^2=0, 0{\cdot}5, 1{\cdot}0$. Inspection of Figure 1 shows clearly that the overall regression line does not follow the threshold segmented regression model when there is measurement error.

Now consider the linear regression model with $E(Y|X)=\alpha_0+\beta_0(X-\tau_0)_+$, with W and X jointly normally distributed. As indicated in the Appendix, the naive estimator is not consistent for the TLV τ_0, but instead it is a consistent estimate of another value, τ_*, which depends on the values of all the parameters. There are only a few other problems where the exact asymptotic biases for the naive estimator in measurement error models have been worked out, namely for linear, quadratic, log-linear and probit regression, see Carroll *et al.*[10] for details. These cases are all rather easy, and so our calculations represent the first explicit asymptotic bias calculations in truly non-linear models.

In Figure 2 we exhibit the asymptotic biases for the naive estimator, for the cases $\tau_0=\pm 1$. Details about the calculations can be found in the Appendix. It is interesting to note that the asymptotic bias is almost exactly a linear function of the measurement error variance σ_u^2 in the case that $\tau_0=-1$, and the direction of the bias is *in the wrong direction*, at least if one uses the traditional correction for attenuation of simple linear regression as a guide. In fact, with $\tau_0=-1$, the effect of measurement error is that the estimator goes away from zero, and not towards it. When $\tau_0=1$, the bias is in the expected direction, but not nearly so extreme.

We have also extended the exact bias calculations to the logistic regression model, using the approximation of the logistic function by the normal distribution function with standard deviation 1·7 (see Appendix). We have performed these calculations for the case $\tau_0=0$ and $\tau_0=1$, with results which are roughly the same as in the linear case.

These calculations can be extended to the mixture of normals model for X, see Section 6.

In summary, in the special case considered here, the bias of the naive estimator of τ_0 behaves differently from what one would expect. In particular, for most (although not all) cases we have considered, the effect of measurement error is to underestimate the TLV τ_0, independent of its sign.

3. MEASUREMENT ERROR METHODS

In this section, we describe two general approximate measurement error methods for correcting for measurement error, regression calibration and simex. They are both functional (semi-parametric), in that neither makes assumptions about the distribution of X. We also discuss parametric maximum likelihood estimators (MLEs).

3.1. Regression Calibration

The regression calibration method regresses Y on an estimate of $E(X|W)$. The method provides consistent estimates of slope parameters in linear and quadratic regression and in log-linear models when the distribution of X given W is homoscedastic. For probit and logistic regression, the resulting parameter estimates are typically very nearly consistent (Rosner *et al.*[5]).

The operational problem is to estimate $E(X|W)$. If it is only known that $W=X+U$ where U has mean zero and variance σ_u^2, two methods have been proposed. The first (Carroll and

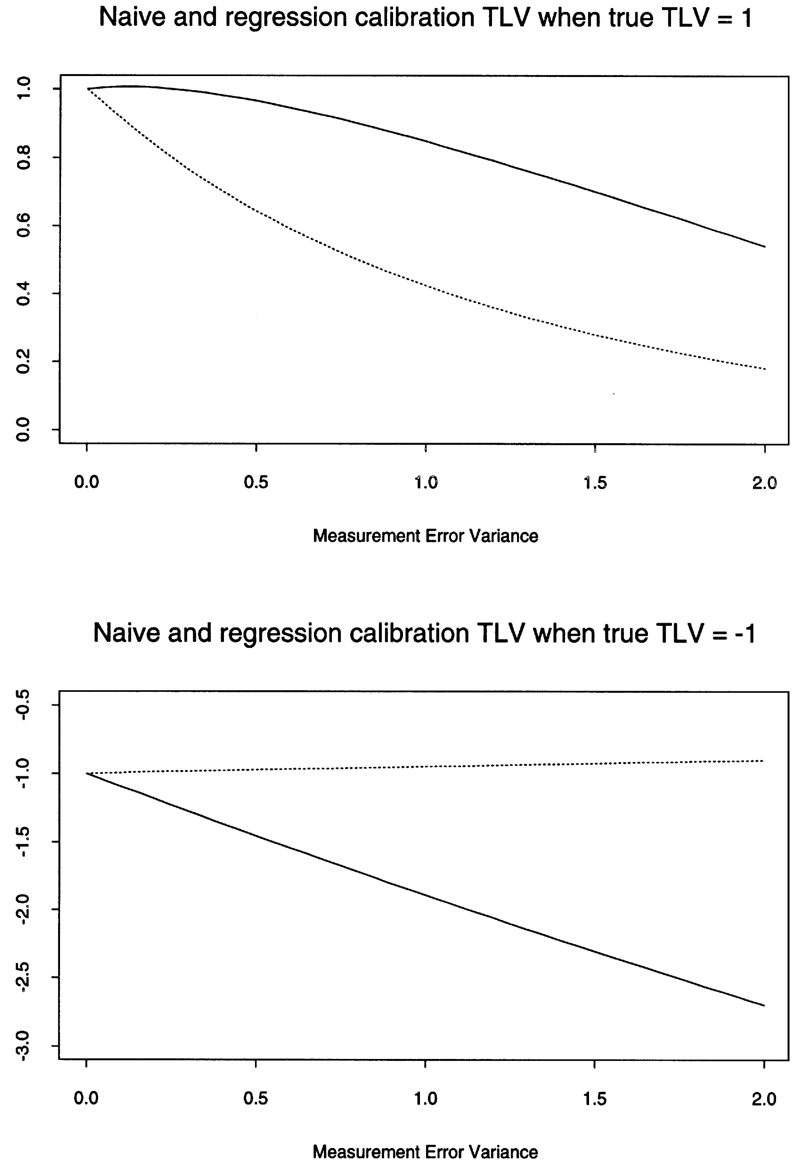


Figure 2. For given amounts of measurement error, the actual limiting value of the naive (solid line) and regression calibration (dashed line) TLV estimates, for $\tau_0 = \pm 1$. Here $\alpha_0 = 0$, $\beta_0 = 2$ and $\sigma_x^2 = 1$

Stefanski,[6] Gleser[7]) is to compute the best linear approximation to $E(X|W)$. Let $\hat{\mu}$ be the sample mean of the W's, and let $\hat{\sigma}_w^2$ be their sample variance. Define $\hat{\sigma}_x^2 = \hat{\sigma}_w^2 - \sigma_u^2$ and let $\hat{\gamma}_1 = \hat{\sigma}_x^2/(\hat{\sigma}_x^2 + \sigma_u^2)$. Then the best linear approximation is $\hat{E}(X|W) = \hat{\mu}(1 - \hat{\gamma}_1) + \hat{\gamma}_1(W - \hat{\mu})$.

There are other methods for estimating $E(X|W)$, see reference 10, Chapter 3, depending on the available data.

3.2. Simulation Extrapolation

A promising alternative to regression calibration is the simex method of Cook and Stefanski.[9] Assume that the errors U are normally distributed; typically, this assumption is not critical in practice. Simex, which stands for *simulation* and *extrapolation*, relies on computer simulation to generate parameter estimates. The idea behind the simex method is most clearly seen in simple linear regression when the independent variable is subject to measurement error. Suppose the regression model is $E(Y|X) = \alpha_0 + \beta_0 X$. It is well known that the ordinary least squares estimate of the slope from regressing Y on W converges to $\beta_0 \sigma_x^2 (\sigma_x^2 + \sigma_u^2)^{-1}$.

For any fixed $\lambda > 0$, suppose one repeatedly 'adds on', via simulation, additional error with mean zero and variance $\lambda \sigma_u^2$ to W, computes the ordinary least squares slope each time and then takes the average. This simulation estimator consistently estimates

$$g(\lambda) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2(1+\lambda)} \beta_0 .$$

Since, formally at least, $g(-1) = \beta_0$, the idea is to plot $g(\lambda)$ against $\lambda \geqslant 0$, fit a model to this plot, and then extrapolate back to $\lambda = -1$. Cook and Stefanski[9] show that this procedure will yield a consistent estimate of β_0 if one fits the model $g(\lambda) = \gamma_0 + \gamma_1(\gamma_2 + \lambda)^{-1}$.

We provide here a brief description of the simex algorithm. Cook and Stefanski[9] or Carroll *et al.*[10] should be consulted for more details and motivation. The algorithm consists of a *simulation* step, followed by an *extrapolation* step. We review each step in turn.

The simulation step works as follows. Fix $B > 0$ ($B = 100$ is standard), and for $b = 1, \ldots, B$, generate via computer independent standard normal random variables $\{\varepsilon_{ib}\}$. The variance of $W_i + \sigma_u \lambda^{1/2} \varepsilon_{ib}$ given X_i is $\sigma_u^2(1+\lambda)$. For each b, define $\hat{\tau}_b(\lambda)$ as the estimated TLV from regressing Y_i on $W_i + \sigma_u \lambda^{1/2} \varepsilon_{ib}$ for $i = 1, \ldots, n$, that is, from maximizing (3) replacing W_i with $W_i + \sigma_u \lambda^{1/2} \varepsilon_{ib}$. One now forms an estimate of the 'centre' of the terms $\hat{\tau}_b(\lambda)$ for $b = 1, \ldots, B$, for example their average $\hat{\tau}_{\mathrm{A}}(\lambda)$ or median $\hat{\tau}_{\mathrm{M}}(\lambda)$. Cook and Stefanski[9] suggest that one compute $\hat{\tau}_{\mathrm{A}}(\lambda)$ or $\hat{\tau}_{\mathrm{M}}(\lambda)$ on a fixed grid $\Lambda = (\lambda_1, \ldots, \lambda_{\mathrm{L}})$, thus yielding an understanding of the behaviour of the estimators for different amounts of measurement error. They suggest that one fit a parametric model $\mathscr{G}(\Gamma, \lambda)$ in a vector parameter Γ to the $\hat{\tau}_{\mathrm{A}}(\lambda)$'s or $\hat{\tau}_{\mathrm{M}}(\lambda)$'s as a function of the λ's, thus resulting in estimates $\hat{\Gamma}$. Finally, the simex estimator of τ_0 is $\hat{\tau}_{\mathrm{simex}} = \mathscr{G}(\hat{\Gamma}, -1)$. The simex estimator based upon $\hat{\tau}_{\mathrm{A}}(\lambda)$ is called the mean-simex estimator, while that based upon $\hat{\tau}_{\mathrm{M}}(\lambda)$ is called median-simex. Various parametric models are possible, among them linear, quadratic, and, with $\Gamma = (\gamma_0, \gamma_1, \gamma_2)^{\mathrm{t}}$, the non-linear model $\mathscr{G}(\lambda, \Gamma) = \gamma_0 + \gamma_1(\lambda + \gamma_2)^{-1}$. One can fit the models by ordinary unweighted least squares. For the most part, Cook and Stefanski[9] use the non-linear extrapolant function for estimation in standard generalized linear models.

We note here that one would expect *a priori* that the linear or quadratic extrapolants are to be preferred over the non-linear extrapolant when estimating τ_0. To see why, suppose that W and X are jointly normally distributed with X centred to have mean zero. If the regression calibration model holds exactly, then the naive estimator does not estimate τ_0 but rather $\tau_0(\sigma_x^2 + \sigma_u^2)/\sigma_x^2$, which is linear in the measurement error variance σ_u^2. Since the bias is approximately linear in σ_u^2, one corrects for bias by linear extrapolation. Note that in Figure 2 the bias of the naive estimator is almost exactly linear in the case $\tau_0 = -1$, although for the harder case that $\tau_0 = 1$, there is some non-linearity, on the order of a quadratic model.

In the threshold segmented regression model, we use median-simex for the following reason. It is not always the case that the naive estimator of τ_0 even exists, in the sense that it might not occur on the range of the observed W's. We call these 'breakdowns' of the naive estimator. In our experience with simex, breakdowns occur for a non-trivial percentage of the simulation steps, happening more frequently with larger λ. If we use ordinary or mean-simex, the final averages will all be badly affected by these breakdowns, while the median is far more robust. From Carroll *et al.*,[14] using the mean-simex or median-simex for regular problems results in methods with nearly identical behaviour for large values of B and n, so that median-simex would appear to be the method of choice.

3.3. Fully Parametric Methods

Simex and regression calibration are approximate methods, in the sense that they give biases which are of order σ_u^6 and σ_u^4, respectively, as $\sigma_u^2 \rightarrow 0$. Consistency, as opposed to approximate consistency, can be obtained, but at the cost of having to make parametric assumptions about the distribution of X. There is a longstanding reluctance in the measurement error literature to making such assumptions, because of the fear of what might happen if the assumptions are wrong. Much of this reluctance, however, is based upon the fact that, in linear regression, the functional (semi-parametric) and structural (parametric) methods differ hardly at all if X is nearly normally distributed.

In a new and hard problem such as our threshold segmented regression, we have already seen that the experience of linear regression on bias calculations is very nearly worthless. It is not too far of a leap to suppose then that efficiency calculations from linear regression will be of similar limited utility. We will explore this issue, and begin by describing the likelihood function.

Let the density function of W given X be $f(w|x, \sigma_u^2, \eta_{01})$, the density of X be $f(x|\eta_{02})$, and the density for Y given X be $f\{y|\alpha_0 + \beta_0(x-\tau_0)_+, \xi_0\}$. Let $\theta_0 = (\alpha_0, \beta_0, \tau_0, \xi_0)$, $\Theta_0 = (\theta_0, \eta_{01}, \eta_{02}, \sigma_u^2)$, $\theta = (\alpha, \beta, \tau, \xi)$ and $\Theta = (\theta, \eta_1, \eta_2, \sigma^2)$. Then the likelihood function of the observed (Y, W) data is

$$f(Y, W|\Theta) = \int f\{Y|\alpha + \beta(x-\tau)_+, \xi\} f(W|x, \sigma^2, \eta_1) f(x|\eta_2) \mathrm{d}x.$$

Typically, the likelihood function is not concave, for example, in linear regression with no measurement error. However, with measurement error the likelihood is differentiable (in contrast to the no-error case), and for fixed values of τ, the likelihood in the remaining free parameters usually is well-behaved, so that one can compute the profile likelihood as a function of τ, and then use search/grid techniques to maximize this function.

4. BIASES OF MEASUREMENT ERROR METHODS

In this section, we discuss the limiting values of the different approximate measurement error methods, when W and X are jointly normally distributed, and X has a standard normal distribution.

In Figure 2 we exhibit the bias of the regression calibration estimator. One can see that in the case $\tau_0 = -1$, the regression calibration estimator has almost no asymptotic bias, but for $\tau_0 = 1$, the bias is even larger than that of the naive estimator. This somewhat surprising behaviour can be explained as follows. If X is normally distributed with mean μ and variance σ_x^2, then $E(X|W) = \mu(1-\gamma_1) + \gamma_1 W$, where γ_1 is the reliability (see Section 2). The regression calibration estimator can be calculated by replacing W_i with $\mu(1-\gamma_1) + \gamma_1 W_i$ in (3). We conclude by some algebra that the regression calibration estimator is related to the naive estimator as follows:

$$\hat{\tau}_{\mathrm{rc}} = \gamma_1 \hat{\tau}_{\mathrm{naive}} + \mu(1-\gamma_1).$$

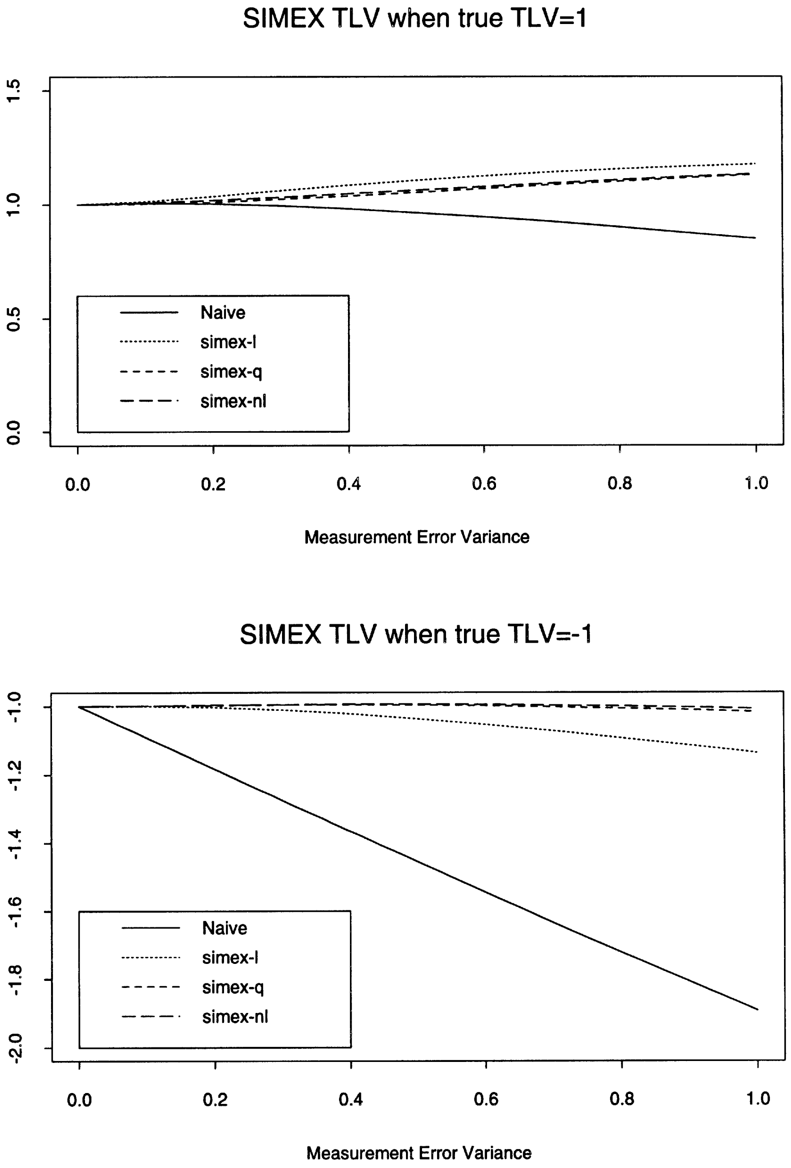


Figure 3. For given amounts of measurement error, the actual limiting value of the naive and simex estimators, for $\tau_0 = \pm 1$. The extensions ‘ – l’, ‘ – q’ and ‘ – nl’ refer to the linear, quadratic and non-linear extrapolants, respectively. Here $\alpha_0 = 0$, $\beta_0 = 2$ and $\sigma_x^2 = 1$

In Figure 2, the calculations were done with $\mu = 0$, so that the regression calibration estimator is always smaller in absolute value than the naive estimator. This works fine if the naive estimator is too large ($\tau_0 = -1$), but is disastrous when the naive estimator is too small ($\tau_0 = 1$).

In effect then, we conclude that the regression calibration estimator is not generally useful in reducing bias in the threshold segmented regression model, although when the TLV τ_0 is smaller than the mean of X, there is some possibility that regression calibration will be acceptable.

The limiting values of the simex TLV estimators are given in Figure 3. With either the linear, quadratic or non-linear extrapolants, the simex estimator is seen to provide estimates which are much closer to the actual value of τ_0 in large samples. Despite our previous analysis, the linear extrapolant has slightly larger limiting bias than the non-linear extrapolant, although the bias is small; the quadratic extrapolant has fairly small bias.

5. LINEAR SIMULATIONS

We performed a small simulation study for the linear regression model with Y given X being normally distributed with variance $\xi_0 = 1$. The measurement error U was normally distributed with mean zero and variance $\sigma_u^2 = 0{\cdot}6$. The predictor X was generated to have mean $\mu = 0$, variance $\sigma_x^2 = 1$, and to follow either the normal, Laplace or a mixture normal distribution with means $(\mu_1, \mu_2) = (-0{\cdot}75, 1{\cdot}125)$, variance $(\sigma_1^2, \sigma_2^2) = (0{\cdot}15625, 0{\cdot}15625)$ and mixing percentage $p = 0{\cdot}6$; in this mixture case, we used $\sigma_u^2 = 0{\cdot}015$. We set $\alpha_0 = 0$, $\beta_0 = 2$, and $\tau_0 = \pm 1$. The sample sizes were $n = 250$ and $n = 500$. There were 200 repetitions of the experiment.

The estimators used were the naive estimator, the simex estimators with linear, quadratic and non-linear extrapolants, the maximum likelihood estimator, and the regression calibration estimator using the best linear approximation to $E(X|W)$. We fixed the last two estimators to be calculated assuming normality in order to obtain some understanding of the effect of model misspecification, especially for the mixture normal distribution where X is significantly non-normally distributed. Because of possible breakdowns of the estimator, we constrained the estimators of τ_0 to the interval $[-2{\cdot}5, 2{\cdot}5]$, which covers most of the range of the X-values.

The results for the estimation of τ_0 are displayed in Tables I, II and III. The sample means and medians of the simulations were similar, so that only the former are reported. We also computed

Table I. Simulation with X and W jointly normally distributed. Estimators are RC (regression calibration) and ML (maximum likelihood). SIM refers to the simex, L Q and NL mean linear, quadratic and non-linear extrapolants. MAD is the median absolute deviation from the median; RMSE is the square root of the mean squared error

	NAIVE	RC	ML	SIML	SIMQ	SIMNL
$\tau = -1$, $n = 250$						
Mean	−1·558	−0·965	−1·031	−1·236	−0·885	−0·719
MAD	0·566	0·185	0·187	0·395	0·485	0·985
RMSE	0·650	0·227	0·347	0·498	0·600	1·355
$\tau = 1$, $n = 250$						
Mean	0·994	0·621	0·975	1·157	1·111	0·965
MAD	0·450	0·437	0·144	0·508	0·701	0·661
RMSE	0·580	0·526	0·181	0·647	0·879	0·909
$\tau = -1$, $n = 500$						
Mean	−1·548	−0·965	−1·013	−1·167	−0·774	−0·423
MAD	0·548	0·118	0·063	0·262	0·403	0·949
RMSE	0·589	0·148	0·079	0·335	0·510	1·405
$\tau = 1$, $n = 500$						
Mean	1·033	0·643	1·004	1·199	1·191	1·102
MAD	0·295	0·379	0·099	0·352	0·536	0·526
RMSE	0·390	0·433	0·133	0·461	0·672	0·811

Table II. Simulation when X has a Laplace distribution. Estimators are RC (regression calibration) and ML (maximum likelihood), both incorrectly assuming normality. SIM refers to the simex estimators: L, Q and NL mean linear, quadratic and non-linear extrapolants. MAD is the median absolute deviation from the median; RMSE is the square root of the mean squared error

	NAIVE	RC	ML	SIML	SIMQ	SIMNL
$\tau = -1$, $n = 250$						
Mean	− 1·510	− 0·929	− 1·010	− 1·349	− 0·987	− 1·139
MAD	0·616	0·254	0·199	0·566	0·665	0·949
RMSE	0·717	0·321	0·309	0·679	0·843	1·228
$\tau = 1$, $n = 250$						
Mean	1·493	0·925	1·150	1·428	1·101	1·310
MAD	0·544	0·247	0·181	0·540	0·621	0·835
RMSE	0·668	0·305	0·227	0·673	0·770	1·056
$\tau = -1$, $n = 500$						
Mean	− 1·539	− 0·962	− 1·048	− 1·304	− 1·033	− 0·954
MAD	0·577	0·188	0·102	0·450	0·526	0·870
RMSE	0·656	0·243	0·153	0·548	0·665	1·234
$\tau = 1$, $n = 500$						
Mean	1·332	0·824	1·058	1·240	0·976	1·034
MAD	0·508	0·244	0·238	0·441	0·540	0·854
RMSE	0·590	0·356	0·347	0·547	0·680	1·134

such summary statistics as the 90th and 10th percentiles of the distributions, but they do not add any significant information to the results, and are not reported.

The results on biases are roughly in accord with our asymptotic calculations for the case that X and U are normally distributed, as well as with results computed in Section 6 when X has a mixture normal distribution. The simex estimators with linear extrapolant have the overall best bias behaviour. The normal theory maximum likelihood estimator has good bias behaviour except at the mixture normal, which is to be expected since this is a case that the normal model is a (perhaps almost ridiculously) poor fit.

The variability results show the bias-variance trade-off, namely that decreased bias is bought at the cost of increased variance. Almost without exception, the simex-type estimators are more variable than the others, so that in terms of mean squared error, the regression calibration and maximum likelihood estimators dominate, while the naive estimator is competitive. Typically, the simex non-linear extrapolants are disasters, in contrast to their behaviour in the ordinary linear and generalized linear models.

The success of the normal distribution maximum likelihood estimator at the normal and Laplace models (Table II) is striking. It has very little bias and is vastly less variable than the simex estimators. The fact of the matter is that parametric maximum likelihood is vastly superior to the functional (semi-parametric) methods as long as the model for X is not too badly specified.

We also investigated the estimated standard errors of the different estimators (Table IV) in the normal–X case using the delta method and the jack-knife method proposed by Stefanski and Cook.[15] It turned out that the estimated standard errors were accurate for the likelihood estimator, but especially inaccurate for the simex estimators. In simulations not reported here, we also compared the bootstrap standard error estimates with the delta method for the naive

Table III. Simulation with U normally distributed and X having a mixture normal density with means − 0·75, 1·125, variances 0·15625 and 0·15625 and mixing proportion 0·6. The measurement error variance equals 0·15. Estimators are RC (regression calibration) and ML (maximum likelihood), both incorrectly assuming normality. SIM refers to the simex estimators: L, Q and NL mean linear, quadratic and nonlinear extrapolants. MAD is the median absolute deviation from the median; RMSE is the square root of the mean squared error

	NAIVE	RC	ML	SIML	SIMQ	SIMNL
$\tau = -1$, $n = 250$						
Mean	− 0·882	− 0·767	− 0·790	− 0·813	− 0·895	− 0·778
MAD	0·137	0·235	0·210	0·198	0·203	0·262
RMSE	0·162	0·253	0·226	0·226	0·252	0·518
$\tau = 1$, $n = 250$						
Mean	0·620	0·537	0·767	0·827	1·020	0·684
MAD	0·404	0·468	0·248	0·292	0·387	0·867
RMSE	0·528	0·562	0·292	0·439	0·594	1·354
$\tau = -1$, $n = 500$						
Mean	− 0·890	− 0·774	− 0·798	− 0·820	− 0·907	− 0·858
MAD	0·118	0·226	0·202	0·182	0·152	0·160
RMSE	0·138	0·237	0·211	0·204	0·184	0·202
$\tau = 1$, $n = 500$						
Mean	0·635	0·552	0·772	0·854	1·066	1·067
MAD	0·372	0·449	0·229	0·210	0·273	0·710
RMSE	0·421	0·483	0·260	0·277	0·368	1·123

Table IV. Simulation with X and W jointly normally distributed. Estimators are RC (regression calibration) and ML (maximum likelihood). SIM refers to the simex estimators: L Q and NL mean linear quadratic and non-linear extrapolant. The first row is the simulated standard error of the estimator, while the second row is the average estimated standard error

	NAIVE	RC	ML	SIML	SIMQ	SIMNL
$\tau = -1$, $n = 500$						
Simulated standard error	0·216	0·144	0·079	0·291	0·458	1·281
Mean estimated standard error	0·167	0·104	0·079	0·758	0·758	0·758

estimator in linear and logistic regression, finding that bootstrap standard errors were much more accurate.

6. BRONCHITIS EXAMPLE

In occupational medicine one important issue is the assessment of the health hazard of specific harmful substances in a working area. One concept of modelling is to assume that there is a threshold concentration, called the threshold limiting value (TLV) under which there is no risk due to the substance. Estimating the TLV is of particular interest in the industrial workplace. We consider here the specific problem of estimating the TLV in a dust burdened mechanical engineering plant in Munich.

The regressor variable X is the logarithm of 1·0 plus the average dust concentration in the working area over the period of time in question. In addition, the duration of exposure Z_1 and smoking status Z_2 are also measured. Following Ulm,[12] we based our analysis upon the segmented logistic model

$$\text{pr}(Y = 1|X, Z) = H\{\alpha_0 + \beta_0(X - \tau_0)_+ + \eta_{01}Z_1 + \eta_{02}Z_2\}. \tag{5}$$

It is impossible to measure X exactly, and instead sample dust concentrations were obtained several times between 1960 and 1977. The resulting measurements are W.

There were 1246 observations: 23 per cent of the workers reported chronic bronchitis, and 74 per cent were smokers. Measured dust concentration had a mean of 1·07 and a standard deviation of 0·72. The durations were effectively independent of concentrations, with correlation 0·093, compare with Ulm's[12] Figure 3. Smoking status is also effectively independent of dust concentration, with the smokers having mean concentration 1·068, and the non-smokers having mean 1·083. Thus, in this example, for likelihood calculations we will treat the Z's as if they were independent of X, and we will condition on the Z's.

A preliminary segmented regression analysis ignoring measurement error suggested an estimated TLV $\hat{\tau} = 1{\cdot}27$. We will call this the naive TLV. In Figure 4 we show the results of such an analysis when regressing bronchitis only on dust concentration. A generalized additive model fit using Splus suggests a threshold in the neighbourhood of the estimated value. Note also that an ordinary logistic regression is sufficiently different from the generalized additive model fit to suggest a changepoint. Obviously, other models can provide a good fit to these error-prone data, for example, a quadratic model.

In Figure 4 we also plot a kernel density estimate of the observed dust concentrations, with a Gaussian kernel and bandwidth equal to 0·25. The dust concentrations appear to be strongly bimodal, with almost no observations in the vicinity of the naive TLV. For purposes of illustration we fit a two-population mixture normal model to the data, that is, one having density function

$$f_W(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p) = (p/\sigma_1)\phi\{(w - \mu_1)/\sigma_1\} + \{(1 - p)/\sigma_2\}\phi\{(w - \mu_2)/\sigma_2\}. \tag{6}$$

The maximum likelihood estimates were $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p}) = (0{\cdot}520, 1{\cdot}927, 0{\cdot}236^2, 0{\cdot}215^2, 0{\cdot}607)$.

We computed the theoretical, asymptotic bias of the naive TLV estimator and that of the simex estimators with linear and quadratic extrapolant functions, in a special case aimed to approximate the Munich data. For the underlying distribution of X, we used a mixture normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p) = (0{\cdot}45, 1{\cdot}90, 0{\cdot}03, 0{\cdot}03, 0{\cdot}60)$, which has mean 1·03 and variance 0·535. We added small amounts of measurement error with variance σ_u^2 ranging from 0·0 to 0·04; at the extreme end of the scale, we have a situation that X comes from two subpopulations, both of which are estimated with large measurement error. We used $\tau_0 = 0{\cdot}4$ and $\tau_0 = 1{\cdot}6$, the latter reflecting the Munich data.

The biases are exhibited in Figure 5. The naive estimator and the simex estimator with linear extrapolant are both considerably more biased than the simex estimator with quadratic extrapolant. Note that the simex estimator with non-linear extrapolant has very poor bias behaviour.

As far as we can ascertain, there is no data available to fit an error model relating W to X. In the absence of such information, for illustration we used an additive error model $W = X + U$, and we assumed that $\sigma_u^2 = 0{\cdot}035$, making $\hat{\sigma}_x^2 = 0{\cdot}489$. While the error variance σ_u^2 is rather small relative to the marginal variance of X, it is fairly large relative to the variance of each of the components of the mixture.

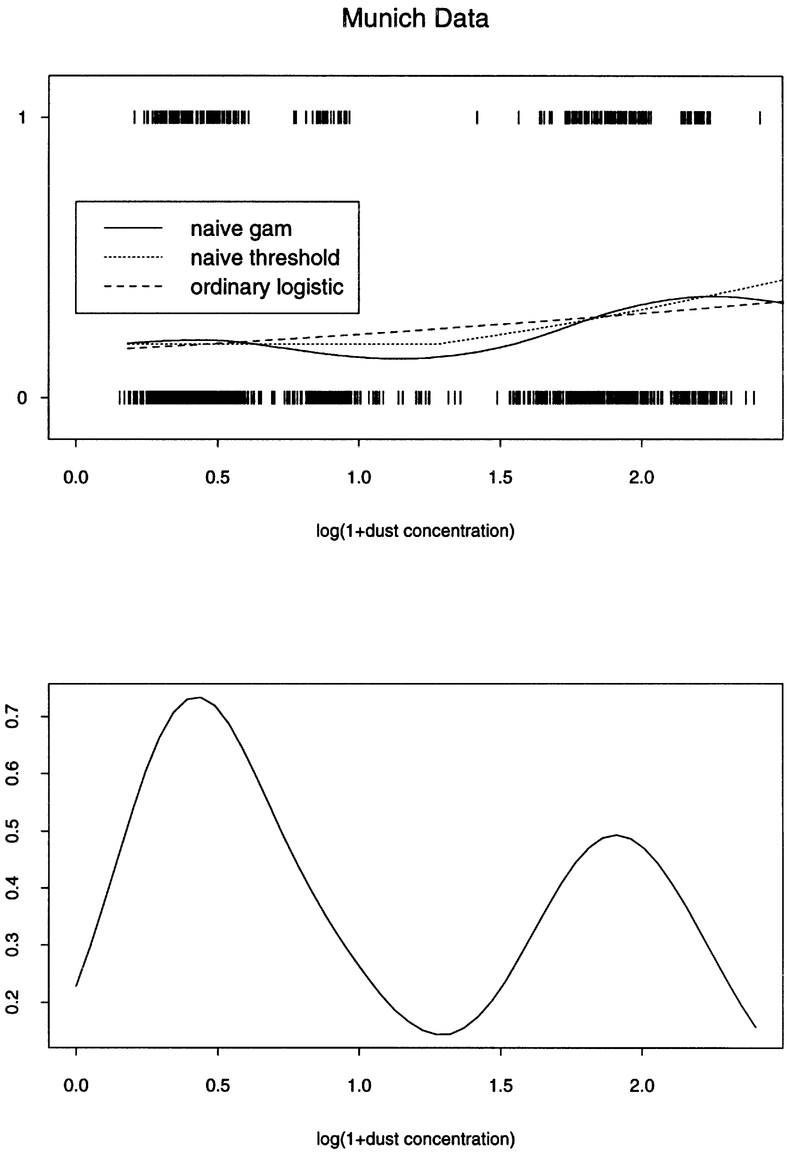
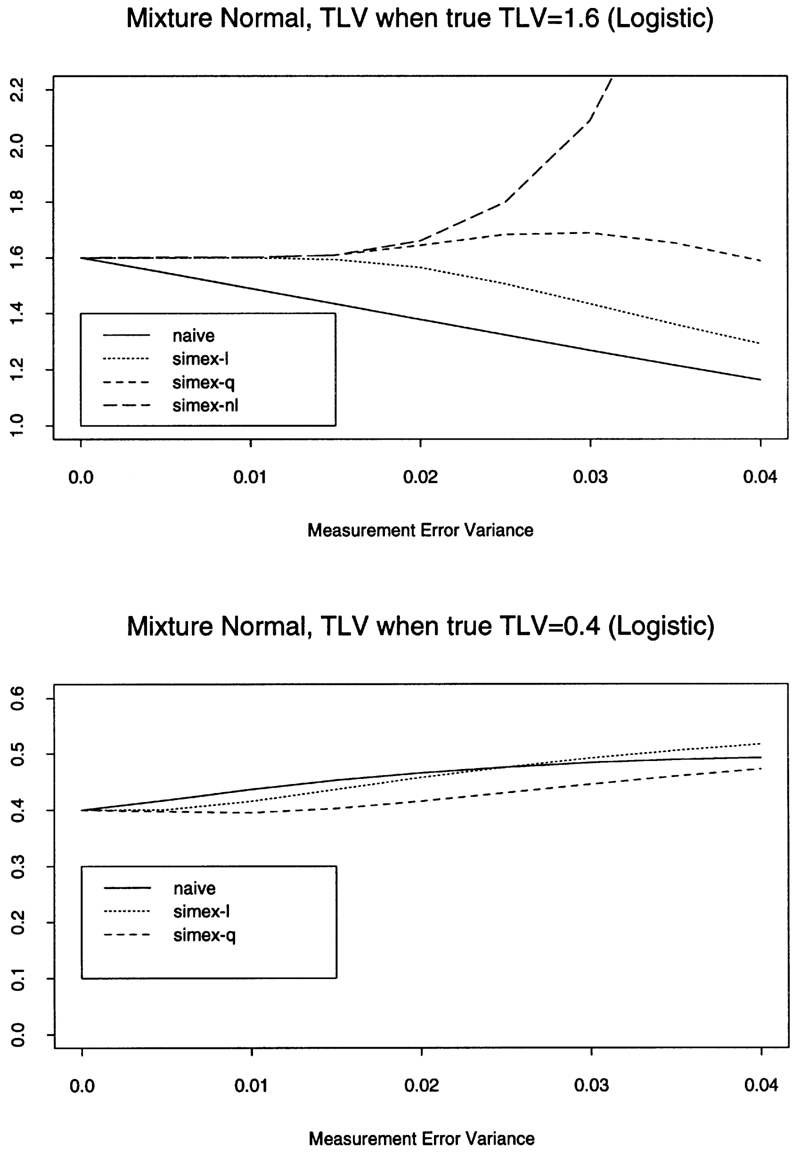


Figure 4. The Munich plant. The top figure are various binary regression models of bronchitis incidence on log(1 + dust concentration). The fitting methods were (i) gam (generalized additive model), (ii) segmented logistic regression and ordinary logistic regression. The bottom figure is a kernel density estimate of the observed concentrations

In context, the additive error model is clearly not *exactly* correct, for two reasons. First, there is one observation which reports zero concentration, although the actual concentration may be non-zero. This observation is not particularly informative for estimating τ_0, and so pretending that it arises from an additive error model is probably benign. A second problem is that the error in reported dust concentration may well be heteroscedastic, although our use of the log transformation has probably corrected for some of this problem.


Mixture Normal, TLV when true TLV=1.6 (Logistic)

Mixture Normal, TLV when true TLV=0.4 (Logistic)

Figure 5. Limit of estimated TLV in a mixture normal model for the naive estimator and various simex estimators. See text for details

We now turn to the analysis of the data, see Table V for details. The likelihood estimator and the regression calibration estimator were computed assuming that X has a mixture normal distribution. We give the results for the linear quadratic and non-linear extrapolants for simex. Instead of a full-likelihood analysis, for computational convenience we used a pseudo-likelihood analysis, in which the parameters of the distribution of X were estimated from the W's only.

The pseudo-maximum likelihood estimate and the regression calibration estimate suggest a large correction to the naive estimate. That the naive estimator should be biased downwards is

Table V. Estimated TLV in the Munich data, when $\sigma_u^2 = 0{\cdot}035$

Method	TLV-τ_0	Nominal standard error	Bootstrap standard error
Naive	1·27	0·41	0·24
Pseudo-MLE	1·76	0·17	0·21
Regression calibration	1·75	0·12	0·19
Simex: linear	1·37	0·23	0·23
Simex: quadratic	1·40	0·23	0·34
Simex: non-linear	1·40	0·23	0·86

clearly seen in Figure 5, where the true TLV is 1·6 while for this amount of measurement error the naive estimator converges to 1·25 approximately, figures which are closely mirrored in the example. The three simex estimates all give smaller corrections, which is perplexing in the light of Figure 5, which suggests that for this amount of measurement error they are estimating very different quantities. Further simulations described at the end of this section explore this issue in more detail. The assumption that X has a mixture normal distribution appears to be playing a crucial role here, providing considerable information about the TLV.

It is important to note that with the mixture normal distribution for X, the function $E(X|W)$ is highly non-linear. Thus the simple relationship described in Section 3.1 between the naive estimator and the regression calibration estimator does not hold here.

The standard errors of the naive estimator differed greatly, with the bootstrap standard error being much smaller. We performed a simulation in the logistic case with normally distributed X, and found that on average the delta-method standard errors were 25 per cent too large, consistent with this example. The delta-method and bootstrap standard errors for the pseudo-maximum likelihood estimate are reasonably close, also in line with our linear simulations. The delta-method and bootstrap standard errors for the simex estimator with linear extrapolant are very nearly the same, but the standard error of Stefanski and Cook[15] is independent of the extrapolant used, and is thus probably too small for the quadratic and non-linear extrapolant.

Though our main interest was in the estimation of the TLV, it should be mentioned that the estimation of the slope parameter is also corrected in a different way. In the example the naive estimate was 0·85, the simex was 0·98 (linear), 1·07 (quadratic), while the pseudo-maximum likelihood estimate was 3·24. In general the slope estimator behaves similarly to that in other generalized linear models and its absolute value has positive correlation to the TLV. The naive estimate of the constant term (− 3·0) remains unchanged by the different methods of correction.

Finally, we have attempted to gain a better understanding of why the simex and mixture-normal maximum likelihood estimators give such different answers with these data. We have performed a series of simulations to investigate this matter, first when X is modelled correctly as a mixture normal. For example, we set $(\alpha_0, \beta_0, \tau_0) = (-1{\cdot}54, 3{\cdot}24, 1{\cdot}76)$ and let X have a mixture-normal distribution with mixing proportion 0·607, means $(0{\cdot}520, 1{\cdot}927)$ and variances $(0{\cdot}236^2 - \sigma_u^2, 0{\cdot}215^2 - \sigma_u^2)$, where $\sigma_u^2 = 0{\cdot}035$. These parameter values basically are the same as obtained by the mixture-normal maximum likelihood analysis. By simulating from this distribution, we hoped to learn whether *for this model* the simex estimator could reasonably be biased for non-infinite sample sizes. Based on a sample of size $n = 1000$, with 200 simulated experiments, the mixture maximum likelihood estimated had mean 1·54 and median 1·83, with standard deviation 0·75. In contrast, the simex estimator with quadratic extrapolant had mean 1·19, median 1·01 and standard deviation 0·79. There are two points to note here: (i) simex has the sort of serious bias we observed in the data; (ii) however, the standard errors of both estimators were much larger than estimated in the data.

One has to be very careful of distributional assumptions on latent variables. Indeed, further study of Figure 5 suggests that a mixture of two normals may not fully reflect the real distribution of the dust concentrations. The first 'subpopulation' in Figure 5 clearly has some skewness, more in fact than can be explained by the model we have fit. For instance, one might hypothesize that the data actually reflect a mixture of *three* normals, or perhaps two gammas. We tested the former by generating X to have a distribution with means $(0{\cdot}40, 0{\cdot}95, 1{\cdot}90)$, variances $(0{\cdot}20^2 - \sigma_u^2, 0{\cdot}20^2 - \sigma_u^2, 0{\cdot}255^2 - \sigma_u^2)$ and mixing percentages $(0{\cdot}450, 0{\cdot}157, 0{\cdot}393)$. The variance of the third component is probably too large for the data, but this model at least exhibits the skewness in the first 'subpopulation'. We found here that when we assumed a mixture of two normals, the resulting incorrectly specified maximum likelihood estimate was only marginally less biased than the simex estimator, and had 50 per cent greater variance. This simulation is a sobering illustration of the need for care in parametric modelling in hard non-linear problems.

While we have assumed here an additive error model, an alternative analysis based on the Berkson model should also be considered. The Berkson model says that the recorded dust concentration is a convenient setting for and not an error-prone version of true dust concentration. In symbols, this means that $X = W + U_{\mathrm{B}}$, where U_{B} has some convenient distribution with mean zero and variance $\sigma_{u\mathrm{B}}^2$. Under the Berkson model, the 'naive' and regression calibration analyses coincide, since $E(X|W) = W$. One reason for considering the Berkson model in this context is the fact that the data themselves are subject to some grouping, which suggests the possibility that all individuals with certain common characteristics were assigned a common concentration; this is precisely the type of situation where the Berkson model is often applied.

An empirical reason for considering the Berkson model here arises from the results of further analysis of the distribution of W itself. Various kernel density estimation using empirically chosen bandwidths, as well as Bayesian analysis of the mixture model, suggests that the distribution of W is made up of at least three components (in this case, two for the first mode), and the component variances for the left-hand mode are considerably smaller than our illustrative measurement error variance $\sigma_u^2 = 0{\cdot}035$. It is possible that these empirically driven density estimators are being misled by the grouping in the data to find additional components in the distribution of W with small variability. However, if the small-variance components are real, then either the additive error model holds but with much smaller error variance than chosen here in our illustration, or the additive error model fails, in which case the Berkson model is an appealing alternative.

We have analysed the Berkson model assuming normality of the Berkson errors, with variances ranging from $\sigma_{u\mathrm{B}}^2 = 0{\cdot}0$ to $\sigma_{u\mathrm{B}}^2 = 0{\cdot}2$. The estimate of the changepoint τ_0 varied from $\hat{\tau} = 1{\cdot}27$ when $\sigma_{u\mathrm{B}}^2 = 0{\cdot}0$ to $\hat{\tau} = 1{\cdot}40$ when $\sigma_{u\mathrm{B}}^2 = 0{\cdot}2$. Likelihood ratio confidence intervals were somewhat more sensitive to the size of the measurement error. We restricted τ to the interval $[0{\cdot}50, 2{\cdot}40]$, and found likelihood ratio confidence intervals of $[0{\cdot}65, 1{\cdot}70]$ when $\sigma_{u\mathrm{B}}^2 = 0{\cdot}0$, to $[0{\cdot}50, 2{\cdot}00]$ when $\sigma_{u\mathrm{B}}^2 = 0{\cdot}2$.

7. CONCLUSIONS

Our analytical calculations, simulations and the practical example have shown that ignoring the measurement error when estimating the (changepoint) TLV τ_0 can yield a severe bias. Although this is a known fact in other models we observe here biases different from that usually encountered in generalized linear models, see Figure 2. In most cases the TLV is biased to a lower value, that is, ignoring measurement error leads to underestimation of the threshold value.

It is especially important to state that one of the standard methods of correcting for measurement error, regression calibration, differs in two ways from what one might expect: (i) it can be less variable than the naive estimator; and (ii) it can be even more biased than the naive estimator.

The recently proposed simex estimators seem to be a promising alternative for the researcher to cope with the problem of estimating the TLV in the presence of measurement error in the predictor. These estimators are less biased than the naive and regression calibration estimators. We recommend the linear extrapolant, which has the best mean squared error behaviour or the quadratic extrapolant having the best bias behaviour in the examples considered. The non-linear extrapolant for simex, which is the method of choice in standard generalized linear models, should be avoided, since it can have very poor bias behaviour, see Figure 5.

Regression calibration and simex are functional (semi-parametric) methods, which make no explicit assumptions about the distribution of the unobservable X. While these two methods have similar behaviour in linear and logistic regression, one surprising finding of our study is that they behave very differently for estimating the TLV.

A second surprising finding concerns the differences between functional and structural (parametric) methods. In linear and logistic regression the differences are typically not great. For estimating the TLV, major differences are observed. This is one of the first measurement error problems where the distinction between functional and structural modelling is critical. The maximum likelihood estimators with correct or at least plausible models for the distribution of X have extremely strong performance, with bias of the same order as the measurement error model estimate, but much smaller variability. Of course, the maximum likelihood estimator does have significant bias when the distribution of X is badly misspecified. The major difficulty with a likelihood analysis here, besides the need for a correct model, is that one has to model the distribution of X given Z. In our example, X and Z were approximately independent, but in other situations this need not be the case.

The finding that the correctly specified maximum likelihood estimator is so much less variable than the standard correction methods comes as a surprise. It appears that the specification of a marginal distribution for X carries considerable information about the changepoint. Structural modelling ought not be a neglected tool in measurement error models.

Summarizing, the estimation of the TLV in threshold models in the presence of additive measurement error is a practically relevant problem, which cannot be treated like common estimation problems in error in variables models. Regression calibration cannot be recommended in general, while the other semi-parametric method – simex – seems to be a good choice. If, however, the distribution of the predictor X and the dependence structure of X and the other covariates can be modelled, this can be used to obtain parametric estimators of the TLV with a much better performance.

APPENDIX

Exact Limit of the Naive Estimator

First consider the linear threshold segmented regression model with mean $E(Y|X) = \alpha_0 + \beta_0(X - \tau_0)_+$ and variance ξ_0. Write $\theta_0 = (\alpha_0, \beta_0, \tau_0)$. The naive estimator converges to the minimizer $(\alpha_*, \beta_*, \tau_*)$ of the function

$$E_{\theta_0}\{Y - \alpha - \beta(W - \tau)_+\}^2$$

where the subscript θ_0 indicates that expectations are taken at the parameter θ_0. Because the error about the line is independent of X and W, for the purposes of estimating $(\alpha_0, \beta_0, \tau_0)$, this problem is the same as finding the minimum $(\alpha_*, \beta_*, \tau_*)$ of the function

$$E_{\theta_0}\{(\alpha_0 - \alpha) + \beta_0(X - \tau_0)_+ - \beta(W - \tau)_+\}^2. \tag{7}$$

If (7) can be computed as a function of (α, β, τ), then it can be minimized for any parameter setting $(\alpha_0, \beta_0, \tau_0, \mu, \sigma_x^2, \sigma_u^2)$.

These calculations can be performed explicitly when X is normally distributed with mean μ and variance σ_x^2 and W given X is normally distributed with mean X and variance σ_u^2. Numerical integration is required only to evaluate the bivariate normal integral, which is available in many places, including in the system GAUSS.[17] The exact formulae are long and their derivation is tedious; details are available from the first author. For this normal case, write (7) as $\mathscr{K}(\theta_0, \theta, \sigma_u^2, \mu, \sigma_x^2)$, where $\theta = (\alpha, \beta, \tau)$.

The bias calculations can also be extended to the case that X has a mixture normal distribution with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and mixing proportions (p_1, p_2) (which sum to 1·0), in which case (7) becomes

$$\sum_{k=1}^{2} p_k \mathscr{K}(\theta_0, \theta, \sigma_u^2, \mu_k, \sigma_k^2).$$

In general, the naive estimator of θ_0 converges to θ_*, which is the maximizer of the expected log-likelihood using W instead of X, namely $E_{\Theta_0}\mathscr{L}(Y, \alpha + \beta(W - \tau)_+, \xi)$, where the expectation is taken at the population parameter Θ_0. Taking derivatives, we see that θ_* satisfies

$$0 = E_{\Theta_0}[(\partial/\partial\theta)\ \mathscr{L}\{Y, \alpha_* + \beta_*(W - \tau_*)_+, \xi_*\}] = \mathscr{H}(\theta_0, \eta_{01}, \eta_{02}, \sigma_u^2, \theta_*) \tag{8}$$

say. For linear normal, linear logistic and log-linear Poisson regression, this takes a simple form. The part of (8) relevant for estimation of $(\alpha_0, \beta_0, \tau_0)$ can be written as

$$E\left([Y - H\{\alpha_* + \beta_*(W - \tau_*)_+\}] \begin{Bmatrix} 1 \\ (W - \tau_*)_+ \\ -\beta_* I(W > \tau_*) \end{Bmatrix} \right) = 0.$$

Conditioning on W and X gives

$$\begin{aligned} &\mathscr{H}(\theta_0, \eta_{01}, \eta_{02}, \sigma_u^2, \theta_*) \\ &\quad = E_{\Theta_0}\left([H\{\alpha_0 + \beta_0(X - \tau_0)_+\} - H\{\alpha_* + \beta_*(W - \tau_*)_+\}] \begin{Bmatrix} 1 \\ (W - \tau_*)_+ \\ -\beta_* I(W > \tau_*) \end{Bmatrix} \right). \end{aligned} \tag{9}$$

We use (9) in computing the limit of the naive estimator for probit regression.

For the logistic model, the limiting value of the naive estimator is the solution $(\alpha_*, \beta_*, \tau_*)$ of (9), with $H(\cdot)$ being the logistic distribution function. The logistic distribution function $H(v)$ is very closely approximated by $\Phi(v/1{\cdot}7)$, and we make this substitution into (9), which now becomes computable explicitly as a function of bivariate and trivariate normal probabilities. An alternative is to use the somewhat finer approximation of Monahan and Stefanski,[16] which is of the form $\sum_{j=1}^{8} p_j \Phi(s_j v)$ for tabulated constants (p_j, s_j). The details in the case that X is normally distributed are available from the first author, while the implementation in the mixture normal case follows the same idea as previously discussed.

Standard Errors

Standard error estimates can be computed via delta method expansions. Because of the threshold, the objective function for the naive estimator is not differentiable, so that the delta method does not enjoy its usual Taylor series justification (the bias calculations are not affected by the lack of differentiability). However, the answers obtained by the delta method are correct in the linear case with no measurement error (Feder[11]), as long as X has a continuous density function. Without measurement error and assuming that the parameter estimates have the usual $n^{1/2}$-rate of convergence, one can justify the formal calculations for the logistic and Poisson cases. With measurement error, and again assuming the usual rate of convergence to the limiting values, one can also use the delta method to compute approximate standard errors.

ACKNOWLEDGEMENTS

Küchenhoff's research was initiated during a visit to Texas A & M University. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030). We wish to thank a referee for the valuable comments.

REFERENCES

1. Shaban, S. A. 'Change point problem and two-phase regression: an annotated bibliography', *International Statistical Review*, **48**, 83–93 (1980).
2. Seber, G. A. F. and Wild, C. J. *Nonlinear Regression*, Wiley, New York, 1989.
3. Gbur, E. E. and Dahm, P. F. 'Estimation of the linear-linear segmented regression model in the presence of measurement error', *Communications in Statistics – Theory and Methods*, **14**, 809–826 (1985).
4. Grimshaw, S. D. 'Estimation of the linear-plateau segmented regression model in the presence of measurement error', *Communications in Statistics – Theory and Methods*, **21**, 2399–2413 (1992).
5. Rosner, B., Willett, W. C. and Spiegelman, D. 'Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error', *Statistics in Medicine*, **8**, 1051–1070 (1989).
6. Carroll, R. J. and Stefanski, L. A. 'Approximate quasilikelihood estimation in models with surrogate predictors', *Journal of the American Statistical Association*, **85**, 652–663 (1990).
7. Gleser, L. J. 'Improvements of the naive approach to estimation in non-linear errors-in-variables regression models', *in* Brown, P. J. and Fuller, W. A. (eds), *Statistical Analysis of Measurement Error Models and Application*, American Mathematics Society, Providence, 1990.
8. Pierce, D. A., Stram, D. O., Vaeth, M. and Schafer, D. W. 'The errors-in-variables problem: Considerations provided by radiation dose-response analyses of the A-bomb survivor data', *Journal of the American Statistical Association*, **87**, 351–359 (1992).
9. Cook, J. R. and Stefanski, L. A. 'Simulation–extrapolation estimation in parametric measurements error models', *Journal of the American Statistical Association*, **89**, 1314–1328 (1994).
10. Carroll, R. J., Ruppert, D. and Stefanski, L. A. *Measurement Error in Nonlinear Models*, Chapman and Hall, New York, 1995.
11. Feder, P. I. 'On asymptotic distribution theory in segmented regression problems – identified case', *Annals of Statistics*, **3**, 49–83 (1975).
12. Ulm, K. 'A statistical method for assessing a threshold in epidemiological studies', *Statistics in Medicine*, **10**, 341–349 (1991).
13. Stasinopoulous, D. M. and Rigby, R. A. 'Detecting break points in generalized linear models', *Computational Statistics and Data Analysis*, **13**, 461–471 (1992).
14. Carroll, R. J., Küchenhoff, H., Lombard, F. and Stefanski, L. A. 'Asymptotics for the SIMEX estimator in structural measurement error models', *Journal of the American Statistical Association*, **91**, 242–250 (1996).
15. Stefanski, L. A. and Cook, J. R. 'Simulation-extrapolation: the measurement error jackknife', *Journal of the American Statistical Association*, **90**, 1247–1256 (1995).
16. Monahan, J. and Stefanski, L. A. 'Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral', in Balakrishnan, N. (ed), *Handbook of the Logistic Distribution*, Marcel Dekker, New York, 1991.
17. GAUSS. Version 3·0. Aptech Systems, Inc. Maple Valley, 1992.