Statistica Sinica **7**(1997), 497-525

ON SEGMENTED MULTIVARIATE REGRESSION

Jian Liu, Shiying Wu* and James V. Zidek

*University of British Columbia and *Statistics Canada*

Abstract: This paper concerns segmented multivariate regression models, models which have different linear forms in different subdomains of the domain of an independent variable. Without knowing that number and their boundaries, we first estimate the number of these subdomains using a modified Schwarz criterion. The estimated number of regions proves to be weakly consistent under fairly general conditions. We then estimate the subdomain boundaries ("thresholds") and the regression coefficients within subdomains by minimizing the sum of squares of the residuals. We show that the threshold estimates converge (at rates, $1/n$ and $n^{-1/2}$, respectively at the model's threshold points of discontinuity and continuity) and that the regression coefficients as well as the residual variances are asymptotically normal. The basic condition on the error distribution required for the veracity of our asymptotic results is satisfied by any distribution with zero mean and a moment generating function (having bounded second derivative around zero). As an illustration, a segmented bivariate regression model is fitted to real data and the relevance of the asymptotic results is examined via simulations.

Key words and phrases: Asymptotic normality, consistency, local exponential boundedness, rate of convergence, segmented multivariate regression.

1. Introduction

Many practical situations involve a response variable which depends on some independent variables through a function whose form cannot be uniformly well approximated by the leading terms of a single Taylor expansion. Consequently, the usual linear regression model is not applicable and the simplicity of the methodology is lost.

However a segmented linear model has much of the simplicity of the classical linear methodology, and more flexibility. It may be regarded as a piecewise linear approximation deriving from different Taylor expansions in different subdomains. We require that a certain independent variable be selected and used to partition the domain of the independent variables into subdomains. The relationship between the response and independent variables is allowed to vary from subdomain-to-subdomain.

The partitioning variable may be suggested by extraneous considerations. The positive correlation of the production rate on the composition of the chemical agents involved reverses when the temperature exceeds a certain threshold (or

limit); Dunicz (1969) provides a specific example in the context of industrial chemistry, where a broken line model arises in a natural way. Examples drawn from the agricultural and biological sciences appear in Sprent (1961).

In economics, government interventions and policy changes at designated times can influence both economic structure and market structure. An example is given by McGee and Carleton (1970). They investigate the effect of the abolition in December, 1968, of commission splitting on the daily dollar volume of sales in regional stock exchanges, specifically the Boston Stock Exchange. These dollar volumes (y, say) are compared with those of the New York and American Stock Exchanges combined (x, say). Their results demonstrate the value of piecewise linear models when the relationship between y and x is nonlinear. Partitioning the data by "time" gives the four periods, Jan 67 - Oct 67, Nov 67 - July 68, Aug 68 - Nov 68, and Dec 68 - Nov 69 and very linear submodels for each period (see McGee and Carleton (1970)). The effect of the abolition of commission splitting is clear and the authors conclude their piecewise linear regression "is a satisfying solution".

Broken line regressions also arise in clustering objects into several groups by some explanatory variable. For example, we may wish to classify athletes by age with a performance index as a dependent variable or to group cars by their weights or engine sizes with fuel efficiency as the response of interest. A natural classification method minimizes the penalized pooled group variations,

$$\sum_{g}\sum_{t\in g}[Y_{tg}-\overline{Y}_{g}]^2+c_n(card(g)),$$

where $\overline{Y}_g$ is the sample mean of group g and $c_n(card(g))$ is the penalty for over-grouping. In the aforementioned examples, we adopt different constant regressions over different age, weight or engine size groups. More detailed discussion of the automobile example is given in Section 3.

This paper deals primarily with situations like those described above where the segmented regression model "explains" a real phenomenon. But it is linked to other paradigms in modern regression theory. Such paradigms concern the situation described above: the regression function of y on x cannot be globally well approximated by the leading terms of its Taylor expansion, ruling out a global linear model (see the references below, for example). Various locally weighted "nonparametric regression" approaches have been proposed (see Friedman (1991), for a recent survey). However in higher dimensions, difficulties confront such approaches as the "curse of dimensionality" (COD) becomes progressively more severe. These difficulties are well described by Friedman (1991) who proposes an alternative methodology called "multivariate adaptive regression splines", or "MARS". His procedure is closely related to another of Breiman

and Meisel (1976) which involves an extension of what Friedman calls "recursive partitioning".

Our methodology may be viewed as adaptive regression using a different method of partitioning than Breiman and Meisel (1976). By placing an upper bound on the number of partitions, we avoid the COD. And we adopt a different stopping criterion in partitioning x–space; it is based on ideas of model selection rather than testing and seems more appealing to us. Finally, and most importantly, we are able to provide a large sample theory for our methodology. This feature of our work seems important to us. Although the MARS methodology appears to be supported by the empirical studies of Friedman (1991), there is an inevitable concern about the general merits of an ad hoc procedure when it lacks a theoretical foundation.

To summarize, our methodology has some of the simplicity of global linear models, and some of the modelling flexibility of nonparametric approaches. Our large sample theory gives precise conditions under which our methodology would work well (with sufficiently large samples). By restricting the number of x–subdomains sufficiently we avoid the COD. And our methodology enjoys the advantage of MARS, that partitioning is data–based.

In this paper we partition the x–domain using an x co-ordinate, x_d, suggested, as in the above examples, by substantive considerations. We have parallel, more complex results for the domain of nonparametric regression where the partitioning variable is not determined in advance. The corresponding papers are in preparation.

We now give a more precise description of the problem addressed in this paper. Let Y be the response and $x_1, \ldots, x_p$, the p regressor variables. The latter may be covariates or design variables; the theory of this paper includes both. For simplicity, they will usually be referred to as "covariates" in the sequel.

As indicated above, the model describing the way Y depends upon the x's is determined entirely by x_d, $d \leq p$. So we may decompose $(-\infty, \infty]$ into nonoverlapping intervals $(\tau_{i-1}, \tau_i], \tau_{i-1} < \tau_i, i = 1, \ldots, l+1$ (with $\tau_0 = -\infty$ and $\tau_{l+1} = \infty$). The intervals correspond to different regression models. Suppose

$$Y = f_i(x_1, \ldots, x_p; \tilde{\theta}_i) + \epsilon^*, \text{ if } x_d \in (\tau_{i-1}, \tau_i], i = 1, \ldots, l+1; \tag{1.1}$$

ϵ^*, with mean zero, represents noise and $\tilde{\theta}_i$ is the model parameter vector, $i = 1, \ldots, l+1$. If $f_i(\cdot)$, defined above were sufficiently smooth we could approximate it by a linear model,

$$Y = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} x_j + \epsilon, \text{ if } x_d \in (\tau_{i-1}, \tau_i], i = 1, \ldots, l+1, \tag{1.2}$$

where ϵ has mean 0 and variance σ_i^2. The result is a segmented regression model. Except in unusual situations, the model parameters including the number of segments must be estimated.

The classical linear regression model achieves maximal simplicity among models in the class characterized by (1.2); there is only one Y on x regression model. The change-point problem gives rise to other members of this class. Yao (1988) solves such a problem in which $\beta_{ij} = 0$ for all $i = 1, \ldots, l$ and $j = 1, \ldots, p$; x_d is the explanatory variable controlling the allocation of measurements associated with the various dependence structures. Our formulation differs from that of Yao in that we introduce an explanatory variable to allocate response measurements. But these two formulations agree when viewed from the perspective of experimental design. To obtain a third special case of (1.2), assume all regressors are known functionals of x_d as in segmented polynomial regression. Feder (1975) discusses at length, the last case with l assumed known. Hinkley (1969, 1970) looks at estimation and inference for segmented regression models under a different setup. Quandt (1960, 1972) and Hudson (1966) also consider segmented regression under various conditions.

Now with the partitioning variable x_d identified, the inferential problem confronting us involves three parts: (i) the specification of the number of pieces in the model, l; (ii) the determination of the boundaries $\{\tau_i\}$ (called "thresholds" hereafter) of intervals over which each of the model pieces applies; (iii) the estimation of the linear model parameters within each interval. If l and the $\{\tau_i\}$ were specified, part (iii) would consist essentially of applying the classical theory, interval–by–interval. Consistency and asymptotic normality of estimates would thus obtain within intervals. Moreover, the residual sum of squares for error would provide an indication of goodness of fit.

However, l and the $\{\tau_i\}$ have to be estimated. The model parameter estimators and residual sum of squares obtained from the interval–by–interval analysis described above are functions of l and the $\{\tau_i\}$. Summing the residual sums of squares for the various intervals yields an overall index of the quality of fit of the segmented model; with l fixed, the $\{\tau_i\}$ may be estimated by minimizing this index. But further minimization of the index to estimate l results in gross overfitting and formally inconsistent estimators since the data demand, through this minimization process, that the maximal allowable value of l be selected. Instead this index must be adapted by adding a penalty term which increases with l.

We present a suitable penalty term in this paper, and this may be considered a principal result of the work reported herein. The penalty must be severe enough to limit the estimate, $\hat{l}$, of l. But it must also force $\hat{l}$ to converge quickly enough to the "true" value of l as to assure the asymptotic properties of all the other estimates which depend on $\hat{l}$. The criterion proposed by Schwarz (1978)

and described below, is not sufficiently stringent. Our proposed alternative (see equation (2.3) below) seems to work well.

This paper has two parts. Section 2 contains the first: a development of a convenient notation and a description of our method for fitting a segmented regression model. The second consists of an exploration of the quality of the proposed methodology. Simulation studies in Section 3 show that in small samples our method correctly identifies the number of pieces and thresholds for a segmented model. An example in this same section shows the feasibility of our approach in realistic situations. Finally, in Section 4 we give an asymptotic theory for our methodology. We prove under appropriate conditions that: (i) the estimate of l is weakly consistent; (ii) estimates of discontinuous thresholds, $\tau_i, i = 1, \ldots, l$, converge at the rate of $O_p(1/n)$ and those of the continuous thresholds at the usual rate of $O_p(1/\sqrt{n})$; (iii) and estimates of the segmented regression coefficients and the residual variances are asymptotically normal. Proofs of these assertions appear in Section 5.

2. The Estimation Procedure

For the segmented linear regression model (1.2), let $(Y_1, x_{11}, \ldots, x_{1p})$, …, $(Y_n, x_{n1}, \ldots, x_{np})$ be the independent observations of the response, Y, and covariates, $x_1, \ldots, x_p$. Let $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tp})'$ for $t = 1, \ldots, n$, and $\tilde{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$, $i = 1, \ldots, l+1$. Then

$$Y_t = \mathbf{x}_t' \tilde{\beta}_i + \epsilon_t, \text{ if } x_{td} \in (\tau_{i-1}, \tau_i], \ i = 1, \ldots, l+1, t = 1, \ldots, n; \tag{2.1}$$

the $\{\epsilon_t\}$ are independent and identically distributed (hereafter i.i.d.) random variables having mean zero and common variance σ^2. It is assumed that the $\{\epsilon_t\}$ are independent of $\{\mathbf{x}_t\}$ and $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{l+1} = \infty$. The $\{\epsilon_t\}$ need not be i.i.d.; they could have different distributions from one modelling interval to another. But we adopt this assumption for convenience.

In the sequel, a superscript or subscript 0 denotes the "true" parameter values. In addition, let the $n \times n$ matrix $I_n(\alpha, \eta)$ be defined by

$$I_n(\alpha, \eta) := \text{diag}(\mathbf{1}_{(x_{1d} \in (\alpha, \eta])}, \ldots, \mathbf{1}_{(x_{nd} \in (\alpha, \eta])}), \forall \ -\infty \leq \alpha < \eta \leq \infty.$$

For simplicity, let

$$X_n := \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}, \ \mathbf{Y}_n := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \tilde{\epsilon}_n := \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \ X_n(\alpha, \eta) = I_n(\alpha, \eta) X_n,$$

and

$$H_n(\alpha, \eta) := X_n(\alpha, \eta)[X_n'(\alpha, \eta) X_n(\alpha, \eta)]^- X_n'(\alpha, \eta);$$

in general, for any matrix A, A^- will denote a generalized inverse while $\mathbf{1}_{(\cdot)}$ represents the indicator function. Finally let the observations and residuals in the intended modeling interval be

$$\mathbf{Y}_n(\alpha,\eta) = I_n(\alpha,\eta)\mathbf{Y}_n,\ \tilde{\epsilon}_n(\alpha,\eta) = I_n(\alpha,\eta)\tilde{\epsilon}_n,$$

and the fitted sum of squares be

$$S_n(\alpha,\eta) := \mathbf{Y}_n'[I_n(\alpha,\eta) - H_n(\alpha,\eta)]\mathbf{Y}_n,$$

$$S_n(\tau_1,\ldots,\tau_l) := \sum_{i=1}^{l+1} S_n(\tau_{i-1},\tau_i), \tau_0 := -\infty, \tau_{l+1} := \infty,$$

$$T_n(\alpha,\eta) := \tilde{\epsilon}_n' H_n(\alpha,\eta)\tilde{\epsilon}_n$$

and the prediction of $\mathbf{Y}_n$ in the modeling interval be

$$\hat{\mathbf{Y}}_n(\alpha,\eta) := H_n(\alpha,\eta)\mathbf{Y}_n.$$

Then, in terms of true parameters, (2.1) can be rewritten in the vector form,

$$\mathbf{Y}_n = \sum_{i=1}^{l^0+1} X_n(\tau_{i-1}^0,\tau_i^0)\tilde{\beta}_i + \tilde{\epsilon}_n. \tag{2.2}$$

The estimation of all the parameters is done primarily in two steps. First we estimate l^0, the number of thresholds, $\tau_1^0,\ldots,\tau_{l^0}^0$, by minimizing the modified Schwarz' criterion (Schwarz (1978)),

$$MIC(l) := \ln[S(\hat{\tau}_1,\ldots,\hat{\tau}_l)/(n-p^*)] + p^*\frac{c_0(\ln n)^{2+\delta_0}}{n}, \tag{2.3}$$

for some constants $c_0 > 0, \delta_0 > 0$, where $p^* = (l+1)p + l \approx (l+1)(p+1)$ is the total number of fitted parameters, and for any fixed l, $\hat{\tau}_1,\ldots,\hat{\tau}_l$ are the least squares estimates which minimize $S_n(\tau_1,\ldots,\tau_l)$ subject to $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{l+1} = \infty$. Recall that the Schwarz criterion (SC) is defined by

$$SC(l) = \ln[S(\hat{\tau}_1,\ldots,\hat{\tau}_l)/(n-l)] + p^*\frac{2\ln(n)}{n}. \tag{2.4}$$

So $MIC(l)$ and $SC(l)$ differ in the severity of their penalty for overspecification; and a severe penalty is essential for the correct specification of a non-Gaussian, segmented regression model, $SC(l)$ being derived under the Gaussian assumption (c.f., Yao (1988)). It must be noted, however, that although a much severer penalty than that in (2.3) assures consistency of $\hat{l}$, underspecification is likely for such sample sizes. Below, we briefly discuss the choice of c_0 and δ_0 for small to moderate sample sizes.

In general in model selection, a relatively large penalty term would be preferable for easily identified models. A large penalty will greatly reduce the probability of overestimation while not unduly risking underestimation. However, if the model is difficult to identify, for example if it were continuous and $\|\tilde{\beta}_{j+1} - \tilde{\beta}_j\|$ were small, the penalty cannot be too large without incurring probable underestimation.

Another factor influencing the choice of the penalty is the error distribution. A distribution with heavy tails is likely to generate extreme values, making it look as though a change in response has occurred. To counter this effect, one needs a heavier penalty.

We recognize the point made by an anonymous referee that there exists no unique choice among the class of alternatives to the Shwarz criterion. Others may also give consistent model parameter estimators.

However, we do not see any reasonable way of making a "best" choice among the possibilities. The unicity of such a choice would be illusory and reflect our substitution of one choice with another, that of a process model. Such a model would need to specify for example, whether both discontinuous and continuous changes at the segmentation points would be admitted (as ours does).

We chose our penalty criterion $p^* c_0 [\ln(n)]^{2+\delta_0}/n$, in accordance with the results of our simulation study and our desire to change the well-established criterion of Schwarz as little as possible. In Section 3, our desires are also reflected in our choices of constants for situations involving small sample sizes.

Given that the best criterion is model dependent and no uniformly optimal choice can be made, the following considerations guide us to a reasonable choice of δ_0 and c_0:

(1) the proof of Lemma 5.2 in Section 5 suggests the exponent $2 + \delta_0$ in the penalty term of MIC may be further reduced, while retaining the consistency of the model selection procedure; and since the Schwarz criterion (where the exponent is 1) obtains from maximizing the posterior likelihood in a model selection paradigm and enjoys widespread use in model selection, it provides a natural baseline. From this perspective, δ_0 should be small to reduce the potential risk of underestimation when the noise is normal and n not too large.

(2) for a small sample, it is practically difficult to distinguish normal, double exponential, and t distributed noise. Hence, one would not expect the choice of penalty criterion to be critical.

(3) for large samples, SC (Schwarz criterion) tends to overestimate l^0 if the noise is not normal (Yao (1988)). We observe such overestimation in our simulations under various model specifications when $n = 50$ (see Section 3.3).

Item (1) above suggests we choose a small value for δ_0; and by (2), with δ_0 chosen, we can choose some moderate n_0, and solve for c_0 by forcing MIC equal

to SC at n_0. By (3), $n_0 < 50$ seems desirable. In the simulation reported in the next section, we (arbitrarily) choose δ_0 to be 0.1 (which is considered small). With such a δ_0, we arbitrarily choose $n_0 = 20$ and solve for c_0: $c_0 = 0.299$.

In summary, since the "best" selection of the penalty is model dependent for finite samples, there exists no optimal pair, (c_0, δ_0). On the other hand, our choices, $\delta_0 = 0.1$ and $c_0 = 0.299$, seem satisfactory in most of our simulation experiments. (The results of these experiments appear in Section 3.3.) However, further study is needed on the choice of δ_0 and c_0 under a variety of assumptions.

With estimates, $\hat{l}$ of l^0, and, $\hat{\tau}_i$ for τ_i^0, $i = 1, \ldots, \hat{l}$ available, we then estimate the other regression parameters $\{\tilde{\beta}_i^0\}$ and the residual variance σ_0^2 by ordinary least squares,

$$\hat{\tilde{\beta}}_i = [X_n'(\hat{\tau}_{i-1}, \hat{\tau}_i) X_n(\hat{\tau}_{i-1}, \hat{\tau}_i)]^- X_n'(\hat{\tau}_{i-1}, \hat{\tau}_i) \mathbf{Y}_n, \ i = 1, \ldots, \hat{l} + 1,$$

and

$$\hat{\sigma}^2 = S_n(\hat{\tau}_1, \ldots, \hat{\tau}_l)/(n - \hat{p^*}),$$

where $\hat{p^*} = (\hat{l}+1)p + \hat{l}$. Under regularity conditions essential for the identifiability of the regression parameters, we shall see in Section 4 that the ordinary least squares estimates $\hat{\tilde{\beta}}_j$ will be unique with probability approaching 1, for $j = 1, \ldots, \hat{l} + 1$, as $n \to \infty$.

3. Simulated and Real Examples

In this section, simulation studies are used to assess the performance of the procedure proposed in the preceding section. Limited by computing power, we study only moderate sample sizes with two to three dependence structures so that $l^0 = 1$ or 2. An example demonstrates its feasibility in realistic situations.

Let: (i) $\{\epsilon_t\}$ be i.i.d. with mean 0 and variance σ^2; (ii) $\mathbf{z}_t = (x_{t1}, \ldots, x_{tp})$ so that $\mathbf{x}_t' = (1, \mathbf{z}_t)$, where $\{x_{tj}\}$ are i.i.d. $N(0, 4)$; and (iii) $DE(0, \lambda)$ denotes the double exponential distribution with mean 0 and variance $2\lambda^2$.

For $d = 1$ and $\tau_1^0 = 1$, the following 5 sets of specifications of the model are used:

(a) $p = 2$, $\tilde{\beta}_1 = (0, 1, 1)'$, $\tilde{\beta}_2 = (1.5, 0, 1)'$, $\epsilon_t \sim N(0, 1)$;
(b) $p = 2$, $\tilde{\beta}_1 = (0, 1, 1)'$, $\tilde{\beta}_2 = (1.5, 0, 1)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$;
(c) $p = 2$, $\tilde{\beta}_1 = (0, 1, 0)'$, $\tilde{\beta}_2 = (1, 1, 0.5)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$;
(d) $p = 3$, $\tilde{\beta}_1 = (0, 1, 0, 1)'$, $\tilde{\beta}_2 = (1, 0, 0.5, 1)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$;
(e) $p = 3$, $\tilde{\beta}_1 = (0, 1, 1, 1)'$, $\tilde{\beta}_2 = (1, 0, 1, 1)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$.

From the theory of Section 4 we know that the least squares estimate $\hat{\tau}_1$, is appropriate if the model is discontinuous at τ_1^0. To explore the behavior of $\hat{\tau}_1$ for moderate sized samples, Models (a)-(d) are chosen to be discontinuous. The noise term in Model (a) is chosen to be normal as a reference, normal noise

being widely used in practice. However, our emphasis is on more general noise distributions. Because the double exponential distribution is commonly used in regression modeling and it has heavier tails than the normal distribution, it is used as the distribution of the noise in all other models. The deterministic part of Model (b) is chosen to be the same as that of Model (a) to make them comparable. Note that Models (a) and (b) have a jump of size 0.5 at $x_1 = \tau_1$ while $\mathrm{Var}\,(\epsilon_1) = 1$, which is twice the jump size. Except for the estimation of the parameter τ_1, our model selection method and estimation procedures work well for both continuous and discontinuous models. Model (e) is chosen to be a continuous model to demonstrate the behavior of the estimates for this type of model.

Let L be an upper bound imposed on l. In all, 100 replications are simulated for sample sizes, 30, 50, 100 and 200. Although $L = 3$ was tried in some experiments, the number of under- and over-estimated l^0's are the same as those obtained when $L = 2$. The result $\hat{l} = 3$ obtains in only 1 or 2 of the 100 replications. This agrees with our intuition for a two-piece model; it is unlikely a four-piece model will be selected over a two-piece one if a two-piece model is selected over a three-piece one. This experience suggests setting $L = 2$ to save some computational effort, and the results reported in Tables 3.1 and 3.2 are obtained in this way. However, much larger values of L are feasible in practice where a single run suffices. The constants δ_0 and c_0 in MIC are chosen as 0.1 and 0.299 respectively, for the reasons given in Section 2.

Our results are summarized in Tables 3.1 and 3.2. Table 3.1 contains the estimates of l^0, τ_1^0 and the standard error of the estimate of τ_1^0, $\hat{\tau}_1$, based on the MIC. The following observations derive from the table:

(i) for sample sizes greater than 30, the MIC correctly identifies l^0 in most of the cases. Hence, for estimating l^0, the result seems satisfactory. Comparing Models (a) and (b), it seems that the distribution of the noise has a significant influence on the estimation of l^0, for sample sizes of 50 or less.

(ii) for small sample sizes, the bias of $\hat{\tau}_1$ is related to the shape of the underlying model. It is seen that the biases are positive for Models (a) and (b), and negative for the others. In an experiment where Models (a) and (b) are changed so that the jump size at $x_1 = \tau_1$ is -0.5, instead of 0.5, negative biases are observed for every sample size. These biases decrease as the sample size becomes larger.

(iii) the standard error of $\hat{\tau}_1$ is relatively large in all cases considered; and, as expected, the standard error decreases as the sample size increases. This suggests that a large sample size is needed to estimate τ_1^0 reliably. An experiment with sample size of 400 for a model similar to Model (e) is reported in Wu (1992). In that experiment the standard error of $\hat{\tau}_1$ is significantly reduced.

(iv) the choice of $\delta_0 = 0.1$ seems adequate for most of our experimental models since it does not lead to any discernible pattern in the results, such as regular

overestimation of l when $n = 30$ and underestimation of l when $n = 50$ or vice versa.

The continuity of Model (e) leads us to our prior expectation that its identification would be the most difficult of all the cases considered. The c_0 chosen above seems too big for this case, the tendency toward underestimating l being obvious when the sample size is small. More plausibly, with the small sample size and the high noise level, there is simply not enough information to reveal the underlying model. Therefore, choosing a lower dimensional model with positive probability may be appropriate by the principle of parsimony.

In summary, since the optimal penalty is model dependent for samples of moderate size, no globally optimal pair of (c_0, δ_0) can be recommended. On the other hand, our choices of δ_0 and c_0 perform reasonably well for our experimental models.

Table 3.2 shows the estimated values of the other model parameters for the models in Table 3.1 and a sample size of 200. The results indicate that, in general, the estimates of the $\tilde{\beta}_j$'s and σ_0^2 are quite close to their true values even when $\hat{\tau}_1$ is inaccurate. So, for the purpose of estimating $\tilde{\beta}_j$'s and σ_0^2, and interpolation when the model is continuous, a moderate sized sample say of size 200 may be sufficient. When the model is discontinuous, interpolation near the threshold may not be accurate due to the inaccurate $\hat{\tau}_1$. A careful comparison of the estimates obtained from Models (a) and (b) shows that the estimation errors are generally smaller with normally distributed errors. The estimates of β_{20} have relatively larger standard errors. This is because a small error in $\hat{\beta}_{21}$ would result in a relatively large error in $\hat{\beta}_{20}$.

Table 3.1. Frequency of correct identification of l^0 in 100 repetitions and the estimated thresholds for segmented regression models.

(m, m_u, m_o are the frequencies of correct, under- and over-estimations of l^0)

$MIC : m(m_u, m_o)$	sample size			
$\hat{\tau}_1$ (SE)	30	50	100	200
Model (a)	79 (18, 3)	95 (4, 1)	100 (0, 0)	100 (0, 0)
	1.168 (1.500)	1.033 (1.353)	1.410 (0.984)	1.259 (0.665)
Model (b)	70 (21, 9)	86 (8, 6)	99 (0, 1)	100 (0, 0)
	1.022 (1.546)	1.220 (1.407)	1.432 (0.908)	1.245 (0.692)
Model (c)	80 (6, 14)	97 (1, 2)	100 (0, 0)	100 (0, 0)
	0.890 (0.737)	0.761 (0.502)	0.901 (0.221)	0.932 (0.151)
Model (d)	85 (8, 7)	99 (0, 1)	100 (0, 0)	100 (0, 0)
	0.791 (1.009)	0.860 (0.665)	0.971 (0.232)	0.963 (0.169)
Model (e)	68 (23, 9)	87 (12, 1)	100 (0, 0)	100 (0, 0)
	0.463 (1.735)	0.708 (1.332)	0.989 (0.923)	0.940 (0.707)

Table 3.2. Estimated regression coefficients and variances of noise and their standard errors with $n = 200$.

(Conditional on $\hat{l} = 1$)

$\hat{\beta}_{ij}$ (SE)	Model (a)	Model (b)	Model (c)	Model (d)	Model (e)
β_{10}	-0.003 (0.145)	-0.018 (0.146)	0.004 (0.143)	-0.008 (0.154)	-0.059 (0.177)
β_{11}	1.001 (0.038)	0.995 (0.037)	1.000 (0.035)	0.995 (0.041)	0.985 (0.045)
β_{12}	1.000 (0.024)	0.996 (0.025)	-0.004 (0.025)	0.000 (0.024)	1.000 (0.025)
β_{13}	—	—	—	0.994 (0.023)	0.995 (0.025)
β_{20}	1.485 (0.345)	1.388 (0.332)	0.962 (0.243)	1.009 (0.225)	0.960 (0.283)
β_{21}	0.005 (0.063)	0.019 (0.067)	0.008 (0.055)	0.000 (0.049)	0.008 (0.057)
β_{23}	1.006 (0.034)	0.998 (0.034)	0.495 (0.032)	0.498 (0.032)	0.998 (0.036)
β_{24}	—	—	—	0.997 (0.034)	0.996 (0.036)
σ^2	0.948 (0.108)	0.950 (0.154)	0.956 (0.156)	0.953 (0.160)	0.944 (0.158)

To assess the performance of the MIC when $l^0 = 2$, and to compare it with the Schwarz Criterion (SC) as well as a criterion proposed by Yao (1989), simulations were done for a very simple special case of our general model with sample sizes n of up to 450. Here we adopt Yao's (1989) set-up where a univariate piecewise constant model is to be estimated. Note that such a model is a special case of Model (1.2). Specifically, Yao's model is

$$y_t = \beta_j^0 + \epsilon_t \quad \text{if } x_t \in (\tau_{j-1}^0, \tau_j^0], \quad j = 1, \ldots, l^0 + 1,$$

where $x_t = t/n$ for $t = 1, \ldots, n$, ϵ_t is i.i.d. with mean zero and finite $2m$th moment for some positive integer m. Yao shows that with $m \geq 3$, the minimizer of $\log \hat{\sigma}_l^2 + l \cdot C_n/n$ is a consistent estimate of l^0 for $l \leq L$, the known upper bound of l^0, where $\{C_n\}$ is any sequence satisfying $C_n n^{-2/m} \to \infty$ and $C_n/n \to 0$ as $n \to \infty$. Four sets of specifications of this experimental model are used:
(f) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 2,\ \beta_{30}^0 = 4,\ \epsilon_t \sim DE(0, 1/\sqrt{2})$;
(g) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 2,\ \beta_{30}^0 = 4,\ \epsilon_t \sim t_7/\sqrt{1.4}$;
(h) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 1,\ \beta_{30}^0 = -1,\ \epsilon_t \sim DE(0, 1/\sqrt{2})$;
(i) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 1,\ \beta_{30}^0 = -1,\ \epsilon_t \sim t_7/\sqrt{1.4}$,
where t_7 refers to the Students-t distribution with 7 degrees of freedom.

In each of these cases the variances of ϵ_t are scaled to 1 to make the noise levels comparable. Note that for $\epsilon_t \sim t_7/\sqrt{1.4}$, $E(\epsilon_t^6) < \infty$ and $E|\epsilon_t^7| = \infty$, so the model barely satisfies the condition of Yao and Au (1989) with $m = 3$ and does not satisfy our exponential boundedness condition. In the paper of Yao and Au (1989), $\{C_n\}$ is not specified, so $\{C_n\}$ must be chosen to satisfy the conditions. The simplest $\{C_n\}$ is $C_n = c_1 n^\alpha$. With $m = 3$, we have $n^{\alpha - 2/3} \to \infty$ implying $\alpha > 2/3$. (We shall call the criterion with such a C_n, YC, hereafter.) To reduce the potential risk of underestimating l^0, we round 2/3 up to 0.7 as our choice of

α. The δ_0 and c_0 in MIC are chosen as 0.1 and 0.299 respectively, for the reasons previously mentioned. c_1 is chosen by the same method we used to choose c_0, that is, forcing $\log n_0 = c_1 n_0^\alpha$ and solving for c_1. With $n_0 = 20$ and $\alpha = 0.7$, we get $c_1 = 0.368$.

The results for model selection are reported in Tables 3.3-3.4. Table 3.3 tabulates the empirical distributions of the estimated l^0 for different sample sizes. From the table, it is seen that for most cases, MIC and YC perform significantly better than SC. And with a sample size of 450, MIC and YC correctly identify l^0 in more then 90% of the cases. For Models (f) and (g), which are more easily identified, YC makes more correct identifications than MIC. But for Models (h) and (i), which are harder to identify, MIC makes more correct identifications. From Theorem 4.1 and the remark after its proof, it is known that both MIC and YC are consistent for the models with double exponential noise. This theory seems to be confirmed by our simulation.

Table 3.3. The empirical distribution of $\hat{l}$ in 100 repetitions by MIC, SC and YC for piecewise constant model.

(n_0, n_1, n_2, n_3 are the frequencies of $\hat{l} = 0, 1, 2, 3$ respectively)

$MIC : n_0, n_1, n_2, n_3$ $YC : n_0, n_1, n_2, n_3$ $SC : n_0, n_1, n_2, n_3$	sample size		
	50	150	450
Model (f)	5, 30, 48, 17	0, 18, 79, 3	0, 0, 98, 2
	5, 36, 45, 14	0, 36, 64, 0	0, 9, 91, 0
	0, 17, 52, 31	0, 1, 64, 35	0, 0, 83, 17
Model (g)	5, 38, 51, 6	0, 23, 72, 5	0, 0, 99, 1
	7, 41, 48, 4	0, 46, 53, 1	0, 7, 93, 0
	3, 18, 56, 23	0, 2, 79, 19	0, 0, 87, 13
Model (h)	0, 3, 81, 16	0, 0, 96, 4	0, 0, 98, 2
	0, 3, 84, 13	0, 0, 100, 0	0, 0, 100, 0
	0, 0, 63, 37	0, 0, 82, 18	0, 0, 87, 13
Model (i)	0, 5, 85, 10	0, 0, 97, 3	0, 0, 100, 0
	0, 7, 86, 7	0, 0, 100, 0	0, 0, 100, 0
	0, 1, 73, 26	0, 0, 83, 17	0, 0, 93, 7

Model selection seems to be little affected by varying the noise distribution. This may be due to the scaling of the noises by their variances, since variance is more sensitive to tail probabilities compared to quantiles or mean absolute deviation. Because most people are familiar with the use of variance as an index of dispersion, we adopt it, although other measures may reveal the tail effect on model identification better for our moderate sample sizes. Table 3.4 shows the estimated thresholds and their standard deviations for Models (f), (g), (h), (i),

conditional on $\hat{l} = l^0$. Overall, they are quite accurate, even when the sample size is 50. For Models (h) and (i), the accuracy of $\hat{\tau}_2$ is much better than that of $\hat{\tau}_1$, since τ_2 is much easier to identify by the model specification. In general, for models which are more difficult to identify, a larger sample size is needed to achieve the same accuracy.

Table 3.4. The estimated thresholds and their standard errors for piecewise constant model.

(Conditional on $\hat{l} = 2$)

$\hat{\tau}_1$, (SE)	sample size		
$\hat{\tau}_2$, (SE)	50	150	450
Model (f)	0.335 (0.078)	0.338 (0.039)	0.334 (0.012)
	0.660 (0.032)	0.666 (0.008)	0.667 (0.003)
Model (g)	0.313 (0.076)	0.332 (0.032)	0.334 (0.013)
	0.656 (0.015)	0.669 (0.009)	0.667 (0.002)
Model (h)	0.316 (0.027)	0.334 (0.007)	0.333 (0.002)
	0.662 (0.030)	0.667 (0.006)	0.667 (0.003)
Model (i)	0.323 (0.023)	0.332 (0.010)	0.334 (0.004)
	0.661 (0.030)	0.666 (0.007)	0.667 (0.003)

A data set used in Henderson and Velleman (1981) has been analyzed, using our proposed method. The data consist of measurements of three variables, miles per gallon (MPG), weight (WT) and horse power (HP), on thirty eight 1978-79 model automobiles. The dependence of MPG on WT and HP is of interest. Graphs of the data reveal a nonlinear dependence of MPG on WT (see Figure 3.1). Four models are fitted to the data set and the MIC procedure is used to select the "best" model. The four candidate models are

$$MPG = \beta_0 + \beta_1 WT + \epsilon, \tag{3.1}$$

$$MPG = \beta_0 + \beta_1 WT + \beta_2 HP + \epsilon, \tag{3.2}$$

$$MPG = \beta_0 + \beta_1 WT + \beta_2 WT^2 + \beta_3 HP + \epsilon, \tag{3.3}$$

$$MPG = \begin{cases} \beta_{10} + \beta_{11} WT + \beta_{12} HP + \epsilon, & \text{if } WT < \tau, \\ \beta_{20} + \beta_{21} WT + \beta_{22} HP + \epsilon, & \text{if } WT \geq \tau. \end{cases} \tag{3.4}$$

As suggested by the previous simulations, the constants c_0 and δ_0 in the penalty term of MIC are chosen, respectively, as 0.2 and 0.05. The MIC values for the four models are 2.24, 2.28, 2.12 and 2.11 respectively. So Model (3.4) is chosen as the "best" model. With this model, $\hat{\sigma}^2 = 4.90$, $\hat{\tau} = 2.7$, and the estimated coefficients are $(\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}) = (48.82, -5.23, -0.08)$, $(\hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}) =$

$(30.76, -1.84, -0.05)$. Needless to say, the selected model is only the "best" among the four models considered; further model reduction may be possible.

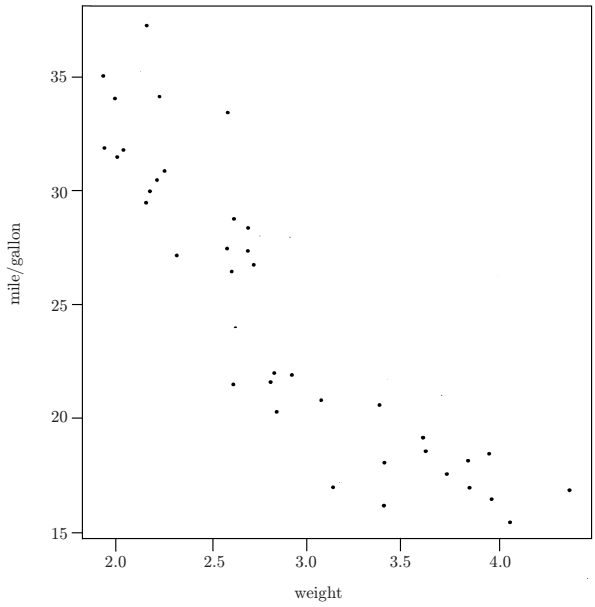

Figure 3.1. Mile per gallon vs. weight for 38 cars.

4. Asymptotic Properties of the Parameter Estimates

Consider the segmented linear regression model, (2.1), or its equivalent vector form (2.2). Let $\hat{l}$ minimize $MIC(l)$ as defined in (2.3). To identify the number of thresholds l^0 consistently, assume:

Assumption 4.1. $\{\mathbf{x}_t\}$ is a strictly stationary, ergodic process with positive definite matrices $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0-\delta,\tau_i^0))}\}$ and $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0,\tau_i^0+\delta))}\}$ in a small δ-neighbourhood of each of the true thresholds $\tau_1^0, \ldots, \tau_{l^0}^0$; or

Assumption 4.1′. If the covariates are not random, $(1/n)\sum_{t=1}^n \mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{[x_{td}\in(\tau_i^0-\delta,\tau_i^0]]}$ and $\frac{1}{n}\sum_{t=1}^n \mathbf{x}_t\ \mathbf{x}_t'\mathbf{1}_{[x_{td}\in(\tau_i^0,\tau_i^0+\delta]]}$ converge to positive definite real matrices for $\delta \in (0, \min_{1\le j\le l^0}\ (\tau_{j+1}^0 - \tau_j^0)/4)$.
(When $l^0 = 1$, this last requirement reduces to $0 < \delta < \infty$ since $\tau_{l^0+1}^0$ is always set to be infinity.)

Assumption 4.1 implies that the design matrix $X_n(\alpha, \eta)$ has full column rank a.s. as $n \to \infty$ for every open interval (α, η) in the small neighbourhood of $\tau_i^0, i = 1, \ldots, l^0$, for which x_d has a positive probability density. When *Assumption* 4.1′ is satisfied, $X_n(\alpha, \eta)$ will have full column rank for large n and for every open interval (α, η) in a small neighbourhood of $\tau_i^0, i = 1, \ldots, l^0$. So $\hat{\hat{\beta}}_i$ will be unique

with probability tending to 1 as $n \to \infty$, for $i = 1, \ldots, \hat{l}$, provided that $\hat{l}$ converges to l^0 in probability.

When the segmented regression model (2.1) reduces to the segmented polynomial or functional segmented regressions discussed by Feder (1975), a condition similar to either *Assumption* 4.1 or *Assumption* 4.1′ is essential for identifying the segmented model parameters (see Feder 1975). In particular, for the segmented polynomial regression model, *Assumption* 4.1 is automatically satisfied if the key covariate x_d has a positive density within a small neighbourhood of each of the thresholds.

In addition, we need to place some restriction on the distribution of the i.i.d. errors $\{\epsilon_t\}$. We will require that they be *locally exponentially bounded* (*LEB*). A random variable Z is said to be LEB if there exist constants, c_0 and T_0, in $(0, \infty)$ such that

$$E(e^{uZ}) \leq e^{c_0 u^2}, \ \forall \ |u| \leq T_0. \tag{4.1}$$

Remark. The LEB condition is satisfied by any distribution with zero mean and moment generating function having bounded second derivative near zero. Many commonly used error distributions such as the symmetrized exponential family are of this type. Hence the theory of this section applies to a wide range of problems.

Since the sample size n is always finite, only bounded $l^{0\prime}s$ can be effectively identified. So we assume an upper bound L of l^0. Another simplification in the nonlinear minimization of $S(\tau_1, \ldots, \tau_l)$ is obtained without loss of generality by limiting the possible values of $\tau_1 < \cdots < \tau_l$ to the discrete set, $\{x_{1d}, \ldots, x_{nd}\}$.

Theorem 4.1. *Suppose the segmented linear regression model* (2.2) *obtains, with* X_n *independent of* $\tilde{\epsilon}_n$. *Assume:*

(i) *the* $\tilde{\epsilon}_n$ *have i.i.d., LEB components with mean zero and variance* σ_0^2;

(ii) $l^0 \leq L$ *for some specified upper bound* $L > 0$;

(iii) *one of Assumptions* 4.1 *or* 4.1′ *is satisfied.*

Then $\hat{l} \to l^0$ *in probability as* $n \to \infty$.

Next, we show that the threshold estimates $(\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{l}})$ converge to the true thresholds, $(\tau_1^0, \ldots, \tau_{l^0}^0)$ at the rate of $O_p(1/n)$; and the least squares estimates of $\tilde{\beta}_j^0$ and σ_0^2 based on the estimated thresholds are asymptotically normal.

Assumption 4.2. (A.4.2.1) The covariates $\{\mathbf{x}_t\}$ are i.i.d. random variables with $E(\mathbf{x}_1'\mathbf{x}_1)^u < \infty$ for some $u > 2$.

(A.4.2.2) Within some small neighborhoods of the true thresholds, x_{1d} has a positive and continuous probability density function $f_d(\cdot)$ with respect to the one dimensional Lebesgue measure.

(A.4.2.3) There exists one version of $E[\mathbf{x}_1\mathbf{x}_1'|x_{1d} = x]$ which is continuous within some neighborhoods of the true thresholds and that version has been adopted.

Remark. Assumptions (A.4.2.1) - (A.4.2.3) are satisfied if $(x_{t1}, \ldots, x_{tp})'$ has a joint exponential distribution in the canonical form.

To obtain a rate of convergence, $O_p(1/n)$, for the threshold estimates, we need the following additional assumption.

Assumption 4.2′. There exist some positive η and δ such that

$$\inf_{|x-\tau_j^0|<\delta}\{P(|\mathbf{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0)| > \eta|x_d = x)\} > 0$$

for some j.

This assumption holds if $(x_{t1}, \ldots, x_{tp})'$ has a joint distribution from the exponential family in canonical form.

Similar assumptions can be made when the covariates are nonrandom by replacing the distributions in Assumption 4.2 and 4.2′ by their empirical counterparts.

Theorem 4.2. *Consider the segmented linear regression model* (2.2) *with* X_n *independent of* $\tilde{\epsilon}_n$. *Assume that its i.i.d., LEB components have mean zero and variance* σ_0^2 *while Assumptions* 4.1, 4.2 *and* 4.2′ *hold for some* $j = 1, \ldots, l^0$. *Then*

$$\hat{\tau_j} - \tau_j^0 = O_p(\frac{1}{n}).$$

Remark. If we replace *Assumption* 4.2′ by the slightly weaker condition $P(\mathbf{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0) \neq 0|x_d = \tau_j^0) > 0$ for some $j = 1, \ldots, l^0$ and maintain the rest of the assumptions of Theorem 4.2, then

$$\hat{\tau_j} - \tau_j^0 = O_p(\frac{\ln^2 n}{n}).$$

A detailed proof can be found in Wu (1992).

Recall that $\hat{\tilde{\beta}}_j$ and $\hat{\sigma}^2$ are the least squares estimates of $\tilde{\beta}_j^0$ and σ_0^2 based on the estimates $\hat{l}$ and $\hat{\tau}_j$'s as defined in Section 2, $j = 1, \ldots, l^0 + 1$.

Theorem 4.3. *Under the conditions of Theorem* 4.2 *except Assumption* 4.2′, *the least squares estimates* $\hat{\tilde{\beta}}_i$ *and* $\hat{\sigma}^2$ *based on the estimated* $\hat{l}$ *and* $\hat{\tau}_i$*'s as defined in Section* 2 *are asymptotically normal estimates of* $\tilde{\beta}_i^0$ *and* σ_0^2, $i = 1, \ldots, l^0 + 1$. *Namely,* $\sqrt{n}(\hat{\tilde{\beta}}_i - \tilde{\beta}_i^0)$ *and* $\sqrt{n}(\hat{\sigma}^2 - \sigma_0^2)$, $i = 1, \ldots, l^0 + 1$, *converge in distribution to normal distributions with zero means and finite variances.*

The asymptotic variances can be computed by first treating l^0 and $(\tau_i^0, i = 1, \ldots, l^0)$ as known so that the usual "estimates" of the variances of the estimates of the regression coefficients and residual variance can be written down explicitly then by substituting $\hat{l}$ and $(\hat{\tau}_i, i = 1, \ldots, \hat{l})$ for l^0 and $(\tau_i^0, i = 1, \ldots, l^0)$ in these variance "estimates".

Though most of the above results are stated for a discontinuous segmented regression model, i.e. $P(\mathbf{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0) \neq 0 | x_d = \tau_j^0) > 0$ for some $j = 1, \ldots, l^0$, similar asymptotic results hold at continuity points of the model except that the rate of convergence of the threshold estimates is reduced to $O_p(1/\sqrt{n})$ (see Wu (1992)).

5. Proofs

The proof of Theorem 4.1 follows a series of preliminary lemmas.

Lemma 5.1. *Assume $Z_1, \ldots, Z_k$ are i.i.d. LEB random variables, i.e. for some $T_0 > 0$ and $0 < c_0 < \infty$, $E(e^{uZ_1}) \leq e^{c_0 u^2}$ for $|u| \leq T_0$. Let $S_k = \sum_{i=1}^k a_i Z_i$, where the a_i's are constants. Then for any $t_0 > 0$ satisfying $|t_0 a_i| \leq T_0$ and all $i \leq k$,*

$$P\{|S_k| \geq x\} \leq 2e^{-t_0 x + c_0 t_0^2 \sum_{i=1}^k a_i^2}. \tag{5.1}$$

Proof. The result is a direct application of Markov's inequality to $P\{S_k \geq x\} = P\{e^{t_0 S_k} \geq e^{t_0 x}\}$ and $P\{S_k \leq -x\} = P\{-S_k \geq x\}$.

Lemma 5.2. *For the segmented regression model (2.2), assume that the i.i.d. errors $\{\epsilon_t\}$ are LEB and independent of X_n. Let $T_n(\alpha, \eta), -\infty \leq \alpha < \eta \leq \infty$, be defined as in Section 2. Then*

$$P\Big\{ \sup_{\alpha < \eta} T_n(\alpha, \eta) \geq \frac{9p_0^3}{T_0^2} \ln^2 n \Big\} \to 0, \qquad \text{as } n \to 0, \tag{5.2}$$

where p_0 is the true order of the model and T_0 is the constant associated with the local exponential boundedness of $\{\epsilon_t\}$.

Proof. Conditioning on X_n, we have

$$\begin{aligned} P\Big\{ \sup_{\alpha<\eta} T_n(\alpha,\eta) \geq \frac{9p_0^3}{T_0^2} \ln^2 n | X_n \Big\} &= P\Big\{ \max_{x_{sd} < x_{td}} \tilde{\epsilon}_n' H_n(x_{sd}, x_{td}) \tilde{\epsilon}_n \geq \frac{9p_0^3}{T_0^2} \ln^2 n | X_n \Big\} \\ &\leq \sum_{x_{sd} < x_{td}} P\Big\{ \tilde{\epsilon}_n' H_n(x_{sd}, x_{td}) \tilde{\epsilon}_n \geq \frac{9p_0^3}{T_0^2} \ln^2 n | X_n \Big\}. \end{aligned}$$

Since $H_n(x_{sd}, x_{td})$ is idempotent, one can write, for $p := rank(H_n(x_{sd}, x_{td})) = rank(\Lambda) \leq p_0$ with Q having full row rank p

$$\tilde{\epsilon}_n' H_n(x_{sd}, x_{td}) \tilde{\epsilon}_n = \tilde{\epsilon}_n' Q' Q \tilde{\epsilon}_n = \sum_{l=1}^p u_l^2,$$

where $Q' = (\mathbf{q}_1, \ldots, \mathbf{q}_p)$ and $u_l = \mathbf{q}_l' \tilde{\epsilon}_n$, $l = 1, \ldots, p$. Since $p \le p_0$ and p_0 is finite, it suffices to show that

$$\sum_{x_{sd}<x_{td}} P\Big\{u_l^2 \ge \frac{9p_0^2}{T_0^2} \ln^2 n \mid X_n\Big\} \to 0, \qquad \text{as } n \to 0,$$

for any l. Noting that $p = trace(H_n(x_{sd}, x_{td})) = \sum_{l=1}^{p} \| \mathbf{q}_l \|^2$, we have $\| \mathbf{q}_l \|^2 = \mathbf{q}_l'\mathbf{q}_l \le p \le p_0$, $l = 1, \ldots, p$. By Lemma 5.1, with $t_0 = T_0/p_0$ we have

$$\begin{aligned} &\sum_{x_{sd}<x_{td}} P\{|u_l| \ge 3p_0 \ln n/T_0 \mid X_n\} \\ &\le \sum_{x_{sd}<x_{td}} 2\exp(-\frac{T_0}{p_0} \cdot \frac{3p_0}{T_0} \ln n) \exp(c_0 (T_0/p_0)^2 p_0) \\ &\le n(n+1)/n^3 \exp(c_0 T_0^2/p_0) \to 0, \end{aligned}$$

as $n \to \infty$, where c_0 is the constant specified in Lemma 5.1. Finally, by appealing to the dominated convergence theorem we obtain the desired result without conditioning.

Lemma 5.3. *For the segmented regression model* (2.2), *assume:* (i) *the design matrix* X_n *satisfies Assumption* 4.1 *or Assumption* 4.1′; (ii) *the i.i.d. errors* $\{\epsilon_t\}$ *are LEB and independent of* X_n. *Then*

$$[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r) - S_n(\tau_r^0, \tau_r^0 + \delta)]/n \xrightarrow{a.s.} C_r \tag{5.3}$$

for any $\delta \in (0, \min_{1 \le j \le l^0}(\tau_{j+1}^0 - \tau_j^0)/4)$, *any* $r = 1, \ldots, l^0$ *and some* $C_r > 0$ *as* $n \to \infty$.

Proof. It suffices to prove the result when $l^0 = 1$. Since the proofs under *Assumption* 4.1 and *Assumption* 4.1′ are essentially the same, for brevity we prove the result only under *Assumption* 4.1. For expository simplicity, we omit the subscripts and superscripts 0 in this proof. Let $X_1^* = X_n(\tau_1 - \delta, \tau_1)$, $X_2^* = X_n(\tau_1, \tau_1 + \delta)$, $X^* = X_n(\tau_1 - \delta, \tau_1 + \delta) = X_1^* + X_2^*$, $\tilde{\epsilon}^* = I_n(\tau_1 - \delta, \tau_1 + \delta)\tilde{\epsilon}_n$ and $\hat{\tilde{\beta}} = (X^{*\prime} X^*)^- X^{*\prime} \mathbf{Y}_n$. As in ordinary regression, we have

$$\begin{aligned} &S_n(\tau_1 - \delta, \tau_1 + \delta) \\ &= \|X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}})\|^2 + \|X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}})\|^2 + \|\tilde{\epsilon}^*\|^2 + 2\tilde{\epsilon^*}' X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + 2\tilde{\epsilon^*}' X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}). \end{aligned}$$

It then follows from the Law of Large Numbers for stationary ergodic stochastic processes that as $n \to \infty$,

$$\hat{\tilde{\beta}} \xrightarrow{\text{a.s.}} \{E\{\mathbf{x}_1 \mathbf{x}_1' \mathbf{1}_{[x_{1d} \in (\tau_1 - \delta, \tau_1 + \delta)]}\}\}^{-1} E\{Y_1 \mathbf{x}_1 \mathbf{1}_{[x_{1d} \in (\tau_1 - \delta, \tau_1 + \delta)]}\} := \tilde{\beta}^*,$$

and hence

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}S_n(\tau_1-\delta,\tau_1+\delta)\\ &=(\tilde\beta_1-\tilde\beta^*)'\cdot E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{[x_{1d}\in(\tau_1-\delta,\tau_1)]})\cdot(\tilde\beta_1-\tilde\beta^*)\\ &\quad+(\tilde\beta_2-\tilde\beta^*)'\cdot E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{[x_{1d}\in(\tau_1,\tau_1+\delta)]})\cdot(\tilde\beta_2-\tilde\beta^*)+\sigma^2P\{x_{td}\in(\tau_1-\delta,\tau_1+\delta)\}.\end{aligned}$$

Similarly, we can show that $\frac{1}{n}S_n(\tau_1-\delta,\tau_1)$ and $\frac{1}{n}S_n(\tau_1,\tau_1+\delta)$ converge to $\sigma^2P\{x_{1d}\in(\tau_1-\delta,\tau_1)\}$ and $\sigma^2P\{x_{1d}\in(\tau_1,\tau_1+\delta)\}$, respectively. Finally, we have

$$\begin{aligned}C_r&=(\tilde\beta_1-\tilde\beta^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{[x_{1d}\in(\tau_1-\delta,\tau_1)]})(\tilde\beta_1-\tilde\beta^*)\\ &\quad+(\tilde\beta_2-\tilde\beta^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{[x_{1d}\in(\tau_1,\tau_1+\delta)]})\cdot(\tilde\beta_2-\tilde\beta^*)\end{aligned}$$

is positive. This completes the proof.

Lemma 5.4. *Assumptions* (i) *and* (ii) *of Lemma* 5.3 *imply*
(i) *for all* $l<l^0$, $P\{\hat\sigma_l^2>\sigma_0^2+C\}\to1,\ n\to\infty$ *for some* $C>0$ (ii) *for all* l *such that* $l^0\le l\le L$, L *being a specified upper bound for* l^0,

$$0\le 1/n\sum_{t=1}^{n}\epsilon_t^2-\hat\sigma_l^2=O_p(ln^2(n)/n),\tag{5.4}$$

where $\hat\sigma_l^2=\frac{1}{n}S_n(\hat\tau_1,\dots,\hat\tau_l)$ *is the least squares estimate* (*LSE*) *of* σ^2 *when* l^0 *is the number of true thresholds.*

Proof. (i) Since $l<l^0$, for $\delta\in(0,\min_{1\le j\le l^0}(\tau_{j+1}^0-\tau_j^0)/4)$, there exists $1\le r\le l^0$, such that $(\hat\tau_1,\dots,\hat\tau_l)\in A_r:=\{(\tau_1,\dots,\tau_l):|\tau_s-\tau_r^0|>\delta,\forall\ s=1,\dots,l\}$. Hence, if we can show that for each $r,\ 1\le r<l^0$, with probability approaching 1,

$$\min_{(\tau_1,\dots,\tau_l)\in A_r}S_n(\tau_1,\dots,\tau_l)/n>\sigma_0^2+C_r,$$

for some $C_r>0$. By choosing $C:=\min_{1\le r\le l^0}\{C_r\}$, we prove the desired result.

For any $(\tau_1,\dots,\tau_l)\in A_r$, let $\xi_1\le\cdots\le\xi_{l+l^0+1}$ be the ordered set $\{\tau_1,\dots,\tau_l,\tau_1^0,\dots,\tau_{r-1}^0,\tau_r^0-\delta,\tau_r^0+\delta,\tau_{r+1}^0,\dots,\tau_{l^0}^0\}$ and let $\xi_0=-\infty$, $\xi_{l+l^0+2}=\infty$. Then it follows from Lemmas 5.2 and 5.3 and the law of large numbers that uniformly in A_r,

$$\begin{aligned}&\frac{1}{n}S_n(\tau_1,\dots,\tau_l)\ge\frac{1}{n}S_n(\xi_1,\dots,\xi_{l+l^0+1})\\ &=\frac{1}{n}\tilde\epsilon_n'\tilde\epsilon_n+O_p(\ln^2(n)/n)+\frac{1}{n}(S_n(\tau_r^0-\delta,\tau_r^0+\delta)-S_n(\tau_r^0-\delta,\tau_r^0)-S_n(\tau_r^0,\tau_r^0+\delta))\\ &=\sigma_0^2+C_r+o_p(1),\end{aligned}\tag{5.5}$$

where C_r is defined in (5.3).

(ii) Let $\xi_1 \le \cdots \le \xi_{l+l^0}$ be the ordered set, $\{\hat\tau_1, \ldots, \hat\tau_l, \tau_1^0, \ldots, \tau_{l^0}^0\}$, $\xi_0 = \tau_0^0 = -\infty$ and $\xi_{l+l^0+1} = \tau_{l^0+1}^0 = \infty$. Using an argument similar to that in (i) and noting that $S_n(\cdot)$ is the residual sum of squares, we can show that

$$\begin{aligned}\tilde\epsilon_n'\tilde\epsilon_n \ge S_n(\tau_1^0, \ldots, \tau_{l^0}^0) \ge\ & n\hat\sigma_l^2 \ge\ S_n(\tau_1^0, \ldots, \tau_{l^0}^0, \hat\tau_1, \ldots, \hat\tau_l) \\ &= \tilde\epsilon_n'\tilde\epsilon_n - \sum_{j=1}^{l^0+1} \sum_{\tau_{j-1}^0 \le \xi_{k-1} < \xi_k \le \tau_j^0} T_n(\xi_{k-1}, \xi_k) = \tilde\epsilon_n'\tilde\epsilon_n + O_p(\ln^2(n)).\end{aligned}$$

This proves (ii).

Proof of Theorem 4.1. It follows from Lemma 5.4 (i) that $P\{\hat l \ge l^0\} \to 1$ as $n \to \infty$. By Lemma 5.4 (ii) and the law of large numbers, for $l^0 < l \le L$,

$$0 \ge [\hat\sigma_l^2 - \frac{1}{n}\tilde\epsilon_n'\tilde\epsilon_n] - [\hat\sigma_{l^0}^2 - \frac{1}{n}\tilde\epsilon_n'\tilde\epsilon_n] = O_p(\ln^2 n/n),$$

and

$$[\hat\sigma_{l^0}^2 - \sigma_0^2] = [\hat\sigma_{l^0}^2 - \tilde\epsilon_n'\tilde\epsilon_n/n] + [\tilde\epsilon_n'\tilde\epsilon_n/n - \sigma_0^2] = O_p(\ln^2 n/n) + o_p(1) = o_p(1).$$

Hence $0 \le (\hat\sigma_{l^0}^2 - \hat\sigma_l^2)/\hat\sigma_{l^0}^2 = O_p(\ln^2(n)/n)$. Note that for $0 \le x < 1/2$, $\ln(1-x) \ge -2x$. Therefore,

$$\begin{aligned}MIC(l) - MIC(l^0) &= \ln(\hat\sigma_l^2) - \ln(\hat\sigma_{l^0}^2) + c_0(l-l^0)(\ln n)^{2+\delta_0}/n \\ &= \ln(1 - (\hat\sigma_{l^0}^2 - \hat\sigma_l^2)/\hat\sigma_{l^0}^2) + c_0(l-l^0)(\ln(n))^{2+\delta_0}/n \\ &\ge -2O_p(\ln^2(n)/n) + c_0(l-l^0)(\ln(n))^{2+\delta_0}/n > 0\end{aligned}$$

for sufficiently large n; so $\hat l \xrightarrow{p} l^0$ as $n \to \infty$.

The proof of Theorem 4.2 will be delayed until after the proof of Theorem 4.3. To simplify the statement of required preliminary results, let $R_j = (\tau_{j-1}^0, \tau_j^0]$, $\hat R_j = (\hat\tau_{j-1}, \hat\tau_j]$, $\tau_0^0 = \hat\tau_0 = -\infty$, $\tau_{l^0+1}^0 = \hat\tau_{l^0+1} = \infty$, and $\Delta_{nj} = |\hat\tau_j - \tau_j^0| = O_p(a_n)$, $j = 1, \ldots, l^0+1$, where $\{a_n\}$ is a sequence of positive numbers.

Lemma 5.5. *Suppose that the assumptions of Theorem* 4.2 *except Assumption* 4.2′ *are satisfied and that* $\{(Z_t, x_{td})\}$ *is a strictly stationary, ergodic sequence. If for some* $u > 1$, $E|Z_1|^u < \infty$, *then* $\frac{1}{n}\sum_{t=1}^n |Z_t\|\mathbf{1}_{(x_{td} \in \hat R_j)} - \mathbf{1}_{(x_{td} \in R_j)}| = O_p(a_n^{1/v})$, *where* $1/v = 1 - 1/u$.

Proof. It suffices to show that for every j,

$$\frac{1}{n}\sum_{t=1}^n |Z_t|\mathbf{1}_{(|x_{td} - \tau_j^0| < \Delta_{nj})} = O_p(a_n^{1/v}).$$

By assumption, $\Delta_{nj} = O_p(a_n)$. So for all $\epsilon > 0$ there exists $M > 0$ such that $P(\Delta_{nj} > a_n M) < \epsilon$ for all n. Thus

$$\begin{aligned} & P(\frac{1}{n}\sum_{t=1}^{n}|Z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{nj})} > a_n^{1/v}M) \\ = \; & P(\frac{1}{n}\sum_{t=1}^{n}|Z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{nj})} > a_n^{1/v}M, \Delta_{nj} \le a_n M) \\ & + P(\frac{1}{n}\sum_{t=1}^{n}|Z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{nj})} > a_n^{1/v}M, \Delta_{nj} > a_n M) \\ \le \; & P(\frac{1}{n}\sum_{t=1}^{n}|Z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<a_n M)} > a_n^{1/v}M) + \epsilon. \end{aligned}$$

Hence it remains only to prove that $S^* := a_n^{-1/v} n^{-1} \sum_{t=1}^{n} |Z_t| \mathbf{1}_{(|x_{td}-\tau_j^0|<a_n M)}$ is bounded in probability. However Hölder's inequality and the assumptions imply that the expected value of this last quantity is bounded above by $(E|Z_1|^u)^{1/u} a_n^{-1/v} (C a_n M)^{1/v}$ for some constant C. This shows that S^* is bounded in L^1 and hence in probability, so the proof is complete.

Proposition 5.1. *Assume for the segmented linear regression model* (2.2) *that the assumptions of Theorem* 4.2 *except Assumption* 4.2′ *hold. Then*

$$\hat{\tau} - \tau^0 = o_p(1),$$

where $\tau^0 = (\tau_1^0, \ldots, \tau_{l^0}^0)$ *and* $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{l}})$ *is the least squares estimator of* τ^0 *based on* $l = \hat{l}$, $\hat{l}$ *being a minimizer of* $MIC(l)$ *subject to* $l \le L$.

Proof. By Theorem 4.1, the problem can be restricted to $\{\hat{l} = l^0\}$. For any sufficiently small $\delta' > 0$, substituting δ' for the δ in (5.5) in the proof of Lemma 5.4 (i), we get the following inequality

$$\begin{aligned} \frac{1}{n} S_n(\tau_1, \ldots, \tau_{l^0}) \ge \; & \frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n + O_p(\ln^2(n)/n) \\ & + \frac{1}{n}[S_n(\tau_r^0 - \delta', \tau_r^0 + \delta') - S_n(\tau_r^0 - \delta', \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta')], \end{aligned}$$

uniformly in $(\tau_1, \ldots, \tau_{l^0}) \in A_r := \{(\tau_1, \ldots, \tau_{l^0}) : |\tau_s - \tau_r^0| > \delta', 1 \le s \le l^0\}$. By Lemma 5.3, the last term on the RHS converges to a positive C_r. For sufficiently large n, this C_r will dominate the term $O_p(\ln^2 n/n)$. Thus, uniformly in A_r, $r = 1, \ldots, l^0$, and with probability tending to 1,

$$\frac{1}{n} S_n(\tau_1, \ldots, \tau_{l^0}) > \frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n + \frac{C_r}{2}.$$

This implies that, with probability approaching 1, no τ in A_r is qualified as a candidate for the role of $\hat{\tau}$, where $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{l^0})$. In other words, $P(\hat{\tau} \in A_r^c) \to 1$ as $n \to \infty$. Since this is true for all r, $P(\hat{\tau} \in \bigcap_{r=1}^{l^0} A_r^c) \to 1$, as $n \to \infty$. Note that for $\delta' \leq \min_{0 \leq i \leq l^0}\{(\tau_{i+1}^0 - \tau_i^0)/2\}$,

$$\bigcap_{r=1}^{l^0}\{|\hat{\tau}_r - \tau_r^0| \leq \delta'\} = \bigcap_{r=1}^{l^0}\{|\hat{\tau}_{i_r} - \tau_r^0| \leq \delta', \text{for some } 1 \leq i_r \leq l^0\} = \{\hat{\tau} \in \bigcap_{r=1}^{l^0} A_r^c\}.$$

Thus we have,

$$P(|\hat{\tau}_r - \tau_r^0| \leq \delta' \text{ for } r = 1, \ldots, l^0) = P(\hat{\tau} \in \bigcap_{r=1}^{l^0} A_r^c) \to 1, \text{ as } n \to \infty,$$

which completes the proof.

Proof of Theorem 4.3. Let $(\tilde{\beta}_1^*, \ldots, \tilde{\beta}_{l^0+1}^*)$ and σ^{2*} be the least squares estimates of $(\tilde{\beta}_1, \ldots, \tilde{\beta}_{l^0+1})$ and σ_0^2 when l^0 as well as $(\tau_1^0, \ldots, \tau_{l^0}^0)$ are assumed known. Then it is clear that $\sqrt{n}[(\tilde{\beta}_1^{*'}, \ldots, \tilde{\beta}_{l^0+1}^{*'})' - (\tilde{\beta}_1', \ldots, \tilde{\beta}_{l^0+1}')']$ and $\sqrt{n}[\sigma^{2*} - \sigma_0^2]$ converge in distribution to normal distributions. So it suffices to show that $\tilde{\beta}_j^* - \hat{\tilde{\beta}}_j = o_p(n^{-1/2})$ and $\sigma^{2*} - \hat{\sigma}^2 = o_p(n^{-1/2})$.

Set for say, $R_j = (\tau_{j-1}^0, \tau_j^0]$, $X_j^* = I_n(R_j)X_n = I_n(\tau_{j-1}^0, \tau_j^0)X_n$ and $\hat{X}_j = I_n(\hat{R}_j)X_n$. Then, with probability tending to 1 as n tends to infinity,

$$\begin{aligned}
&\tilde{\beta}_j^* - \hat{\tilde{\beta}}_j \\
&= [(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}\hat{X}_j'Y_n] + [(\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
&=: (I)(II) + (III)(IV).
\end{aligned}$$

By the law of large numbers, both (II) and (III) are $O_p(1)$; and the order of $o_p(n^{-1/2})$ for (I) and (IV) follows from Lemma 5.5 by taking $Z_t = (\mathbf{a}'\mathbf{x}_t)^2$ for any real vector $\mathbf{a}$ and $u > 2$.

Similarly, we can show that $\sigma^{2*} - \hat{\sigma}^2 = o_p(n^{-1/2})$. This completes the proof.

Proof of Theorem 4.2. The already proven result that $\hat{\tilde{\tau}} \to \tilde{\tau}$ implies we need only consider those τ's in the neighborhoods of the true threshold parameters the τ^0's. For simplicity, we may assume without loss of generality that $l^0 = 1$. Furthermore, we assume that $d = 1$ and $p \geq d = 1$. Also, for notational convenience, we let $\tau^0 = {\tau_1}^0 = 0$.

By combining Proposition 5.1 and Theorem 4.3, we get $\hat{\tau} \xrightarrow{p} \tau^0$, $\hat{\tilde{\beta}}_1 \xrightarrow{p} \tilde{\beta}_1^0$, and $\hat{\tilde{\beta}}_2 \xrightarrow{p} \tilde{\beta}_2^0$. So it suffices to consider those $\tau' s$ and $\tilde{\beta}_j$, $j = 1, 2$, satisfying $(\tau, \tilde{\beta}_1, \tilde{\beta}_2) \in \omega(\Delta)$, where $\Delta > 0$, and

$$\omega(\Delta) = \{(\tau, \tilde{\beta}_1, \tilde{\beta}_2) : \ \|\tilde{\beta}_j - \tilde{\beta}_j^0\| < \Delta, j = 1, 2, |\tau - \tau^0| < \Delta\}.$$

Claim 1. For any $\epsilon > 0$ there exist $K > 0$ and $\Delta > 0$, such that for all sufficiently large n

$$P\{S_n^*(\tau, \tilde{\beta}_1, \tilde{\beta}_2) - S_n^*(\tau^0, \tilde{\beta}_1, \tilde{\beta}_2)\} > 1 - \epsilon,$$

uniformly in $(\tau, \tilde{\beta}_1, \tilde{\beta}_2) \in \omega(\Delta)$ and $|\tau| > K/n$ where

$$S_n^*(\tau, \tilde{\beta}_1, \tilde{\beta}_2) := S_n(-\infty, \tau; \tilde{\beta}_1) - S_n(\tau, \infty; \tilde{\beta}_2),$$

$$S_n(\alpha, \eta; \tilde{\beta}) = \sum_{t=1}^{n} (Y_t - \mathbf{x}_t{}'\tilde{\beta})^2 \mathbf{1}_{(x_{t1} \in (\alpha, \eta))}.$$

To this end, we note that

$$\begin{aligned}
&S_n^*(\tau, \tilde{\beta}_1, \tilde{\beta}_2) - S_n^*(\tau^0, \tilde{\beta}_1, \tilde{\beta}_2) \\
&= S_n(\tau^0, \tau; \tilde{\beta}_1) - S_n(\tau^0, \tau; \tilde{\beta}_2) \\
&= \sum_{t=1}^{n} [(\mathbf{x}_t{}'\tilde{\beta}_2^0 + \epsilon_t - \mathbf{x}_t{}'\tilde{\beta}_1)^2 - (\mathbf{x}_t{}'\tilde{\beta}_2^0 + \epsilon_t - \mathbf{x}_t{}'\tilde{\beta}_2)^2] \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} \\
&= 2\sum_{t=1}^{n} \epsilon_t \mathbf{x}_t{}'(\tilde{\beta}_2 - \tilde{\beta}_1) \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} + \sum_{t=1}^{n} \{\mathbf{x}_t{}'(2\tilde{\beta}_2^0 - \tilde{\beta}_1 - \tilde{\beta}_2) \mathbf{x}_t{}'(\tilde{\beta}_2 - \tilde{\beta}_1)\} \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} \\
&= (I) + (II). \qquad (5.6)
\end{aligned}$$

In the following, we show that after appropriate renormalization, (I) is arbitrarily small while (II) is positive.

By choosing Δ sufficiently small and the Law of Large Numbers we get

$$\begin{aligned}
\frac{1}{n}(II) &= \frac{1}{n} \sum_{t=1}^{n} [\mathbf{x}_t'(\tilde{\beta}_2^0 - \tilde{\beta}_1^0)]^2 \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} + O(\Delta) \cdot \frac{1}{n} \sum_{t=1}^{n} \|\mathbf{x}_t\|^2 \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} \\
&\geq \frac{\eta^2}{n} \sum_{t=1}^{n} \mathbf{1}_{(|\mathbf{x}_t'(\tilde{\beta}_2^0 - \tilde{\beta}_1^0)| > \eta)} \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} + O(\Delta) \cdot \frac{1}{n} \sum_{t=1}^{n} \|\mathbf{x}_t\|^2 \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))}, \qquad (5.7)
\end{aligned}$$

where $O(\Delta)$ denotes the term of the same order as Δ for small Δ.

Set $\bar{\tilde{\beta}}^0 = 2(\tilde{\beta}_2^0 - \tilde{\beta}_1^0)$. Then the first term of (5.6) can be written as

$$\frac{1}{n}(I) = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))} \mathbf{x}_t' \bar{\tilde{\beta}}^0 + O(\Delta) \cdot \frac{1}{n} \sum_{t=1}^{n} |\epsilon_t| \|\mathbf{x}_t\| \mathbf{1}_{(x_{t1} \in (\tau^0, \tau))}. \qquad (5.8)$$

For any $\eta > 0$ let:

$$\begin{aligned}
&\omega(\eta, \tau) := \{(1, \mathbf{x}')' : \ x_1 \in [\tau^0, \tau]\} \bigcap \{(1, \mathbf{x}')' : \ |(1, \mathbf{x}')'(\tilde{\beta}_2^0 - \tilde{\beta}_1^0)| > \eta\}; \\
&Q(\tau) = E\{\mathbf{1}_{(x_{t1} \in (\tau^0, \tau))}\}; \ \tilde{Q}(\tau) = E\{\mathbf{1}_{(\mathbf{x}_t \in \omega(\eta, \tau))}\}. \qquad (5.9)
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{S_n(\tau,\tilde\beta_1,\tilde\beta_2)-S_n(\tau^0,\tilde\beta_1,\tilde\beta_2)}{nQ(\tau)}\\
&\geq \eta^2\frac{1}{nQ(\tau)}\sum_{t=1}^{n}\mathbf{1}_{(\mathbf{x}_t\in\omega(\eta,\tau))}+\frac{1}{nQ(\tau)}\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t'\bar{\tilde\beta}^0\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}\\
&\quad+O(\Delta)\frac{1}{nQ(\tau)}\sum_{t=1}^{n}[\|\mathbf{x}_t\|^2+|\epsilon_t|\|\mathbf{x}_t\|]\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}. \qquad (5.10)
\end{aligned}$$

It now suffices to show that on the RHS of (5.10), the first term is positive while the second and the third are $o_p(1)$ and $O_p(\Delta)$, respectively, uniformly in $\tilde\beta' s$.

Claim 2. For any $\epsilon>0$ and $\gamma>0$, there exists $K>0$ such that for any $n>0$:

(a) $$P\Big\{\sup_{K/n<\tau-\tau^0\leq\Delta}[|\frac{1}{n\tilde Q(\tau)}\sum_{t=1}^{n}\mathbf{1}_{(\mathbf{x}_t\in\omega(\eta,\tau))}-1|]<\gamma\Big\}>1-\epsilon;$$

(b) $$P\Big\{\sup_{K/n<\tau-\tau^0\leq\Delta}[|\frac{1}{nQ(\tau)}\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t'\bar{\tilde\beta}^0\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}|]<\gamma\Big\}>1-\epsilon;$$

(c) $$\sup_{K/n<\tau-\tau^0\leq\Delta}\Big\{\frac{1}{nQ(\tau)}\sum_{t=1}^{n}[\|\mathbf{x}_t\|^2+|\epsilon_t|\|\mathbf{x}_t\|]\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}\Big\}=O_p(1);$$

(d) $$\inf_{K/n<\tau-\tau^0\leq\Delta}\Big\{\frac{\tilde Q(\tau)}{Q(\tau)}\Big\}>0.$$

Clearly if Claim 2 holds, for arbitrarily small $\epsilon>0$ the RHS of (5.10) is positive with probability exceeding $1-\epsilon$ uniformly in $(\tau,\tilde\beta_1,\tilde\beta_2)\in\omega(\Delta)$.

Proof of Claim 2. (d) Assumption 4.2′ tells us that for some $\eta>0$, there exists a $\delta>0$ such that

$$\delta^*=\inf_{|x-\tau^0|<\delta}[P\{|\mathbf{x}_t'(\tilde\beta_2^0-\tilde\beta_1^0)|>\eta\mid x_{t1}=x\}]>0.$$

Hence

$$\begin{aligned}
\tilde Q(\tau)&=E[\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}\cdot\mathbf{1}_{(|\mathbf{x}_t'(\tilde\beta_2^0-\tilde\beta_1^0)|>\eta)}]\\
&=E\{E[\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}\mathbf{1}_{(|\mathbf{x}_t'(\tilde\beta_2^0-\tilde\beta_1^0)|>\eta)}\mid x_{t1}]\}\\
&\geq E[\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}\inf_{|x_{t1}-\tau^0|<\delta}[P\{|\mathbf{x}_t'(\tilde\beta_2^0-\tilde\beta_1^0)|>\eta\mid x_{t1}\}]]\\
&\geq\delta^*E[\mathbf{1}_{(x_{t1}\in(\tau^0,\tau))}]=\delta^*Q(\tau),
\end{aligned}$$

where we have assumed that for the constant provided by Claim 1, $\Delta \in (0, \delta)$ (otherwise we can select a new smaller Δ with that property). Thus

$$\inf_{K/n<\tau-\tau^0\le\Delta} \Big\{\frac{\tilde{Q}(\tau)}{Q(\tau)}\Big\} > \delta^* > 0.$$

By Assumption 4.2,

$$Q(\tau) = \int_{\tau^0}^{\tau} f_1(x)dx = f_1(\xi)(\tau - \tau^0), \quad \xi \in (\tau^0, \tau),$$

which implies that $Q(\tau)$ is equivalent to $\tau - \tau^0$ or simply τ when $\tau^0 = 0$ is assumed. Thus we may assume without loss of generality that $Q(\tau) = \tau$, at least for small τ.

As for $\tilde{Q}(\tau)$, we note that for any $\tau > 0$, $b > 1$ with $b\tau < \Delta$,

$$|\frac{\tilde{Q}(b\tau)}{\tilde{Q}(\tau)} - 1| = \frac{|\tilde{Q}(b\tau) - \tilde{Q}(\tau)|}{\tilde{Q}(b\tau)} \le \frac{E[\mathbf{1}_{(x_{t1}\in[\tau,b\tau])}]}{\delta^* Q(\tau)} \le \frac{K^*(b-1)\tau}{\delta^*\tau} =: K^*(b-1), \tag{5.11}$$

where the constant K^* used here and in the sequel does not depend on τ.

Now let:

$$\begin{aligned}
\tilde{Q}_n(\tau) &= \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{(\mathbf{x}_t\in\omega(\eta,\tau))};\\
Q_n(\tau) &= \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{(x_{t1}\in(0,\tau))};\\
R_n(\tau) &= \frac{1}{n}\sum_{t=1}^{n} [\|\mathbf{x}_t\|^2 + |\epsilon_t|\|\mathbf{x}_t\|]\mathbf{1}_{(x_{t1}\in(0,\tau))};\\
R(\tau) &= E[R_n(\tau)].
\end{aligned} \tag{5.12}$$

We now need to replace the interval $K/n < \tau < \Delta$ by a countable subset $\{z_i = b^i K/n,\ i = 0, 1, \ldots\}$. To that end let $H(\cdot)$ and $H_n(\cdot)$ generically represent any member of $\{\tilde{Q}(\cdot), Q(\cdot),\ R(\cdot)\}$ and $\{\tilde{Q}_n(\cdot), Q_n(\cdot), R_n(\cdot)\}$, respectively. Note that for $0 < x \le y \le bx \le \Delta$, when

$$|\frac{H_n(x)}{H(x)} - 1| < \gamma \ \text{ and } \ |\frac{H_n(bx)}{H(bx)} - 1| < \gamma,$$

we have

$$-O(|b-1|) - \gamma \le \frac{H_n(y)}{H(y)} - 1 \le O(|b-1|) + \gamma, \tag{5.13}$$

where $O(|b-1|) \to 0$ as $b \to 1$.

To verify (5.13), noting that both $H(\cdot)$ and $H_n(\cdot)$ are non-decreasing functions of τ we have

$$\frac{H_n(y)}{H(y)} - 1 \le \frac{H_n(bx)}{H(x)} - 1 = \frac{H(bx)}{H(x)} \cdot \frac{H_n(bx)}{H(bx)} - 1.$$

By an argument similar to that used for (5.11) we can show that the last quanitity obtained just above cannot exceed

$$\begin{aligned} &|\frac{H(bx)}{H(x)} - 1| \cdot (1+\gamma) + \gamma \\ &\le O(|b-1|)(1+\gamma) + \gamma = O(|b-1|) + \gamma, \end{aligned} \tag{5.14}$$

while

$$\frac{H_n(y)}{H(y)} - 1 \ge \frac{H_n(x)}{H(bx)} - 1 = \frac{H(x)}{H(bx)} \cdot \frac{H_n(x)}{H(x)} - 1 \ge -O(|b-1|) - \gamma. \tag{5.15}$$

To establish (a) and (c) of Claim 2 by taking (5.13) into account, we need in addition that for any $\epsilon > 0$ and $\gamma > 0$, there exist $b > 1$ and $K > 0$, such that

$$P\Big\{\sup_{i\ge 0} |\frac{H_n(b^i K/n)}{H(b^i K/n)} - 1| > \gamma\Big\} \le \sum_{j=0}^{\infty} \frac{K^*}{b^j K} = \frac{b \cdot K^*}{K(b-1)} < \epsilon. \tag{5.16}$$

To prove (5.16), observe that its LHS cannot exceed

$$\begin{aligned} \sum_{i=0}^{\infty} P\Big\{|\frac{H_n(b^i K/n)}{H(b^i K/n)} - 1| > \gamma\Big\} &\le \sum_{i=0}^{\infty} \frac{1}{\gamma^2 n^2} \frac{\mathrm{Var}\,[nH_n(b^i K/n)]}{[H(b^i K/n)]^2} \\ &\le \sum_{i=0}^{\infty} \frac{K^*}{n} \frac{1}{H(b^i K/n)} \le \sum_{i=0}^{\infty} \frac{K^*}{K b^i}. \end{aligned}$$

It remains to verify (b). Let:

$$\begin{aligned} \tilde{U}_n(\alpha,\xi) &= \frac{1}{n}\sum_{t=1}^{n} |\epsilon_t||\mathbf{x}_t' \bar{\tilde{\beta}}^0| \mathbf{1}_{(x_{t1}\in(\alpha,\xi))}; \\ U_n(\tau) &= \frac{1}{n}\sum_{t=1}^{n} \epsilon_t(\mathbf{x}_t' \bar{\tilde{\beta}}^0) \mathbf{1}_{(x_{t1}\in(0,\tau))}; \\ \tilde{U}(\alpha,\xi) &= E[\tilde{U}_n(\alpha,\xi)]; \\ U(\tau) &= E[U_n(\tau)]; \end{aligned} \tag{5.17}$$

and $z_j = b^j K/n$, $j = 0, 1, \ldots$, $K > 0$, $b > 1$, $n \ge 1$. Observe that for $z_j < \tau \le z_{j+1}$,

$$\begin{aligned} |\frac{U_n(\tau)}{\tau}| \le |\frac{U_n(\tau)}{z_j}| \le \frac{|U_n(\tau) - U_n(z_j)|}{z_j} + |\frac{U_n(z_j)}{z_j}| &\le \frac{\tilde{U}_n(z_j,\tau)}{z_j} + |\frac{U_n(z_j)}{z_j}| \\ &\le \frac{\tilde{U}_n(z_j, z_{j+1})}{z_j} + |\frac{U_n(z_j)}{z_j}|. \end{aligned}$$

Hence,

$$\sup_{K/n<\tau\le\Delta}\Big\{|\frac{U_n(\tau)}{Q(\tau)}|\Big\} = \sup_{j\ge 0}\Big\{\sup_{K/n<\tau\le\Delta, z_j<\tau\le z_{j+1}}[|\frac{U_n(\tau)}{Q(\tau)}|]\Big\}$$
$$\le \sup_{j\ge 0, z_j\le\Delta}\Big\{\frac{\tilde{U}_n(z_j,z_{j+1})}{z_j}\Big\} + \sup_{j\ge 0, z_j\le\Delta}\Big\{|\frac{U_n(z_j)}{z_j}|\Big\}. \quad (5.18)$$

Thus, it suffices to show that the two terms on the RHS of (5.18) are of the order $o_p(1)$. To this end let $\gamma > 0$. Then

$$P\Big\{\sup_{j\ge 0}[|\frac{U_n(z_j)}{z_j}|] > \gamma\Big\} \le \sum_j P\Big\{|\frac{U_n(z_j)}{z_j}| > \gamma\Big\} \le \sum_j \frac{1}{\gamma^2 z_j^2}\mathrm{Var}\,[U_n(z_j)]$$
$$= \sum_j \frac{\sigma^2}{n\gamma^2 z_j^2} E[(\mathbf{x}_t'\bar{\bar{\beta}}^0)^2 \mathbf{1}_{(x_{t1}\in(0,z_j))}] \le \sum_j \frac{K^*}{nz_j^2} z_j \le \frac{b}{b-1}\frac{K^*}{K} \to 0, \quad (5.19)$$

as $K \to \infty$, by using Assumptions 4.2 and 4.2′.

To show that the first term of the RHS of (5.18) is $o_p(1)$, we need the following results:

$$\sup_{j\ge 0, z_j\le\Delta}\Big\{\frac{\tilde{U}(z_j,z_{j+1})}{z_j}\Big\} \le K^*(b-1), \quad (5.20)$$

and

$$\sup_{j\ge 0, z_j\le\Delta}\Big\{\frac{|\tilde{U}_n(z_j,z_{j+1}) - \tilde{U}(z_j,z_{j+1})|}{z_j}\Big\} = o_p(1). \quad (5.21)$$

We shall first establish (5.20). Observe that

$$\frac{\tilde{U}(z_j,z_{j+1})}{z_j} = \frac{K^*}{z_j} E[|\mathbf{x}_t'\bar{\bar{\beta}}^0|\mathbf{1}_{(x_{t1}\in(z_j,z_{j+1}))}]$$
$$\le \frac{K^*}{z_j} E\{E[|\mathbf{x}_t'\bar{\bar{\beta}}^0| \mid x_{t1}]\cdot \mathbf{1}_{(x_{t1}\in(z_j,z_{j+1}))}\}$$
$$\le \frac{K^*}{z_j} E\{\{E[|\mathbf{x}_t'\bar{\bar{\beta}}^0|^2 \mid x_{t1}]\}^{1/2}\cdot \mathbf{1}_{(x_{t1}\in(z_j,z_{j+1}))}\}$$
$$\le \frac{K^*}{z_j}(z_{j+1}-z_j) = K^*(b-1).$$

Finally, for (5.22), we note that for any $\gamma > 0$,

$$P\Big\{\sup_{j\ge 0, z_j\le\Delta}\{\frac{|\tilde{U}_n(z_j,z_{j+1}) - \tilde{U}(z_j,z_{j+1})|}{z_j}\} > \gamma\Big\}$$
$$\le \sum_j \frac{1}{\gamma^2 z_j^2}\mathrm{Var}\,[\tilde{U}_n(z_j,z_{j+1})]$$

$$= \sum_j \frac{K^*}{nz_j^2} \mathrm{Var}\left[|\epsilon_t||\mathbf{x}_t'\bar{\tilde{\beta}}^0|\mathbf{1}_{(x_{t1}\in(z_j,z_{j+1}))}\right]$$
$$= \sum_j \frac{K^*}{nz_j^2} E\{\{E[|\mathbf{x}_t'\bar{\tilde{\beta}}^0|^2 \mid x_{t1}]\} \cdot \mathbf{1}_{(x_{t1}\in(z_j,z_{j+1}))}\}$$
$$\leq \sum_j \frac{K^*}{nz_j^2}(z_{j+1} - z_j) \leq K^*/(Kb).$$

Thus, choosing K sufficiently large will make the RHS of this last term arbitrarily small. This completes the proof of Theorem 4.2.

6. Concluding Remarks

Section 2 offers a segmented regression methodology for fitting a set of independent variables to a response variable. Our method retains much of the simplicity of the linear model. And it gains some of the flexibility of a nonparametric model. Yet it avoids the difficulties faced by standard nonparametric methods in higher dimensions when the "curse of dimensionality"is encountered.

The method looks promising. This promise is indicated by the empirical results of Section 3 and the theoretical results indicated in Section 4 and proved in Section 5.

The number of pieces, l, for our piecewise model is chosen by the data. But an upper bound is imposed on l to ensure computational feasibility and to ensure sufficient data so that each piece of the model can be well estimated even when x is a vector of high dimension. And it is shown that our estimate of l is consistent under reasonably general conditions.

The data also select the subdomain boundaries for the piecewise linear model; the results are shown to be consistent. The rate of convergence of these boundary estimates is surprisingly fast at $(\log n)^2/n$.

Once the boundaries are estimated, conventional parameter estimates for parameters of the component linear models are used. Moreover, their asymptotic theory is the same of that of conventional (global) linear models, permitting the calculation of confidence intervals and so on. However, we do not have at this time the asymptotic distributions of the threshold estimates and estimator of the number of subdomains.

The constant c_0 in display (2.3) is chosen in accordance with the guidelines in Section 2. It allows us, even in small samples, to adhere to the philosophy underlying our methodology: there is an underlying function of x which is being approximated differently in different subdomains by the leading terms of different Taylor expansions and hence linear models. At the same time, our choice allows for the possibility of choosing the quadratic model and a single domain.

Work currently underway addresses some of the questions left open by this work, such as how to partition using more than one independent variable and how to deal with correlated errors.

References

Breiman, L. and Meisel, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71**, 301-307.

Dunicz, B. L. (1969). Discontinuities in the surface structure of alcohol-water mixtures. *Kolloid-Zeitschr. u. Zeitschrift f. Polymere* **230**, 346-357.

Feder, P. I. (1975). On asymptotic distribution theory in segmented regression problems – identified case. *Ann. Statist.* **3**, 49-83.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Ann. Statist.* **19**, 1-141.

Henderson, H. V. and Velleman, P. F. (1981). Building multiple regression model interactively. *Biometrics* **37**, 391-411.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56**, 495-504.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57**, 1-17.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Amer. Statist. Assoc.* **61**, 1097-1129.

McGee, V. E. and Carleton, W. T. (1970). Piecewise regression. *J. Amer. Statist. Assoc.* **65**, 1109-1124.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53**, 873-880.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *J. Amer. Statist. Assoc.* **67**, 306-310.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Sprent, P. (1961). Some hypotheses concerning two phase regression lines. *Biometrics* **17**, 634-645.

Wu, S. Y. (1992). Asymptotic inference of segmented regression models. *Ph.D. thesis, Department of Statistics, University of British Columbia*.

Yao, Y. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6**, 181-189.

Yao, Y. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser.* A **51**, 370-381.

Department of Statistics, University of British Columbia, 6356 Agriculture Road, Vancouver, British Columbia, Canada V6T 1Z2.

Statistics Canada, JRH Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

(Received February 1993; accepted March 1996)