Statistica Sinica **19** (2009), 597-609

SELECTING THE NUMBER OF CHANGE-POINTS IN SEGMENTED LINE REGRESSION

Hyune-Ju Kim, Binbing Yu and Eric J. Feuer

Syracuse University, National Institute of Aging, and National Cancer Institute

Abstract: Segmented line regression has been used in many applications, and the problem of estimating the number of change-points in segmented line regression has been discussed in Kim et al. (2000). This paper studies asymptotic properties of the number of change-points selected by the permutation procedure of Kim et al. (2000). This procedure is based on a sequential application of likelihood ratio type tests, and controls the over-fitting probability by its design. In this paper we show that, under some conditions, the number of change-points selected by the permutation procedure is consistent. Via simulations, the permutation procedure is compared with such information-based criterior as the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), and Generalized Cross Validation (GCV).

Key words and phrases: Change-points, model selection, permutation test, segmented line regression.

1. Introduction

The problem of model selection in linear regression has long been of interest to both applied and theoretical statisticians. Much of the literature is concerned with the problem of determining the "best" subset of independent variables, and Hocking (1976) summarizes various selection criteria that can be classified into two major approaches: hypothesis testing and information criteria. Stepwise selection procedures based on the hypothesis testing approach sequentially apply classical F-tests that test whether some of the regression parameters are zero. The other approach utilizes information criteria such as the C_p criterion (Mallows (1973)), the Akaike information criterion (AIC; Akaike (1974)), and the Bayes information criterion (BIC; Schwarz (1978)) to choose a best model from a set of competing ones. These selection criteria are usually the sum of two measures: a measure of the goodness-of-fit of a candidate model and a measure of penalty for model complexity. As discussed in Burnham and Anderson (2002, Sec. 6.3), it is known in the literature that AIC tends to over-fit the true model and that BIC is consistent if there is a true model among the candidates. Zheng and Loh (1995)

proposed a new consistent criterion that generalizes these well-known criteria in the linear model case. Their criterion chooses the model dimension, κ, among M_n possible covariates as $\hat{\kappa} = \arg\min_{0 \le k \le M_n}\{RSS(k) + h_n(k)\hat{\sigma}^2\}$, where $RSS(k)$ is the residual sum of squares for the model with the first k independent variables, $\hat{\sigma}^2 = RSS(M_n)/(n - M_n)$, and $h_n(k)$ is a non-decreasing penalty function. When $h_n(k) = 2k$ or $k \ln n$, we obtain the AIC or BIC, respectively, and Zheng and Loh (1995) showed the consistency of $\hat{\kappa}$ for the BIC, under some conditions on M_n and $h_n(k)$. For additional details and recent developments in information theoretic criteria, and also for other approaches including those based on cross validation and Bootstrapping, see George (2000), Hastie, Tibshirani and Friedman (2001), Rao and Wu (2001) and Miller (2002).

In the context of change-point problems, Yao (1988) proposed to use the Schwarz criterion (BIC) to estimate the number of change-points in an independent normal sequence, and established its weak consistency. Liu, Wu and Zidek (1997) used a modified Schwartz criterion to estimate the number of change-points in segmented multivariate regression. The approach of Liu et al. can be applied to segmented linear models with or without the continuity constraints at the change-points, and the estimated number of change-points is weakly consistent under some conditions. For multiple structural change models, Bai and Perron (1998, 2003) proposed a sequential method to estimate the number of change-points, studied asymptotic properties of the least square estimate, and compared its performance with those based on other criteria. Recently, Tiwari, Cronin, Davies, Feuer, Yu and Chib (2005) developed Bayesian procedures to select a segmented line regression model among a set of candidate models, and compared its performance with the performance of the permutation procedure of Kim, Fay, Feuer and Midthune (2000). Also in the context of multivariate adaptive splines regression (MARS), Friedman (1991) used a selection criterion based on generalized cross validation (GCV) to estimate the number of knots.

Kim et al. (2000) proposed a series of permutation tests to determine the unknown number of change-points in segmented line regression; this tends to be conservative, from the nature of hypothesis testing. In choosing a model between the one with i change-points and the alternative with j change-points ($i < j$), the model with i change-points is selected against the model with j change-points if

$$RSS(i) \le (1 + c_n(i, j; \alpha))RSS(j) = \hat{\sigma}_j^2 h_n(i, j), \tag{1.1}$$

where $\hat{\sigma}_j^2 = RSS(j)/(n - 2j - 2)$ and $c_n(i, j; \alpha)$ is a critical value obtained under the null model with i change-points. The procedure of Kim et al. estimates

$c_n(i, j; \alpha)$, equivalently the p-value of the test, by using the permutation distribution of the test statistic, motivated by the non-conventional asymptotic behavior of the test statistic, and sequentially conducts a series of permutation tests to a conclusion. Although there is a similarity between the formulation of $\hat{\kappa}$ of Zheng and Loh (1995) and (1.1), they are not directly comparable since $h_n(i, j)$ in (1.1) depends on the alternative model as well, and the procedure requires a series of tests to be conducted sequentially.

Our aim in this paper is to study asymptotic properties of the number of change-points selected by the permutation procedure, and to compare the permutation procedure with information-based-criteria. In Section 2, we review asymptotic results in segmented line regression. In Section 3, we review the permutation procedure of Kim et al. (2000) and we prove that the number of change-points selected by the permutation procedure converges to the true number of change-points as the sample size increases. Section 4 discusses some generalizations of the main results in Section 3, including their extension to multiple regression and modified significance levels. In Section 5, we review some information-based criteria such as BIC, AIC and GCV, and these criteria are compared to the permutation procedure via simulations. Concluding remarks are made in Section 6.

2. Review of Asymptotic Results in Segmented Line Regression

Consider the segmented line regression model, $y_i = \beta_0 + \beta_1 x_i + \delta_1 (x_i - \tau_1)^+ + \cdots + \delta_\kappa (x_i - \tau_\kappa)^+ + \epsilon_i$, where τ_j's $(j = 1, \ldots, \kappa)$ are the unknown change-points, β_j's $(j = 0, 1)$ and δ_j's $(j = 1, \ldots, \kappa)$ are the regression parameters, and $a^+ = a$ for $a > 0$ and 0 otherwise. We assume that the $y_i's$ are independently distributed with constant variance σ_0^2. When κ is known, say k_0, asymptotic properties of the least square estimates of $(\boldsymbol{\beta}^T, \boldsymbol{\tau}^T) = (\beta_0, \beta_1, \delta_1, \ldots, \delta_{k_0}, \tau_1, \ldots, \tau_{k_0})$ are studied in Feder (1975) in great detail. Under some conditions about the mean functions associated with design points $x_1, \ldots, x_n$, Feder showed the consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\tau}}$, and proved the asymptotic normality of these estimators. Feder (1975) considered only one predictor variable, but allowed a general mean function $\mu(x) = \sum_{k=1}^{k_0+1} f_k(x) I_{[\tau_{k-1}, \tau_k)}(x)$, where $I_A(x) = 1$ if $x \in A$ and 0 otherwise, with $\tau_0 = -\infty$, $\tau_{k_0+1} = \infty$. To prove the consistency and asymptotic normality of the estimators, Feder (1975) removed observations in a small neighborhood of the true change-points to generate the pseudo-sample, and established the consistency and asymptotic normality of the estimators based on the pseudo-sample under some conditions on the mean functions. Then, the asymptotic equivalence between the pseudo-sample estimators and the full sample estimators was established to obtain the desired results. Special applications of Feder's results are found in Hinkley (1971) and Hušková (1998). For

$k_0 = 1$, Hinkley (1971) studied asymptotic behavior of the maximum likelihood estimators under normal theory, and provided the asymptotic variances of the estimated regression slopes and $\hat{\tau}_1$. Hušková (1998) considered a two-phase segmented line regression model with zero slope in the first phase and equally spaced design points, finding the asymptotic distribution of $\hat{\beta}$, the estimated slope of the second phase, and of $\hat{\tau}_1$. Liu, Wu and Zidek (1997) considered a multiple regression model with p independent variables, one of which was the partitioning variable associated with the change-points. There the main goal was to prove the consistency of $\hat{\kappa}_L$ which is based on a modified BIC defined as $\hat{\kappa}_L = \arg\min_{0\le k\le M}\left\{\ln[RSS(k)/(n-p^*)] + p^* c_0 (\ln n)^{2+\delta_0}/n\right\}$, for some $c_0 > 0$, $\delta_0 > 0$ and a pre-determined value M, where $RSS(k)$ is the residual sum of squares for the model with k change-points, and $p^* = (k+1)p + k$. The criterion of Liu et al. uses a stronger penalty than that of Yao (1988), and this is justified for non-Gaussian models.

3. Permutation Test and Its Asymptotic Properties

3.1. Permutation test

Suppose that we have n-pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, and consider the model,

$$E(y|x) = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_\kappa (x - \tau_\kappa)^+,$$

with the unknown number of change-points, κ, and the change-points, τ_j's ($j = 1, \ldots, \kappa$). We assume that the $y_i's$ are independently distributed with constant variance. The problem of fitting the model with a given value of κ has been discussed in Kim et al. (2000), where Lerman's grid search (1980) is used to fit a segmented line regression model, and the permutation test is used to see if there is enough evidence to select a model with a larger number of change-points than the one in the null hypothesis. The permutation procedure begins by testing the hypotheses $H_0 : \kappa = 0$ versus $H_1 : \kappa = M$, where M is a predetermined maximum number of change-points. If the null hypothesis is rejected, we test $H_0 : \kappa = 1$ versus $H_1 : \kappa = M$. Otherwise, we test the null hypothesis $\kappa = 0$ against the alternative hypothesis that $\kappa = M - 1$. We repeat this process until we test the null hypothesis of $\kappa = i$ against the alternative of $\kappa = i + 1$ for some $0 \le i < M$, and we denote the selected number of change-points as $\hat{\kappa}$, where $\hat{\kappa} = i + 1$ if we reject the null hypothesis in the last test, and $\hat{\kappa} = i$ otherwise. Kim et al. (2000) proposed to use α_0/M as a significance level at each stage of the permutation tests, arguing that the sequential application of

the proposed tests would maintain the overall significance level under α_0 and control the over-fitting probability, since $P(\hat{\kappa} > k^* | \kappa = k^*) \leq (1 - k^*/M)\alpha_0$ for $k^* = 0, \ldots, M-1$. Through simulation studies, they also indicated that the under-fitting probability, $P(\hat{\kappa} < k^* | \kappa = k^*)$, would be small if the procedure is powerful enough. In this paper, our goal is to prove that $\hat{\kappa}$ is consistent under conditions similar to those of Liu, Wu and Zidek (1997), Zheng and Loh (1995), and Feder (1975).

3.2. Asymptotic properties

First, we note that the procedure of Kim et al. (2000) estimates κ as j when we reject j of the null hypotheses and do not reject $M-j$ of the null hypotheses in conducting the M tests sequentially; also that there are $R_j = \binom{M}{j}$ sequences of such M tests where the j rejections and $M-j$ acceptances can occur. So,

$$P(\hat{\kappa} = j | \kappa = k^*) = \sum_{r=1}^{R_j} P(E_{j,r} | \kappa = k^*),$$

where $E_{j,r}$ is the r-th event where the j rejections and $M-j$ acceptances of the null hypotheses are observed. For notational simplicity, we denote the significance level of the individual permutation test α_0/M as α. Let $A_{k_0,k_1;\alpha}$ denote the event that $H_0 : \kappa = k_0$ is not rejected at level α against $H_1 : \kappa = k_1$, while $R_{k_0,k_1;\alpha}$ denotes the event that $H_0 : \kappa = k_0$ is rejected at level α in favor of $H_1 : \kappa = k_1$. Then we note that for a given value of k^* and $j < k^*$, $E_{j,r}$ occurs when $A_{k_0,k^*;\alpha}$ occurs for some $k_0 = 0, \ldots, j$, and furthermore $A_{k_0,k^*;\alpha}$ occurs d_{k_0} times where $d_{k_0} = d_{M,j,k^*}(k_0) = \binom{L_{k_0}-1}{k_0}\binom{M-L_{k_0}}{j-k_0}$ for $L_{k_0} = M - (k^* - k_0) + 1$. For example, with $M = 4$, $k^* = 4$ and $j = 2$, Figure 1 shows that $R_j = 6$:

$$\begin{aligned}
E_{2,1} &= R_{0,4;\alpha} \cap R_{1,4;\alpha} \cap \mathbf{A}_{2,4;\alpha} \cap A_{2,3;\alpha},\\
E_{2,2} &= R_{0,4;\alpha} \cap \mathbf{A}_{1,4;\alpha} \cap R_{1,3;\alpha} \cap A_{2,3;\alpha},\\
E_{2,3} &= R_{0,4;\alpha} \cap \mathbf{A}_{1,4;\alpha} \cap A_{1,3;\alpha} \cap R_{1,2;\alpha},\\
E_{2,4} &= \mathbf{A}_{0,4;\alpha} \cap R_{0,3;\alpha} \cap R_{1,3;\alpha} \cap A_{2,3;\alpha},\\
E_{2,5} &= \mathbf{A}_{0,4;\alpha} \cap R_{0,3;\alpha} \cap A_{1,3;\alpha} \cap R_{1,2;\alpha},\\
E_{2,6} &= \mathbf{A}_{0,4;\alpha} \cap A_{0,3;\alpha} \cap A_{0,2;\alpha} \cap R_{1,2;\alpha},
\end{aligned}$$

where the events $A_{k_0,k^*;\alpha}$ $(k_0 = 0, 1, 2)$ are denoted in bold, and $d_0 = \binom{0}{0}\binom{3}{2} = 3$, $d_1 = \binom{1}{1}\binom{2}{1} = 2$, and $d_2 = \binom{2}{2}\binom{1}{0} = 1$. In Figure 1, $R_{k_0,k_1;\alpha}$ and $A_{k_0,k_1;\alpha}$ are denoted as R_{k_0,k_1} and A_{k_0,k_1} for notational simplicity.

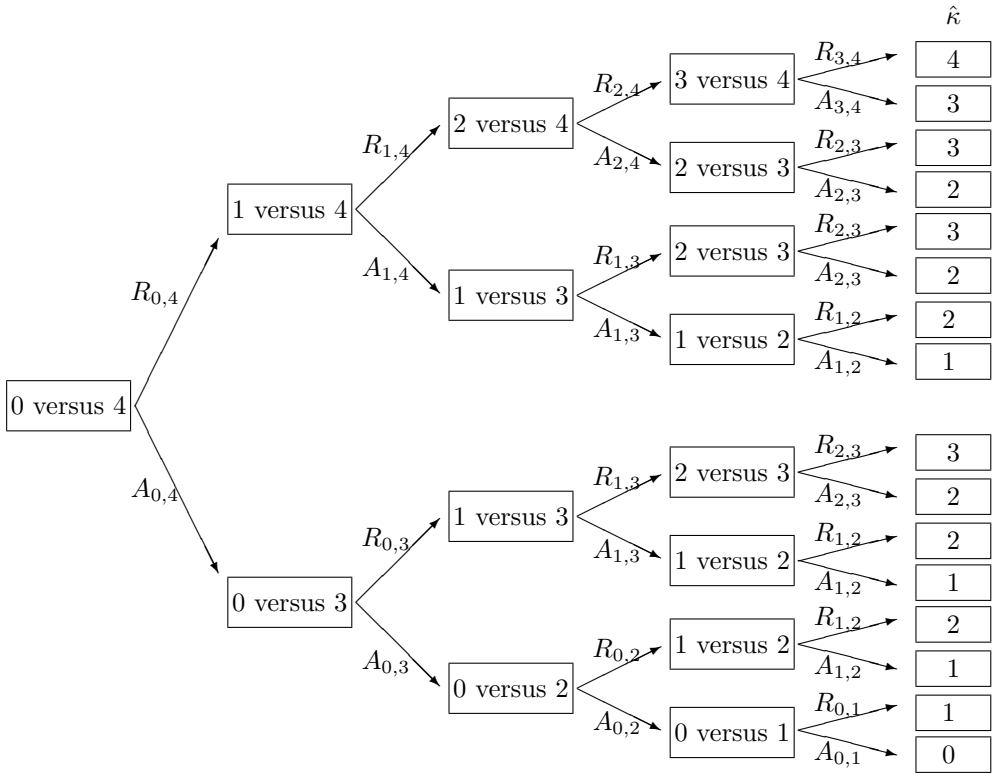

Figure 1. Permutation tests with $M_0 = 0$ and $M = 4$.

Then, for $j < k^*$,

$$P(\hat{\kappa} = j|\kappa = k^*) = \sum_{r=1}^{R_j} P(E_{j,r} \ |\kappa = k^*) \leq \sum_{k_0=0}^{j} d_{k_0} P(A_{k_0,k*;\alpha} \ |\kappa = k^*). \quad (3.1)$$

Similarly, we note that for $j > k^*$, $E_{j,r}$ occurs when $R_{k^*,k_1;\alpha}$ occurs for some $k_1 = j, \ldots, M$, and furthermore $R_{k^*,k_1;\alpha}$ occurs d_{k_1} times where $d_{k_1} = d_{M,j,k^*}(k_1) = \binom{L_{k_1}-1}{k^*}\binom{M-L_{k_1}}{j-k^*-1}$ for $L_{k_1} = M - (k_1 - k^*) + 1$. Thus, for $j > k^*$,

$$P(\hat{\kappa} = j|\kappa = k^*) = \sum_{r=1}^{R_j} P(E_{j,r} \ |\kappa = k^*) \leq \sum_{k_1=j}^{M} d_{k_1} P(R_{k^*,k_1;\alpha} \ |\kappa = k^*). \quad (3.2)$$

In order to show that $\hat{\kappa}$ is consistent, or that the right hand sides of the above inequalities converge to zero as n goes to infinity, we need some technical conditions that provide the consistency of the estimators under the true model with

k^* change-points. These include the conditions that the δ_i's $(i = 1, \ldots, k^*)$, are big enough and/or the number of observations between $[\hat{\tau}_i, \hat{\tau}_{i+1})$ $(i = 0, \ldots, k^*)$ is large enough. In this section, we first consider the case where M is fixed, and then we discuss the case where M depends on n. In order to avoid a lengthy introduction of notations and assumptions, and also since our main interest is on the analysis of cancer rates measured annually, we focus on the case with fixed independent variable, x, as in Assumption 4.1' of Liu, Wu and Zidek (1997). Note that the random x case can be handled by using an assumption similar to Assumption 4.1 of Liu, Wu and Zidek (1997).

Assumption 3.2.1.

(A1) $n^{-1}\sum_{i=1}^{n}(1, x_i)(1, x_i)^T I_{(\tau_j^*-\Delta, \tau_j^*]}(x_i)$ and $n^{-1}\sum_{i=1}^{n}(1, x_i)(1, x_i)^T I_{(\tau_j^*, \tau_j^*+\Delta]}(x_i)$ converge to positive definite real matrices for $\Delta \in (0, \min_{1\le j\le k^*}(\tau_{j+1}^* - \tau_j^*)/4)$.

(A2) The ϵ_i are independent and identically distributed with mean zero and variance σ_0^2, and for some constants B_0 and T_0 in $(0, \infty)$, $E(e^{t\epsilon_i}) \le e^{B_0 t^2}$ for all $|t| \le T_0$.

Theorem 3.2.1. *Suppose that Assumption* 3.2.1 *is satisfied and that M is fixed. Then $\hat{\kappa}$ converges to k^* in probability as $n \to \infty$.*

See Appendix A at `http://www.stat.sinica.edu.tw/statistica` for the proof.

In the remaining part of this section, we consider the case where M increases as n increases, as in Zheng and Loh (1995). When we allow M to depend on n, $M = M_n$, (A1) in Assumption 3.2.1 is not always satisfied, the significance level of each individual test depends on M, and the result on c in Lemma A.1 need not be satisfied. In this case, we assume conditions motivated by Feder (1975). Feder obtained the consistency and asymptotic normality of the estimators under technical assumptions on the spacings of the x-variable, and on the mean function, some of which can be simplified for a segmented line regression model with equally spaced design points.

We first introduce some notation. Without loss of generality, suppose the x_j's are scaled so that $x_j \in [0, 1]$, $\tau_0 = 0$, and $\tau_{k^*+1} = 1$. For observed $(x_1, \ldots, x_n)^T$ and $\boldsymbol{t}_k = (t_1, \ldots, t_i, t_{i+1}, \ldots, t_k)^T$ for any given k, let

$$X_k(\boldsymbol{t}_k) = \begin{pmatrix} 1 & x_1 & (x_1 - t_1)^+ & \cdots & (x_1 - t_i)^+ & (x_1 - t_{i+1})^+ & \cdots & (x_1 - t_k)^+ \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_n & (x_n - t_1)^+ & \cdots & (x_n - t_i)^+ & (x_n - t_{i+1})^+ & \cdots & (x_n - t_k)^+ \end{pmatrix}$$

and $H_k(\boldsymbol{t}_k) = X_k(\boldsymbol{t}_k)(X_k^T(\boldsymbol{t}_k)X_k(\boldsymbol{t}_k))^{-1}X_k^T(\boldsymbol{t}_k)$. For $i = 0, \ldots, k$, let $X_i(\boldsymbol{t}_k)$ denote the matrix composed of the first $i+2$ columns of $X_k(\boldsymbol{t}_k)$ and $H_i(\boldsymbol{t}_k) = X_i(\boldsymbol{t}_k)(X_i^T(\boldsymbol{t}_k)X_i(\boldsymbol{t}_k))^{-1}X_i^T(\boldsymbol{t}_k)$. For a true model with k^* change-points at $\boldsymbol{\tau}_{k^*} = (\tau_1, \ldots, \tau_{k^*})^T$ and $\boldsymbol{\beta}^* = (\beta_0, \beta_1, \delta_1, \ldots, \delta_{k^*})^T$, we define $\boldsymbol{\mu}^* = \boldsymbol{\mu}(\boldsymbol{\tau}_{k^*}) = E[\mathbf{y}|\kappa = k^*] = X_{k^*}(\boldsymbol{\tau}_{k^*})\boldsymbol{\beta}^*$ and $\eta_i = \boldsymbol{\mu}^{*T}(I - H_i(\boldsymbol{\tau}_{k^*}))\boldsymbol{\mu}^*$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$.

Assumption 3.2.2.

(C1) There are at least $n/\ln n$ observations in each segment of $[\hat{\tau}_j, \hat{\tau}_{j+1})$ and so of $[\tau_j, \tau_{j+1})$, for $j = 0, \ldots, k^*$.

(C2) The ϵ_i are independently and identically distributed with $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma_0^2$, and $E|\epsilon_i|^{2(1+\delta)} < \infty$ for some $\delta > 0$.

(C3) $\limsup_n M_n/n < 1$, and M_n is increasing slowly such that $\lim_{n\to\infty} M_n/\sqrt{\eta^*} = 0$, where $\eta^* = \boldsymbol{\mu}^{*T}(1 - H_{k^*-1}(\boldsymbol{\tau}_{k^*}))\boldsymbol{\mu}^*$.

Theorem 3.2.2. *If Assumption* 3.2.2 *holds, then $\hat{\kappa}$ converges to k^* in probability as $n \to \infty$.*

See Appendix B at `http://www.stat.sinica.edu.tw/statistica` for the proof.

Remark 1. If one starts the permutation procedure by testing $H_0 : \kappa = M_0$ versus $H_1 : \kappa = M$ for $0 < M_0 < M$, then $R_j = \binom{M-M_0}{j-M_0}$, $L_{k_0} = (M - M_0) - (k^* - k_0) + 1$, and $d_{k_0} = \binom{L_{k_0}-1}{k_0-M_0}\binom{M-M_0-L_{k_0}}{j-k_0}$ in (3.1).

Remark 2. Note that conditions (A1) and (A2) in Assumption 3.2.1 are the same ones as those used by Liu, Wu and Zidek (1997) to establish the consistency of the number of change-points estimated by a modified BIC. For an equally spaced x case, (A1) basically states that the number of observations in each segment is proportional to n, which does not allow the possibility of M increasing as a function of n. Sometimes in practice, however, it could be more reasonable to increase M as n increases. For such a case, (C3) in Assumption 3.2.2 states the condition on M as in Zheng and Loh (1995), and we use (C1) and (C2) for the conditions on the independent variable x and the error variable ϵ, as in Feder (1975).

4. Generalization

4.1. Multiple regression

In Section 3.2, we focused on the segmented line regression with one independent variable in order to make the presentation simple and also to make a

direct connection to our previous work: Kim et al. (2000), and Joinpoint software available at `http://srab.cancer.gov/joinpoint/index.html`. The main results can be generalized to cases where we have more than one covariate, as in Liu, Wu and Zidek (1997). As in Liu et al., we assume that there is a partitioning variable, say x_d, among p independent variables, $x_1, \ldots, x_p$, with which the change-points are defined, and we consider condition (A1) of Assumption 3.2.1 in the context of matrices. See Assumptions 4.1 and 4.1' in Liu et al. Then, it is straightforward to generalize the results of Theorem 3.2.1.

4.2. Modified procedure

The idea behind using the significance level α_0/M for each permutation test is to control the over-fitting probability under α_0. We make a modification that produces a more powerful procedure. Let κ denote the true number of change-point, $\hat{\kappa}$ denote the number of change-points selected by the permutation procedure, and consider the example in Figure 1 that starts with testing the null hypothesis of zero change-points versus the alternative of $M = 4$ change-points. With a significance level of $\alpha_j(k_0, k_1)$ at the j-th stage ($j = 1, \ldots, M$) to test the null hypothesis of k_0 change-points versus the alternative hypothesis of k_1 change-points, the procedure has the following property:

$$\begin{aligned}
&\Pr(\hat{\kappa} > 0 | \kappa = 0) \le \alpha_1(0, M) + \alpha_2(0, M-1) + \alpha_3(0, M-2) + \cdots + \alpha_M(0, 1),\\
&\Pr(\hat{\kappa} > 1 | \kappa = 1) \le \alpha_2(1, M) + \alpha_3(1, M-1) + \cdots + \alpha_M(1, 2),\\
&\Pr(\hat{\kappa} > 2 | \kappa = 2) \le \alpha_3(2, M) + \cdots + \alpha_M(2, 3),\\
&\vdots\\
&\Pr(\hat{\kappa} > M-1 | \kappa = M-1) \le \alpha_M(M-1, M).
\end{aligned}$$

If we like to bound these over-fitting probabilities by α, then we can assign different values for each α_j, for example

$$\begin{aligned}
&\alpha_1(0, M) = \alpha_2(0, M-1) = \alpha_3(0, M-2) = \cdots = \alpha_M(0, 1) = \frac{\alpha_0}{M},\\
&\alpha_2(1, M) = \alpha_3(1, M-1) = \cdots = \alpha_M(1, 2) = \frac{\alpha_0}{M-1},\\
&\alpha_3(2, M) = \cdots = \alpha_M(2, 3) = \frac{\alpha_0}{M-2},\\
&\vdots\\
&\alpha_M(M-1, M) = \alpha_0.
\end{aligned}$$

This modification ensures higher power of the overall test, and the main results in the previous section also hold.

5. Other Criteria

The problem of determining the number of unknown change-points in segmented line regression shares some similarity with the problem of determining linear model dimension in that one's goal is to determine the dimension of the regression matrix, but the regression matrix in segmented line regression with unknown number of change-points includes unknown parameters. However, information-based criteria can still be applied and Liu, Wu and Zidek (1997) showed the consistency of the estimators obtained by the modified BIC. In the context of spline regression, Friedman (1991) proposed multivariate adaptive regression splines (MARS) that chooses the final model with the smallest GCV value, $GCV(k) = (1/n)\sum_{i=1}^{n}(y_i - \hat{\mu}_k(x_i))^2/\left(1 - C(k)/n\right)^2$, where k is the number of basis functions, $\hat{\mu}_k$ is the regression mean value estimated under the model with k basis functions, and $C(k)$ is the cost complexity measure of a model containing k basis functions.

Table 1 summarizes a simulation study to compare the modified permutation test with the model selection procedures based on the BIC and GCV. From further simulations it was found that the AIC over-estimates κ even further, these results are not included. Since the simulation of the power of the permutation test requires extensive computing, we use the efficient simulation method proposed by Boos and Zhang (2000); the number of simulations and permutations conducted are 1,600 and 319, respectively. In all the simulations we conducted, we considered only equally spaced integer values of x, as in annually observed cancer incidence and mortality rates, and the model parameters were selected based on some data. We used two or three different values of σ_0, the smaller one in Cases 0-1 and 4-1 and the middle values in Cases 1-2, 2-2, 3-2, 5-2 and 6-2, representing the standard deviations of the incidence rates for the selected sites. The other values of σ_0 are chosen to study the behavior of the selection procedures according to various effect sizes, δ_i/σ_0. For the modified permutation test, the overall significance levels were chosen to be 0.05 or 0.15; the proportions of correct selections among the 1,600 simulated data sets are denoted in bold.

From the table, we observe that the criterion based on the GCV method tends to considerably over-fit the number of unknown change-points, and the permutation tests tend to be conservative. The permutation test with $\alpha = 0.05$ shows the highest probability of correct selection when there is no change, or when the minimum effect size, $\min_i \delta_i/\sigma_0$, and the number of observations in each segment are reasonably large. See Cases 0-1, 0-2, 1-2, 1-3, 2-2, 2-3, 3-1, 3-2 and 3-3. This indicates that the permutation test tends to pick up the correct model more often when the amount of change is substantial, and the BIC performs better when there are subtle changes, either with a relatively smaller effect size or with a small number of observations in a segment, as in Cases 1-1,

Table 1. Comparison of the permutation procedure, BIC and GCV

	Perm ($\alpha = 0.05$)				Perm ($\alpha = 0.15$)				BIC				GCV			
Case	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0-1	**0.955**	0.027	0.012	0.006	**0.899**	0.051	0.028	0.022	**0.890**	0.073	0.028	0.009	**0.712**	0.157	0.084	0.047
0-2	**0.969**	0.020	0.008	0.003	**0.911**	0.050	0.024	0.016	**0.906**	0.071	0.019	0.004	**0.711**	0.156	0.086	0.047
1-1	0.241	**0.728**	0.025	0.006	0.116	**0.794**	0.058	0.032	0.082	**0.817**	0.082	0.019	0.025	**0.717**	0.183	0.075
1-2	·	**0.963**	0.029	0.008	·	**0.909**	0.058	0.033	·	**0.879**	0.096	0.024	·	**0.700**	0.208	0.092
1-3	·	**0.954**	0.033	0.013	·	**0.893**	0.067	0.040	·	**0.848**	0.117	0.036	·	**0.663**	0.219	0.118
2-1	0.698	**0.271**	0.025	0.006	0.526	**0.392**	0.057	0.025	0.463	**0.456**	0.071	0.011	0.233	**0.524**	0.179	0.064
2-2	0.088	**0.871**	0.032	0.009	0.035	**0.859**	0.076	0.031	0.022	**0.853**	0.105	0.020	0.005	**0.714**	0.199	0.082
2-3	·	**0.959**	0.030	0.011	·	**0.877**	0.072	0.051	·	**0.848**	0.114	0.038	·	**0.659**	0.220	0.121
3-1	·	0.001	**0.954**	0.046	·	·	**0.859**	0.141	·	·	**0.874**	0.126	·	·	**0.764**	0.236
3-2	·	·	**0.953**	0.047	·	·	**0.861**	0.139	·	·	**0.862**	0.138	·	·	**0.735**	0.265
3-3	·	·	**0.954**	0.046	·	·	**0.856**	0.144	·	·	**0.887**	0.113	·	·	**0.795**	0.205
4-1	·	0.747	**0.226**	0.027	·	0.545	**0.359**	0.096	·	0.456	**0.472**	0.072	·	0.232	**0.579**	0.189
4-2	·	0.073	**0.880**	0.047	·	0.028	**0.833**	0.139	·	0.017	**0.853**	0.130	·	0.007	**0.733**	0.260
5-1	·	0.144	0.053	**0.803**	·	0.052	0.022	**0.926**	·	0.055	0.031	**0.914**	·	0.014	0.021	**0.966**
5-2	·	0.001	0.001	**0.998**	·	·	0.001	**0.999**	·	·	0.001	**0.999**	·	·	·	**1.000**
5-3	·	·	·	**1.000**	·	·	·	**1.000**	·	·	·	**1.000**	·	·	·	**1.000**
6-1	·	0.408	0.162	**0.430**	·	0.209	0.131	**0.660**	·	0.232	0.172	**0.596**	·	0.081	0.149	**0.770**
6-2	·	0.001	0.004	**0.996**	·	·	0.001	**0.999**	·	0.001	0.001	**0.998**	·	·	0.001	**0.999**
6-3	·	·	·	**1.000**	·	·	·	**1.000**	·	·	·	**1.000**	·	·	·	**1.000**

where the model parameters for each case are as follows:

Case	Cancer cite	κ	$\mu(x)$	σ
0-1	Hodgkins	0	$1.1 - 0.003x$	0.040
0-2				0.020
1-1	Brain/ONS	1	$1.8 + 0.014x - 0.019(x - 13)^+$	0.060
1-2				0.030
1-3				0.015
2-1	Kidney/RP	1	$2.0 + 0.024x - 0.011(x - 15)^+$	0.054
2-2				0.029
2-3				0.0145
3-1	Breast	2	$4.1 - 0.003x + 0.039(x - 5)^+ - 0.034(x - 12)^+$	0.020
3-2				0.010
3-3				0.005
4-1	NonHodkins	2	$2.4 + 0.036x - 0.020(x - 15)^+ - 0.021(x - 20)^+$	0.023
4-2				0.0115
5-1	Colorectal	3	$4.1 + 0.008x - 0.026(x - 10)^+ + 0.030(x - 20)^+ - 0.042(x - 23)^+$	0.024
5-2				0.012
5-3				0.006
6-1	Corpus/Uterus	3	$3.0 - 0.057x + 0.041(x - 4)^+ + 0.021(x - 13)^+ - 0.022(x - 23)^+$	0.034
6-2				0.017
6-3				0.0085

2-1, and 4-1. When we relax the overall over-fitting probability of α to 0.15, it is observed that the performance of the permutation procedure is close to that of the BIC method in most of cases, and with reasonable probabilities of correct selection. That is, the over-fitting probability of the BIC method is larger than 0.05 and less than 0.15 in all of Table 1, except for Cases 6-1, 6-2

and 6-3 where over-fitting is not possible. Tiwari et al. (2005) observed similar simulation results in comparing the performance of the BIC with that of the permutation procedure. However, the BIC approach used in Tiwari et al. (2005) is not directly comparable to the one used in this paper, since it was based on regression coefficients estimated by the posterior modes, while the BIC is based on the least squares estimates of the regression coefficients, here.

6. Concluding Remarks

In this paper, we examined the asymptotic efficiency of a permutation procedure in selecting the number of change-points in segmented line regression, and compared its performance with those of the methods based on Bayesian information criterion and the generalized cross validation. Although the problem of selecting the number of change-points involves a regression matrix with some unknown parameters, which distinguishes the problem from the classical regression model selection problem, we find that the large sample theory can yield consistency of model selection procedures such as the BIC and the permutation procedure. As indicated in a simulation study, BIC works better in picking up small changes while the permutation procedure tends to be conservative. Thus, for cancer rate analyses, where the goal is parsimonious models rather than picking up all possible changes, this paper supports the application of the permutation procedure and establishes its asymptotic accuracy. In situations where the implementation of the permutation procedure is not practical due to computational limitations, BIC can be an appropriate method with an over-fitting probability between 0.05 and 0.15 in cases similar to those considered in Table 1.

Acknowledgements

Kim's research was partially supported by NIH Contract 263-MQ-413149. The authors thank the associate editor and the referees for many valuable comments and suggestions.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716-723.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrics* **70**, 9-38.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18**, 1-22.

Boos, D. D. and Zhang, J. (2000). Monte Carlo evaluation of resampling-based hypothesis tests. *J. Amer. Statist. Assoc.* **95**, 486-492.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference:A Practical Information-Theoretic Approach*. Springer, New York.

Feder, P. (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**, 1-67.

George, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **95**, 1304-1308.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer-Verlag, New York.

Hinkley, D. V. (1971). Inference in two-phase regression. *J. Amer. Statist. Assoc.* **66**, 736-743.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1-49.

Hušková, M. (1998). Estimation in location model with gradual changes. *Comment. Math. Univ. Carolinae* **39**, 147-157.

Kim, H.-J., Fay, M., Feuer, E. J. and Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statist. Medicine* **19**, 335-351.

Lerman, P. M. (1980). Fitting segmented regression models by grid search. *Appl. Statist.* **29**, 77-84.

Liu, J., Wu, S. and Zidek, J. V. (1997). On segmented multivariate regression. *Statist. Sinica* **7**, 497-525.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

Miller, A. (2002). *Subset Selection in Regression*. 2nd edition. Chapman & Hall, London.

Rao, C. R. and Wu, Y. (2001). On model selection. In *Model Selection* (Edited by P. Lahiri). Institute of Mathematical Statistics Lecture Notes-Monograph Series **38**, 1-64. Institute of Mathematical Statistics, Hayward, CA.

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

Tiwari, R. C., Cronin, K. A., Davies, W., Feuer, E. J., Yu, B. and Chib, S. (2005). Bayesian model selection for joinpoint regression with application to age-adjusted cancer rates. *J. Roy. Statist. Soc. Ser. C* **54**, 919-939.

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

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* **90**, 151-156.

Department of Mathematics, Syracuse University, Syracuse, NY 13244-1150, U.S.A.

E-mail: hjkim@syr.edu

Laboratory of Epidemiology, Demography and Biometry, National Institute of Aging, 7210 Wisconsin Avenue, Suite 3C309, Bethesda, MD 20892, U.S.A.

E-mail: yubi@mail.nih.gov

Division of Cancer Control and Population Sciences, National Cancer Institute, 6116 Executive Boulevard, Suite 504, MSC 8317, Bethesda, MD 20892-8317, U.S.A.

E-mail: feuerr@mail.nih.gov

(Received January 2007; accepted October 2007)