Comparability of Segmented Line Regression Models

Hyune-Ju Kim,[1,*] Michael P. Fay,[2,†] Binbing Yu,[3] Michael J. Barrett,[3] and Eric J. Feuer[2]

[1]Department of Mathematics, Syracuse University, Syracuse, New York 13244-1150, U.S.A.
[2]Division of Cancer Control and Population Sciences, National Cancer Institute, 6116 Executive Boulevard, Suite 504, MSC 8317, Bethesda, Maryland 20892-8317, U.S.A.
[†]*Current address:* National Institute of Allergy and Infectious Diseases, 6700B Rockledge Drive, MSC 7609, Bethesda, Maryland 20892-7609, U.S.A.
[3]Information Management Services, Inc., 12501 Prosperity Drive, Suite 200, Silver Spring, Maryland 20904, U.S.A.
[*]*email:* hjkim@syr.edu

Summary. Segmented line regression models, which are composed of continuous linear phases, have been applied to describe changes in rate trend patterns. In this article, we propose a procedure to compare two segmented line regression functions, specifically to test (i) whether the two segmented line regression functions are identical or (ii) whether the two mean functions are parallel allowing different intercepts. A general form of the test statistic is described and then the permutation procedure is proposed to estimate the p-value of the test. The permutation test is compared to an approximate F-test in terms of the p-value estimation and the performance of the permutation test is studied via simulations. The tests are applied to compare female lung cancer mortality rates between two registry areas and also to compare female breast cancer mortality rates between two states.

Key words: Change point; Comparability; Joinpoint regression; Permutation test; Segmented regression; Spline regression.

1. Introduction

One of the important questions in cancer vital statistics is the identification of the timing and extent of changes in rate trend patterns. Segmented line regression, also known in the literature as multiphase regression with the continuity constraint, piecewise linear regression, and broken line regression, has been successfully applied to describe these trend data, and various statistical techniques have been developed for inference problems. Recently, Kim et al. (2000) applied segmented line regression to describe continuous changes in cancer mortality and incidence rates, for which the rate of a linear trend is an important measure, and proposed a permutation procedure to determine the number of change points and to estimate their locations and relevant regression parameters. In Kim et al. (2000), change points are called joinpoints as in Hudson (1966) and Gallant and Fuller (1973) to emphasize the continuity of the linear phases at the change points, and the procedure is implemented in Joinpoint 2.6, which is now available at `http://srab.cancer.gov/joinpoint/index.html`. Some other statistical methods to study trend data such as spline regression and Bayesian methods have been discussed in Kim et al. (2000) and the procedures such as MARS of Friedman (1991) or MASAL of Zhang (1997) can also be applied to fit the segmented line regression model. In the spline literature, where change points are called knots, the emphasis is usually on approximating and interpolating regression functions, while one of the important aims in Kim et al. (2000) was to make an inference on change points. See Feder (1975a) and Seber and Wild (1989, chapter 9) for further discussion on segmented regression versus spline regression.

While the method for properly identifying change points has been well described in Kim et al. (2000), only a single series is considered. In practice, however, we often handle several comparable series where we are interested in determining whether these series share common segmented line regression models. An example of ongoing concern is the comparison of segmented line regression models among different age groups with a goal of minimizing the number of different models. Other concerns in addition to the multiple age group case include the comparison of cancer incidence/mortality rates between two different regions, which can be considered a simple two-group comparison. Figure 1 shows the natural logarithms of age-adjusted breast cancer mortality rates for white females aged 50 or older for both New York and Michigan for the period from 1975 to 1999. The scatter plots and also the fits made by Joinpoint 2.6 indicate similar patterns for the two states. One of our interests would be to see whether the annual percentage change rates are comparable between the two states.

1.1 *Related Work and Outline*

The main objective of this article is to propose a procedure to compare two segmented line regression functions. Our specific interest is on testing (i) whether two segmented line regression

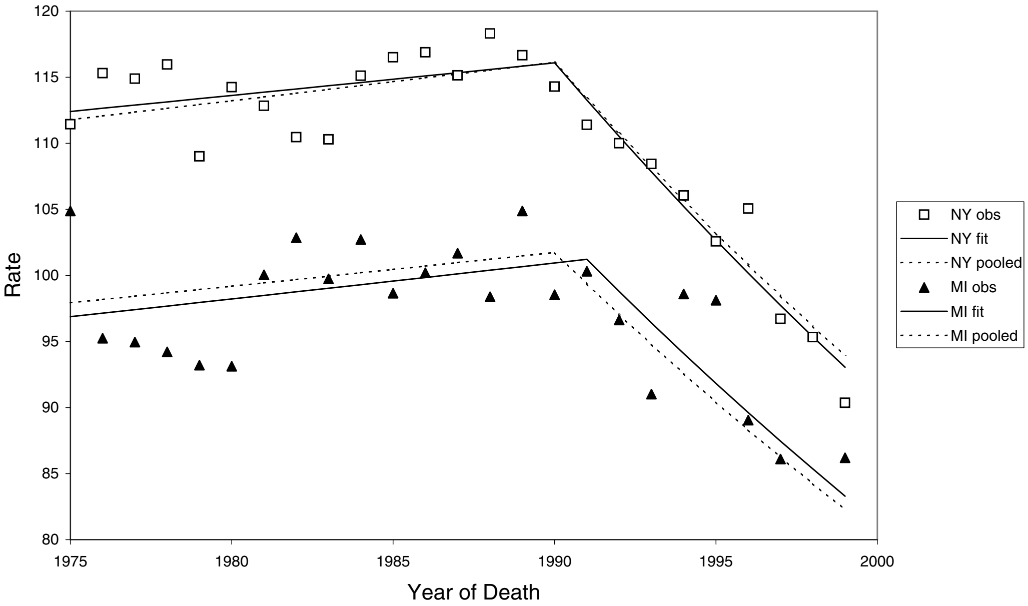

Figure 1. Breast cancer mortality for white females of age 50+.

functions are identical or (ii) whether the two mean functions are parallel allowing different intercepts. When the locations of the change points are known, it is reduced to a problem of comparing ordinary regression mean functions, which can be handled by a traditional F-test. The comparison of two general regression curves, linear or nonlinear, is discussed by several authors who mostly used nonparametric approaches or tests based on asymptotic normality. Hall and Hart (1990) proposed a nonparametric approach to the problem using a bootstrap procedure to detect a difference between two mean functions of nonparametric regression. Their test statistic is essentially a squared difference between two kernel estimates of two regression functions, averaged over the observed sample points, and the bootstrap method is used to determine critical points of the test statistic. Härdle and Marron (1990) also proposed a method to compare nonparametric regression curves by parametrically transforming one and then by comparing it with the other one. Several tests were developed to test the equivalence and parallelism of curves in more general situations with longitudinal data, non-Gaussian data, and general configurations of covariates (e.g., see Fan and Lin, 1998; Hunsberger and Follman, 2001; Zhang and Lin 2003). Closely related ideas are also found in profile analysis where the profiles are compared typically by using asymptotic chi-square or F-tests.

In this article, we focus on the problem of comparing segmented line regression models with unknown change points. In Section 2, we describe the idea of comparing two groups, which can be used in general change-point models, with/without continuity constraints. We first propose a test statistic in Sections 2.1 and 2.2 along with some discussion on fitting the segmented line regression model, and study its asymptotic behavior as well as the problem of how to handle the unknown number of change points. Then we describe a general permutation procedure in Section 2.3 to estimate the p-value of the test. Simulation studies to assess the performance of the permutation test are summarized in Section 3. The permutation procedure is compared to a classical approximate F-test in terms of the p-value estimation, and the power of the permutation test is studied via simulations. Examples are then presented in Section 4. The final section summarizes the results and discusses future research directions.

2. Test Statistic and Permutation Procedure

2.1 *Test Statistic and Basic Properties*

Let (x_{ij}, y_{ij}) denote the pair of the jth observations for the ith group ($i = 1, 2$; $j = 1, \ldots, n_i$). For example, y_{ij} can be a cancer incidence rate at the jth year, year x_{ij}, for the age group i. Suppose that the regression mean of y_{ij}, given $x_{ij} = x$, $\mu_i(x)$, is

$$\mu_i(x) = \beta_{i,0} + \beta_{i,1}x + \delta_{i,1}(x - \tau_{i,1})^+ + \cdots + \delta_{i,\kappa_i}(x - \tau_{i,\kappa_i})^+,$$

where κ_i is the unknown number of change points, the $\tau_{i,l}$'s $(l = 1, \ldots, \kappa_i)$ are the unknown change points, the $\beta_{i,l}$'s $(l = 0, 1)$ and the $\delta_{i,l}$'s $(l = 1, \ldots, \kappa_i)$ are the regression parameters, and $a^+ = a$ for $a > 0$ and 0 otherwise. The problem of fitting the model has been discussed in Kim et al. (2000), where the permutation test is used to see whether there is enough evidence to require a model with a larger number of change points than the one in the null hypothesis. Our goal in this article is to test whether the first group shares the common segmented line regression model with the second group.

Let the parameter of the model be $(\kappa_i, \boldsymbol{\theta}_i) = (\kappa_i, \tau_{i,1}, \ldots, \tau_{i,\kappa_i}, \beta_{i,0}, \beta_{i,1}, \delta_{i,1}, \ldots, \delta_{i,\kappa_i})$ for the ith group. The first test we consider, "Test 1," tests H_0: $(\kappa_1, \boldsymbol{\theta}_1) = (\kappa_2, \boldsymbol{\theta}_2)$ against H_1: $(\kappa_1, \boldsymbol{\theta}_1) \neq (\kappa_2, \boldsymbol{\theta}_2)$ for the groups 1 and 2. One of the main difficulties that distinguishes this problem from other change-point problems is that κ_i, the number of change points, is unknown. If κ_1 and κ_2 are known to be k and if we consider $\mathrm{H}_{0,k}$: the two segmented line models are comparable with k change points versus $\mathrm{H}_{1,k}$: the two models with k change points are not comparable, then a natural choice of statistic is the F-type statistic defined as

$$F_k = \frac{(\mathrm{RSS}_{H_{0,k}} - \mathrm{RSS}_{H_{1,k}})/d_{1,k}}{\mathrm{RSS}_{H_{1,k}}/d_{2,k}},$$

where RSS denotes the residual sum of squares and $d_{1,k}$ and $d_{2,k}$ are appropriate degrees of freedom. If k is the true number of change points, then the least-square estimators are consistent and asymptotically normal under some regularity assumptions imposed in Feder (1975a,b) and it can be shown that the null distribution of F_k can be approximated by a classical F distribution. However, if the model is not correctly specified, the least-square estimators are biased and are not consistent in general. See the Appendix and Feder (1975b) for details. With κ_1 and κ_2 unknown, our approach in this article is to consider the least-square estimators of the parameters including κ_1 and κ_2 under the restriction that κ_1 and κ_2 are less than or equal to some prespecified number $k_{\max}$. That is, we consider a larger model with a known maximum number of change points, $k_{\max}$, that contains the models described under H_0 and H_1. The method to be described in the following sections is practical for $k_{\max} \leq 4$, and for data sets of our interests, for example, cancer incidence rates in the last 30 years, the $k_{\max}$ value of 4 is reasonable.

For the second test we consider, "Test 2," our interest is only in the comparability of the slopes and change points, but not in that of the intercepts. We follow the same strategy of assuming that the number of change points is fixed at a known maximum, $k_{\max}$, except now the parameter of interest would be $\boldsymbol{\theta}_i^* = (\tau_{i,1}, \ldots, \tau_{i,k_{\max}}, \beta_{i,1}, \delta_{i,1}, \ldots, \delta_{i,\kappa_{\max}})$, (i.e., $\boldsymbol{\theta}_i$ without $\beta_{i,0}$) and we test $\mathrm{H}_0 : (\kappa_1, \boldsymbol{\theta}_1^*) = (\kappa_2, \boldsymbol{\theta}_2^*)$ versus $\mathrm{H}_1 : (\kappa_1, \boldsymbol{\theta}_1^*) \neq (\kappa_2, \boldsymbol{\theta}_2^*)$.

To conduct Tests 1 and 2, we assume that the number of change points for the group i, κ_i, is estimated as a known maximum number of change points, $k_{\max}$, fit the models with $k_{\max}$ change points both under the null and under the alternative hypotheses, and then assess the comparability of the two models with the test statistic $F_{k_{\max}}$. The asymptotic null distribution of $F_{k_{\max}}$ is not a classical F distribution anymore, and we propose to use a permutation procedure to estimate the p-value of the test in Section 2.3. In Section 3 and in the Appendix, we will discuss the effect of underfitting, when $k_{\max}$ is smaller than κ, and overfitting, when $k_{\max}$ is larger than κ, on the permutation procedure.

2.2 *Fitting*

Consider the following test statistic:

$$\begin{aligned} F &= F_{k_{\max}} \\ &= \frac{\left\{\displaystyle\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\tilde{y}_{ij})^2 - \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_{ij})^2\right\} \Big/ d_1}{\left\{\displaystyle\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_{ij})^2\right\} \Big/ d_2}, \end{aligned} \qquad (1)$$

where the $\hat{y}_{ij}$ values are the predicted values for the ith group fit separately from the other group and the $\tilde{y}_{ij}$ values are the predicted values under the null hypothesis, and d_1 and d_2 are appropriate degrees of freedom. The degrees of freedom for our statistic are set as $d_1 = 2 + 2k_{\max}$ and $d_2 = n_1 + n_2 - 4 - 4k_{\max}$ for Test 1, and $d_1 = 1 + 2k_{\max}$ and $d_2 = n_1 + n_2 - 4 - 4k_{\max}$ for Test 2. To estimate $\hat{y}_{ij}$ we use the grid search method as in Lerman (1980) for each group separately. Estimation of $\tilde{y}_{ij}$ will be done by applying the same method for the reduced null model. The grid search method first fits the regression parameters by the usual least-squares method given the location of the change points, and then estimates the change points by simultaneously searching for the locations that minimize the residual sum of squares. Since the grid search algorithm is computationally intensive, the method is tractable for $k_{\max} \leq 4$, but it becomes impractical as $k_{\max}$ gets larger. The details of fitting are as follows:

(i) For a given set of change points, $\boldsymbol{t} = (t_1, \ldots, t_k)$, where $k = k_{\max}$ is fixed, estimate the regression parameters by the least-squares method, $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}} = (X_{\boldsymbol{t}}'X_{\boldsymbol{t}})^{-1}X_{\boldsymbol{t}}'\mathbf{y}$, where for Test 2, $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2})', \boldsymbol{\beta} = (\beta_{1,0}, \beta_{2,0}, \beta_1, \delta_1, \ldots, \delta_k)'$, and

$$X_{\boldsymbol{t}} = \begin{pmatrix} 1 & 0 & x_{11} & (x_{11}-t_1)^+ & \cdots & (x_{11}-t_k)^+ \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & x_{1n_1} & (x_{1n_1}-t_1)^+ & \cdots & (x_{1n_1}-t_k)^+ \\ 0 & 1 & x_{21} & (x_{21}-t_1)^+ & \cdots & (x_{21}-t_k)^+ \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & x_{2n_2} & (x_{2n_2}-t_1)^+ & \cdots & (x_{2n_2}-t_k)^+ \end{pmatrix},$$

while for Test 1, $\boldsymbol{\beta} = (\beta_0,\ \beta_1,\ \delta_1, \ldots, \delta_k)'$, and $X_{\boldsymbol{t}}$ is the same except the first two columns are replaced by a column of ones.

(ii) For a given value of $\boldsymbol{t} = (t_1, \ldots, t_k)$, compute the residual sum of squares as $\mathrm{RSS}(\boldsymbol{t}) = (\mathbf{y} - X_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}})'(\mathbf{y} - X_{\boldsymbol{t}}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}})$.

(iii) Let $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_k)$ be the value of $\boldsymbol{t}$ which minimizes $\mathrm{RSS}(\boldsymbol{t})$ over the grid points, and let $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\tau}}}$ and $(\tilde{y}_{11}, \ldots, \tilde{y}_{1n_1}, \tilde{y}_{21}, \ldots, \tilde{y}_{2n_2})' = X_{\tilde{\boldsymbol{\tau}}}\tilde{\boldsymbol{\beta}}$.

Once we obtain the value of the test statistic, our goal is to estimate its p-value, which will be discussed in the following section.

Remark. Often in practice $\mathrm{Cov}\,(\boldsymbol{y}) = V = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} \neq \sigma^2 I_{(n_1+n_2)\times(n_1+n_2)}$. For such data, the weighted least-squares method can be used where the weighted least-squares estimator of $\boldsymbol{\beta}$ can be obtained as $\hat{\boldsymbol{\beta}}_w = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$. This general approach can be used to handle heteroscedastic errors as well as correlated errors. To incorporate the weighted least-square regression in the program, we need to estimate V, which can be done following similar arguments discussed in Section 2.2 of Kim et al. (2000).

2.3 *Estimation of the p-Value*

As discussed in Section 2.1, it can be shown that if the model is correctly specified in terms of the number of the unknown change points, κ, then the null distribution of our test statistic (1) can be approximated by a classical F distribution under the regularity conditions imposed in Feder (1975a,b). As it will be shown in the following section via simulations, however, the accuracy of the F approximation is not satisfactory for moderate sample sizes of our interests, and effects of under/overspecification of κ on the approximation still need to be studied to apply the asymptotic analysis. To estimate the p-value more accurately, we propose a permutation procedure, whose ideas are described below and justified in the Appendix. We first consider a situation where $n_1 = n_2 = n$ and $x_{1j} = x_{2j}$, which is our main interest, and then discuss a general situation later.

Step 1. Choose a test statistic, T, such that larger values of the test statistic denote less likely events under the null hypothesis. We will use the statistic, $T = F$ in (1).

Step 2. Compute the test statistic for the original data, say $T(\mathbf{y})$, by calculating $\tilde{y}_{ij}$, fitted under the null model, and $\hat{y}_{ij}$, fitted under the alternative hypothesis, for $i = 1$, 2 and $j = 1, \ldots, n$. Let the residuals under the null model be $\tilde{\epsilon}_{ij} = y_{ij} - \tilde{y}_{ij}$. Recall that the fits are made by the grid search method, which limits the value of $k_{\max}$ to be small.

Step 3. For $a = 1, \ldots, N_p - 1$, where N_p is a fixed integer, we permute two residuals, $\tilde{\epsilon}_{1j}$ and $\tilde{\epsilon}_{2j}$, at each $j = 1, \ldots, n$, to obtain $\tilde{\epsilon}_{1j}^{(a)}$ and $\tilde{\epsilon}_{2j}^{(a)}$, and then add them back onto the means from the null model to generate a new pair of responses, $y_{1j}^{(a)} = \tilde{y}_{1j} + \tilde{\epsilon}_{1j}^{(a)}$ and $y_{2j}^{(a)} = \tilde{y}_{2j} + \tilde{\epsilon}_{2j}^{(a)}$. For each set of permuted data, we compute the test statistic, $T(\mathbf{y}_{(a)})$. The approximate exchangeability of the residuals, needed to motivate the permutation procedure, is proved in the Appendix. In this step, we take Monte Carlo samples of size $N_p - 1$ because it is not practically efficient to obtain all of the 2^n permutations.

Step 4. The empirical distribution of $T(\boldsymbol{y}_{(0)}), T(\boldsymbol{y}_{(1)}), \ldots, T(\boldsymbol{y}_{(N_p-1)})$, where $\mathbf{y}_{(0)} = \mathbf{y}$ is used to estimate the permutation distribution of the test statistic under the null hypothesis. Since the null hypothesis would be rejected for larger values of the test statistic, we estimate the p-value of the test as

$$p = \frac{\text{number of times that } [T(\mathbf{y}_{(a)}) \geq T(\mathbf{y})] \text{ for a} \in \{0, 1, \ldots, N_P - 1\}}{N_P}.$$

We also consider some more complicated test statistics, T. For Test 1, it is reasonable to reject H_0 if data provide enough evidence against $\mathrm{H}_{0,k}$ for all k, and not to reject H_0 if data do not provide enough evidence against $\mathrm{H}_{0,k}$ for any k. Similar arguments hold for Test 2 after appropriately changing the definitions of $\mathrm{H}_{0,k}$ accordingly. This suggests $T = \min_k F_k$ or $T = \max_k F_k$ as possible test statistics in Step 1 above and these statistics are explored in Section 3. Another possibility that was not explored is to use either $T = \max_k P_k$ or $T = \min_k P_k$ as possible test statistics, where the P_k are the estimates of the p-values in testing $\mathrm{H}_{0,k}$ versus $\mathrm{H}_{1,k}$ by using the F statistic. In the following section, the performance of the min/max type statistics and the effects of under/overfitting by $k_{\max}$ will be studied along with the performance of the permutation test.

Remark. When the design points are not common, the proposed permutation procedure can be extended as follows. If similar assumptions as in Hall and Hart (1990) are satisfied for the design points, i.e., $\max_{1 \leq j \leq n} |x_{1j} - x_{2j'}| = O(n^{-1+\Delta})$ as $n \to \infty$, for each $\Delta > 0$, where j' is the index value for which $|x_{1j} - x_{2j'}|$ is minimized, then the matched residuals are asymptotically exchangeable and can be permuted to generate the permutation data set. In general, any set of residuals for the x values such that $|x_{ij} - x_{i'j'}| = O(n^{-1})(i, i' = 1, 2; j = 1, \ldots, n_i; j' = 1, \ldots, n_{i'})$ are asymptotically exchangeable. When the two samples have different sample sizes, n_1 and n_2, the Appendix shows that the residuals are asymptotically exchangeable if $k_{\max}$ is equal to the true number of change points.

3. Simulations

3.1 *Introduction*

To assess and compare the performance of the permutation procedure proposed in Section 2.3, simulation studies are conducted for a variety of parameter settings. For all of these simulations, the values of the predictor variable, x, were chosen as $69, \ldots, 95$ denoting the yearly time points. Since Monte Carlo estimation of the power of permutation tests is computationally intensive, we used the recommendation of Boos and Zhang (2000), who studied the designs of Monte Carlo evaluations of resampling-based hypothesis tests. We followed their rule of thumb and let the number of simulations, N_{SIM}, and the number of permutations, $N_P - 1$, meet the following two conditions: $N_P - 1 \approx 8(N_{\mathrm{SIM}})^{1/2}$ and αN_P equals an integer. In our study, $N_{\mathrm{SIM}} = 1600$ and $N_P - 1 = 319$. We considered a model, $\log y = \mu(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ and $\mu(x) = \beta_0 + \beta_1 x + \delta\,(x - \tau)^+$. Following examinations of some cancer mortality data, σ was chosen as 0.03, and the mean of Group 1 was chosen to be $\mu(x) = 5 + [\log(1 + 0.03)](x - 69) + [\log(1 + 0.02) - \log(1 + 0.03)]\,(x - 80)^+$, so that $\mu(69) = 5$. Since the annual percentage change (APC) for the segment with the regression slope of β is $100(e^{\beta} - 1)$, the regression coefficients of $\log(1 + 0.03) \approx 0.03$ and $\log(1 + 0.02) \approx 0.02$ correspond to APC of 3 and 2 for each linear segment, respectively.

3.2 *Size and F Approximation*

Table 1 includes the estimated size and power of the test, which are the relative frequencies of the simulation runs where the p-value is less than 0.05 under the null model and under the alternative model, respectively. The size and power are estimated for $k_{\max} = 1$ and $k_{\max} = 3$, for which the grid search is practical. The size of the approximate F-test is recorded in the parentheses for the first three cases in Table 1, and indicates that the F distribution generally underestimates the p-value. Further simulations to compare the performance of the approximate F-test and the permutation test in estimating the p-value were conducted and are summarized in Figure 2. Figure 2 shows the histograms of the p-values of Test 1, estimated by the approximate F-test as well as by the permutation test, for the null-parameter settings of $\mu(x) = 5 + \log(1 + 0.03)(x - 69) + [\log(1 + 0.02) - \log(1 + 0.03)](x - 80)^+$ for both groups, and for two different choices of $k_{\max}$. In Figure 2, the p-values approximated by the F distributions indicate considerable departures from a uniform distribution in general, and the overfitting with $k_{\max} = 3$ seems to underestimate the p-values even further, resulting in larger sizes. We note that even the correct specification of κ as $k_{\max} = 1$ does not significantly improve the uniformity of a p-value distribution. For Test 2, similar null behavior of the approximate F-test is observed and the report is omitted. Although the complete results are not reported here, further simulations were conducted to study whether a larger sample size improves the uniformity of the distribution. For medium-sized

Table 1
Empirical size and power of 0.05 *level permutation test*

	Group 2 parameters			Test 1		Test 2	
	β_0^*	APC_1	APC_2	$k_{\max} = 1$	$k_{\max} = 3$	$k_{\max} = 1$	$k_{\max} = 3$
Size	5	3	2	0.0444 (0.0640)[a]	0.0406 (0.0725)	0.0519 (0.0675)	0.0556 (0.0763)
	4.97[b]	3	2			0.0500 (0.0675)	0.0541 (0.0835)
	4.94	3	2			0.0481 (0.0800)	0.0688 (0.1007)
Power	**4.97**	3	2	0.7644	0.5755		
	4.94	3	2	1.0000	0.9986		
	5	**3.6**	2	0.9994	0.9923	0.4213	0.3047
	5	**3.9**	2	1.0000	1.0000	0.7906	0.6069
	4.94	**3.6**	2	0.3900	0.2157	0.4269	0.2998
	4.94	**3.9**	2	0.8538	0.6848	0.7963	0.5577
	5	3	**1.4**	0.9294	0.7230	0.7913	0.5872
	5	3	**1.1**	0.9994	0.9687	0.9888	0.9435
	4.94	3	**1.4**	1.0000	1.0000	0.7900	0.6093
	4.94	3	**1.1**	1.0000	1.0000	0.9888	0.9287
	5	**3.6**	**1.4**	0.8600	0.6827	0.6181	0.3735
	5	**3.6**	**1.1**	0.9369	0.7947	0.9413	0.8010
	4.94	**3.6**	**1.4**	0.9700	0.8455	0.6094	0.4324
	4.94	**3.6**	**1.1**	0.9994	0.9889	0.9325	0.7808

Note: The number of simulations is 1600 and the number of permutations is 319. Test 1 is designed to test for the common mean function against different mean functions, while Test 2 tests for the null hypothesis of the mean functions with the same slopes against a general alternative. Log $y = \mu(x) + \epsilon$, where $\epsilon \sim N(0,\, (0.03)^2)$ and $\mu(x) = \beta_0^* + [\log(1 + APC_1/100)](x - 69) + [\log(1 + APC_2/100) - \log(1 + APC_1/100)](x - \tau)^+$ for $x = 69, 70, \ldots, 95$. Group 1: $\tau = 80$, $\beta_0^* = 5$, $APC_1 = 3$, and $APC_2 = 2$. Group 2: $\tau = 80$, and (β_0^*, APC_1, APC_2) are chosen as above.

[a] The values in the parentheses are the estimated sizes of the approximate F-test.

[b] The boldface value indicates the parameter that has changed from the parameter value of Group 1.

samples with $x = 69, 69.5, 70, 70.5, \ldots, 94.5, 95$ (i.e., $n = 53$), no significant improvement was observed and departures from a uniform distribution for $n = 53$ seems even worse than for the case of $n = 27$ when the model is overfitted with $k_{\max} = 3$. For a large sample with $x = 69, 69.1, 69.2, \ldots, 69.9, 70, 70.1, \ldots, 94.9, 95$ ($n = 261$), we were able to estimate the p-values only for $k_{\max} = 1$ due to computational limitations. In that case, the results showed some improvement in the uniformity of p-values of the approximate F-tests, but the null distribution of the approximate F-test is still not comparable to that of the permutation test.

3.3 *Power and the Choice of* $k_{\max}$

The simulation study on the power of the permutation test is also summarized in Table 1. In Table 1, the change point for Group 2 was set to be the same as in Group 1, and various changes in the regression parameters are considered. As Test 1 tests for any differences between the two groups, one can see in Table 1 that the power increases as the differences between the groups increase. For Test 2, we would expect power to increase for groups that are less parallel, and in general we observe that in Table 2. Comparing the results for $k_{\max} = 3$ to those for $k_{\max} = 1$, we see in Table 1 that the tests lose power when κ is overfitted, and the percentage of power lost is generally larger in situations where the test with $k_{\max} = 1$ has lower power. Simulations were also done to determine empirical power in similar situations to Table 1 except with changes in the change points between groups, $\tau = 80$ for Group 1 and $\tau = 90$ for Group 2. These simulations showed powers greater than 70% in all of parameter choices used in Table 1, and in many cases gave empirical power of 100% (results not shown).

We conducted further simulation studies to compare the effects of using various values of $k_{\max}$ and it is summarized in Table 2. The table includes the estimated power for various $k_{\max}$ and various changes in the model parameters. The values of $k_{\max}$ considered are $0, 1, 2, 3, \hat{\kappa}_0, \tilde{\kappa}$, and $\hat{\kappa}^+$, where $\hat{\kappa}_0$ is the number of change points estimated under the null hypothesis of common segmented line models, $\tilde{\kappa} = \max(\hat{\kappa}_1, \hat{\kappa}_2)$ is the larger number of change points estimated separately for each group, and $\hat{\kappa}^+ = \max(\hat{\kappa}_0, \tilde{\kappa})$.

The table also includes the power of the min/max type statistics, discussed in Section 2.3. We found that the min/max type statistics improve the power, but the test based on min F breaks down for some cases such as Case 2f in this study. Case 2f causes the same kind of trouble for the test based on $\hat{\kappa}_0$ and on $k_{\max} = 0$. The test based on $\tilde{\kappa}$ is not recommended in general because this test tends to underestimate the p-value in general. In summary, some loss of power is expected if we overfit the model, and so we recommend to choose $k_{\max}$ as small as possible or use the test statistic of max F if the computation is manageable. As a data-based choice of $k_{\max}$, we recommend the use of $\hat{\kappa}^+$, which retains power for situations like Case 2f, and improves $\tilde{\kappa}$ in maintaining the size of the test.

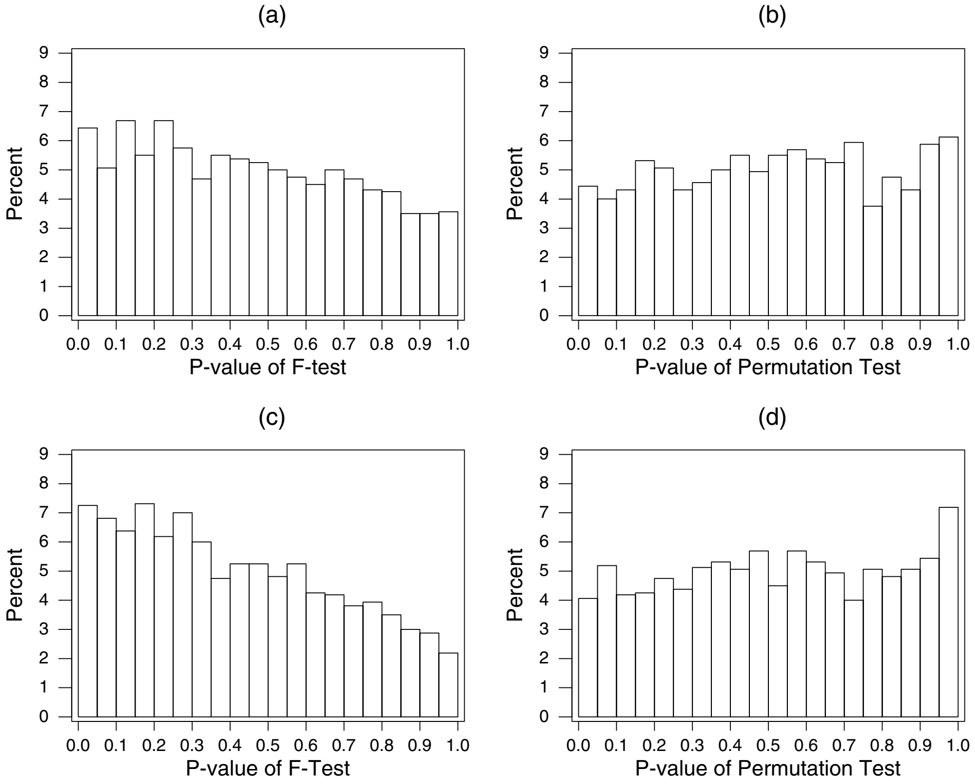

Figure 2. Histogram of p-values of the approximate F- and permutation tests for Test 1 based on 1600 simulations. The number of permutations for the permutation test is 319. The common null model for both groups is $\log y = \mu(x) + \epsilon$, where $\epsilon \sim N(0, (0.03)^2)$ and $\mu(x) = 5 + [\log(1 + 0.03)](x - 69) + [\log(1 + 0.02) - \log(1 + 0.03)](x - 80)^+$ for $x = 69, 70, \ldots, 95$: (a) F-test with $k_{\max} = 1$, (b) permutation test with $k_{\max} = 1$, (c) F-test with $k_{\max} = 3$, and (d) permutation test with $k_{\max} = 3$.

3.4 *Comparison with Hall and Hart's Test*

Tables 1–4 of Hall and Hart (1990) show the simulated power of the bootstrap test for various combinations of error distributions, bandwidths, sample sizes, and effect sizes. To detect a change in the two mean functions, $f(x)$ and $g(x)$, of size $f(x) - g(x) = c$ or cx for an effect size of $c/\sigma = 0.5$, 1, and 2, the best power was obtained for the bandwidth of $p = 1/(2)^{1/2}$ among the three bandwidths investigated in their simulation studies. Hall and Hart (1990), however, indicated that there is no straightforward relationship between power and the bandwidth, and proposed a data-driven choice of the bandwidth, $\hat{p}$, whose performance is summarized in Table 6 of their paper. In Table 3 below, we compare our test with $k_{\max} = 1$, 3, and $\hat{\kappa}^+$ to Hall and Hart's test, whose power for $f(x) - g(x) = c$ or cx is quoted from Tables 1–4 and 6 of Hall and Hart (1990). We find that the power of Hall and Hart's test with $\hat{p}$ is higher than those of ours in situations with small power, but our test, even with $k_{\max} = 3$, outperforms Hall and Hart's test with $\hat{p}$ in situations with reasonable power. The permutation test with $\hat{\kappa}^+$ performs considerably better than Hall and Hart's test with the data-driven choice of bandwidth. For the situations with $f(x) - g(x) = x - 1/2$ or $2x - 1$ or $3x - 1.5$, which were indicated in Hall and Hart (1990) as possible alternatives where the bandwidth choice of $1/(2)^{1/2}$ may not work, we simulated the power of Hall and Hart's test for $p = 1/\pi, 1/2, 1/(2)^{1/2}$, and compared to that of the permutation test. For these mean function choices, the permutation test with $\hat{\kappa}^+$ has higher power than the test of Hall and Hart for any choice of their bandwidths considered in their paper. For a practically detectable difference of $3x - 1.5$, the permutation test shows higher power even with $k_{\max} = 3$.

4. Examples

A recent publication (U.S. Cancer Statistics Working Group, 2002) reports 1999 cancer-incidence data for 78% of the U.S. population from registries meeting specified quality standards in the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program and the Centers for Disease Control and Preventions National Program of Cancer Registries (NPCR). These pooled registries are not a random sample of the United States, and it is therefore of interest to determine how representative they are. Mortality is available for the entire nation, and for lung cancer, mortality tracks very closely with incidence. Because the South is underrepresented in the pooled areas and smoking rates are known to be higher in the South, especially for males (Shopland et al., 1996), we examine male and female lung-cancer mortality trends from 1969 to 1999 in the pooled and nonpooled areas. For males, the rates in the pooled areas are lower than in the nonpooled areas (data not shown), but here for illustrative purposes we focus on levels and trends for females.

Table 2
Empirical size and power comparison for different choices of k_{max}

			$k_{\max}$ in $T = F_{k_{\max}}$							T	
A. Test		Case	$\tilde{\kappa}$	$\hat{\kappa}_0$	$\hat{\kappa}^+$	0	1	2	3	max F	min F
Test 1	Size	2a	0.057	0.046	0.057	0.040	0.055	0.045	0.050	0.050	0.040
	Power	2b	0.878	0.822	0.882	0.866	0.770	0.660	0.550	0.776	0.736
		2c	0.732	0.707	0.780	0.875	0.720	0.650	0.535	0.615	0.850
		2d	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2e	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2f	1.000	0.003	1.000	0.000	1.000	1.000	1.000	1.000	0.015
Test 2	Size	2a	0.062	0.046	0.063	0045	0.033	0.042	0.048	0.036	0.021
		2b	0.061	0.039	0.049	0.039	0.048	0.048	0.051	0.033	0.039
		2c	0.088	0.066	0.071	0.045	0.057	0.060	0.060	0.048	0.030
	Power	2d	0.973	0.962	0.968	0.985	0.979	0.905	0.827	0.982	0.952
		2e	0 216	0.180	0.203	0.259	0.167	0.176	0.155	0.137	0.214
		2f	1.000	0.000	1.000	0.000	1.000	1.000	1.000	1.000	0.003

B. Case	Group	τ_1	τ_2	τ_3	β_0^*	APC_1	APC_2	APC_3	APC_4
2a	1	80			5.00	3.0	2.0		
	2	80			5.00	3.0	2.0		
2b	1	80			5.00	3.0	2.0		
	2	80			4.97	3.0	2.0		
2c	1	80	87	94	5.00	3.0	2.0	−1.0	−2.0
	2	80	87	94	4.97	3.0	2.0	−1.0	−2.0
2d	1	80			5.00	3.0	2.0		
	2	80			4.94	3.6	1.4		
2e	1	80	87	94	5.00	3.0	2.0	−1.0	−2.0
	2	80	87	94	4.94	3.6	1.4	−1.0	−2.0
2f	1	88			5.00	−1.0	1.0		
	2	88			4.85	1.0	−1.0		

Note: The number of simulations is 1600 and the number of permutations is 319. $\tilde{\kappa} = \max(\hat{\kappa}_1, \hat{\kappa}_2)$ where $\hat{\kappa}_i$ is the number of change points estimated for Group 1. $\hat{\kappa}_0$ = the number of change points estimated under the null hypothesis of common change-point models. $\hat{\kappa}^+ = \max(\tilde{\kappa}, \hat{\kappa}_0)$. Log $y = \mu(x) + \epsilon$, where $\epsilon \sim N(0, (0.03)^2)$ and $\mu(x) = \beta_0^* + [\log(1 + APC_1/100)](x - 1975) + [\log(1 + APC_2/100) - \log(1 + APC_1/100)] \times (x - \tau_1)^+ + [\log(1 + APC_3/100) - \log(1 + APC_2/100)](x - \tau_2)^+ + [\log(1 + APC_4/100) - \log(1 + APC_3/100)](x - \tau_3)^+$ for $x = 1975, 1976, \ldots, 2001$, where the parameters for each case are as given in B.

Table 3
Power comparison with Hall and Hart's test

Hall and Hart's test					Permutation test					
	Bandwidth p							$k_{\max}$ of Test 1		
$f(x) - g(x)$	$1/\pi$	$1/2$	$1/\sqrt{2}$	$\hat{p}$	Δ_0	Δ_1	Δ_2	1	3	$\hat{\kappa}^+$
0	0.051	0.054	0.052	0.052	0	0	0	0.060	0.046	0.056
0.5	0.315	0.412	0.460	0.352	0.5σ	0	0	0.301	0.184	0.368
1	0.887	0.946	0.962	0.717	σ	0	0	0.841	0.652	0.916
2	1.000	1.000	1.000	0.879	2σ	0	0	1.000	1.000	1.000
0.5x	0.137	0.152	0.157	0.146	0	0.5σ	0.5σ	0.110	0.078	0.158
x	0.403	0.467	0.492	0.356	0	σ	σ	0.353	0.249	0.471
2x	0.943	0.973	0.974	0.703	0	2σ	2σ	0.921	0.770	0.973
x − 0.5	0.110	0.050	0.050	*	-0.5σ	σ	σ	0.107	0.095	0.143
2x − 1	0.214	0.136	0.064	*	$-\sigma$	2σ	2σ	0.311	0.204	0.449
3x − 1.5	0.495	0.364	0.126	*	-1.5σ	3σ	3σ	0.622	0.530	0.779

Note: Hall and Hart's test: Group 1: $y_i = g(x_i) + \epsilon_i$, where the ϵ_i's are independent N(0, 1) variables. Group 2: $z_i = f(x_i) + \eta_i$, where the η_i's are independent N(0, 1) variables. $x_i = i/n$, for $i = 1, \ldots, 30$. $\hat{p}$ is the data-driven choice of bandwidth used in Table 6 of Hall and Hart (1990).

Permutation test: Log $y_{ij} = \beta_{i,0} + \beta_{i,1}x + [\beta_{i,2} - \beta_{i,1}](x_{ij} - 11.5/30)^+ + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ with $\sigma = 0.03$ and $x_{ij} = (j - 0.5)/30$ for $j = 1, \ldots, 30$ and $i = 1, 2$. Group 1: $\beta_{1,0} = 5$, $\beta_{1,1} = \log(1.03)$ with $APC_1 = 3$, and $\beta_{1,2} = \log(1.02)$ with $APC_2 = 2$. Group 2: $\beta_{2,0} = \beta_{1,0} + \Delta_0$, $\beta_{2,1} = \beta_{1,1} + \Delta_1$, $\beta_{2,2} = \beta_{1,2} + \Delta_2$.

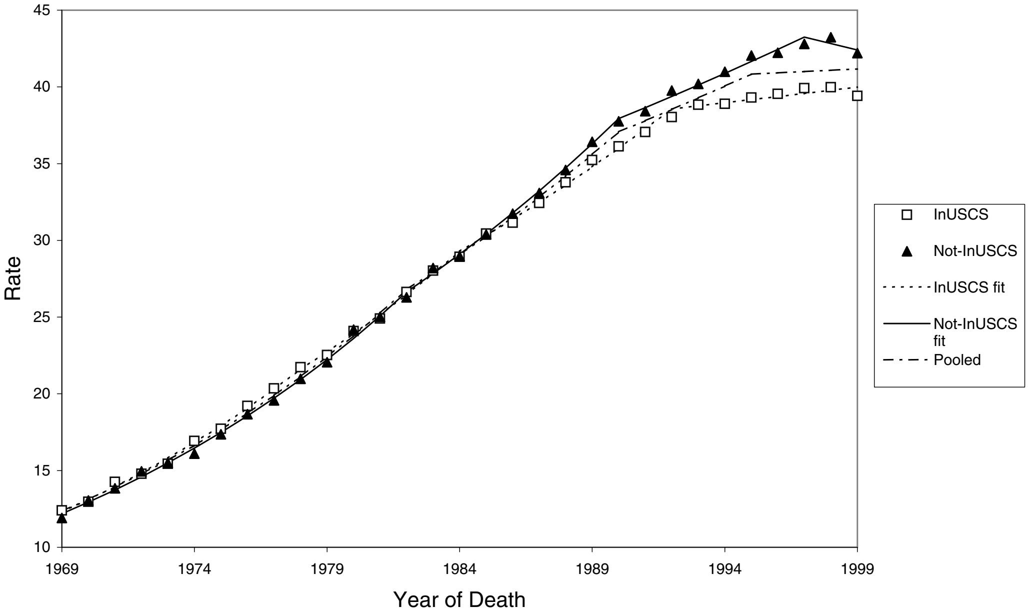

Figure 3. Female lung cancer mortality.

The response variable for the analysis of mortality is the natural logarithms of age-adjusted lung cancer mortality rates, and the independent variable is the yearly data points from 1969 to 1999. Fitting uncorrelated errors model with $k_{\max} = 3$ to lung cancer mortality data, two regions separately, Joinpoint 2.6 estimated the $\hat{\mu}(x) = -120.6114 + 0.0625x - 0.0119(x - 1978)^{+} - 0.0162(x - 1984)^{+} - 0.0292(x - 1992)^{+}$ for the data from the pooled area, reported in the U.S. Cancer Statistics 1999 (InUSCS), and $\hat{\mu}(x) = -115.9385 + 0.0602x - 0.0159(x - 1982)^{+} - 0.0255(x - 1990)^{+} - 0.0286(x - 1997)^{+}$ for the data from the nonpooled area (Not-InUSCS). In Figure 3, we see the trends are quite similar until the mid-1980s and then the pooled group (InUSCS) started to decrease faster compared to the nonpooled population (Not-InUSCS). In the recent few years, however, the nonpooled area indicated a sharper decrease in female lung cancer mortality than in the pooled area, although the level for the nonpooled group (Not-InUSCS) was higher. Running a comparability Test 1 with 4499 permutations, we find the permutation p-value of 0.0002, which provides strong evidence to conclude that the regression parameters in the two groups do not remain the same. Comparability Test 2 to compare the slopes also yielded a very small p-value of 0.0002, concluding that the annual percentage change rates in female lung cancer mortality are different in the two groups. Fitting the autocorrelated errors models by the weighted least-squares method, we get a p-value of 0.0002 for the comparability Test 1, which leads to the same conclusion. See Kim et al. (2000) for further discussion on fitting with uncorrelated or correlated errors. When the comparability tests were run for the first half of the data from 1969 to 1984, we find that there is no significant difference in the two regression functions with the permutation p-values of 0.0973 and 0.0893 for Test 1 and Test 2, respectively, which supports a considerable deviation from each other observed in the latter part of the study period.

As the second example to highlight the use of Test 2, we run the permutation comparability test for the breast cancer mortality rates for white females of age 50 or older between New York and Michigan presented in Section 1. Considering that $\hat{\kappa}^{+} = 1$, we used $k_{\max} = 1$ for the following analysis. Fitting the uncorrelated errors model to breast cancer mortality data with one change point and with the weighted least-squares method, New York and Michigan separately, Joinpoint 2.6 estimated $\hat{\mu}(x) = 0.4854 + 0.0021x - 0.0267(x - 1990)^{+}$ for New York and $\hat{\mu}(x) = -0.8381 + 0.0027x - 0.0271(x - 1991)^{+}$ for Michigan. The scatter plots in Figure 1 show similar patterns for the two states and the comparability Test 2 with 4499 permutations estimated the p-value as 0.7589. Thus the analysis does not provide evidence to conclude that the annual percentage change rates in the two states differ. As one might guess from the graphs, Test 1 shows a significant difference between the intercepts.

5. Discussion

Segmented line regression has been a useful tool to describe changes in trend data such as cancer incidence data, and in this article the permutation test is proposed to compare two segmented line regression functions. Our proposed approach requires choosing some maximum number of change points, $k_{\max}$, usually less than or equal to 4. Our simulations have shown choosing $k_{\max}$ too large results in a loss of power, and we recommend an estimation of $k_{\max}$ from the data, for which the tests have simulated sizes close to the nominal size. The examples and simulation studies indicate that the proposed approach is effective to test the equality or the parallelism of the two segmented line regression functions. The proposed procedure utilizes and extends the regression fitting implemented in Joinpoint 2.6, and the program to run the comparability test will be available at `http://srab.cancer.gov/joinpoint` in the future.

As indicated in Section 1.2, the hypotheses of parallel profiles and common profiles are often of interest in profile analysis. In profile analysis, the test statistics are formed as F statistics in the context of MANOVA. In addition to that the hypotheses in this article are formulated in the context of segmented line regression, the permutation test proposed in this article can handle nonnormal errors as long as the residuals are asymptotically exchangeable, can accommodate unequal design points, and, furthermore, can handle general types of hypotheses such as the ones described below.

Suppose we wish to test the equality of two segmented line regression functions given that they are known to be parallel, which is often the hypothesis of interest in profile analysis. Then, we can compute the test statistic, $F_{k_{\max}}$, for which the practical value of $k = k_{\max} \leq 4$, by getting $\mathrm{RSS}_{H_0,k}$ as in Test 1 and $\mathrm{RSS}_{H_1,k}$ with $\boldsymbol{\beta} = (\beta_{1,0},\ \beta_{2,0},\ \beta_1,\ \delta_1, \ldots, \delta_k)'$, and then estimate its p-value by permuting residuals, whose asymptotic exchangeability is justified for the Test 1 null hypothesis. We also note that the hypotheses discussed in this article can be expressed in the following general form of hypotheses: $\mathrm{H}_0 : \boldsymbol{H\beta} = \boldsymbol{h},\ \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$ versus $\mathrm{H}_1 : \boldsymbol{H\beta} \neq \boldsymbol{h}$ or $\boldsymbol{\tau}_1 \neq \boldsymbol{\tau}_2$, where for the models with κ_1 and κ_2 change points, respectively, $\boldsymbol{\tau}_1 = (\tau_{1,1}, \ldots, \tau_{1,\kappa_1})', \boldsymbol{\tau}_2 = (\tau_{2,1}, \ldots, \tau_{2,\kappa_2})'$, and $\boldsymbol{\beta} = (\beta_{1,0}, \beta_{1,1}, \delta_{1,1}, \ldots, \delta_{1,\kappa_1}, \beta_{2,0}, \beta_{2,1}, \delta_{2,1}, \ldots, \delta_{2,\kappa_2})'$. This generalization helps us to extend the test to more general situations. For example, one might be interested in testing whether the two regression models share the same slopes only in the last phase. That is, suppose we wish to test $\mathrm{H}_0 : \beta_{1,1} + \delta_{1,1} + \cdots + \delta_{1,k_1} = \beta_{2,1} + \delta_{2,1} + \cdots + \delta_{2,k_2}$ against a general alternative for appropriately chosen k_1 and k_2. Then, the use of $\boldsymbol{H} = (0,\ 1, \ldots, 1,\ 0, \ldots, -1, \ldots, -1)$ and $\boldsymbol{h} = 0$ and the same arguments used in Section 2 would provide an estimate of the p-value.

A natural extension of the results obtained here is to compare more than two groups, say m groups, where the regression mean function for each group is presented as a segmented line regression model. One method is to use the p-values obtained from the permutation tests for a two-group comparison and a modified Bonferroni procedure, such as a sequentially rejective Bonferroni (SRB) procedure proposed by Holm (1979). The modified SRB procedure, which has greater power than the Bonferroni procedure, can be used to determine p-values significant enough to distinguish two groups. In a situation where we have to determine whether adjacent age groups can be combined together, our concern is on the comparison of the adjacent two groups, and so we can use the $(m - 1)$ p-values resulting from the $(m - 1)$ permutation tests of adjacent groups. In a situation where one is interested in the comparison of all possible pairs of groups, the SRB can be further improved as suggested in Shaffer (1986).

Another interesting question is to compare the locations of the change points regardless of the regression parameters. It was found that the observed residuals, even pairwise, are not asymptotically exchangeable under the null hypotheses of only the same change-point locations, and thus the method proposed in this article cannot be directly applied to test for the same locations of change points. If the model is correctly specified in terms of the number of change points, large sample normal approximations for the distributions of the estimated change points can be applied to compare the locations of change points between the two groups, but its performance is not expected to be great and this problem requires further research.

Résumé

Les modèles de régression linéaire segmentés, composés de phases linéaires continues, ont été appliqués pour décrire des changements de tendance de taux. Dans cet article nous proposons une procédure pour comparer deux fonctions de régression linéaires segmentées, afin de tester soit a) si les deux fonctions de régression linéaires segmentées sont identiques ou ii) si les deux fonctions moyennes sont parallèles et admettent des intercepts différents. Une forme générale de la statistique de test est décrite, et on propose alors une procédure par permutation pour estimer la p-value du test. Le test de permutation est comparé à un test F approché en terme d'estimation de la p-value, et les performances du test de permutation sont étudiées par simulation. Les tests sont appliqués pour comparer les taux de mortalité par cancer du poumon chez les femmes, entre deux régions de recensement, et également pour comparer les taux de mortalité par cancer du sein chez la femme entre deux stades.

References

Boos, D. D. and Zhang, J. (2000). Monte Carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association* **95,** 486–492.

Fan, J. and Lin, S.-K. (1998). Test of significance when data are curves. *Journal of the American Statistical Association* **93,** 1007–1021.

Feder, P. (1975a). On asymptotic distribution theory in segmented regression Problems-Identified Case. *The Annals of Statistics* **3,** 49–83.

Feder, P. (1975b). The log likelihood ratio in segmented regression. *The Annals of Statistics* **3,** 84–97.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19,** 1–67.

Gallant, A. R. and Fuller, W. A. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *Journal of the American Statistical Association* **73,** 144–147.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* **85,** 1039–1049.

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *The Annals of Statistics* **18,** 63–89.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

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

Hunsberger, S. and Follmann, D. A. (2001). Testing for treatment and interaction effects in semi-parametric analysis of covariance. *Statistics in Medicine* **20,** 1–19.

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

Lerman, P. M. (1980). Fitting segmented regression models by grid search. *Applied Statistics* **29,** 77–84.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression.* New York: John Wiley.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* **81,** 826–831.

Shopland, D. R., Hartman, A. M., Gibson, J. T., Mueller, M. D., Kessler, L. G., and Lynn, W. R. (1996). Cigarette smoking among U.S. adults by state and region: Estimates from the current population survey. *Journal of the National Cancer Institute* **88,** 1748–1758.

U.S. Cancer Statistics Working Group. (2002). *United States Cancer Statistics: 1999 Incidence.* Atlanta: Department of Health and Human Services, Centers for Disease Control and Prevention, National Cancer Institute.

Zhang, H. (1997). Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics* **6,** 74–91.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.

Received April 2003. *Revised April* 2004.
Accepted May 2004.

Appendix

Exchangeability of Residuals

In this section, we examine the exchangeability of residuals obtained under the null model, especially when the model is either under- or overfitted. We discuss a simple case of a segmented line regression model with known change points, and then the asymptotic results can be extended for the case of unknown change points when the estimators of the change points are consistent.

Suppose that the true model has k^* change points at $(\tau_1, \ldots, \tau_{k^*})$ and let $k = k_{\max}$. To describe the null model with k change points at $\boldsymbol{t} = (t_1, \ldots, t_k)$, let $\boldsymbol{\beta}_k = (\beta_0,\ \beta_1,\ \delta_1,\ \delta_2, \ldots, \delta_k)'$ for Test 1 and $\boldsymbol{\beta}_k = (\beta_{1,0},\ \beta_{2,0},\ \beta_1,\ \delta_1,\ \delta_2, \ldots,\ \delta_k)'$ for Test 2. Also let $\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2})'$ and define the design matrices as $X_{\boldsymbol{t}_k} = (X_0, A_{\boldsymbol{t}_k})$, where X_0 is the first two or three columns of $X_{\boldsymbol{t}}$ in Section 2.2, and

$$A_{\boldsymbol{t}_k} = \begin{pmatrix} (x_{11}-t_1)^+ & \cdots & (x_{11}-t_k)^+ \\ \hdotsfor{3} \\ (x_{1n_1}-t_1)^+ & \cdots & (x_{1n_1}-t_k)^+ \\ (x_{21}-t_1)^+ & \cdots & (x_{21}-t_k)^+ \\ \hdotsfor{3} \\ (x_{2n_2}-t_1)^+ & \cdots & (x_{2n_2}-t_k)^+ \end{pmatrix}.$$

Case 1. $k = k^*$.

If $k = k^*$, the residual in fitting $\mathbf{y} = X_{\boldsymbol{t}_k}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$,

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_k &= \left(\tilde{\boldsymbol{\epsilon}}_1', \tilde{\boldsymbol{\epsilon}}_2'\right)' = \left(I_{(n_1+n_2)\times(n_1+n_2)} - X_{\boldsymbol{t}_k}\left(X_{\boldsymbol{t}_k}'X_{\boldsymbol{t}_k}\right)^{-1}X_{\boldsymbol{t}_k}'\right)\mathbf{y} \\ &= (I - P_{\boldsymbol{t}_k})\mathbf{y}, \end{aligned}$$

where $P_{\boldsymbol{t}_k} = X_{\boldsymbol{t}_k}(X_{\boldsymbol{t}_k}'X_{\boldsymbol{t}_k})^{-1}X_{\boldsymbol{t}_k}'$. Then,

$$\begin{aligned} E_{H_0,k^*}[\tilde{\boldsymbol{\epsilon}}_k \,|\, \boldsymbol{t}] &= (I - P_{\boldsymbol{t}_k})X^*\boldsymbol{\beta}^* \\ &= (I - P_{\boldsymbol{t}_k})(X_{\boldsymbol{t}_k} + X^* - X_{\boldsymbol{t}_k})\boldsymbol{\beta}^* \\ &= (I - P_{\boldsymbol{t}_k})(X^* - X_{\boldsymbol{t}_k})\boldsymbol{\beta}^*, \end{aligned}$$

and $\mathrm{Var}_{H_0,k^*}[\tilde{\boldsymbol{\epsilon}}_k \,|\, \boldsymbol{t}] = \sigma^2(I - P_{\boldsymbol{t}_k})$, where $\boldsymbol{\beta}^*$ and $X^* = X^*(\tau_1, \ldots, \tau_{k^*})$ are the regression parameters and the design matrix under the true null model with k^* change points, respectively. When the model is correctly specified, i.e., $k = k^*$, the least-square estimators of the change points, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k^*})$ are consistent as proved in Feder (1975a), and thus $E_{H_0,k^*}[\tilde{\epsilon}_{ij}] \to 0$ for $i = 1, 2; j = 1, 2, \ldots, n_i$. In our situation of the equally spaced x's, $\mathrm{Var}_{H_0,k^*}[\tilde{\boldsymbol{\epsilon}}]$ can be approximated by $\sigma^2 I_{(n_1+n_2)\times(n_1+n_2)}$, which can be justified by that the elements of $X_{\boldsymbol{t}_k}'X_{\boldsymbol{t}_k}$ and $X_{\boldsymbol{t}_k}(X_{\boldsymbol{t}_k}'X_{\boldsymbol{t}_k})^{-1}X_{\boldsymbol{t}_k}'$ are $O(n)$ and $O(1/n)$, respectively, if $x_i = i/n$. Thus, all the residuals are approximately exchangeable if the model is correctly specified.

Case 2. $k \neq k^*$, but $n_1 = n_2$ and $x_{1j} = x_{2j}$.

We show that the residuals are approximately exchangeable if the data are matched, so that $x_{1j} = x_{2j} \equiv x_j$ for all j. Consider first the null hypothesis under Test 1 so that $\mu_1(x) = \mu_2(x) = \mu(x)$ for all x. We write a very general model as $y_{ij} = \mu(x_j) + \epsilon_{ij}$, where $\epsilon_{ij} \sim F_j(\mu(x_j))$, and F_j is an unknown distribution that may depend on $\mu(x_j)$. Now suppose we estimate $\mu(x)$ with some model $\tilde{\mu}(x)$ which need not be consistent nor correctly specified. Then the residuals are given by $\tilde{\epsilon}_{ij} = y_{ij} - \tilde{\mu}(x_j) = (\mu(x_j) - \tilde{\mu}(x_j)) + \epsilon_{ij}$ and the distribution of the residuals does not depend on group membership, so $\tilde{\epsilon}_{1j}$ and $\tilde{\epsilon}_{2j}$ have the same distribution and are exchangeable.

The situation under Test 2 is more complicated. We write the model as $y_{ij} = \mu_i(x_j) + \epsilon_{ij}$, where $\epsilon_{ij} \sim F_j(\mu_i(x_j))$, and F_j is an unknown distribution that may depend on $\mu_i(x_j)$. Under the null hypothesis of Test 2, $\mu_2(x) = (\beta_{2,0} - \beta_{1,0}) + \mu_1(x)$, so that the residuals under the null hypothesis are given by

$$\begin{aligned} \tilde{\epsilon}_{1j} &= y_{1j} - \tilde{\mu}_1(x_j) = (\mu_1(x_j) - \tilde{\mu}_1(x_j)) + \epsilon_{1j}, \\ \tilde{\epsilon}_{2j} &= y_{2j} - \tilde{\mu}_2(x_j) = (\mu_2(x_j) - \tilde{\mu}_2(x_j)) + \epsilon_{2j}, \\ &= (\mu_1(x_j) - \tilde{\mu}_1(x_j)) + (\beta_{2,0} - \beta_{1,0}) - (\tilde{\beta}_{2,0} - \tilde{\beta}_{1,0}) + \epsilon_{2j}. \end{aligned}$$

Thus, if F_j does depend on $\mu_i(x_j)$, the residuals will not have the same distribution and may not be exchangeable. If F_j does not depend on $\mu_i(x_j)$, then as long as $(\beta_{2,0} - \beta_{1,0}) \approx (\tilde{\beta}_{2,0} - \tilde{\beta}_{1,0})$, the residuals $\tilde{\epsilon}_{1j}$ and $\tilde{\epsilon}_{2j}$ will approximately have the same distribution and be approximately exchangeable. Lengthy but straightforward matrix algebra shows that $E_{H_{0,k^*}}[\tilde{\beta}_{1,0} - \beta_{1,0} \,|\, \boldsymbol{t}] = E_{H_{0,k^*}}[\tilde{\beta}_{2,0} - \beta_{2,0} \,|\, \boldsymbol{t}]$, which will imply the asymptotic exchangeability of the matched residuals. This is true for any k, but the nonzero biases, $E_{H_{0,k^*}}[\tilde{\beta}_{i,0} - \beta_{i,0} \,|\, \boldsymbol{t}], (i = 1, 2)$, vary depending on whether the model is overfitted, i.e., $k_{\max} > k^*$, or is underfitted, i.e., $k_{\max} < k^*$.