The Statistician (2001)
50, *Part* 1, *pp.* 51–61

Inference and estimation in a changepoint regression problem

Steven A. Julious

SmithKline Beecham, Harlow, UK

[Received September 1999. Final revision September 2000]

Summary. The two-line model when the location of the changepoint is known is introduced, with an *F*-test to detect a change in the regression coefficient. The situation when the changepoint is unknown is then introduced and an algorithm proposed for parameter estimation. It is demonstrated that when the location of the changepoint is not known the *F*-test does not conform to its expected parametric distribution. Nonparametric bootstrap methods are proposed as a way of overcoming the problems encountered. Finally, a physiology example is introduced where the regression change represents the change from aerobic to anaerobic energy production.

Keywords: Bootstrapping; Changepoint regression; Least squares estimates; Piecewise regression; Split lines; Two-phase regression

1. Introduction

The changepoint regression problem was described by Quandt (1958, 1960) since when an extensive literature has developed (Shaban, 1980; Krisnaiah and Miao, 1988). It can be applied to physiological situations where the regression slope is not expected to be constant but to change suddenly at a given point. It is this 'changepoint' which is of primary interest, as it may be a marker for a change in some physiological response, such as age at the menopause in a plot of bone density against age in a study of female bones (Lees *et al*., 1983) or anaerobic thresholds in patients exercising to exhaustion (Bennett, 1988). If the location of the changepoint is known then the estimation of the parameters in the model is straightforward; however, if it is not known an extra parameter (the changepoint) must be estimated. Furthermore, the problem is no longer linear and the only way to estimate the parameters is through numerical optimization.

2. Location of changepoint known

2.1. Estimation of model

For any interval (X_0, X_1) on the real line the problem is defined as

$$f(x_i) = f_1(x_i; \beta_1) \qquad X_0 \leqslant x_i \leqslant \delta,$$

$$= f_2(x_i; \beta_2) \qquad \delta \leqslant x_i \leqslant X_1$$

such that $f_1(\delta; \hat{\boldsymbol{\beta}}_1) = f_2(\delta; \hat{\boldsymbol{\beta}}_2)$, i.e. the slope of the relationship between y and x is constant until

Address for correspondence: Steven A. Julious, Clinical Pharmacology Statistics, SmithKline Beecham, New Frontiers Science Park (South), Third Avenue, Harlow, Essex, CM19 5AW, UK.
E-mail: Steven_A_Julious@sbphrd.com

 0039–0526/01/50051

a point along the x-axis, δ, when it suddenly changes with no discontinuity in the regression relationship. For a simple two-line linear regression this is equivalent to

$$\begin{aligned} f(x_i) &= \alpha_1 + \beta_1 x_i \qquad && X_0 \leqslant x_i \leqslant \delta, \\ &= \alpha_2 + \beta_2 x_i \qquad && \delta \leqslant x_i \leqslant X_1, \end{aligned}$$

where the parameters are constrained so that $\alpha_1 + \beta_1\delta = \alpha_2 + \beta_2\delta$, such that the function $f(x)$ is continuous, although not differentiable at the changepoint. The least squares estimates of the regression parameters can be derived from the normal equations. The design matrix X for parameters $\boldsymbol{\beta} = (\alpha_1, \beta_1, \beta_2)'$ is

$$X = \begin{pmatrix} 1 & x_1 - \delta & \delta \\ 1 & x_2 - \delta & \delta \\ \vdots & \vdots & \vdots \\ 1 & x_t - \delta & \delta \\ 1 & 0 & x_{t+1} \\ 1 & 0 & x_{t+2} \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_T \end{pmatrix}$$

where x_t is the point corresponding to the changepoint on the x-axis (α_1 is estimated from $\alpha_1 = \alpha_2 + (\beta_2 - \beta_1)\delta$), giving the estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Alternatively the parameters for each half of the model can be estimated from

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{pmatrix} - \frac{s}{t} C^{-1} q \tag{1}$$

where

$$\begin{aligned} \boldsymbol{\beta}_1^{*\prime} &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1, \\ \boldsymbol{\beta}_2^{*\prime} &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}_2, \end{aligned}$$

the unconstrained maximum likelihood estimates of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, i.e. a two-line model that is not constrained to meet at a known changepoint in the range of the data. A full description and the derivation of the parameters is given in Appendix A. From equation (1), it is evident that one way of deriving the least squares estimates of the parameters is first to estimate the parameters of the two-line model where the lines are not constrained to meet at a known changepoint, and then to adjust these unconstrained estimates so that the two lines are constrained to meet at a known changepoint. The advantage of deriving the parameters by using equation (1) will become evident later in the paper.

2.2. Testing for a regression change when the changepoint is known

A two-line regression model will have residual sums of squares (RSSs) that are not larger than those for the corresponding one-line model. Therefore, to test whether the two-line model has a statistically better fit the total RSS can be used to see whether the more complicated two-line model significantly reduces the error. This leads to an F-test:

$$F = \frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{\mathrm{RSS}_2/(T-3)}. \tag{2}$$

Here the RSS_1 and RSS_2 are the RSSs for the one- and two-line models respectively and T is the number of observations. The statistic has an F-distribution with 1 and $T - 3$ degrees of freedom.

3. Location of changepoint unknown

3.1. Estimation of model

When the location of the changepoint is unknown the problem is no longer linear. The only way to estimate the parameters is through numerical optimization. However, the numerical optimization is simplified through the use of equation (1) and the following three considerations (Hudson, 1966). Fit a two-line unconstrained model for points $x_1, \ldots, x_t$ and $x_{t+1}, \ldots, x_T$.

(a) If the two fitted lines meet between the adjacent extreme points of each model (x_t, x_{t+1}), then this model will have an RSS that is no larger than that for any other constrained model for these two sets of points constrained to meet between (x_t, x_{t+1}).
(b) If the two lines do not meet between x_t and x_{t+1}, then the constrained model with the smallest RSS will have a changepoint at either x_t or x_{t+1}.
(c) Constraining a model to meet at a required point will not decrease the RSS.

Thus, an algorithm can be easily derived (Fig. 1) that can estimate all the parameters in the model. This algorithm uses these three considerations, with equation (1) used in the programming. All unconstrained two-line models are fitted and the algorithm determines whether each of these models meets within the required region of x: (x_t, x_{t+1}).

The unconstrained models that meet within the required region are recoded as constrained models. The algorithm then determines whether the residual error from the best fitting constrained model is smaller than the residual error from the best fitting unconstrained model. If so, then the

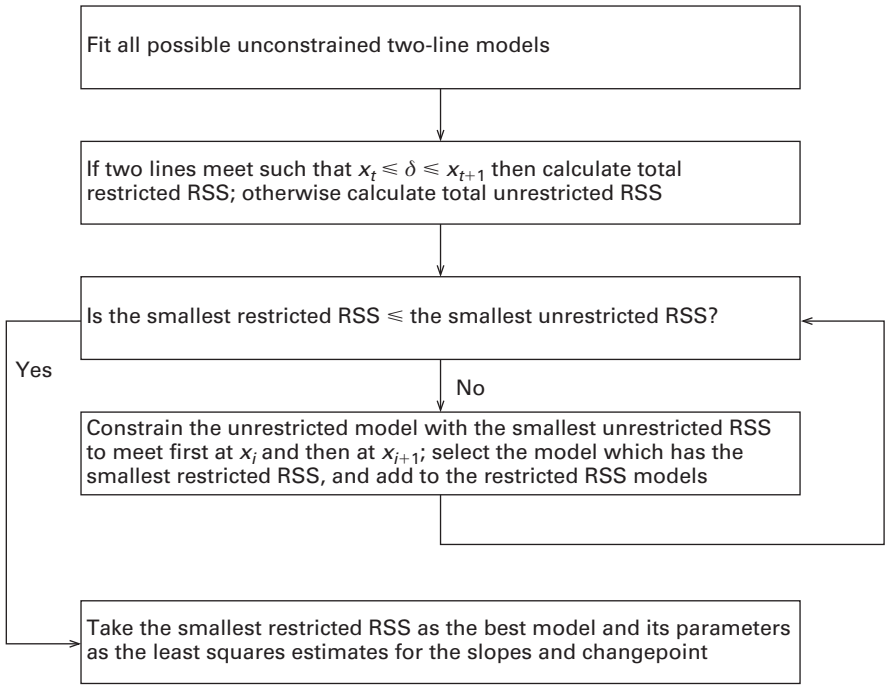

Fig. 1. Algorithm to obtain an estimate of the changepoint

algorithm stops and takes the constrained model with the smallest RSS as the least squares estimates. If not, then the best fitting unconstrained model is constrained, using equation (1), to meet at either x_t or x_{t+1} and added to the constrained models. This process is repeated until we obtain the least squares estimates. Fig. 1 more clearly explains the iterative process.

An alternative algorithm to estimate the parameters of the changepoint regression is the Gauss–Newton algorithm (Thisted, 1988). However, Fig. 2 highlights the main issue in using the Gauss–Newton algorithm for this particular problem. Fig. 2 is a plot of the RSS of 98 two-line models constrained to meet at each possible point on the x-axis ($x_2 = 2$, $x_3 = 3$, ..., $x_{99} = 99$). The original data were simulated under a null model of a common slope of 2 (intercept 0), with a variance of 100. As is evident from Fig. 2, local minima can occur in the RSS along the x-axis which can cause problems when using the Gauss–Newton algorithm (Draper and Smith, 1981). However, as the algorithm in Fig. 1 is tailored for this specific changepoint problem, it is not affected by local minima and is thus more efficient at parameter estimation.

3.2. Testing for a change in slope

An F-statistic can be derived (Worsley, 1983) that again uses the ratio of the sum of squares between the one- and two-line models:

$$F = \frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/2}{\mathrm{RSS}_2/(T-4)}. \tag{3}$$

If the changepoint has to be estimated, this no longer has an exact F-distribution under the null hypothesis (Hinkley, 1988). If the statistic had an exact F-distribution it would be with 2 and $T - 4$ degrees of freedom.

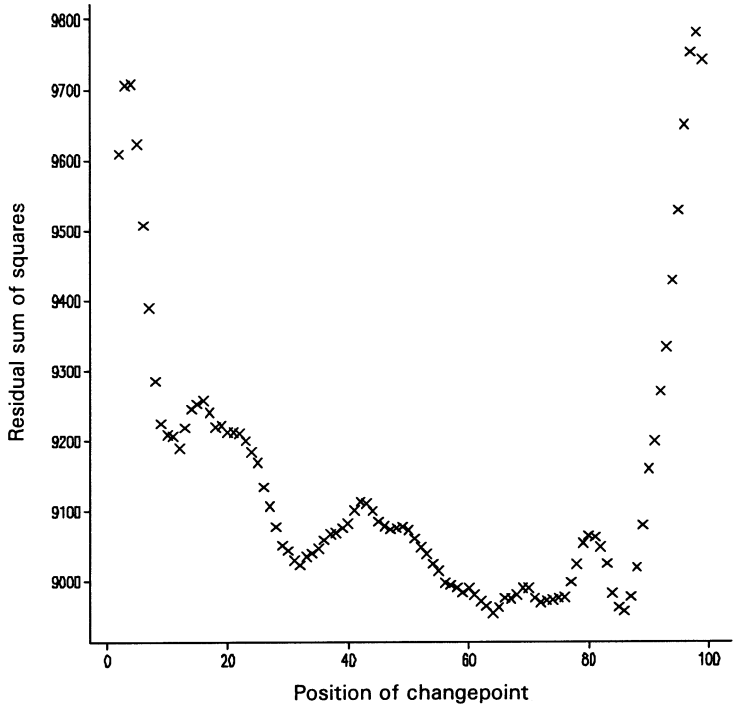

Fig. 2. RSS for the two-line model *versus* the position of the change

3.2.1. Simulation

To investigate how well the distribution of the 'F'-statistic is approximated by the F-distribution, simulations were performed using the interactive matrix language in SAS (SAS Institute, 1985). Simulated F-values were generated by fitting a two-line model to a set of data simulated from a one-line model. The null model was assumed to have a common slope of 2 (intercept 0), with a variance of 100. The simulation was repeated 1000 times. If regular asymptotic theory could be applied then the 'F'-statistic would have an F-distribution on 2 and $T-4$ degrees of freedom. 100 points were fitted by each model ($x_1 = 1$, $x_2 = 2$, ..., $x_{100} = 100$), giving an F-distribution on 2 and 96 degrees of freedom. Fig. 3 gives a probability plot of simulated F-values, against ranked deviates distributed as F on 2 and 96 degrees of freedom. This plot looks fairly straight, except that there is a slight kink in the line at the beginning and at the end of the plot, but the slope does not seem to be 1.

An F-distributed random variable with m and n degrees of freedom has expected mean and variance (Mood *et al*., 1974)

$$\text{mean} = \frac{n}{n-2} \qquad \text{for } n > 2,$$

$$\text{variance} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad \text{for } n > 4.$$

Therefore, for an F-test on $m = 2$ and $n = 96$ degrees of freedom the expected mean and variance are 1.021 and 1.088 respectively. The mean and variance of the simulated F-values were 1.687

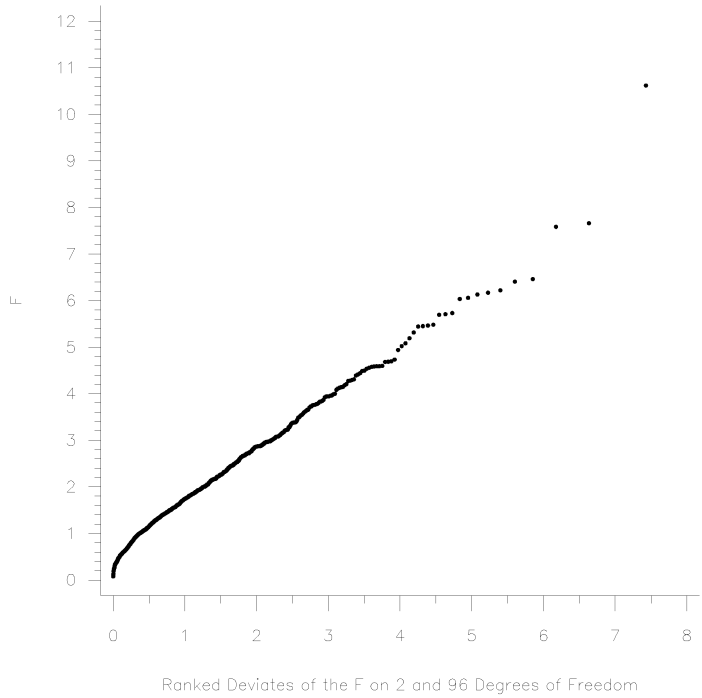

Fig. 3. Probability plot of ranked simulated F-values against ranked deviates distributed as F on 2 and 96 degrees of freedom

and 1.405, 65% and 30% bigger than expected. Thus, the probability plot and the deviation from the expected mean and variance suggest that asymptotic theory cannot be applied to the F-test.

3.3. Bootstrapping

The F-test mentioned previously relies on assumptions regarding the distribution of the parameters. It is these assumptions which cause the test to fail. Efron and Gong (1983) proposed nonparametric bootstrap methods to overcome problems when using parametric tests. Bootstrap methods have been recommended for linear regression analysis (Bunke and Droge, 1984; Wu, 1986) and for the extension of linear regression, changepoint regression (Hinkley, 1988), as well as a situation analogous to changepoint regression, mean shift models (Hinkley and Schechtman, 1987). These bootstrap methods are now investigated in detail for the changepoint problem.

The methodology in applying bootstrap methods to the changepoint problem is quite straightforward.

Step 1: for a given set of data obtain the best fitting two-line and one-line models and calculate the F-statistic.
Step 2: calculate the residuals for the two-line case.
Step 3: using the original x-values, recalculate the new y-values, by using the values from the best fitting one-line model and adding an error term, sampled with replacement from the set of residuals from the best fitting two-line model.
Step 4: to this new set of data, fit a two-line and a one-line model and calculate the F-statistic.
Step 5: repeat steps 3 and 4 a large number of times, each time using the one-line parameters and two-line residuals from the original data.

A bootstrap distribution for the F-test can be derived and a P-value can thus be calculated. The methodology is quite computer intensive although the algorithm in Fig. 1 speeds up the estimation of the parameters.

3.3.1. Simulation

To investigate the properties of the bootstrap for the changepoint regression problem a simulation exercise was undertaken. Simulated results were again generated by using the interactive matrix language in SAS (SAS Institute, 1985).

Simulations were initially undertaken to investigate the influence of the location of a changepoint on the x-axis on the power of a test for various changes in slope. The data were simulated for regression changes at all the points along the x-axis from x_2 to x_{49} on a 50-point scale. 100 simulations were carried out for each point on the x-axis to estimate the empirical power. For each simulation a bootstrap distribution of 100 points was generated and a bootstrap significance level of 5% was chosen (thus $48 \times 100 \times 100$ simulations were done). Fig. 4 gives the empirical power from the simulations of regression changes at various points along the x-axis, for various values of a standardized difference d, where d is defined as $(\beta_1 - \beta_2)/\sigma$, i.e. the difference in slopes before and after the changepoint, standardized by dividing by σ, the standard deviation about the two-line model. The lines are jagged owing to the noise in the simulations. The power is greatest for a changepoint that is near the centre of the x-axis and falls towards each end of the range. The power also increases with increasing sizes of the standardized difference d. The implication of these results is that, when designing a study to investigate a possible regression change, it should if possible be ensured that there are the same numbers of points before the changepoint as after it, to guarantee the appropriate power.

An equivalent simulation was undertaken to assess the effect that the number of points in the

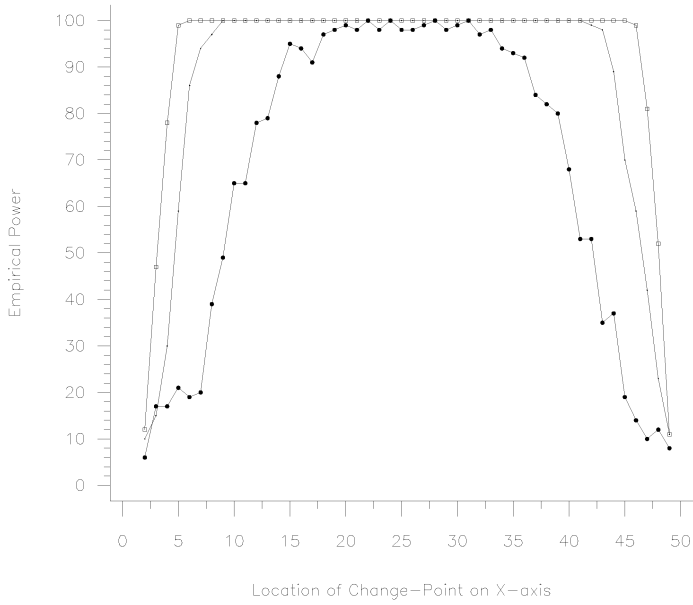

Fig. 4. Empirical power *versus* the location of the regression change for various slope differences: ●, standardized difference 0.2; ———, standardized difference 0.5; □, standardized difference 1.0

regression analysis had on the power. Data were simulated for regression changes at the midpoint along the x-axis for various numbers of points in the regression (from 10 to 50) for various standardized differences d. 100 simulations were carried out for each number of points to estimate the empirical power. For each simulation a bootstrap distribution of 100 points was generated and a bootstrap significance level of 5% was chosen (thus $40 \times 100 \times 100$ simulations were done). Fig. 5 gives the empirical power for a regression change for various numbers of points. From these results it seems that for a large regression change at the midpoint of the x-axis the number of points required is quite small (14 for $d = 1$ at 80% power) with an increasing number of points required for smaller standardized differences.

4. Worked example

When people exercise they need to produce energy and there are different metabolic pathways by which this energy is obtained (aerobic and anaerobic). For a given individual it is important to know whether a given pathway changes during exercise and, if so, when. One way of detecting this is through examining the relationship between two metabolic variables over time while the person is exercising. In this specific example a rower was connected to measuring equipment that reads certain physical responses over time. The workload was increased over time, i.e. the resistance of the rowing machine to the rower was increased.

The variables considered here are those of volume of oxygen inhaled and carbon dioxide exhaled in 1 min. The measurements were taken every 30 s up to a maximum of 17.5 min. What is of interest is whether there is an approximately linear relationship between the two variables or whether there is a change in slope once a critical level of oxygen inhalation is reached. The changepoint represents the point at which a subject switches metabolic pathways, from aerobic to anaerobic.

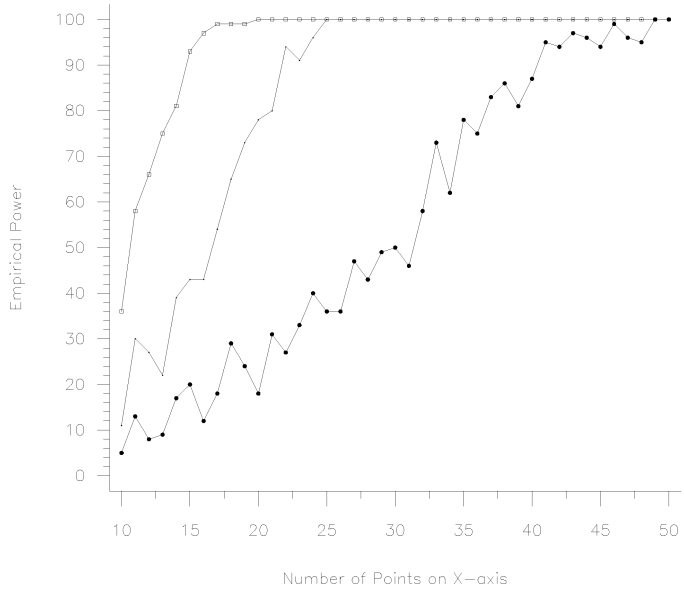

Fig. 5. Empirical power *versus* the number of points used in the regression change for various slope differences: ●, standardized difference 0.2; ———, standardized difference 0.5; □, standardized difference 1.0

The data are given in Table 1 and are plotted in Fig. 6. From looking at the data it seems that the variables increase over time and that there is some fluctuation due to random variation.

The best fitting single-line model with an RSS of 1.072 for carbon dioxide exhaled (Y_i) against oxygen inhaled (X_i) is

$$Y_i = -0.659 + 0.067X_i$$

Table 1. Data collated from measurements over time: volume of oxygen inhaled per minute and volume of carbon dioxide exhaled

Time ordering	*Volume of oxygen (X) (l min^{-1})*	*Volume of carbon dioxide (Y) (l min^{-1})*	*Time ordering*	*Volume of oxygen (X) (l min^{-1})*	*Volume of carbon dioxide (Y) (l min^{-1})*
1	12.5	0.75	19	44.2	2.12
2	26.2	1.12	20	47.9	2.35
3	24.8	0.98	21	49.9	2.50
4	27.4	1.13	22	48.1	2.48
5	31.1	1.31	23	48.4	2.49
6	34.6	1.47	24	51.7	2.71
7	21.5	0.93	25	51.8	2.74
8	27.9	1.34	26	55.5	3.00
9	29.2	1.36	27	54.9	3.02
10	35.2	1.60	28	57.0	3.21
11	32.6	1.47	29	57.9	3.30
12	34.9	1.57	30	58.3	3.37
13	34.9	1.59	31	58.2	3.42
14	37.6	1.73	32	59.5	3.53
15	36.3	1.68	33	59.7	3.55
16	40.1	1.88	34	61.8	3.76
17	42.7	2.01	35	48.4	2.96
18	43.4	2.07			

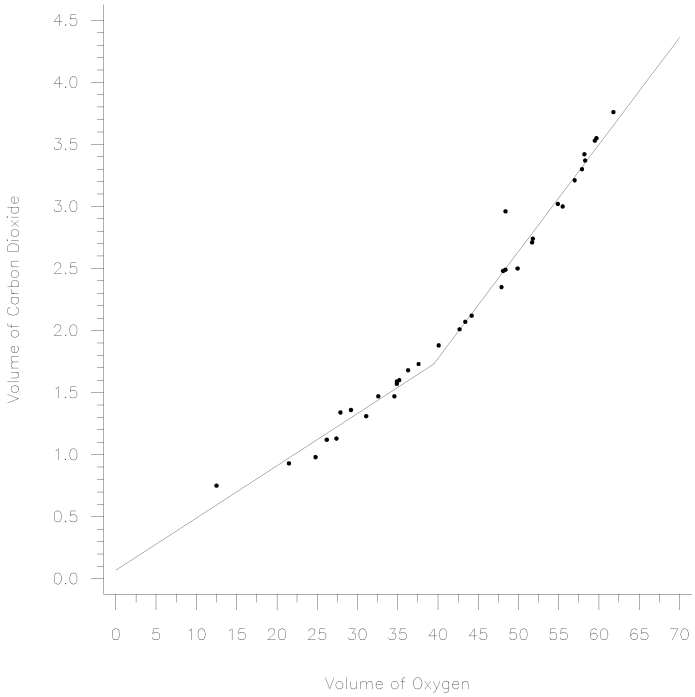

Fig. 6. Volume of carbon dioxide exhaled (litres per minute) *versus* volume of oxygen inhaled (litres per minute)

and the best fitting two-line model with an RSS of 0.389 is

$$\begin{aligned} Y_i &= 0.076 + 0.042X_i \qquad 12.5 \leqslant X_i \leqslant 39.46, \\ Y_i &= -1.659 + 0.086X_i \qquad 39.46 \leqslant X_i \leqslant 61.8. \end{aligned} \tag{4}$$

A comparison of the two- and one-line models gives an F-statistic of

$$\frac{\frac{1}{2}(1.072 - 0.389)}{0.389/31} = 27.21.$$

If the statistic had an exact F-distribution it would be with 2 and 31 degrees of freedom. The bootstrap P-value (on 1000 simulations) is 0.001. The two-line model of best fit is highlighted in Fig. 6. A visual inspection of Fig. 6 gives the impression that this model represents the data well. Although there is more than a twofold increase in the slope between the two halves of the model, the standardized difference, at just $0.071 = (0.086 - 0.042)/0.624$, is quite small.

There is thus strong evidence to suggest that the linear relationship between the amount of carbon dioxide exhaled and oxygen inhaled changes once the amount of oxygen exceeds about $39\,\mathrm{l\,min^{-1}}$. This could be due to the fact that at the start of the exercise, during the aerobic production of energy, oxygen is used, but as the exercise becomes more difficult the rower's energy requirement exceeds the quantity that can be produced through the aerobic pathway alone. At this point the rower starts to make use of anaerobic energy production and this causes the sudden change in the linear relationship between the volumes of carbon dioxide and oxygen.

These data could also be fitted by a curve. Indeed for this particular example an exponential curve would have a slightly better fit with an RSS of 0.337, with one fewer parameter. However, a

curve would not represent either the physiology or the objectives of this pharmacological model. In practice the rowing model would be used in healthy volunteers, in early phase drug development, to investigate the possible pharmacological activity of a new chemical entity and more than one subject would undertake the challenge, maybe in a crossover trial with a number of regimens or doses. New chemical entities that could be investigated in this pharmacological model are therapies which increase glycogenolysis, increasing hepatic and muscle glycogen stores or therapies that reduce the production of lactic acid, such as creatinine-containing products. These types of therapies would be expected to delay the changepoint from aerobic to anaerobic production.

5. Discussion

This paper has proposed an algorithm for parameter estimation where there is an unknown changepoint. The algorithm is recommended for changepoint regression as it is tailored for this specific problem and overcomes issues associated with local minima. Finally, it is proposed that nonparametric bootstrap methods are used to assess changepoint models.

Appendix A

For any interval (X_0, X_1) on the real line the model is defined as

$$\begin{aligned} f(x_i) &= f_1(x_i; \beta_1) \qquad X_0 \leqslant x_i \leqslant \delta, \\ &= f_2(x_i; \beta_2) \qquad \delta \leqslant x_i \leqslant X_1. \end{aligned}$$

Given a set of observations $(x_1, x_2, \ldots, x_n) = (\mathbf{X}_1', \mathbf{X}_2')$, where $\mathbf{X}_1' = (x_1, x_2, \ldots, x_t)'$ and $\mathbf{X}_2' = (x_{t+1}, x_{t+2}, \ldots, x_T)'$, and corresponding values of Y, $(\mathbf{Y}_1', \mathbf{Y}_2')$, to obtain the least squares estimates of the parameters, S needs to be minimized, where S is defined as

$$S = (\mathbf{Y}_1 - \mathbf{X}_1\boldsymbol{\beta}_1)'(\mathbf{Y}_1 - \mathbf{X}_1\boldsymbol{\beta}_1) + (\mathbf{Y}_2 - \mathbf{X}_2\boldsymbol{\beta}_2)'(\mathbf{Y}_2 - \mathbf{X}_2\boldsymbol{\beta}_2),$$

where $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, subject to a linear constraint of the form

$$g(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2') = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')q = d. \tag{5}$$

For the situation of two straight lines, q would be $(1, \delta, -1, -\delta)'$ and

$$g(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2') = \alpha_1 + \hat{\beta}_1\delta - \alpha_2 - \hat{\beta}_2\delta = d.$$

For the special case of a continuous regression with no break, $d = 0$, whereas, for the general case, $d \neq 0$.

The minimum of S can be obtained by using the Lagrange multiplier argument, from

$$\frac{\mathrm{d}S}{\mathrm{d}\beta_i} + \lambda\frac{\mathrm{d}g}{\mathrm{d}\beta_i} = 0. \tag{6}$$

Hence, the maximum likelihood estimates of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ can be derived from

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1^* \\ \hat{\boldsymbol{\beta}}_2^* \end{pmatrix} + \frac{d - s}{t} C^{-1} q \tag{7}$$

where

$$\begin{aligned} \beta_1^{*\prime} &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{Y}_1, \\ \beta_2^{*\prime} &= (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2\mathbf{Y}_2, \end{aligned}$$

the unconstrained maximum likelihood estimates of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, and

$$C^{-1} = \begin{pmatrix} C_1^{-1} & 0 \\ 0 & C_2^{-1} \end{pmatrix}$$

with $\mathbf{C}_1 = \mathbf{X}_1'\mathbf{X}_1$ and $\mathbf{C}_2 = \mathbf{X}_2'\mathbf{X}_2$.

Using the notation of Hudson (1966),

$$s = (\beta_1^{*\prime}, \beta_2^{*\prime})q,$$
$$t = q'C^{-1}q;$$

thus, the solution for least squares estimates for the parameters for the general case where $(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')q = d$ given by Seber (1977) has been derived. By setting $d = 0$ such that $(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')q = 0$ the least squares estimates for the special case $\alpha_1 + \hat{\boldsymbol{\beta}}_1\delta = \alpha_2 + \hat{\boldsymbol{\beta}}_2\delta$ is derived, i.e.

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{pmatrix} - \frac{s}{t} C^{-1}q. \tag{8}$$

Thus, the parameters in the two-line regression model can be obtained via equation (8), which uses the unconstrained estimates for the slopes and adjusts them so that the two lines meet at the required point.

References

Bennett, G. W. (1988) Determination of anaerobic threshold. *Can. J. Statist.*, **16**, 307–316.

Bunke, O. and Droge, B. (1984) Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.*, **12**, 1400–1424.

Draper, N. R. and Smith, H. (1981) *Applied Linear Regression*, 2nd edn. Chichester: Wiley.

Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jacknife and the cross-validation. *Am. Statistn*, **37**, 36–48.

Hinkley, D. V. (1988) Bootstrap methods. *J. R. Statist. Soc.* B, **50**, 321–337.

Hinkley, D. V. and Schechtman, E. (1987) Conditional bootstrap methods in the mean shift model. *Biometrika*, **74**, 85–94.

Hudson, D. J. (1966) Fitting segmented curves whose join points have to be estimated. *J. Am. Statist. Ass.*, **61**, 1097–1129.

Krisnaiah, P. K. and Miao, B. Q. (1988) Review about estimation of change-points. In *Handbook of Statistics* (eds P. K. Krisnaiah and C. R. Rao), vol. 7, pp. 375–402. Amsterdam: North-Holland.

Lees, B., Molleson, T., Arnett, T. R. and Stevenson, J. C. (1993) Differences in proximal femur bone density over two centuries. *Lancet*, **341**, 673–675.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974) *Introduction to the Theory of Statistics*. London: McGraw-Hill.

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

———(1960) Tests of the hypothesis that a linear regression system obeys two separate regimes. *J. Am. Statist. Ass.*, **55**, 324–330.

SAS Institute (1985) *SAS/IML™ User's Guide for Personal Computers*, version 6 edn. Cary: SAS Institute.

Seber, G. A. F. (1977) *Linear Regression Analysis*. Chichester: Wiley.

Shaban, S. A. (1980) Change-point problem and two-phase regression: an annotated bibliography. *Int. Statist. Rev.*, **48**, 83–93.

Thisted, R. A. (1988) *Elements of Statistical Computing*. London: Chapman and Hall.

Worsley, K. J. (1983) Testing for a two-phase multiple regression. *Technometrics*, **25**, 35–42.

Wu, C. F. J. (1986) Jackknife, bootstrap and other re-sampling methods in regression analysis. *Ann. Statist.*, **14**, 1261–1295.