Neuroepidemiology

Neuroepidemiology 2006;27:188–200
DOI: 10.1159/000096149

Published online: October 11, 2006

Tutorial in Biostatistics: Analyzing Associations between Total Plasma Homocysteine and B Vitamins Using Optimal Categorization and Segmented Regression

Heejung Bang[a] Madhu Mazumdar[a] J. David Spence[b]

[a]Division of Biostatistics and Epidemiology, Department of Public Health, Weill Medical College of Cornell University, New York, N.Y., USA; [b]Stroke Prevention and Atherosclerosis Research Centre, Robarts Research Institute, London, Ont., Canada

Key Words
Vitamin B_{12} · Changepoint · Folate · Homocysteine · Optimal categorization · Segmented regression

Abstract
Data analysts consider standard regression models (e.g., generalized linear model) or nonparametric smoothing techniques (e.g., loess or splines) when examining the association between two variables. Before this step, a quantile-based summarization is typically used for exploring the exposure-response relationship. Unfortunately, these exploratory approaches may not be optimal or efficient for guiding the formal analysis in many biological and nutritional data settings. We suggest a recently developed method for selection of cutpoints as a tool of data summary and segmented regression as a modeling approach in the analysis of plasma total homocysteine and related vitamins. These methods are often complementary in discovering the underlying complex pattern of association.

Introduction

When a bivariate association is examined for two factors such as an explanatory variable and a response, a crude summary table or figure is commonly used. Researchers conventionally construct such a presentation mode by categorizing a prognostic factor by quantiles of its distribution and then providing descriptive statistics of the response variable within each category. This approach is primarily used because of its simplicity and ease in understanding but may misrepresent the true relationship.

Finding the optimal rule for variable categorization is an important statistical undertaking, because widely accepted ad hoc methods are known to be inefficient [1]. Identifying too few categories may not be very informative, while too many categories tend to yield an unstable model fit with the risk of excessive fluctuations. The issue of estimating 'optimal cutpoint' has been extensively studied and widely utilized in prognostic factor modeling, in deciding cutoff of normal values for laboratory markers, and medical decision making in patient management. Mazumdar and Glassman [2] and Mazumdar et al. [3] reviewed various statistical solutions for this problem in bivariate and multivariate settings with bi-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0251–5350/06/0274–0188$23.50/0

Accessible online at:
www.karger.com/ned

Dr. Heejung Bang
Department of Public Health, Weill Medical College of Cornell University
411 East 69th Street
New York, NY 10021 (USA)
Tel. +1 212 746 1667, Fax +1 212 746 8544, E-Mail heb2013@med.cornell.edu

nary and censored survival endpoints. These methods are based on adjusting the p values due to multiple testing. However, the authors restricted their attention to one fixed cutpoint, which is useful for risk stratification (e.g., high vs. low risk), but not as useful for summary table construction. No suggestion was made for determining the number of categories.

Recently, O'Brien [4] proposed an efficient and reliable method for estimating the number of categories and the placement of associated cutpoints when a continuous covariate needs to be categorized in relation to the response. In this nonparametric approach, the optimum categorization is defined as the partition that minimizes a distance measure between the expected value of the outcome for each subject and the corresponding estimated average outcome among subjects in the same category. This method is mainly useful for creating tabular and/or graphical summaries of the exposure effect on the response.

Following the above exploratory and descriptive approaches, regression analysis is well accepted to be the standard tool in establishing the potential risk factors that affect the response. The type of regression model to be used depends on the research goal, data type (e.g., continuous, binary, censored outcome), study design (e.g., cross-sectional, longitudinal), appropriate summary measure (e.g., mean, median), type of hypothesized relationship (e.g., linear vs. nonlinear; simple vs. multiple regression), and various statistical approaches (e.g., parametric vs. nonparametric; fixed vs. time-varying regression coefficients).

In biological and nutritional data, multiple cutpoints often exist, meaning the relationships are expected to be different in the segments created by the cutpoints. The methodology called 'segmented (polynomial) regression modeling', in which above and below fixed but unknown critical point(s), the regression lines/curves are expected to change substantially seems to be an appropriate choice for this setting [5–8]. This model utilizes the full power of the continuous exposure data and does not make the unrealistically simple (log) linear model assumption. There are many possible scenarios, but a common one is that the exposure (or dose) increases the response linearly up to some level and the risk remains constant for greater exposure. The opposite scenario is also possible (i.e., there is no influence of the exposure on the response below a certain limiting value, while it increases when the exposure exceeds that threshold). In these situations, estimating transition points correctly is often regarded as the main goal of research. This problem has been extensively studied in (bio)statistics, but is not widely adopted in analyzing clinical data.

We bring together these two statistical methods (i.e., optimal cutpoints and segmented regression), and revisit the relationship of plasma total homocysteine (tHcy) and serum levels of vitamin B_{12} (cobalamin) and folate as agents that are involved in its metabolism. This clinical observation has been discussed in many papers [9–13]. Robertson et al. [12] used a quartile plot that displayed the negative linear association between serum B_{12} and tHcy, while some nonlinear patterns were shown in decile plots by Selhub et al. [11] and Spence et al. [13]. It was concluded that these findings are not necessarily convergent and therefore definite statements are difficult to make. It was also acknowledged that the study populations are not homogeneous, but comparable patterns were observed in some common ranges of B_{12} level.

In this paper, we perform a more rigorous investigation of this issue utilizing the nationally representative National Health and Nutrition Examination Survey (NHANES) data. We begin by reviewing various methods for categorization of a continuous covariate in relation to a continuous outcome, and also contrast the segmented regression model with other traditional competing approaches in the context of the NHANES data analysis. We find that using the O'Brien method for finding the optimal 'number of categories' followed by segmented regression (with multiple cutpoints in bivariate and multivariate settings) illustrates the tHcy-vitamin association more clearly than what was achieved by standard methods. In the 'Appendix', we provide simple codes that can be implemented in SAS and S-plus/R along with the suggested steps for performing similar data analysis.

Methods

The context of our problem is discovering the relationship between two numeric variables (dose/response; exposure/response; covariate/outcome) through finding changepoint(s) for the exposure variable. We review available statistical methods for (1) categorization of the covariate and (2) regression-based approaches for modeling the association. The first part is more for exploratory analysis and the latter is for statistical estimation and inference, guided by the results from the first part. We first describe methods commonly used by data analysts currently, followed by approaches that are either new or not utilized commonly to their fullest extent.

Commonly Utilized Methods

In any research that studies a covariate effect on an outcome, most would agree that a two-dimensional scatter plot should be

the initial step for visualization of the crude pattern of the raw data and the potential outliers. Without this simple but essential step, the whole effort to be followed could be misleading. Then one may attempt a tabular or graphical summary of the covariate-outcome relationship. When the covariate is continuous, it is not meaningful to make a table for all distinct values for the covariate. It is common practice to divide the covariate into a certain number of groups for creating straightforward summarization of the data. Groups with nearly equal numbers of observations are generally used (for example, tertiles to quintiles). Statistically oriented researchers may use tree-based methods (e.g., classification and regression tree and recursive partitioning) [14, 15], but these methods also are not necessarily optimal [4].

The next step in data analysis typically is to try various modeling approaches. In this stage, a generalized linear model (GLM) using polynomial (often, up to the cubic) terms of a covariate is most likely to be the starting point. If the relation seems to be more complex than polynomials, a smoothed regression model can be a better alternative (e.g., loess and splines). Although this approach is quite complex and advanced, procedures built into standard statistical packages make possible its widespread use. The 'loess' method has become part of statistical training and common practice for many nonstatisticians. A variety of specialized smoothing techniques are also widely available (for example, 'sm' option in GPLOT, GAM and TRANSREG procedures in SAS). 'Loess' in particular combines much of the simplicity of linear least squares regression with the flexibility of nonlinear regression. Although implementation of these elegant techniques is becoming ever easier, operating mechanisms under the attractive graphics are still difficult to grasp for many users. So, when threshold effects or breakpoints are expected, loess provides some sense about where they are located, but formal evaluation and inference are still not easy.

The exact number and locations of changepoints, if they are present, are difficult to identify unless strong biological rationale or previous consolidated research supports the theory. Changepoints are not immediately apparent in a scatter plot, particularly for large datasets. In this situation, a 'parametric' segmented regression model is appealing, giving an explicit mathematical expression of the equation and formal estimation and inference with confidence interval (CI) about changepoints. Often in real applications, a piecewise linear regression model with one or two breaks is sufficient, although an extension to higher-order polynomials and/or an increased number of cutpoints does not entail appreciably more effort. Using a correct number of breakpoints is important. For example, if two segmentations are needed but the data are actually modeled with only one changepoint, the statistical analysis could be problematic, as we will demonstrate in our example below.

Novel Methods and Not Commonly Utilized Methods

Suppose that an explanatory variable x_i is 'continuous' but we want to categorize it, while the response variable y_i can be discrete, ordinal or continuous, where *i* indexes the total *n* subjects. We assume that the first two moments of y_i conditional on x_i are $E(y_i|x_i) \equiv \mu_i \equiv \mu_i(x_i)$ and $var(y_i|x_i) = \sigma^2 v_i \equiv \sigma^2 v(x_i)$, where 'E' denotes expectation and 'var' denotes variance.

O'Brien [4] proposed a new method that determines the 'asymptotically' optimal choice of categories of a predictor by minimizing a newly invented measure of distance called average expected distance (AED):

$$AED(\lambda) = E\{1/n \Sigma_i (\bar{y}_{\lambda i} - \mu_i)^2/v_i\} = 1/n \Sigma_i (\bar{\mu}_{\lambda i} - \mu_i)^2/v_i + \sigma^2/n \Sigma_i \bar{v}_{\lambda i}/(v_i n_{\lambda i})$$

where $\lambda = \{\lambda_0 < \lambda_1 < \ldots < \lambda_{k-1} < \lambda_k\}$ represents a partition of k ($k \leq n$) categories and $\bar{y}_{\lambda i}$, $\bar{\mu}_{\lambda i}$, and $\bar{v}_{\lambda i}$ are the sample average of y's, μ's and v's, respectively, in the category to which subject i is assigned with the corresponding sample size $n_{\lambda i}$. Here, $\bar{y}_{\lambda i}$ can be granted as an estimator of μ_i and two summations represent systematic error (information loss) and sampling variability. The performance of this method has been shown to be superior to and more reliable than most existing methods. Interested readers are referred to the original paper by O'Brien [4].

Next, we intend to briefly explain segmented regression models, which represent a form of nonlinear regression or changepoint model. O'Brien's [4] optimal categorization, outlined above, can also be viewed as piecewise constant regression within this rich class of changepoint models. Furthermore, these two techniques tend to complement each other because a figure obtained from optimal categorization often successfully reveals meaningful changepoints that are not noticeable in a busy plot with raw data points.

A general 'changepoints' model can be formulated as follows: a function of a predictor and a response can have a different analytic form, as well as parameters in different subdomains of the x axis. There exist qualitatively different scenarios – a function can be continuous or discontinuous at changepoints, and the locations, as well as the number of changepoints, can be known or unknown. In this paper, we restrict our attention only to the models with unknown but 'continuous-at-the-join points' [see ref. 6, 7, 16–19 for a wide application of such models and mathematical and numerical properties]. Let us introduce an equation for simple linear regression but the slope changes after a certain level of a covariate:

$$E(y_i|x_i) = \psi\{\beta_0 + \beta_1 x_i + \beta_2 (x_i - \tau)_+\} \quad (1)$$

where ψ^{-1} is a known link function for GLM, and $(a)_+$ is defined to be 0 if $a < 0$ and a if $a \geq 0$. Here, $\beta = \{\beta_0, \beta_1, \beta_2\}$ are the regression parameters and τ is the changepoint parameter. Additional changepoints as well as other covariates (e.g., confounders or effect modifiers) can also be introduced in Eq. (1).

Parameter estimates can be obtained by maximizing a judiciously selected likelihood function or nonlinear least squares. For instance, a maximizer of τ by utilizing the following profile likelihood can be sought:

$$f(\tau) = \max_{\beta, \eta} \Sigma_i \, \text{loglik}\{y_i, \beta_0 + \beta_1 x_i + \beta_2 (x_i - \tau)_+, \eta\}$$

where 'loglik' denotes the log-likelihood and η is a vector of nuisance parameters. Most statistical software (e.g., SAS, S-plus/R and SPSS) provides a procedure for nonlinear modeling, so 'segmented regression' can be flexibly programmed, although the method of optimal categorization is not yet a part of the statistical packages.

An important issue in the use of segmented regression, often overlooked in practice, is that we should have knowledge about the optimal number of segmentations. Akaike Information Criterion (AIC) or Bayesian IC (BIC), among others, can decide the optimal number by testing whether the broken line offers a sig-

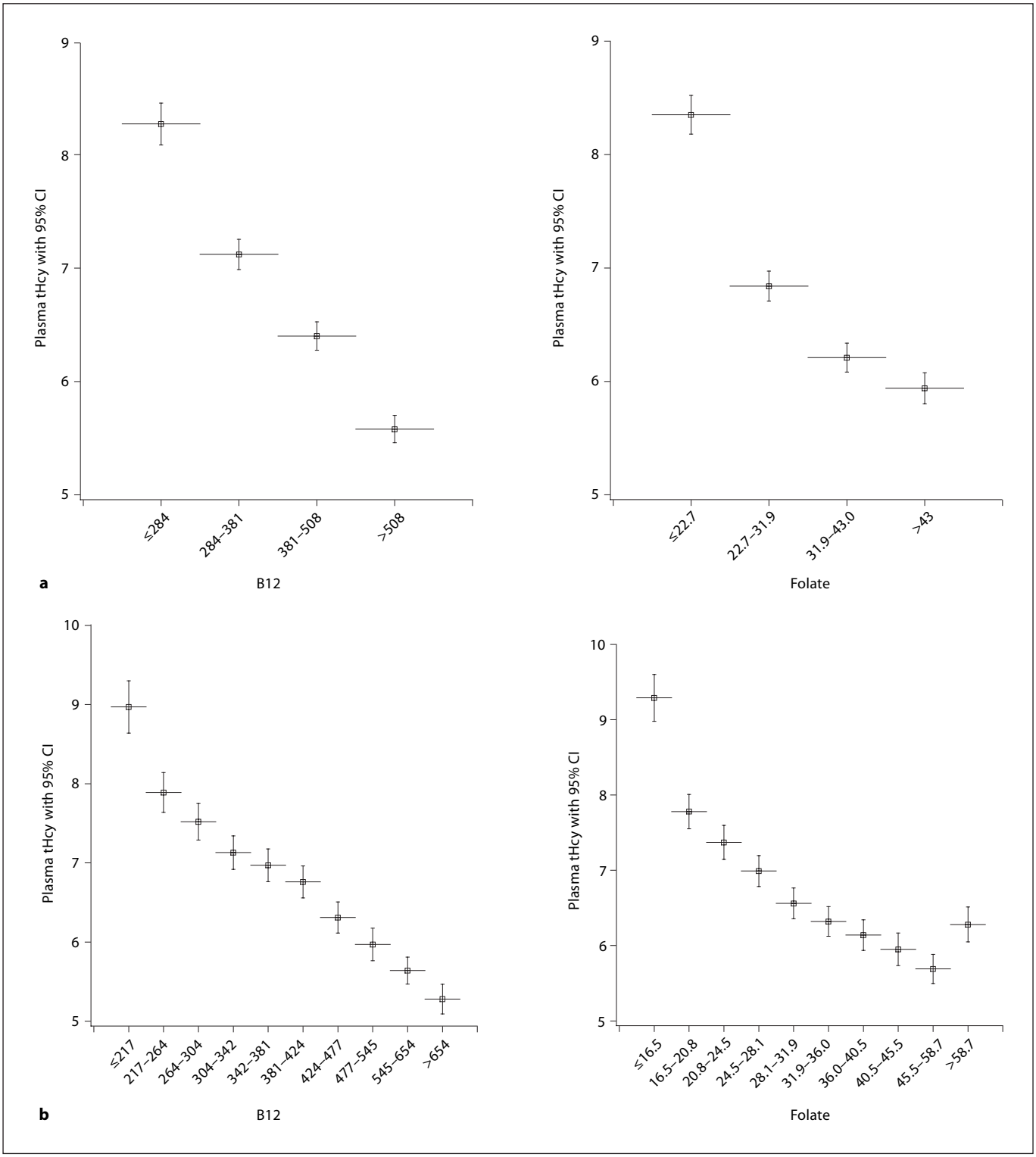

Fig. 1. The association of plasma tHcy with B_{12} and folate using quartiles (**a**) and deciles (**b**) – NHANES data (n = 7,260).

nificantly better fit to the data than a single straight line (i.e., with no changepoint), etc. in a stepwise fashion. Classical AIC assumes large sample size and may cause overfitting, so some suggested a modified version of AIC [20, 21]. Here are the formulae:

AIC = –2 loglik + 2 p,
BIC = –2 loglik + log(n) p,
a modified AIC = –2 loglik + {n(n + p)/(n – p – 2)},

where p = the number of estimated parameters. In the model of linear fits, p = 2*(the number of changepoints + 1). There are also some computational packages specializing in various changepoints models [17, 20, 22, 23].

We apply all methods discussed here to discover the relationship between tHcy and some B vitamins and find that the novel O'Brien method provided the best view of possible cutpoints for data summarization. Guided by the result from this method, a segmented regression with extension to bivariate and multivariable settings and use of a modified AIC for model selection finds the optimal setting.

Results

In this section, we analyzed the data in a part of the national survey, NHANES, that was conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, in several phases over a decade. This survey was designed by stratified probability samples of the civilian, noninstitutionalized US population with some underrepresentative subgroups oversampled; all the data are available in the public domain (http://www.cdc.gov/nchs/nhanes.htm). Comprehensive medical information and laboratory and interview data were ascertained.

For this investigation, 8,832 participants in the nutritional biochemistry examination during 1999–2000 were included. We extracted three variables: serum vitamin B_{12}, serum folate and plasma tHcy, and analyzed cross-sectional associations. Fifteen percent of the data were deleted due to missing values, resulting in 7,511 complete observations. In our main analysis, we also excluded 3% of the extreme values (to be specific, tHcy > 30.3, B_{12} > 1,096 or folate >94.3) so the final sample size was 7,260. We also repeated the same analysis using the entire dataset (n = 7,511) to examine the influence of outliers, and report the findings later.

We first analyzed the data with commonly used methods. Standard descriptive methods of categorizing the covariate (B_{12} and folate) using quartiles and deciles and plotting the midpoints corresponding to mean (and 95% CI) of the corresponding response (tHcy) are shown in figure 1. A summary table for deciles is provided in table 1

Table 1. Summary[1] of relationship between plasma tHcy with B_{12} (upper) and folate (lower) using deciles

Deciles	tHcy		
	sample size	mean	SD
B_{12} range			
1 (≤217)	733	8.97	4.57
2 (217–264)	717	7.89	3.44
3 (264–304)	728	7.52	3.17
4 (304–342)	732	7.13	2.92
5 (342–381)	734	6.97	2.85
6 (381–424)	704	6.76	2.74
7 (424–477)	725	6.31	2.71
8 (477–545)	739	5.97	2.86
9 (545–654)	726	5.64	2.35
10 (>654)	722	5.28	2.57
Folate range			
1 (≤16.5)	738	9.29	4.31
2 (16.5–20.8)	717	7.78	3.12
3 (20.8–24.5)	749	7.37	3.15
4 (24.5–28.1)	740	6.99	2.86
5 (28.1–31.9)	700	6.56	2.76
6 (31.9–36.0)	716	6.32	2.69
7 (36.0–40.5)	752	6.14	2.84
8 (40.5–45.5)	699	5.95	2.92
9 (45.5–58.7)	727	5.69	2.66
10 (>58.7)	722	6.28	3.18

SD = Standard deviation.
[1] Corresponds to the lower panels in figure 1.

as well. A small number of categories such as quartiles are generally insufficient to elucidate sophisticated curvature. The plot based on deciles reveals nonlinearity but fails to show a plateau pattern. Both are indexed by the number/order of observations and lose the original scale of a covariate. Tree analysis produced exceedingly large numbers of terminal nodes, 183 for B_{12} and 157 for folate.

In contrast, the O'Brien method yields 13 cutpoints for B_{12} and 12 cutpoints for folate (fig. 2). Interestingly, these two vitamins behave quite similarly in their effects on tHcy, with the patterns implying floor effects.

Next, we implemented various commonly utilized regression modeling strategies such as two polynomial regression models with (1) linear, quadratic and cubic terms of the covariate; (2) the same terms of the inverse of the covariate; (3) loess, and (4) smooth splines. In the parametric models of (1) and (2), each term is statistically significant with all p values <0.0025. Smoothing parameter in the loess fit was estimated as 0.236 for B_{12} and 0.112 for

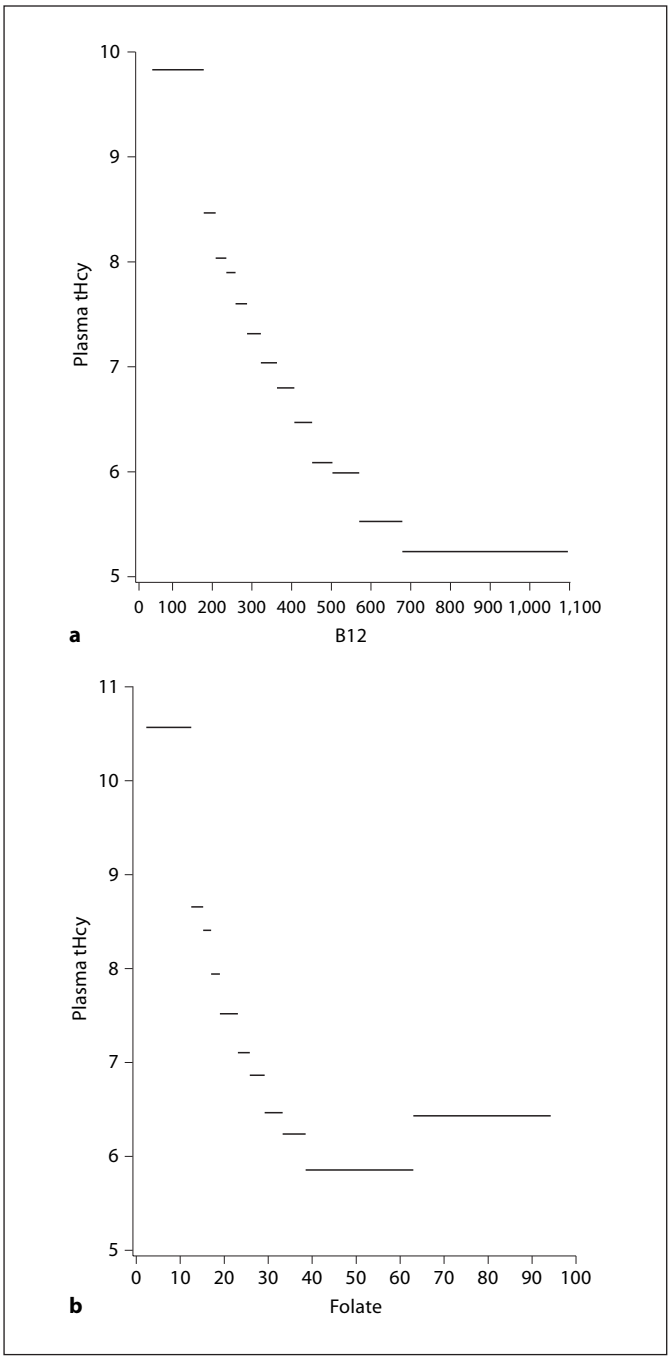

Fig. 2. The association of plasma tHcy with B_{12} (**a**) and folate (**b**) using O'Brien's optimal categorization method.

folate. Figure 3 overlaid the fitted curves from these four models. We find these methods to be suboptimal for our purpose as none of them prompt us to the number and the location of potential changepoints and, more importantly, the estimation and inference about changepoints are difficult, if not impossible.

Table 2. Model selection criteria for segmented regression model for plasma tHcy with B_{12} and folate

Number of changepoints	AIC[1] in B_{12} model	AIC[1] in folate model
0 (i.e., linear term)	23,671	23,984
1	23,546	23,586
2	**23,498**	23,474
3	23,498	**23,455**
4	23,503	23,458

[1] Using a modified version of AIC by Jones and Dey [20], where the smaller value indicates improved model fit.

Based on the result using the O'Brien method in figure 2, we obtained some evidence for the existence of (more than one) changepoints and the plateau pattern. We therefore reanalyzed the same data using nonsmooth segmented regression. AIC [20] presented in table 2 determined the optimal number of changepoints to be 2 for B_{12} and 3 for folate in relation to tHcy. Segmented regression fits (fig. 4 and the upper part in table 3) clearly capture the steep rise in plasma tHcy as serum B_{12} falls before the first breakpoint and the regression line becomes more flat after each changepoint thereafter: 184 (95% CI: 169–200) and 497 (95% CI: 450–544) pmol/l. In contrast, folate seems to convey three breakpoints of 11.6 (95% CI: 10.7–12.5), 26.6 (95% CI: 23.3–29.9) and 50.0 (95% CI: 46.0–54.0) nmol/l – an extremely steep slope seems to be true up to the first breakpoint and the lowering effects of serum folate on tHcy continue approximately up to the last breakpoint with intermediate slope changes.

Next, multivariate regression models adjusting key confounders such as age, sex and the status of the other nutrient were analyzed. Table 3 (the lower part) summarizes regression results that are somewhat different from the previous ones in bivariate counterparts (i.e., using one covariate). Models were dramatically better explained by introducing these potential confounders as reflected in sum of squares (i.e., 67,840–39,241 for the B_{12} model and 67,393–37,455 in the folate model). Two prominent changes emerged: in the B_{12} model, the second changepoint significantly moved downward (497 to 426); in the folate model, three breakpoints did not change materially, while the apparent increase in tHcy at very high folate (slope estimate = 0.04) almost disappeared (slope estimate = –1.00 + 0.88 + 0.09 + 0.04 = 0.01). This change may affect recommendations for vitamin intake.

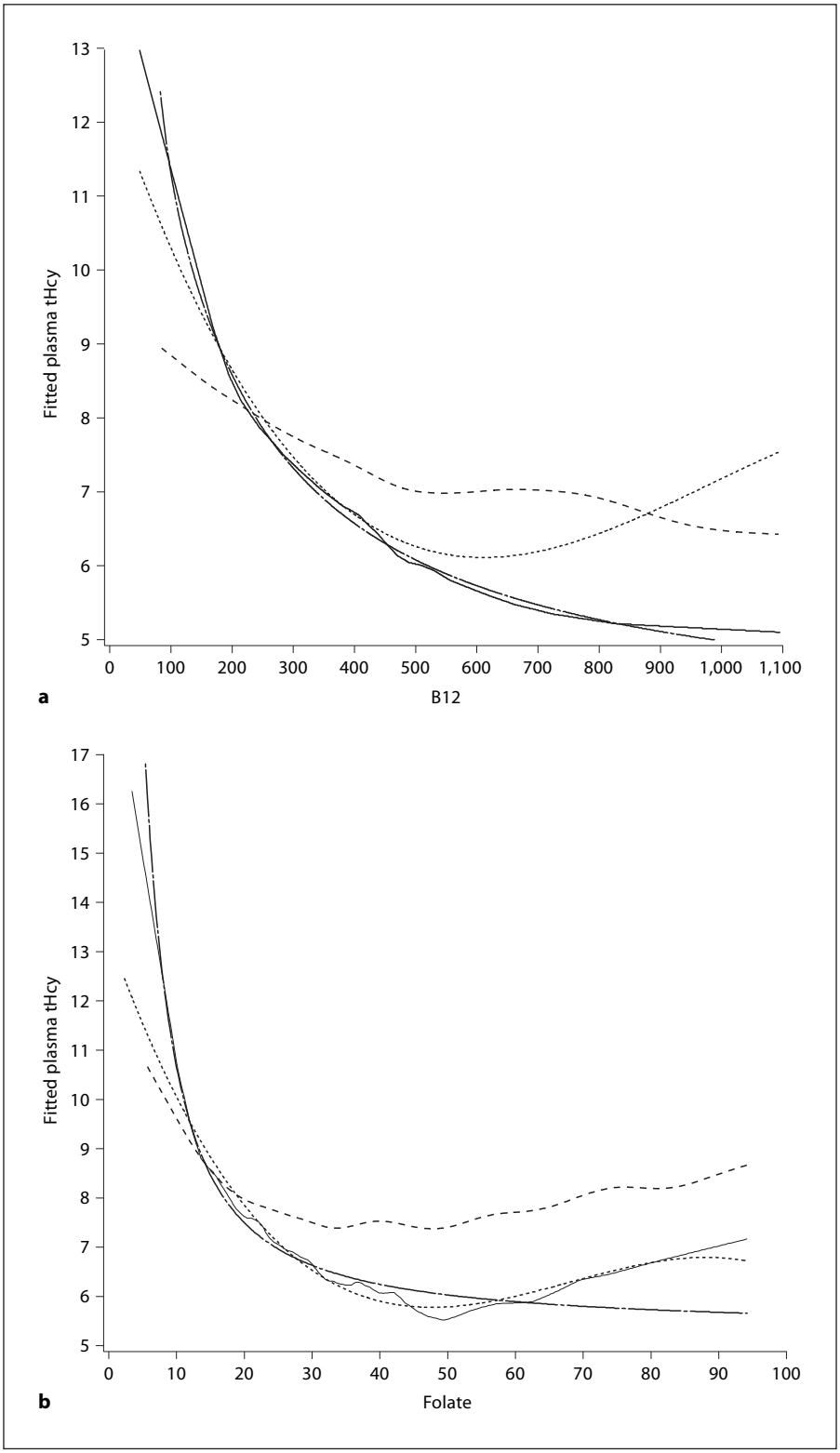

Fig. 3. Polynomial and inverse regression, spline, loess fits for plasma tHcy with B_{12} (**a**) and folate (**b**) unadjusted analysis. Solid, dotted, long dashed and short dashed lines represent loess, cubic polynomials, cubic polynomials of the inverse covariate and smooth splines, respectively.

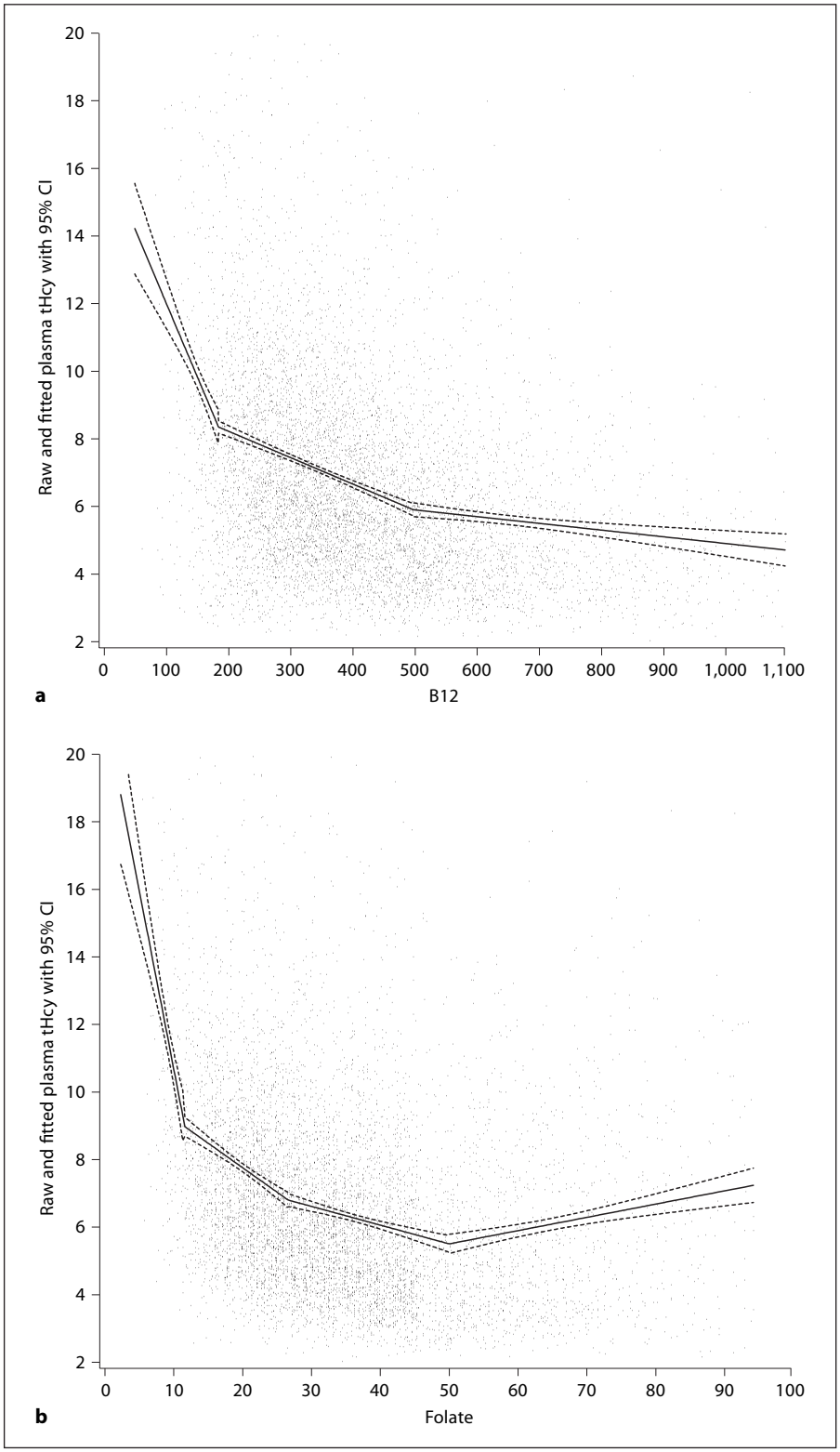

Fig. 4. Segmented regression fit for plasma tHcy with B_{12} (**a**) and folate (**b**) along with raw data – unadjusted analysis. Fitted regression functions are tHcy = 16.3443 – 0.0435 * B12 + 0.0358 * $(B12 - 184)_+$ + 0.00581 * $(B12 - 497)_+$ and tHcy = 21.2327 – 1.0567 * folate + 0.9118 * $(folate - 11.6)_+$ + 0.0893 * $(folate - 26.6)_+$ + 0.0947 * $(folate - 50)_+$. Note that wider y axis than the one in figure 3 was used to show entire data points.

Table 3. Segmented regression model for plasma tHcy with B_{12} and folate-unadjusted (above) vs. multivariate-adjusted (below) analyses

	B_{12} model	Folate model
Unadjusted analysis[a]		
Parameter	estimate (95% CI)	estimate (95% CI)
1st changepoint	184 (167, 199)	11.6 (10.7, 12.5)
2nd changepoint	497 (450, 544)	26.6 (23.3, 29.9)
3rd changepoint	not needed	50 (46, 54)
Initial slope	–0.04 (–0.06, –0.03)	–1.06 (–1.33, –0.79)
Slope change after		
1st changepoint	0.04 (0.02, 0.05)	0.91 (0.64, 1.18)
2nd changepoint	0.006 (0.004, 0.007)	0.09 (0.05, 0.12)
3rd changepoint	not needed	0.09 (0.07, 0.12)
Multivariate-adjusted analysis		
Age, years	0.083 (0.081, 0.085)	0.081 (0.079, 0.084)
Female gender	–1.01 (–1.12, –0.91)	–0.97 (–1.07, –0.86)
Nutrient[b]	–0.04 (–0.043, –0.036)	–0.002 (–0.0023, –0.0016)
1st changepoint	200 (187, 214)	11.5 (10.8, 12.2)
2nd changepoint	426 (367, 485)	24.5 (21.9, 27.1)
3rd changepoint	not needed	50.1 (43.8, 56.4)
Initial slope	–0.03 (–0.04, –0.02)	–1.00 (–1.22, –0.79)
Slope change after		
1st changepoint	0.026 (0.02, 0.03)	0.88 (0.66, 1.09)
2nd changepoint	0.003 (0.002, 0.005)	0.09 (0.06, 0.12)
3rd changepoint	not needed	0.04 (0.03, 0.06)

CI = Approximate confidence interval.
[a] Corresponds to the regression fits presented in figure 4.
[b] For the B_{12} model, it is folate and for the folate model, it is B_{12}.

Discussion

The methodological and computational advancements in the last decade have been remarkable in statistics and applied mathematics, but the complexity of the methodology has kept them from being widely used in biomedical applications. Some methods are not understood well, while others are not accessible because user-friendly computer programs are not available. In this paper, we highlight two statistical methods with important potential for applications in epidemiology and biomedicine: (1) how to categorize a continuous covariate optimally to describe its underlying relationship to a study outcome and (2) how to address potential changepoints if they exist.

Thresholds and changepoints are key features in latency and dose-response analyses in virtually the entire realm of epidemiology (e.g., environmental, occupational, nutritional, clinical and neuroepidemiology). All categorization procedures, including those illustrated here, are inherently data-driven. Scientifically or biologically meaningful cutpoints should be adopted whenever possible. However, when such information is not available a priori, which is indeed often the case in new or emerging research, the methods for finding 'optimal' cutpoints are a valuable component of the statistical toolbox. For the resulting cutpoints to be reasonable, they should be derived from data that are large enough and reliable. Otherwise, such results are seldom reproducible in independent investigations and cannot avoid the criticism of being 'data-dependent'.

Strictly speaking, statistical analysis for NHANES data should account for complex survey design. However, we analyzed this dataset assuming simple random sampling. To the best of our knowledge, nonlinear programming and the associated 'correct' statistical inference fully accounting for complex sampling design are not yet available, although standard weighted regression can be conducted easily. This should be an important topic for statistical research.

The utility and efficacy of cost-effective vitamin regimens on various medical endpoints such as vertical transmission of HIV, carotid plaque, cardiovascular diseases, and cognitive function have recently received growing

attention [12, 13, 24–30]. tHcy is not only known to be an independent risk factor for vascular events, but is also believed to lie in the pathway between B vitamins and medical outcomes [26, 28]. In the era of mandated folate fortification of the grain supply in North America (as of March 1996), it has been suggested that B_{12} is likely to play a key role in vitamin therapy for tHcy and the term of 'folate therapy' is no longer meaningful [12, 13, 29]. Previously, the covariate-outcome association was presented in a graphic or tabular form after a vitamin variable was naively categorized using quartiles, quintiles or deciles, and statistical modeling was not attempted in general. Some authors implied the plausibility of inherent nonlinear effects, but did not address this issue more rigorously [11, 13]. Our objective was to better understand the role of these vitamins in reduction of tHcy. We confirmed that segmented regression is an essential analytic tool in this type of efficacy study. Furthermore, the estimated changepoints may offer useful information for determining treatment dosage in designing clinical trials, and defining what are 'adequate' as opposed to 'normal' (generally meaning within 95% CI) levels of vitamins, and what should be the goal of therapeutic interventions in outcome studies.

The NHANES comprises a general population in the US. Therefore, the range of tHcy is considerably lower than that of subjects in clinical settings [12, 26]. It would be interesting to reanalyze these data utilizing the methodologies discussed here and also to study different populations or other subgroups of interest. Higher or lower doses of vitamin than are usually regarded as normal or adequate may be needed to maintain tHcy low in certain groups of people. For example, Rajan et al. [24] reported that elderly patients with serum B_{12} levels below 221 pmol/l (still within the 'normal' range) require 1,000 μg/day of B_{12} to obtain adequate absorption, whereas the generally accepted recommended daily intake is only 6 μg/day.

Segmented regression methodology is highly underutilized for data scenarios where it could be ideal, although its use is increasing [31–36]. One weakness of this methodology is that even when it is adequately coded, if the model is fit under an incorrect number of changepoints, the output obtained can be erroneous without any proper error message, or without prompting the need for changing the number of changepoints. All the publications in clinical applications cited above used a single cutoff value for fitting. Some might have used one changepoint informed by adequate model selection criteria or only for convenience. If these models were fitted without taking into account the possibility of multiple segments, reexamination of these data would be desirable. In our dataset, when we fitted a one-changepoint model, parameter estimates tended to converge to different midpoints depending on the starting values (results not shown). If such unstable estimation occurs, it can be seen as informal evidence that the model chosen (and the number of changepoints) is probably incorrect.

It is also important to keep in mind the effect of outliers on these methodologies, and to perform sensitivity analysis. We replicated our main analyses performed on a sample of n = 7,260 using the full dataset of n = 7,511 (also containing all implausibly large values). The resulting numbers of cutpoints were minutely changed. This is expected because the optimal cutpoints method by O'Brien is nonparametric, so it is robust to outliers. However, segmented regression is more vulnerable to undue influence due to outliers as expected from any parametric methods. Therefore, we recommend exclusion of those outlying observations before final analysis. Transformation (of response and/or explanatory variables) is a valid alternative for handling highly skewed data although neither method we advocate is invariant to variable-transformation and results should be interpreted with more care. Lastly, one should use reasonable starting values or, more realistically, try different starting values to ensure that sensible convergence (not to local minima) is reached. These are general recommendation for any nonlinear modeling.

In this tutorial, we illustrate how to assess the number of cutpoints needed for data presentation and to fit a segmented regression model in bivariate and multivariate settings with the possibility of multiple changepoints for a continuous outcome. More effort to popularize this methodology for binary, quantal or survival data is still needed [37–42].

It is important to mention that any optimal cutpoint procedure should not be used when one categorizes covariates in the regression model without proper consideration of multiplicity adjustment – it is well documented that maximally selected test statistics greatly inflate type-I error, so p values should be adjusted [2, 15, 43].

Sample size calculation is an important issue in the design of clinical and epidemiological studies. However, limited studies have been done on this topic. This issue is not as relevant for the O'Brien method as it is used for exploratory purposes. The issue that warrants thinking is that what value of sample size makes this method work so that the optimal number of cutpoints found and the trend remain stable. We may think in terms of competing

methods such as the ones based on quantiles or regression tree. Sample size issues in the context of both of these methods also are not clearly specified. The O'Brien method is found to be more efficient than the competing methods and hence we would need a similar (or smaller) sample size than that needed for them. Sample size determination for segmented regression is justified but requires substantive thoughts with respect to 'hypothesis' to be tested and 'effect size' to be detected, which can lead several different scenarios even in the one-changepoint scenario [35, 44]. In general, the location of the changepoint is the parameter of interest in segmented regression so the main interest lies in the estimation itself and we tend to not have a quantitative hypothesis about it being within some specific interval that reflects the desired precision and thereby required power.

Nowadays, novel changepoint models have been developed in more advanced statistical frameworks such as longitudinal data analysis. Naumova et al. [45] suggested a piecewise mixed effects model to address the subject-specific 'critical period' (e.g., the impact of menarche in obesity), while Wu et al. [46] proposed a changepoint mixed model for the analysis of a nonrandomized time-varying treatment. These advanced approaches are implemented easily in standard statistical software with minor programming. Therefore, we encourage adopting these valuable methods in application if needed instead of persisting with old (but not necessarily wise) methods.

Acknowledgments

We want to thank Professor O'Brien for providing the 'cutpoints' function. We also thank Professors Richard Jones and Douglas Hawkins for advice on computation.

Appendix: Suggested Procedure and Statistical Programming

A. Suggested Procedure for Optimal Categorization and Changepoints Model

Step 1: Apply O'Brien's cutpoints method to raw X and Y data (in addition to standard quantile-based method).

Step 2: Check if any changepoints are detected in the plot generated from step 1 (in addition to the raw scatter plot).

Step 3: If changepoints are clearly visible in step 2, then you may go to step 4. If changepoints seem to be present but the number (and locations) is not clear, determine that by model selection criteria (e.g., AIC/BIC). If there is no abrupt change, use standard regression. Delete outliers in this step.

Step 4: Fit a segmented regression model using the selected number of changepoints in step 3 (in simple or multiple regression). Use your guess for initial values of the parameters. Repeat the model run with different starting values and make sure that consistent results are achieved.

Step 5: Perform formal statistical estimation and inference.

B. Optimal Categorization for $y = tHcy$ and $x = B_{12}$

```
data1<-read.table("C:/Research/OBrien/nhanesdata.txt")
#assuming that nhanesdata.txt consists of two columns with the 1st column
   for (non-missing) tHcy and the 2nd column for (non-missing) b12;

b12.unsorted<-data1[,2]
thcy<-data1[order(b12.unsorted),1]
b12<- sort(b12.unsorted)

b12.rank<-rank(b12)
#for continuous normal outcome;
smooth.fit<-gam(thcy~s(b12.rank),family=gaussian)
fitted.vals<-fitted(smooth.fit)
dispersion<-summary(smooth.fit)$dispersion
result<-cutpoints(fitted.vals,y=thcy,x=b12,sig2=dispersion,exact=T)
print(result)
```

Remark: An S-plus function 'cutpoints' can be obtained by contacting Dr. O'Brien. The program is only available in S-Plus currently.

C. Linear Segmented Regression with y = tHcy and x = B_{12}

a. SAS [model with one changepoint (upper) and two changepoints (lower)]

```
proc nlin;
  parameters b0=13 b1=0.01 b2=0.01 tau=400;
  if x<=tau then do;                  *left linear segment;
        model y=b0+b1*x;
        der.b0=1;
        der.b1=x;
        der.b2=0;
        der.tau=0;
        end;
  else do;                            *right linear segment;
        model y=b0+b1*x+b2*(x-tau);
        der.b0=1;
        der.b1=x;
        der.b2=x-tau;
        der.tau=-b2;
        end;
output out=segout predicted=segpred u95m=u95mseg l95m=l95mseg;
run;

proc nlin;
  parameters b0=11.2 b1=-0.01 b2=0.009 tau=200 b3=0.001 tau2=700;
  if x<=tau then do;                  *left linear segment;
        model y=b0+b1*x;
        der.b0=1;
        der.b1=x;
        der.b2=0;
        der.tau=0;
        der.b3=0;
        der.tau2=0;
        end;
  else if tau<x<=tau2 then do;        *mid linear segment;
        model y=b0+b1*x+b2*(x-tau);
        der.b0=1;
        der.b1=x;
        der.b2=x-tau;
        der.tau=-b2;
        der.b3=0;
        der.tau2=0;
        end;
  else do;                            *right linear segment;
        model y=b0+b1*x+b2*(x-tau)+b3*(x-tau2);
        der.b0=1;
        der.b1=x;
        der.b2=x-tau;
        der.tau=-b2;
        der.b3=x-tau2;
        der.tau2=-b3;
        end;
 run;
```

b. S-Plus/R [model with one changepoint (upper) and two changepoints (lower)]

```
summary(nls(y~b0+b1*x+b2*(x-tau)*(x>tau),
          start=list(b0=13,b1=0.1,b2=0.1,tau=400), trace=T))
summary(nls(y~b0+b1*x+b2*(x-tau)*(x>tau)+b3*(x-tau2)*(x>tau2),
       start=list(b0=13,b1=0.1,b2=0.1,b3=0.01,tau=200,tau2=550),
       trace=T))
```

Remark: Some data may need to specify some options, for example, for convergence criteria, iteration and tuning methods. The examples above use default specifications. Check SAS and R/S-plus documents for more details.

References

1 Greenland S: Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995;6:356–365.
2 Mazumdar M, Glassman JR: Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. Stat Med 2000;19:113–132.
3 Mazumdar M, Smith A, Bacik J: Methods for categorizing a prognostic variable in a multivariable setting. Stat Med 2003;22:559–571.
4 O'Brien SM: Cutpoint selection for categorizing a continuous predictor. Biometrics 2004;60:504–509.
5 Feder PI: On asymptotic distribution theory in segmented regression problems – identified case. Ann Stat 1975;3:49–83.
6 Shaban SA: Change point problem and two-phase regression: an annotated bibliography. Int Stat Rev 1980;48:83–93.
7 Seber GA, Wild CJ: Nonlinear Regression. New York, Wiley, 1989.
8 Gallant AR, Fuller WA: Fitting segmented polynomial regression models whose join points have to be estimated. J Am Stat Assoc 1973;68:144–147.
9 Selhub J, Jacques PF, Bostom AG, et al: Relationship between plasma homocysteine and vitamin status in the Framingham Study population: impact of folic acid fortification. Public Health Rev 2000;28:117–145.
10 Brattström LE: Vitamins as homocysteine-lowering agents. J Nutr 1996;126:S1276–S1280.
11 Selhub J, Jacques PF, Rosenberg IH, et al: Serum total homocysteine concentrations in the third National Health and Nutrition Examination Survey (1991–1994): population reference ranges and contribution of vitamin status to high serum concentrations. Ann Intern Med 1999;131:331–339.
12 Robertson J, Iemolo F, Stabler SP, Allen RH, Spence JD: Vitamin B_{12}, homocysteine and carotid plaque in the era of folic acid fortification of enriched cereal grain products. Can Med Assoc J 2005;172:1569–1573.
13 Spence JD, Bang H, Chambless LE, Stampfer MJ: Vitamin intervention for stroke prevention trial: an efficacy analysis. Stroke 2005;36:2404–2409.
14 Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and Regression Trees. New York, Chapman & Hall, 1984.
15 Hawkins DM, Kass GV: Automatic interaction detection; in Hawkins DM (ed): Topics in Applied Multivariate Analysis. Cambridge, Cambridge University Press, 1982, pp 269–302.
16 Motulsky H, Christopoulos A: Fitting Models to Biological Data Using Linear and Nonlinear Regression. New York, Oxford University Press, 2004.
17 Hawkins DM: Fitting multiple change-point models to data. Comput Stat Data Anal 2001;37:323–341.
18 Hawkins DM: A note on continuous and discontinuous segmented regression. Technometrics 1980;22:443–444.
19 Liu J, Wu S, Zidek JV: On segmented multivariate regression. Statistica Sinica 1997;7:497–525.
20 Jones RH, Dey I: Determining one or more change points. Chem Phys Lipids 1995;76:1–6.
21 Hurvich CM, Tsai CL: Regression and time series model selection in small samples. Biometrika 1989;76:297–307.
22 Zeileis A, Kleiber C, Krämer W, Hornik K: Testing and dating of structural changes in practice. Comput Stat Data Anal 2003;44:109–123.
23 Muggeo VMR: Estimating regression models with unknown break-points. Stat Med 2003;22:3055–3071.
24 Rajan S, Wallace JI, Brodkin KI, et al: Response of elevated methylmalonic acid to three dose levels of oral cobalamin in older adults. J Am Geriatr Soc 2002;50:1789–1795.
25 Fawzi WW, Msamanga GI, Hunter D, Renjifo B, Antelman G, Bang H, Manji K, Kapiga S, Mwakagile D, Essex M: Randomized trial of vitamin supplements in relation to transmission of HIV-1 through breastfeeding and early child mortality. AIDS 2002;16:1–10.
26 Toole JF, Malinow MR, Chambless LE, et al: Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction and death: The Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial. JAMA 2004;291:565–575.
27 Elias MF, Sullivan LM, D'Agostino RB: Homocysteine and cognitive performance in the Framingham offspring study: age is important. Am J Epidemiol 2005;162:644–653.
28 Spence JD, Howard VJ, Chambless LE, Malinow MR, Pettigrew LC, Stampfer M, Toole JF: Vitamin Intervention for Stroke Prevention (VISP) trial: Rationale and design. Neuroepidemiology 2001;20:16–25.
29 Spence JD: Homocysteine: call off the funeral. Stroke 2006;37:282–283.
30 Pfeiffer CM, Caudill SP, Gunter EW, et al: Biochemical indicators of B vitamin status in the US population after folic acid fortification: results from the National Health and Nutrition Examination Survey 1999–2000. Am J Clin Nutr 2005;82:442–450.
31 Lamberson WR, Firman JD: A comparison of quadratic versus segmented regression procedures for estimating nutrient requirements. Poult Sci 2002;81:481–484.
32 Ansari F, Gray K, Nathwani D, et al: Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis. J Antimicrob Chemother 2003;52:842–848.
33 Baker K, Firman JD, Blair E, et al: Digestible lysine requirements of male turkeys during the 12 to 18 week period. Int J Poultry Sci 2003;2:229–233.
34 Shuai X, Zhou Z, Yost RS: Using segmented regression models to fit soil nutrient and soybean grain yield changes due to liming. J Agric Biol Environ Stat 2003;8:240–252.
35 Berman N, Wong WK, Bhasin S, Ipp E: Application of segmented regression models for biomedical studies. Am J Physiol 1996;270:E723–732.
36 Oktay K, Hourvitz A, Sahin G, Oktem O, Safro B, Cil A, Bang H: Letrozole reduces estrogen and gonadotropin exposure in women with breast cancer undergoing ovarian stimulation before chemotherapy. J Clin Endocrinol Metab, in press.
37 Cox C: Threshold dose-response models in toxicology. Biometrics 1987;43:511–523.
38 Milton RC, Kim J: Does vitamin E supplementation with 400 IU/d significantly increase all-cause mortality? Electronic letters published at Ann Intern Med – Rapid Responses for Miller et al., 2005;142:37–46.
39 Stasinopoulous DM, Rigby RA: Detecting break points in generalized linear models. Comput Stat Data Anal 1992;13:461–471.
40 Küchenhoff H, Ulm K: Comparison of statistical methods for assessing threshold limiting values in occupational epidemiology. Comput Stat 1996;12:249–264.
41 Pastor R, Guallar E: Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. Am J Epidemiol 1998;7:631–642.
42 Dupuy JF: Estimation in a change-point hazard regression model. Stat Prob Letter 2006, in press.
43 Betensky RA, Rabinowitz D: Maximally selected χ^2 statistics for k × 2 tables. Biometrics 1999;55:317–320.
44 Quandt RE: Tests of the hypothesis that a linear regression system obeys two separate regimes. J Am Stat Assoc 1960;55:324–330.
45 Naumova EN, Must A, Laird NM: Tutorial in biostatistics: evaluating the impact of 'critical periods' in longitudinal studies of growth using piecewise mixed effects models. Int J Epidemiol 2001;30:1332–1341.
46 Wu CO, Tian X, Bang H: Varying-coefficient mixed-effects approach for the analysis of intervention effects in longitudinal studies. Biometrics, submitted.