Statistica Sinica 3(1993), 329-350

MODELING PULMONARY FUNCTION GROWTH WITH REGRESSION SPLINES

David Wypij, Marian Pugh and James H. Ware

Harvard School of Public Health

Abstract: This paper describes methods for modeling the dependence of the level and rate of growth of pulmonary function during childhood on two physiologic variables, age and height, and for assessing the effects of individual and environmental risk factors on measures of pulmonary function. Descriptive analyses stratified by age suggest that the relation between pulmonary function and height in children is linear but age-dependent. Thus, we consider models in which pulmonary function level (or rate of growth) depends linearly on height (or change in height), but with an age-dependent intercept and slope. Regression splines are used to describe the change in intercept and slope with age. To accommodate repeated measures and heterogeneity of variance, robust variance estimates are derived for the estimated regression coefficients. The methods presented provide a flexible family of growth curves, quantify the effects of covariates on level and rate of growth, and have attractive clinical and epidemiological interpretations.

Key words and phrases: Growth curve models, regression splines, successive differences, robust variance estimation, pulmonary function growth.

1. Introduction

The extensive literature on mathematical models for human growth is concerned primarily with models for the dependence of a single outcome variable (such as height) on age. The earliest work focused on the modeling of height growth as a function of age. Many different models were proposed, including the Jenss curve and more recently, double and triple logistic functions. The statistical literature initially emphasized the theory of polynomial growth curves, but has, in recent years, contained many papers on nonlinear models for growth. These papers have discussed inference on both individual and population growth curves, goodness of fit, assessing the influence of covariates, and prediction (see, for example, Bock and Thissen (1980), Berkey and Laird (1986), Rao (1987), and the references therein). Recently there has been much interest in nonparametric and semiparametric approaches for modeling a response variable as a function of a single time metameter, including book length treatments of smoothing splines and kernel estimators by Eubank (1988) and Müller (1988), respectively. To model growth velocities, the analysis of successive differences has been suggested by several authors, including Hills (1968) and Schwertman and Heilbrun (1986). The methods discussed in these reports are not, however, relevant to the setting in which the physiologic variable of primary interest depends not only on age but also on a second physiologic variable that is also age-dependent.

The measurement and modeling of pulmonary function has gained wide acceptance as a simple method of monitoring for chronic respiratory disease and assessment of risk factors (Bates (1989)). Lung function in a healthy individual increases with age until the early to midtwenties, when a slow, natural decline begins (Ferris et al. (1981), Sherrill et al. (1991)). About 20% of the population (mostly, but not all smokers) reach a level of pulmonary function associated with disability. Beaty et al. (1985) and Speizer et al. (1989) have shown that subjects with lower levels of pulmonary function have an excess risk of mortality.

This paper focuses on models for the development of lung function in children and adolescents, rather than the decline in adults. Because a single simple model for lung function does not give accurate predictions in all situations, researchers have introduced a variety of models for pulmonary function measurements. Since pulmonary function (PF) is strongly affected by age in years (AGE) and height in centimeters (HT), these terms are commonly included in models. Other measurements such as body surface area or body mass index can replace or supplement AGE or HT as growth metameters. We briefly review models for lung function in children and adults before focussing on our methodology.

Since the paper by Kory et al. (1961), when modeling adult pulmonary function it has become common practice to use a gender-specific multiple regression model of the form

$$E(\mathrm{PF}) = \beta_0 + \beta_1\mathrm{AGE} + \beta_2\mathrm{HT},$$

such as used in the prediction models of Knudson et al. (1983). Cole (1975) compares a variety of models on independent data sets and finds support for the model

$$E(\mathrm{PF}) = \mathrm{HT}^2 \times (\beta_0 + \beta_1\mathrm{AGE}),$$

which allows pulmonary function to vary proportionally, rather than linearly, to height. Dockery et al. (1985) and Ware et al. (1990) modeled age-related changes in adult pulmonary function levels using

$$E(\mathrm{PF}/\mathrm{HT}^2) = \beta_0 + \beta_1\mathrm{AGE} + \beta_2\mathrm{AGE}^2.$$

The division by HT^2 adjusts for body size and makes the residuals more homoscedastic, while the quadratic age term reflects more rapid pulmonary function decline in older adults. Nonparametric and semiparametric approaches have also recently been suggested. For example, Sherrill et al. (1991) used polynomial smoothing splines to model lung function as a flexible function of single metameter (AGE), but their approach does not easily accommodate a second growth metameter (HT) and estimation of the effects of pulmonary risk factors.

Models proposed for pulmonary function in young children have been similar. For preadolescent children aged 6 to 11, Dockery et al. (1983) proposed gender-specific models of the form

$$E(\log(\mathrm{PF})) = \beta_0 + \beta_1 \log(\mathrm{HT}).$$

Here, any "age effect" is effectively controlled by the log(HT) metameter. Kauffmann et al. (1989) applied models similar to those of Kory et al. (1961) to children aged 6 to 10, with an additional term for the child's weight. Because simple linear models do not fit well throughout the whole childhood range, a common theme has been to break up the age range into disjoint pieces. For example, Tashkin et al. (1984) used models containing body surface area (BSA) measurements as in

$$E(\mathrm{PF}) = \beta_0 + \beta_1\mathrm{AGE} + \beta_2\mathrm{HT} + \beta_3\mathrm{BSA}$$

fit separately to children aged 7 to 11 and 12 to 17. Burchfiel et al. (1986) applied models similar to those of Kory et al. (1961) to children aged 10 to 15, but found no significant AGE effects in a model for 16 to 19 year olds. Schwartz et al. (1988) extended the model of Dockery et al. (1983) to include body mass index (BMI). The resulting model

$$E(\log(\mathrm{PF})) = \beta_0 + \beta_1 \log(\mathrm{HT}) + \beta_2 \log(\mathrm{AGE}) + \beta_3 \log(\mathrm{BMI})$$

is fit separately to children aged 6 to 11, boys aged 12 to 20, and girls aged 12 to 17.

During adolescence, pulmonary function grows at an age-dependent rate that is closely related to the height changes associated with the adolescent growth spurt (Wang et al. (1993)). We have found the simple models above to be inadequate in describing lung function growth from childhood through age 18. There are several important challenges in obtaining pulmonary function models for children and adolescents. Models must be flexible enough to follow the changing relationship between level and growth of lung function and other growth metameters from childhood through the adolescent growth spurt and into young adulthood. Since the data consist of repeated measurements on individual subjects, within-subject correlations must be considered. The large size of epidemiological data sets and the number of different risk factors, including air pollution, asthma status, and personal and parental smoking, makes computational demands a concern.

The approach we use is to model the relationship between pulmonary function, height, and age using regression splines. Regression splines allow the fitted curves to be smooth and flexible, yet their parametric form permits use of familiar techniques for model-based inference and assessment of covariates. Robust methods of variance estimation (Liang and Zeger (1986)) are used to adjust the estimated standard errors for repeated measures.

In Section 2 we describe our data and motivate our approach. Section 3 presents the specifics of the model and estimation methods and applies the methods to our data. Section 4 discusses the advantages and limitations of our methodology, and makes brief comparisons with alternative approaches.

2. The Motivating Example

The Six Cities Study of Air Pollution and Health is a longitudinal study of the natural history of respiratory health and the health effects of air pollution. As part of this study, cohorts of school children were enrolled in first or second grade and examined annually to determine changes in respiratory symptom status and growth of pulmonary function. During the 15 years of the study, a cohort of 13,737 children, born in 1967 or later, was examined in six communities across the United States (Watertown, MA; Kingston and Harriman, TN; Steubenville, OH; a geographically defined section of St. Louis, MO; Portage, WI; and Topeka, KS). The design of the study and the selection of the communities have been previously described (Ferris et al. (1979)).

Pulmonary function can be measured by spirometric testing, in which the subject takes the largest inspiration possible and exhales as rapidly as possible into the spirometer, a recording device. The spirogram obtained from this maneuver can be interpreted as a volume-time curve (see Figure 1), and several outcome measures are used by epidemiologists. The forced vital capacity (FVC) is the total volume of air exhaled, usually after, at most, six seconds. The forced expiratory volume in t seconds (FEV$_t$) is the volume exhaled during the first t seconds (often $t = 1$ second is used). Maximal effort occurs from the start of expiration and the FEV$_1$ is approximately 95% of the FVC in children. The mean forced expiratory flow (FEF) is measured between two designated percentages of the forced vital capacity. Thus, FEF$_{25\text{-}75}$ is the average flow or slope measured between the times at which 25% and 75% of the vital capacity have been exhaled.

In this paper we analyze 28,473 FEF$_{25\text{-}75}$ observations on 5,030 Caucasian boys between the ages of 10 and 18 for whom information on height, asthma status, and personal smoking was available. In recent years, there has been increased focus on FEF$_{25\text{-}75}$ in respiratory epidemiology studies, and we have found this variable to be a sensitive indicator of the adverse effects of asthma/wheeze status. Analogous methods can be used to model FVC or FEV$_1$. Table 1 presents descriptive statistics in our sample. The median number of observations per subject is six, regardless of asthma/wheeze status at entry, though it appears that asthmatic children drop out of the study slightly earlier than non-asthmatic children. As we might expect, $\mathrm{FEF}_{25\text{-}75}$ and $\log(\mathrm{FEF}_{25\text{-}75})$ values are significantly different at entry for the asthma/wheeze groups, while AGE, HT, and log(HT) values are not.

Age-specific means of the level and annual growth velocities of log(HT) and $\log(\mathrm{FEF}_{25\text{-}75})$ are plotted in Figure 2. The growth velocities are defined to be the successive differences of the $\log(\mathrm{FEF}_{25\text{-}75})$ and log(HT) between annual examinations. The pulmonary outcomes have a complicated dependence on height and age, with the HT growth spurt occurring about six months prior to the $\mathrm{FEF}_{25\text{-}75}$ growth spurt. Preliminary analyses showed that simple modifications of the "adult" or the "pre-adolescent" models described above failed to provide a good fit to the data through the adolescent growth spurt and into early adulthood. However, plots of age-specific means of $\log(\mathrm{FEF}_{25\text{-}75})$ versus log(HT), given in Figure 3, suggest that the relationship between $\log(\mathrm{FEF}_{25\text{-}75})$ and log(HT) looks approximately linear within age groups, but the intercepts and slopes vary with age. This motivates an age-dependent model of the form

$$E(\log(\mathrm{FEF}_{25\text{-}75})) = f_1(\mathrm{AGE}) + f_2(\mathrm{AGE}) \times \log(\mathrm{HT}), \tag{2.1}$$

for particular functions $f_1(\cdot)$ and $f_2(\cdot)$. We have found that logarithmic transformation of both HT and $\mathrm{FEF}_{25\text{-}75}$ gives models that are more linear and homoscedastic. Model (2.1) allows a more complicated dependence on AGE than the simpler models proposed by Kory et al. (1961) and Dockery et al. (1983).

An alternate parameterization, suggested by a reviewer, would center the log(HT) term by an age-dependent smoothing of log(HT) on AGE. This would allow the intercept function $f_1(\cdot)$ to be interpreted as the smoothed $\log(\mathrm{FEF}_{25\text{-}75})$ on AGE, which is more useful than the uncentered intercept. However, this would require an additional smoothing of the log(HT) values, and could also affect standard error estimates.

In other contexts, Hills (1968) and Schwertman and Heilbrun (1986) modeled growth velocities for a single outcome variable using successive differences. Modeling successive differences can simplify the characterization of the mean and covariance structure and may be helpful for assessing the effects of time-varying risk factors on growth. In our setting, preliminary age-specific analyses (see Figure 3) show approximately linear relationships between $\Delta\log(\mathrm{FEF}_{25\text{-}75})$ and $\Delta\log(\mathrm{HT})$, suggesting models of the form

$$E(\Delta\log(\mathrm{FEF}_{25\text{-}75})) = f_3(\mathrm{AGE}) + f_4(\mathrm{AGE}) \times \Delta\log(\mathrm{HT}), \tag{2.2}$$

for particular functions $f_3(\cdot)$ and $f_4(\cdot)$, where Δ denotes annual differences. When modeling velocities, we required successive exams to be nine to fifteen months apart. For convenience, we use the age at the end of the interval as the metameter. Similar results are obtained when using the age from the beginning or middle of the interval or standardizing by ΔAGE.

3. Estimation Methods

3.1. Regression splines

A plethora of parametric and nonparametric smoothing methods have been suggested recently, including smoothing splines (Wegman and Wright (1983), Silverman (1985), Eubank (1988)), kernel estimators (Müller (1988)), and generalized additive models (Hastie and Tibshirani (1990)). We use polynomial regression splines for functions $f_1(\cdot)$ and $f_2(\cdot)$ in model (2.1) or $f_3(\cdot)$ and $f_4(\cdot)$ in model (2.2). Regression splines use piecewise polynomials with continuity conditions imposed at the knot points to smoothly approximate a functional relationship between a single response and a single metameter, such as AGE. In our application, the "response variables" are age-dependent intercepts and slopes of a more complicated functional relationship.

A spline is completely characterized by the order of the spline r, which, by convention, is one more than the order of the polynomial, an ordered sequence of knot points, $k_0 < k_1 < \cdots < k_M < k_{M+1}$, where the interval $[k_0, k_{M+1}]$ encompasses the range of metameter values (i.e., AGE), and a vector $(n_1, \cdots, n_M)$ specifying the number of continuity conditions at each interior knot. We choose to impose no continuity conditions at the end knots. In particular, $n_i = 0$ if the spline is not required to be continuous at $k_i; n_i = 1$ if the spline is required to be continuous at k_i, but no condition is placed on the first derivative; and n_i may range up to $r - 1$, which requires the derivatives up to order $r - 2$ be continuous. For example, a cubic spline of order $r = 4$ may have up to $r - 1 = 3$ continuity conditions imposed at each interior knot, on the function value and the first and second derivatives.

Any spline function $S(\cdot)$ can be written as a unique linear combination of piecewise polynomials of the same order as $S(\cdot)$. Two possible choices of basis functions are the truncated power basis and the B-spline basis. The truncated power basis provides a simple framework for hypothesis testing (Smith (1979)), but the basis functions are highly correlated, leading to ill-conditioned design matrices. In contrast, B-splines require recursive evaluation, making their interpretation more complicated, but are more nearly orthogonal, leading to a well-conditioned design matrix. This is because each B-spline basis function is positive only over part of the metameter range, and has limited nonzero overlap with other basis functions. DeBoor (1978) presents algorithms for constructing the B-spline basis functions, $B_i(\cdot)$, and derives their properties. We can write any spline as a linear combination of B-splines $S(x) = \sum_{i=1}^{d} \beta_i B_i(x)$, where the dimension of the B-spline basis is given by

$$d = (M+1) \times r - \sum_{i=1}^{M} n_i.$$

The B-spline basis allows us to construct very general splines. In practice, splines of order higher than cubic ($r = 4$) are rarely used because matching second derivatives produces a curve which is smooth to the eye. If smoothness is not an issue, a linear spline basis with $r = 2$ or a point spline basis with $r = 1$ (consisting of $M + 1$ indicator functions) may be useful, giving piecewise linear or piecewise constant spline functions. It is possible to construct splines with less than full continuity, but we have not found much use for these.

For regression spline modeling of pulmonary function level we use models of the form

$$E(\log(\mathrm{FEF}_{25\text{-}75})) = \sum_{i=1}^{d} \beta_i B_i(\mathrm{AGE}) + \sum_{i=1}^{d} \beta_{d+i} B_i(\mathrm{AGE}) \times \log(\mathrm{HT}), \tag{3.1}$$

with $2 \times d$ columns in the design matrix. The first d columns contain the B-spline basis vectors $B_1(\cdot), \ldots, B_d(\cdot)$, and the next d columns contain these same vectors multiplied by the log(HT) values. The regression spline framework is convenient since it gives a parametric (in fact, linear) model, allowing standard techniques for model-based inference. Models for successive differences, i.e., $\Delta \log(\mathrm{FEF}_{25\text{-}75})$ values, are defined in an analogous manner using $\Delta \log(\mathrm{HT})$.

Figure 4 plots age-dependent intercepts and slopes, together with a stratified (point), linear, and cubic spline model of the form (3.1) using integer knot points $(10, 11, \ldots, 19)$ for $\log(\mathrm{FEF}_{25\text{-}75})$ and the analogous model for $\Delta \log(\mathrm{FEF}_{25\text{-}75})$. Age-dependent mean fitted values are also plotted. The three orders of the spline lead to similar fits.

3.2. Inclusion of covariate effects in the mean structure

The effects of individual or environmental risk factors can be modeled with additive or multiplicative adjustments to (3.1). As an example, we compare outcomes for five categories of asthma/wheeze status. The "active asthma" group consisted of subjects who reported a history of asthma and current wheeze symptoms. The "inactive asthma" group consisted of subjects who reported a history of asthma but no current wheeze symptoms. The "active wheeze" group consisted of subjects who reported current wheeze symptoms, but no history of asthma. The "inactive wheeze" group consisted of subjects who reported no current wheeze symptoms and no history of asthma, but who had a history of wheeze. The baseline comparison group consisted of subjects who never reported asthma or wheeze. These five categories are mutually exclusive, although an individual could move between categories in different years (e.g., from never reporting asthma or wheeze, to active wheeze, then to inactive wheeze).

Figure 5 consists of two separate plots. First, we plot age-dependent means of the log transformed data separately by asthma status (with five categories). We also plot the mean fitted values versus age for model (3.1) estimated separately for observations from each of the five categories, using a linear spline with integer knot points $(10, 11, \ldots, 19)$. The predicted values of the stratified model follow the observed data quite well, lending support to our regression spline models. The predicted values are also remarkably close to being parallel, giving justification for a simple additive effect for asthma status. In practice, we have found additive effects for asthma status to fit the data well and to be easy to interpret, since additive effects on the logged scale correspond to multiplicative effects on the unlogged scale. More complicated covariate effects, such as asthma by age interactions, could be easily incorporated into the model.

3.3. Adjustments for the covariance structure

Due to the within-subject correlations, assuming independence of repeated measures on a subject could lead to incorrect inferences. When the true correlation structure is not known or when heteroscedasticity is present, the generalized estimating equations (GEE) of White (1980) and Liang and Zeger (1986) can be used to obtain more efficient estimates of the model parameters and robust estimates of their variances. For the linear model, their method assumes an independence or other "working" covariance model and uses generalized least squares to estimate the regression parameters. Let Y_i denote the $n_i \times 1$ vector of responses for the ith subject and $E(Y_i) = X_i\beta$, where X_i is the $n_i \times p$ design matrix for the ith subject (including the B-spline basis vectors), β is a $p \times 1$ parameter vector, and $i = 1, \ldots, N$, where N is the number of subjects. If $\hat{V}_i$ is the current "working" covariance matrix for the ith subject, then

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} X_i \right\}^{-1} \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} Y_i \tag{3.2}$$

is the generalized least squares estimator of β. An iterative procedure is used, alternately estimating β using (3.2) and estimating $\hat{V}_i$ using the method of moments on the residuals. The regression parameter estimates are consistent under mild conditions, although they need not be efficient, with the efficiency rising as the working covariance structure approaches the true covariance structure.

White (1980) and Liang and Zeger (1986) also give a consistent "sandwich" estimator of the variances of the regression parameters, valid even if the working correlation model is misspecified. We estimate the variances of the regression parameter estimates by

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \Big\{ \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} X_i \Big\}^{-1} \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} \hat{e}_i \hat{e}_i' \hat{V}_i^{-1} X_i \Big\{ \sum_{i=1}^{N} X_i' \hat{V}_i^{-1} X_i \Big\}^{-1}, \quad (3.3)$$

where $\hat{e}_i$ denotes the residuals from the ith subject. The estimated GEE standard errors for the parameter vector are given by the square roots of the diagonal elements of (3.3).

This method for calculating variances is not likelihood based, and inferences will be valid only if any missing data are missing completely at random (Little and Rubin (1987)). Missing data are an important concern in any longitudinal study, as non-randomly missing data can affect not only variance estimation, but can also result in biased point estimation. In the pulmonary function study, the population was initially healthy and the reasons for missingness (vacations, moving from the study area, etc.) were not expected to be associated with pulmonary function outcome. Thus we expect that the robust variance methods should be valid in this setting.

As our primary goal is to estimate the effects of covariates on pulmonary function level or growth velocity, we view the within-subject correlations as nuisance parameters. To increase efficiency, an approximate working covariance structure is required. To motivate particular choices, we analyzed the residuals from a linear regression spline model with integer knot points, assuming independence of all the observations (see Table 2). For modeling pulmonary function level, the autoregressive working assumption is approximately valid, in which the correlation between residuals from the same subject t years apart is given by $\rho^t, t = 1, 2, \ldots$ In fact, the correlations drop off more slowly than the autoregressive assumption, and more complicated working structures could be used. For simplicity we used the autoregressive working assumption, and estimated ρ using the method of moments on the "lag one" correlations. For modeling pulmonary function growth velocities, the one-step dependence working assumption is approximately valid, where the correlation of residuals from growth in adjacent years is assumed to be τ, and residuals from rates of growth in intervals more than one year apart are assumed to have correlation zero. Table 2 also suggests some mild age-dependent heteroscedasticity in the $\Delta\log(\mathrm{FEF}_{25\text{-}75})$ residuals.

3.4. Comparison of pulmonary function outcome fits

Table 3 reports regression results for six models corresponding to level of $\log(\mathrm{FEF}_{25\text{-}75})$. Based on the results of Figure 5, indicator variables corresponding to the four asthma/wheeze categories were added to model (3.1), with the baseline comparison group reporting no history of asthma or wheeze. The spline intercept and height slope terms are not shown. Comparisons can be made between different orders of the spline and different working covariance assumptions.

As we would expect if the autoregressive working assumption was closer to the truth than assumed independence, the autoregressive working assumption has higher efficiency and gives slightly smaller standard error estimates than the independence working assumption. There are only minor changes in the asthma estimates as the order of the spline changes. It is remarkable that the regression results are also almost identical when the number or placement of knot points changes (results not shown). For example, a linear spline with only half of the knot points of the linear spline model in Table 3 gives parameter estimates and standard errors equal to those shown (to the accuracy of the displayed values). This is due, in part, to the relatively large size of the data set. The fitted pulmonary function values are also quite consistent across models (results not shown).

The regression results have a simple interpretation. The additive asthma effects on log pulmonary function imply that the lung function deficit associated with asthma or wheeze remained constant in percent throughout childhood. In absolute terms, however, the children with asthma continue to lose ground throughout childhood, since pulmonary function increases with age. For example, the linear and cubic spline models with the autoregressive working covariance structure suggest that an asthmatic child is predicted to have $\exp(-0.101) = 0.90$, i.e., 90%, of the $\mathrm{FEF}_{25\text{-}75}$ of a child reporting no history of asthma or wheeze.

Modeling pulmonary function growth velocities is similar to modeling level, and only two such models are included in Table 3. The models for growth velocity do not show as strong a covariate effect as our models for level of pulmonary function, due to the relatively large variability in growth of PF and HT from year to year. A more sophisticated growth velocity analysis would require refinement in the asthma categories (e.g., perhaps separating currently inactive asthmatics who reported asthma symptoms in the previous year from those who had not reported asthma symptoms for several years).

4. Discussion

The methods presented in this paper are attractive from a statistical, epidemiological, and clinical perspective. The basic structure of the models is motivated by the data, and the regression spline methodology offers a flexible way to model the complex dependence of pulmonary function on two growth metameters, age and height. Staying in the linear model framework allows relatively simple adjustments for heteroscedasticity and repeated measures. We prefer the computational advantages of regression splines over other nonparametric or semiparametric approaches when analyzing large data sets with multiple outcome variables ($\mathrm{FEF}_{25\text{-}75}$, FVC, FEV_1, and the ratio $\mathrm{FEV}_1/\mathrm{FVC}$ are standard respiratory response variables), several stratification variables (gender, race), and many covariates (air pollution, asthma status, personal and passive smoking, etc.). The models for annual change are inherently longitudinal, although "cross-sectional" methods are used in the analysis.

The literature on smoothing splines and kernel estimators emphasizes the selection of smoothing parameters or bandwidths using cross-validation or other methods, while work on regression splines has concentrated on the choice of number and placement of knot points. For regression spline modeling of pulmonary function, the estimation of risk factors was insensitive to the particular order of spline, knot points, or continuity conditions prespecified. Thus, the extensive computational demands of a smoothing spline or other method were not expected to alter the epidemiologic findings.

In a study of pulmonary function decline in adults, Sherrill et al. (1991) give graphical presentations of polynomial smoothing splines for different strata (males/females, smokers/nonsmokers, and asymptomatic/symptomatic subjects). However, they were not able to obtain simple summary measures of covariate effects. Further work is needed with more general smoothing methods to account for complicated covariate patterns (including continuous covariates) and to model complex dependencies of the response variable on more than one metameter.

Regression splines are the natural extension of stratified models, used so often in epidemiology. For example, Wang et al. (1993) present linear models for log(PF) on log(HT) stratified by integer age, which are effectively linear spline formulations of (3.1) without imposing continuity conditions at the (integer) knot points. In this case, the fitted values are discontinuous at each knot point. With the regression spline approach, we can easily enforce continuity in the intercept and height slopes of (3.1) at the knot points, to ensure that the fitted values will be a continuous function of both HT and AGE.

The regression spline approach is flexible, allowing stratification and subject-specific or time-varying categorical or continuous predictors (e.g., pack-years of cigarette smoking or annual levels of air pollution). The spline approach offers a convenient way to model PF growth from childhood into adulthood, and model (3.1) applied to PF level or growth velocity has a longitudinal focus. Our models for level of PF are age-dependent functions of log(HT), and children who mature earlier (i.e., those whose HT values are larger at an earlier age) will have a higher predicted log(PF) value (since the log(HT) slopes are positive). Similarly, models for $\Delta \log(\mathrm{PF})$ depend on the $\Delta \log(\mathrm{HT})$ changes between examinations. Lagged effects can also be implemented. For example, PF growth for a particular interval can be modeled as a function of HT growth for the same length interval six months previous to the PF interval (as possibly suggested by Figure 2), although our data would only allow an interpolation to approximate the HT growth for lagged intervals. Derivatives of the spline also can be calculated if necessary. We did not detect any influence of asthma or wheeze on height or height growth, though extensions of our methodology could be made to accommodate such a situation.

The random effects models of Laird and Ware (1982), Berkey and Laird (1986), and Lindstrom and Bates (1988) offer an alternate approach to the analysis of this type of data. For modeling pulmonary function, two approaches were tried and rejected. First, we fitted various nonlinear random effects models to PF, PF/HT^2 (a transformation suggested by data on adults), or similar outcomes. Two-stage estimation methods had convergence difficulties due to the relatively small number of observations per subject, and no parametric form seemed adequate to model the effects of the adolescent growth spurt on PF. Random effects spline models were also tried, but could only be fitted for the simplest of cases. We were unable to estimate random effects models as complex as (3.1) with integer knot points due to convergence difficulties.

Using successive differences to model pulmonary function growth velocities as a function of height growth is convenient and more readily interpretable than including previous responses as independent variables in the model, as suggested by Rosner et al. (1985) (see also Stanek et al. (1989)). As successive differences from an individual are almost independent, the independence working assumption gives high efficiency compared to more complicated models. Our enthusiasm for this approach is tempered, however, by the fact that height is measured with small, though definite, errors, which impact the $\Delta \log(\mathrm{HT})$ values more significantly. An approach which accounts for this measurement error may be appropriate.

The regression spline approach offers a convenient way to accommodate the dependence of pulmonary function on two metameters. Similar methodology may be beneficial in other growth curve and longitudinal data settings, such as in modeling weight as a function of height and age, or blood pressure as a function of age and weight. We hope our techniques will stimulate the use of spline methodology and provide further insights into the development of lung function and in the study of individual and environmental risk factors.

Acknowledgements

This research was supported by Grant GM29745 from the NIGMS. Dr. Wypij was the recipient of a Mellon Foundation Faculty Development Grant. The Six Cities Study was supported, in part, by Grants ES01108 and ES00002 from the NIEHS, Cooperative Agreement CR811650 from the EPA, and Contract RP1001 from the Electric Power Research Institute.

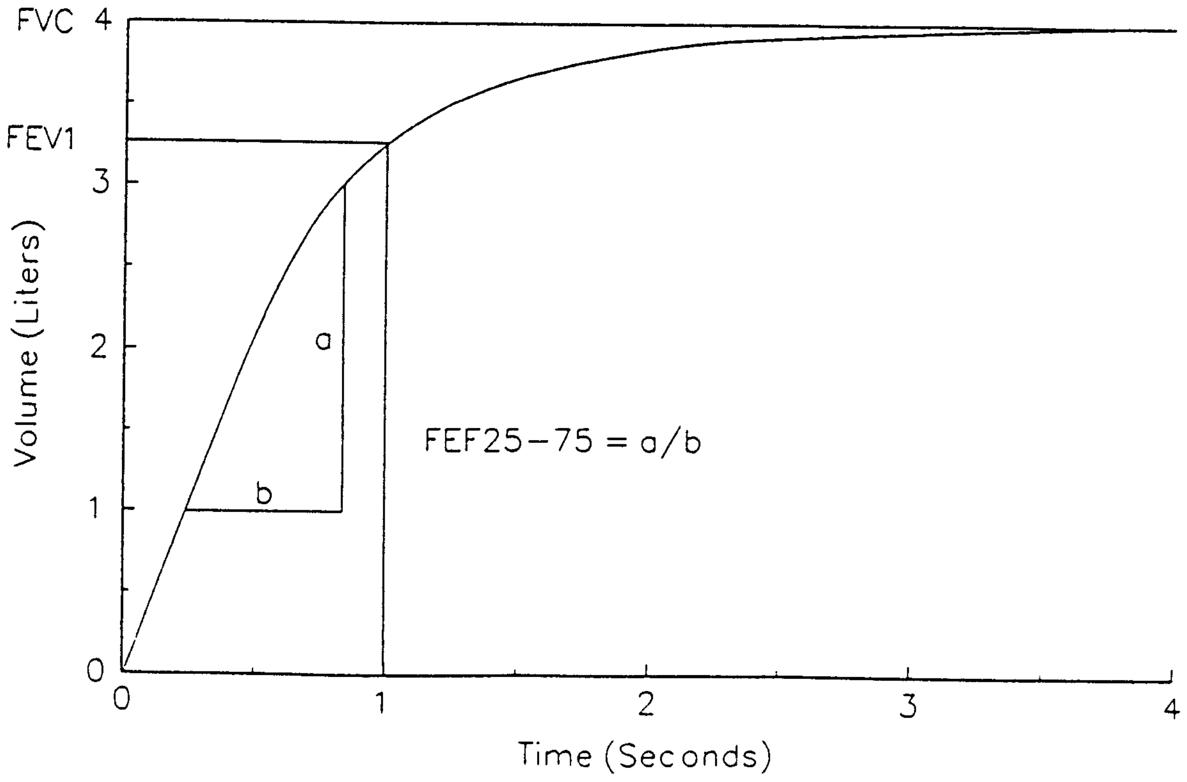

Figure 1. Derived volume-time trace from a forced vital capacity (FVC) maneuver from a normal subject. FEV_1, Forced expiratory volume in 1 second; $FEF_{25\text{-}75}$, mean forced expiratory flow between the times at which 25% and 75% of the vital capacity have been exhaled.

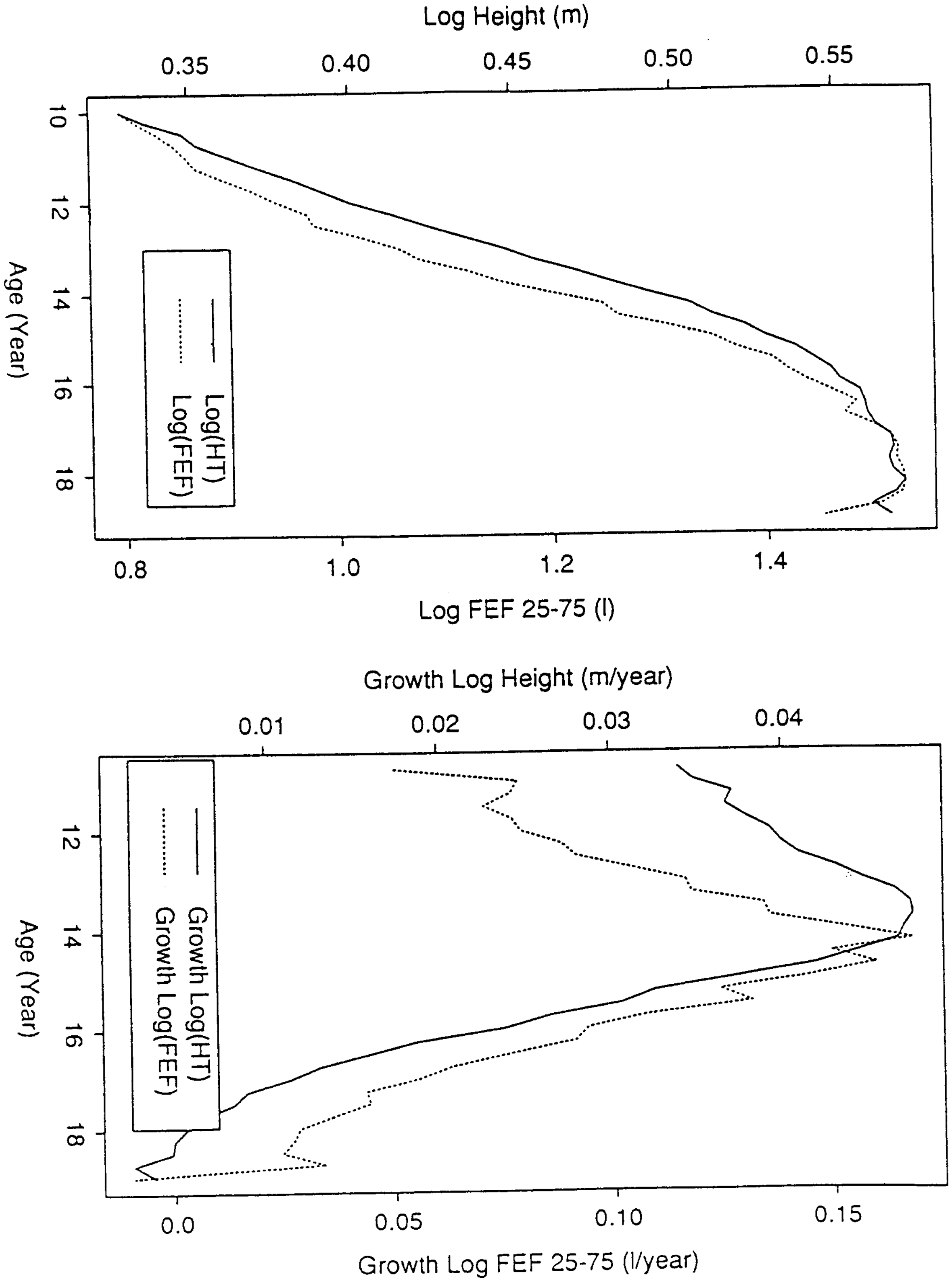

Figure 2. Age-specific means of level and annual growth velocities of log(HT) and log($FEF_{25\text{-}75}$) grouped by quarter year of age. Left panel plots means of log(HT) and log($FEF_{25\text{-}75}$) as a function of age. Right panel plots means of Δ log(HT) and Δ log($FEF_{25\text{-}75}$) as a function of age at the end of the interval.

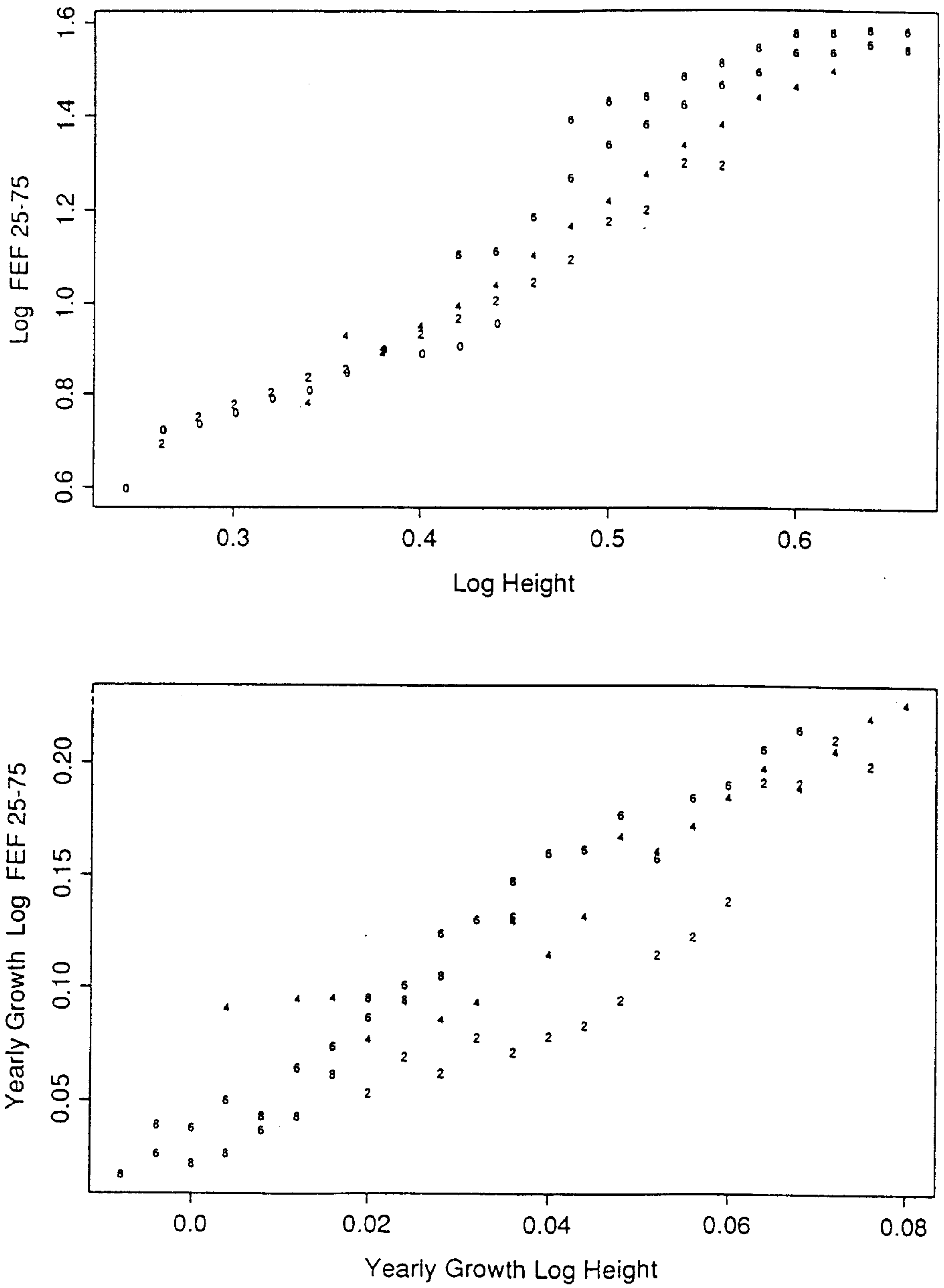

Figure 3. Age-specific means of level and annual growth velocities of $\log(\mathrm{FEF}_{25\text{-}75})$ as a function of height. Top panel plots means of $\log(\mathrm{FEF}_{25\text{-}75})$ as a function of $\log(\mathrm{HT})$. Bottom panel plots means of $\Delta \log(\mathrm{FEF}_{25\text{-}75})$ as a function of $\Delta \log(\mathrm{HT})$. In the panels, 0 denotes subjects aged 10, 2 denotes subjects aged 11 to 12, 4 denotes subjects aged 13 to 14, 6 denotes subjects aged 15 to 16, and 8 denotes subjects aged 17 to 18.

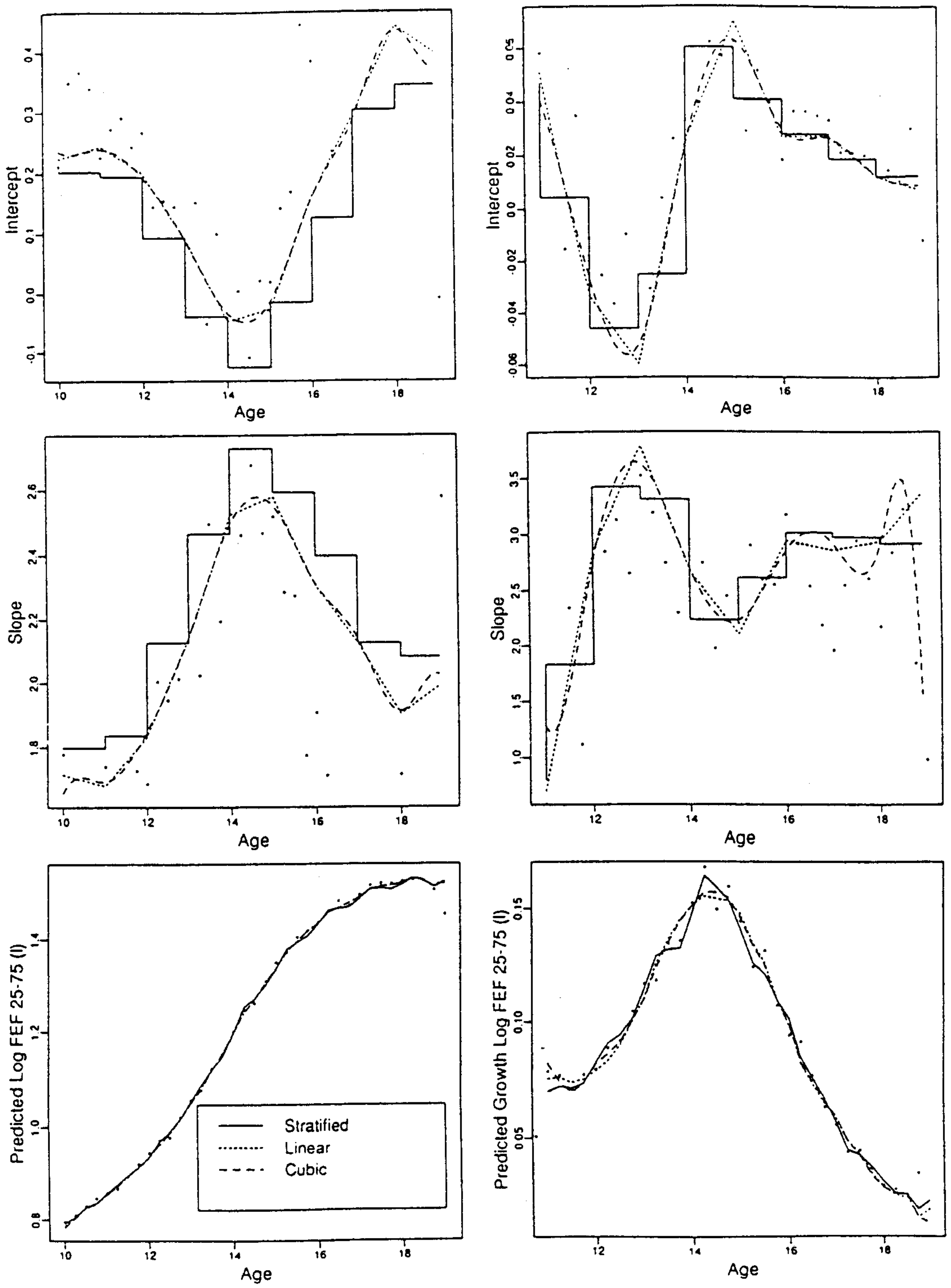

Figure 4. Age-specific intercepts, height slopes, and mean predicted values of level and annual growth velocities of $\log(\text{FEF}_{25\text{-}75})$. Left panel is for level of $\log(\text{FEF}_{25\text{-}75})$ as a function of age. Right panel is for $\Delta \log(\text{FEF}_{25\text{-}75})$ as a function of age. The three curves correspond to stratified (point), linear, and cubic spline formulations using knot points (10,11,. . . ,19). Asterisks correspond to age-specific intercepts and height slopes for observations grouped together by quarter year of age.

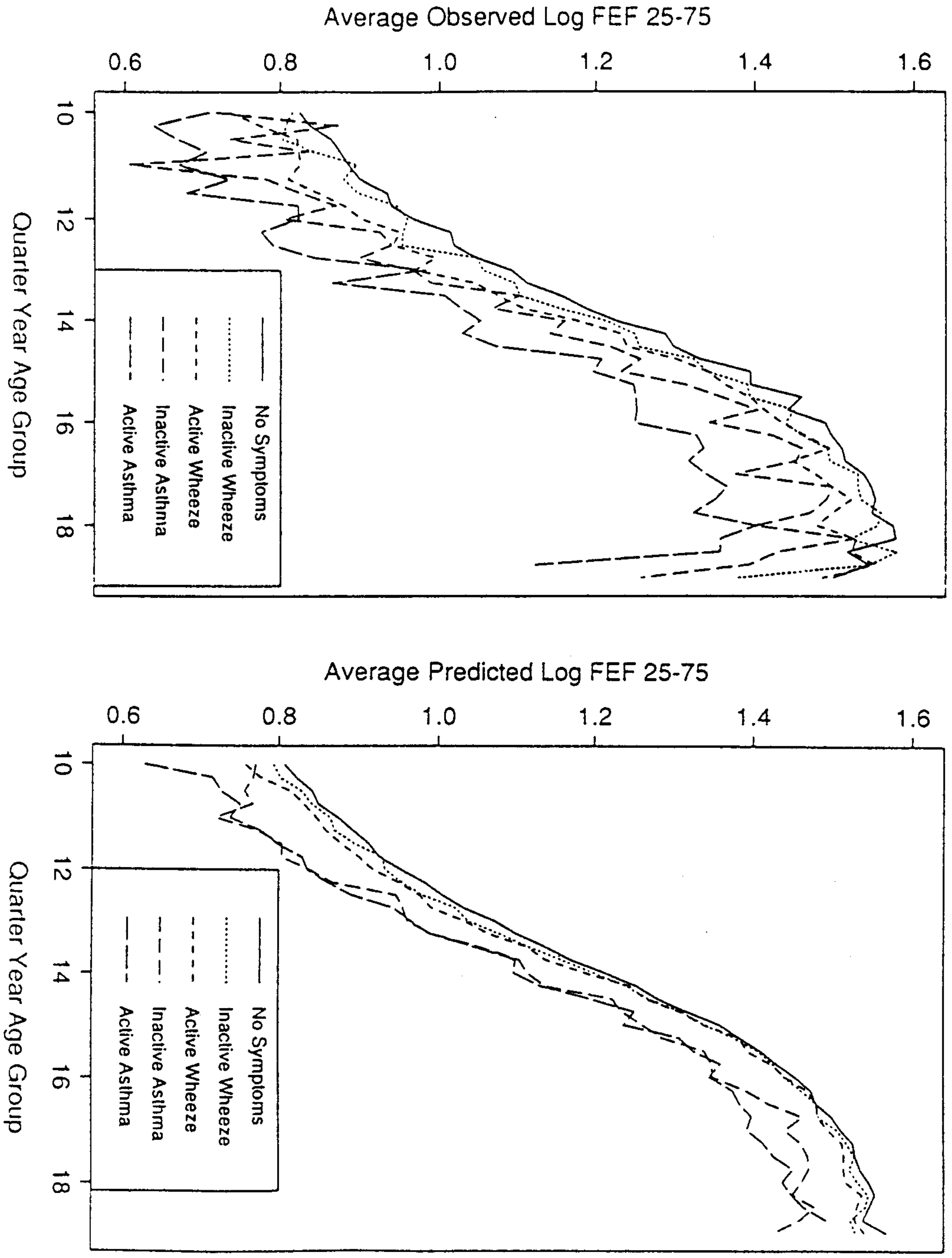

Figure 5. Age-specific means and predicted values of log($FEF_{25\text{-}75}$) grouped by quarter year of age and asthma status. Left panel plots means of log($FEF_{25\text{-}75}$) as a function of age for five groups of children: Children never having reported asthma or wheeze symptoms, children with no current wheeze symptoms but with a history of wheeze but not asthma, currently wheezing children with no history of asthma, children with a history of asthma but who were not currently reporting wheeze, and children with a history of asthma who were currently reporting wheeze symptoms. Right panel plots means of fitted values for separate linear spline formulations for each asthma status group, each using knot points (10,11,...,19).

Table 1. Characteristics of sample of 5,030 Caucasian boys

	Asthma/wheeze status at entry[a]		
	Never reporting asthma or wheeze symptoms ($n = 2920$)	Having active or inactive wheeze, but no asthma history ($n = 1685$)	Having active or inactive asthma ($n = 425$)
No. of observations			
1–3	646 (22.1%)	378 (22.4%)	108 (25.4%)
4–6	872 (29.9%)	507 (30.1%)	156 (36.7%)
7–9	1402 (48.0%)	800 (47.5%)	161 (37.9%)
Age at entry			
	11.46 ± 1.56[b]	11.54 ± 1.79	11.62 ± 1.67
HT at entry (in meters)			
	1.47 ± 0.11	1.47 ± 0.12	1.47 ± 0.12
log(HT) at entry			
	0.38 ± 0.07	0.38 ± 0.08	0.38 ± 0.08
$FEF_{25\text{-}75}$ at entry (in liters/second)[c]			
	2.64 ± 0.77	2.59 ± 0.84	2.34 ± 0.88
log($FEF_{25\text{-}75}$) at entry[c]			
	0.93 ± 0.28	0.90 ± 0.31	0.79 ± 0.36

[a] At entry refers to the first available observation between ages 10 and 18.

[b] Mean ± Standard deviation.

[c] Using ANOVA test, $p < 0.001$.

Table 2. Standard deviations (in parentheses) and within-subject correlations of residuals grouped by integer age, based on linear spline models of the form (3.1) with knot points $(10, 11, \ldots, 19)$, and assuming independence of all observations.

For level of log(FEF$_{25\text{-}75}$)

						Age				
		10	11	12	13	14	15	16	17	18
	10	(.23)	.75	.74	.72	.70	.66	.68	.67	.60
	11		(.23)	.77	.77	.74	.71	.72	.70	.68
	12			(.23)	.79	.76	.73	.72	.72	.65
Age	13				(.23)	.82	.78	.77	.74	.69
	14					(.23)	.83	.80	.76	.73
	15						(.22)	.83	.79	.75
	16							(.23)	.84	.77
	17								(.23)	.82
	18									(.23)

For growth velocity or Δ log(FEF$_{25\text{-}75}$)

					Age				
		11	12	13	14	15	16	17	18
	11	(.16)	−.42	.00	−.01	−.03	.00	−.05	−.01
	12		(.15)	−.48	.01	.01	−.05	.04	.01
	13			(.15)	−.39	−.03	.05	−.09	−.03
Age	14				(.13)	−.39	−.04	−.01	.03
	15					(.13)	−.43	−.05	−.08
	16						(.13)	−.38	−.03
	17							(.13)	−.37
	18								(.13)

Table 3. Estimated asthma coefficients and GEE standard errors (in parentheses) from regression results for various spline orders and working covariance assumptions, with knot points chosen to be $(10, 11, \ldots, 19)$. The asthma/wheeze categories are mutually exclusive. The baseline comparison group consists of subjects never having reported asthma or wheeze.

	For level of log($\text{FEF}_{25\text{-}75}$) with independence working assumption			For growth velocity or Δ log($\text{FEF}_{25\text{-}75}$) with independence working assumption
	Point spline	Linear spline	Cubic spline	Linear spline
Active asthma	−.190 (.014)	−.190 (.014)	−.190 (.014)	.0019 (.0034)
Inactive asthma	−.098 (.014)	−.099 (.014)	−.099 (.014)	.0051 (.0043)
Active wheeze	−.056 (.006)	−.057 (.006)	−.057 (.006)	−.0031 (.0021)
Inactive wheeze	−.028 (.007)	−.029 (.007)	−.029 (.007)	.0009 (.0015)

	For level of log($\text{FEF}_{25\text{-}75}$) with autoregressive working assumption			For growth velocity or Δ log($\text{FEF}_{25\text{-}75}$) with one-step dependence working assumption
	Point spline	Linear spline	Cubic spline	Linear spline
Active asthma	−.099 (.009)	−.101 (.009)	−.101 (.009)	.0010 (.0025)
Inactive asthma	−.087 (.010)	−.089 (.010)	−.089 (.010)	.0042 (.0032)
Active wheeze	−.019 (.004)	−.021 (.004)	−.021 (.004)	−.0010 (.0017)
Inactive wheeze	−.014 (.004)	−.017 (.004)	−.017 (.004)	.0011 (.0012)
	$\hat{\rho} = .804$	$\hat{\rho} = .803$	$\hat{\rho} = .803$	$\hat{\tau} = -.399$

References

Bates, D. V. (1989). *Respiratory Function in Disease*, 3rd edition. W.B. Saunders, Philadelphia.

Beaty, T. H., Newill, C. A., Cohen, B. H., Tockman, M. S., Bryant, S. H. and Spurgeon, H. A. (1985). Effects of pulmonary function on mortality. *J. Chron. Dis.* **38**, 703–710.

Berkey, C. S. and Laird, N. M. (1986). Nonlinear growth curve analysis: Estimating the population parameters. *Ann. Human Biology* **13**, 111–128.

Bock, R. D. and Thissen, D. (1980). Statistical problems of fitting individual growth curves. In *Human Physical Growth and Maturation* (Edited by F. E. Johnston, A. F. Roche and C. Susanne), 265-290. Plenum Press, New York.

Burchfiel, C. M., Higgins, M. W., Keller, J. B., Howatt, W. F., Butler, W. J. and Higgins, I. T. T. (1986). Passive smoking in childhood. Respiratory conditions and pulmonary function in Tecumseh, Michigan. *Amer. Rev. Resp. Dis.* **133**, 966–973.

Cole, T. J. (1975). Linear and proportional regression models in the prediction of ventilatory function (with discussion). *J. Roy. Statist. Soc. Ser.A* **138**, 297–337.

DeBoor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

Dockery, D. W., Berkey, C. S., Ware, J. H., Speizer, F. E. and Ferris, B. G., Jr. (1983). Distribution of forced vital capacity and forced expiratory volume in one second in children 6 to 11 years of age. *Amer. Rev. Respir. Dis.* **128**, 405–412.

Dockery, D. W., Ware, J. H., Ferris, B. G., Jr., Glicksberg, D. V., Fay, M. E., Spiro III, A. and Speizer, F. E. (1985). Distribution of forced expiratory volume in one second and forced vital capacity in healthy, white, adult never-smokers in six U.S. cities. *Amer. Rev. Respir. Dis.* **131**, 511-520.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel-Dekker, New York.

Ferris, B. G., Jr., Speizer, F. E., Spengler, J. D., Dockery, D. W., Bishop, Y. M. M., Wolfson, M. and Humble, C. (1979). Effects of sulfur oxides and respirable particles on human health: Methodology and demography of populations in study. *Amer. Rev. Respir. Dis.* **120**, 767–779.

Ferris, B. G., Jr., Speizer, F. E. and Ware, J. H. (1981). Use of tests of pulmonary function to measure effects of air pollutants. In *Measurement of Risks* (Edited by G. G. Berg and H. D. Maillie), 211–230. Plenum Publishing Corporation, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hills, M. (1968). A note on the analysis of growth curves. *Biometrics* **24**, 192–196.

Kauffmann, F., Tager, I. B., Muñoz, A. and Speizer, F. E. (1989). Familial factors related to lung function in children aged 6–10 years. Results from the PAARC epidemiologic study. *Amer. J. Epidemiol.* **129**, 1289–1299.

Knudson, R. J., Lebowitz, M. D., Holberg, C. J. and Burrows, B. (1983). Changes in the normal maximal expiratory flow-volume curve with growth and aging. *Amer. Rev. Respir. Dis.* **127**, 725–734.

Kory, R. C., Callahan, R., Boren, H. G. and Syner, J. C. (1961). The Veterans Administration-Army cooperative study of pulmonary function. I. Clinical spirometry in normal men. *Amer. J. Med.* **30**, 243–258.

Laird, N. M and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.* **83**, 1014–1022.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley, New York.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Springer-Verlag, Berlin.

Rao, C. R. (1987). Prediction of future observations in growth curve models (with discussion). *Statist. Sci.* **2**, 434–471.

Rosner, B., Muñoz, A., Tager, I., Speizer, F. and Weiss, S. (1985). The use of an autoregressive model for the analysis of longitudinal data in epidemiologic studies. *Statist. in Med.* **4**, 457–467.

Schwartz, J. D., Katz, S. A., Fegley, R. W. and Tockman, M. S. (1988). Analysis of spirometric data from a national sample of healthy 6- to 24-year-olds (NHANES II). *Amer. Rev. Resp. Dis.* **138**, 1405–1414.

Schwertman, N. C. and Heilbrun, L. K. (1986). A successive differences method for growth curves with missing data and random observation times. *J. Amer. Statist. Assoc.* **81**, 912–916.

Sherrill, D. L., Lebowitz, M. D., Knudson, R. J. and Burrows, B. (1991). Smoking and symptom effects on the curves of lung function growth and decline. *Amer. Rev. Resp. Dis.* **144**, 17–22.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser.B* **47**, 1–52.

Smith, P. L. (1979). Splines as a useful and convenient statistical tool. *Amer. Statist.* **33**, 57–62.

Speizer, F. E., Fay, M. E., Dockery, D. W. and Ferris, B. G., Jr. (1989). Chronic obstructive pulmonary disease mortality in six U.S. cities. *Amer. Rev. Resp. Dis.* **140**, S49–S55.

Stanek, E. J. III, Shetterley, S. S., Allen, L. H., Pelto, G. H. and Chavez, A. (1989). A cautionary note on the use of autoregressive models in analysis of longitudinal data. *Statist. in Med.* **8**, 1523–1528.

Tashkin, D. P., Clark, W. A., Simmons, M., Reems, C., Coulson, A. H., Bourque, L. B., Sayre, J. W., Detels, R. and Rokaw, S. (1984). The UCLA population studies of chronic obstructive respiratory disease. VII. Relationship between parental smoking and children's lung function. *Amer. Rev. Resp. Dis.* **129**, 891–897.

Wang, X., Dockery, D. W., Wypij, D., Fay, M. and Ferris, B. G., Jr. (1993). Pulmonary function between 6 and 18 years of age. *Pediatric Pulmonology* **15**, 75–88.

Ware, J. H., Dockery, D. W., Louis, T. A., Xu, X., Ferris, B. G., Jr. and Speizer, F. E. (1990). Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *Amer. J. Epidemiol.* **132**, 685–700.

Wegman, E. J. and Wright, I. W. (1983). Splines in statistics. *J. Amer. Statist. Assoc.* **78**, 351–365.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

(Received March 1992; accepted February 1993)