

Confidence Intervals for Semiparametric Regression Curves Based on the Truncated Spline Estimator for Longitudinal Data

Nur Kholifah ^{1*}, Afiqah Saffa Suriaslan ²

¹ Muğla Sıtkı Koçman Üniversitesi, Turkey, email: nurkholifah11@gmail.com,

² Institut Teknologi Sepuluh Nopember, Indonesia, email: afiqahfq75@gmail.com

Abstract

Confidence intervals are required to quantify uncertainty in an estimated regression curve, but their construction becomes more complex when the model combines parametric and nonparametric components for repeated measurements. This study develops a pointwise confidence interval for a semiparametric regression curve based on a truncated spline estimator for continuous longitudinal data. The parametric component was specified as linear, whereas the unknown age effect was represented by a truncated-power spline. Model parameters were estimated using weighted least squares, spline order and knots were selected through generalized cross-validation, and a Student's *t* pivotal quantity was used when the common variance scale was unknown. The procedure was applied to 252 height measurements from 16 infants aged 0-24 months in Surabaya, comprising eight male and eight female infants. The selected model was a second-order truncated spline with four knots at 4, 8, 12, and 16 months. It produced a GCV value of 8.8505, an MSE of 8.2975, and a coefficient of determination of 91.0469%. The fitted curve indicated the largest average monthly increase during 0-4 months and the smallest during 16-24 months. The sex coefficient indicated an estimated constant mean-height difference of 0.4926 cm, rather than a difference in growth velocity. The proposed interval is conditional on the selected spline specification, normal-error assumption, and diagonal working covariance. Consequently, the empirical results are illustrative and should not be interpreted as simultaneous-band coverage or as a population growth standard.

Keywords: *Semiparametric Regression; Longitudinal Data; Truncated Spline; Pointwise Confidence Interval; Weighted Least Squares; Infant Growth*

Introduction

Regression analysis is one of the statistical methods used to determine the functional relationship between a response variable and one or more predictor variables. Depending on the information available about the functional form, regression curves can be estimated through parametric, nonparametric, or semiparametric approaches. Parametric regression specifies the shape of the relationship in advance and is efficient when that specification is correct, but an incorrectly specified functional form can produce biased estimates and misleading conclusions. Nonparametric regression imposes fewer restrictions on the curve and allows the data to reveal nonlinear patterns, although its flexibility must be controlled to avoid excessive variability. Semiparametric regression combines these two perspectives by representing predictors with a known or scientifically interpretable functional form through a parametric component while accommodating an unknown relationship through a nonparametric component (Green & Silverman, 1994; Ruppert et al., 2003; Wood, 2017).

The estimation problem in semiparametric regression arises primarily from the presence of the unknown function in the nonparametric component. Rather than forcing that relationship into a single global polynomial, the function is estimated from the observed data using an appropriate smoothing technique. This structure is useful when a study includes both predictors whose effects can reasonably be summarized by a small number of parameters and predictors

whose effects may change across the observation domain. The parametric component retains interpretability, whereas the nonparametric component protects the analysis against an overly restrictive functional assumption. Consequently, a semiparametric model can offer a useful balance between interpretability and flexibility, provided that the basis representation, smoothing mechanism, covariance structure, and inferential procedure are clearly specified.

Several smoothing techniques are available in nonparametric regression, including kernel estimators, local polynomials, smoothing splines, regression splines, and penalized splines. Spline-based methods are particularly attractive because they approximate a complex curve by joining low-degree polynomial pieces at locations called knots. The resulting curve can adapt to changes in slope or curvature without requiring a high-degree global polynomial. Classical developments in spline smoothing emphasize the relationship among the choice of basis, the number and location of knots, and the control of curve roughness (Wahba, 1990; Green & Silverman, 1994). Later developments have demonstrated that spline representations can also be incorporated into mixed-effects and semiparametric models, thereby extending their use to clustered and repeated-measurement settings (Brumback & Rice, 1998; Durban et al., 2005; Gu & Ma, 2005; Verbyla et al., 1999).

The truncated spline is a regression-spline representation constructed from ordinary polynomial terms and truncated power functions. The word truncated refers to basis functions that contribute only after a specified knot, so the fitted curve is allowed to change its local behavior across subintervals (Fadhilah et al., 2016). This representation is mathematically transparent because the contribution of each knot can be written explicitly, and the resulting model can be estimated using linear-model calculations once the order and knot locations have been fixed. A major advantage of splines is their ability to accommodate smooth functions and data patterns that change across particular subintervals (Budiantara, 2014). At the same time, flexibility is not automatic evidence of model adequacy: an insufficient number of knots can mask important features, whereas too many knots can increase variability and reduce interpretability. Knot selection must therefore be linked to an objective model-selection criterion and to the scientific meaning of the observation scale.

The need for flexible curve estimation becomes more pronounced when the response is recorded repeatedly over time. Longitudinal data consist of repeated observations on the same experimental units, in contrast to cross-sectional data, in which each unit is generally observed at a single time point. Repeated measurements make it possible to study within-subject change, differences between subjects, and the timing of changes in a response trajectory. They can also provide greater efficiency than a cross-sectional design for investigating change because each subject contributes information at several time points (Diggle et al., 2002; Fitzmaurice et al., 2011). This advantage, however, is accompanied by a central statistical complication: measurements collected from the same subject are usually correlated and cannot be treated as independent observations.

Correlation within subjects affects both estimation and statistical inference. If the mean structure is estimated without acknowledging the covariance among repeated measurements, coefficient estimates may remain interpretable under restrictive conditions, but their standard errors and confidence intervals can be incorrect. Established approaches to longitudinal analysis therefore model or accommodate within-subject covariance through marginal models, generalized estimating equations, random-effects models, or other covariance-weighted procedures (Liang & Zeger, 1986; Diggle et al., 2002). Weighted least squares provides one such framework for continuous responses because it incorporates a weight matrix related to the inverse covariance of the observations. Its inferential validity depends on how that covariance matrix is defined or estimated. Thus, in a semiparametric longitudinal model, the construction of the truncated spline curve and the representation of repeated-measurement dependence must be considered jointly rather than as separate technical details.

The literature has developed several flexible approaches for longitudinal trajectories. Zeger and Diggle (1994) proposed semiparametric models for longitudinal responses and demonstrated their application to CD4-cell trajectories. Hoover et al. (1998) developed nonparametric estimators for time-varying coefficient models with longitudinal data. Zhang et al. (1998) formulated semiparametric stochastic mixed models, whereas Lin and Zhang (1999) established inference for generalized additive mixed models using smoothing splines. Guo (2002) formulated functional mixed-effects models, Rice and Wu (2001) introduced nonparametric mixed-effects models for unequally sampled noisy curves, and Huang et al. (2002) studied basis-function approximations for varying-coefficient models with repeated measurements. These contributions show that flexible modeling of mean trajectories must be accompanied by an explicit treatment of subject-specific variation or within-subject correlation.

Research has also examined the comparative and computational properties of nonparametric estimators in longitudinal settings. Welsh et al. (2002) investigated the locality and efficiency of spline and kernel methods in marginal longitudinal nonparametric regression. Wu and Zhang (2002) developed local polynomial mixed-effects models, while Wang (2003) examined marginal kernel regression that accounts for within-subject correlation. Morris (2015) subsequently reviewed functional regression as a broader framework for curve-valued and longitudinal data. These studies provide a broad foundation for flexible longitudinal modeling, but they employ different estimands, covariance assumptions, smoothing devices, and inferential targets. Their results cannot be transferred automatically to a truncated-power spline estimated by weighted least squares. In particular, the sampling distribution of the fitted curve and the variance used in its confidence interval must be derived under the assumptions of the specific estimator used in the study.

An important inferential issue in spline regression is the confidence interval. A point estimate alone presents the fitted trajectory but does not show the uncertainty associated with estimating the curve from a finite sample. A confidence interval adds lower and upper limits for the unknown mean response at a specified predictor value and confidence level. This interpretation differs from a prediction interval, which concerns a future individual observation and therefore includes both uncertainty in the estimated mean and individual-level variability. It also differs from a simultaneous confidence band, which is designed to cover an entire function over a domain with a stated joint probability. Because pointwise intervals are generally narrower than simultaneous bands, an article must identify precisely which inferential object has been constructed and avoid interpreting the proportion of raw observations inside a mean-curve interval as its nominal coverage probability.

Constructing confidence intervals for spline curves is more complicated than reporting intervals for ordinary regression coefficients. The uncertainty depends on the design matrix, the covariance structure, the estimated error variance, and the chosen spline basis. When the order and knots are selected from the same data, the reported interval is commonly conditional on the selected model unless the selection uncertainty is incorporated explicitly (Claeskens et al., 2009; Crainiceanu et al., 2005). In addition, repeated measurements reduce the amount of independent information relative to a sample of unrelated observations of the same total size. For these reasons, a confidence interval for a semiparametric longitudinal curve should be derived from the distribution of the estimator under clearly stated assumptions and should use degrees of freedom consistent with the number of estimated parameters and independent sampling units.

The selection of spline complexity is also central to the quality of the resulting interval. Generalized cross-validation (GCV) is widely used to balance goodness of fit and model complexity in spline estimation. In penalized B-splines, for example, the number of basis functions and the penalty jointly control smoothness (Eilers & Marx, 1996; Wood, 2004, 2011). In a truncated regression spline without an explicit roughness penalty, the order, number of knots, and knot positions directly determine the flexibility of the fitted curve. GCV can provide a

reproducible criterion for selecting among candidate specifications, but the candidate set and the way repeated observations enter the criterion must be reported. A small GCV value supports a model relative to the alternatives considered; it does not by itself establish that the covariance structure, residual assumptions, or confidence-interval procedure is correct.

Longitudinal growth data provide a natural application for this framework because physical growth is continuous but rarely follows a single linear trend throughout infancy. Height increases rapidly during early infancy and then generally continues at a slower rate, so the relationship between age and height may exhibit changes in slope or curvature over clinically meaningful periods. International child-growth standards were developed from longitudinal and cross-sectional data to describe expected growth patterns by age and sex (WHO Multicentre Growth Reference Study Group, 2006). Nevertheless, a model fitted to a local sample has a different purpose from a population growth standard: it summarizes the trajectory observed in that sample and quantifies its uncertainty. Comparisons with external standards can strengthen interpretation, but conclusions must remain proportional to the sample size, sampling design, and geographical scope of the data.

In the present application, infant height is treated as the response, age is modeled through the nonparametric component, and sex is included in the parametric component. This specification allows the age-height relationship to change across intervals while representing the adjusted difference between male and female infants through a regression coefficient. The coefficient for sex describes a difference in expected height under the assumed model; it should not automatically be interpreted as a difference in growth velocity unless an interaction between sex and age is included. The truncated spline component is appropriate for identifying intervals in which the estimated age trajectory changes, while the longitudinal structure requires the uncertainty calculation to reflect repeated observations from the same infants.

Previous Indonesian studies have applied truncated spline regression to longitudinal data and have discussed semiparametric spline modeling in applied contexts (Budiantara, 2014; Fadhilah et al., 2016). However, the cited literature provides limited exposition of a closed-form confidence interval for a fitted semiparametric truncated-spline curve when the variance is unknown and the estimator is obtained through a covariance-weighted criterion. The methodological gap addressed in this article is therefore not the absence of all inferential work on longitudinal splines, but the need to state the estimator, its distribution, the pivotal quantity, and the resulting interval coherently for this particular model representation. Clarifying this scope places the proposed derivation within the broader literature and avoids overstating its novelty.

Accordingly, this study aims to construct pointwise confidence intervals for semiparametric regression curves based on a truncated spline estimator for continuous longitudinal data. The parametric component is approximated by a linear function, whereas the unknown age effect is represented using a truncated spline whose order and knots are selected using the minimum GCV criterion. The model parameters are estimated through weighted least squares, and a pivotal quantity is used to derive the interval when the population variance is unknown. The proposed procedure is then applied to longitudinal height measurements of infants aged 0-24 months in Surabaya. The study contributes an explicit matrix-based derivation, an implementable computational algorithm, and an empirical illustration, while recognizing that the resulting inference is conditional on the selected spline specification and the assumed covariance structure.

Method

This study employed a methodological and applied quantitative design. The methodological component derived a point estimator and a pointwise confidence interval for a semiparametric truncated-spline regression curve under a longitudinal-data structure. The applied component used the resulting estimator to model infant height as a function of age and sex. The inferential

target was the conditional mean curve at a specified combination of predictor values, rather than an interval for a future individual measurement or a simultaneous confidence band over the entire age domain. The derivation was conducted conditionally on the selected spline order and knot configuration. Consequently, uncertainty arising from the data-driven selection of the spline specification was not included in the reported interval and is treated as a limitation of the procedure.

The empirical data consisted of repeated height measurements for 16 infants aged 0-24 months in Surabaya in 2015. Eight infants were male and eight were female. The available records comprised 119 observations for male infants and 133 observations for female infants, giving a total of $M = 252$ measurements. Height in centimeters was the response variable, age in months was assigned to the nonparametric component, and sex was included as a binary parametric predictor. For interpretation of the fitted equations, sex was coded as $x = 1$ for male infants and $x = 0$ for female infants. Because the numbers of observations differed between the two groups and potentially among subjects, the notation allowed subject-specific numbers of repeated measurements.

The analytical procedure consisted of six connected stages: specification of the longitudinal semiparametric model; representation of the unknown age effect by a truncated-power spline; construction of the matrix form and working covariance matrix; estimation by weighted least squares; selection of spline order and knots using generalized cross-validation; and derivation of a pointwise confidence interval through a pivotal quantity. The same fitted specification was then used to calculate sex-specific curves and confidence limits. This sequence ensures that the estimator, its covariance, and the interval refer to the same design matrix and covariance assumption.

- a. The confidence intervals for the semiparametric regression curves based on the truncated spline estimator for longitudinal data were constructed through the following steps:

- 1) Let the paired observations be denoted by (x_{ij}, t_{ij}, y_{ij}) , where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m_i$. These observations are assumed to follow the semiparametric regression model for longitudinal data given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + g(t_{ij}) + \varepsilon_{ij}; \varepsilon_{ij} \sim N(0, \sigma_i^2) \quad (1)$$

For subject i , let m_i denote the number of available measurements, y_{ij} the height measured at occasion j , x_{ij} the parametric predictor, and t_{ij} the age associated with that measurement. The total number of observations is $M = \sum_{i=1}^n m_i$. Measurements from different infants were assumed to be independent. Repeated measurements within the same infant were organized in a subject-level response vector so that their variance-covariance structure could be represented in block form. The mean structure contains a parametric component for sex and a smooth component for age, while the random error captures departures of the observed height from the conditional mean.

The parametric component is approximated by a linear function, whereas the nonparametric component $g(t_{ij})$ is approximated using a truncated spline of order p with k knots. Thus, the model can be written as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_0 + \sum_{s=1}^p \alpha_s t_{ij}^s + \sum_{r=1}^k \alpha_{(p+r)} (t_{ij} - \tau_r)_+^p + \varepsilon_{ij}$$

The age function was represented by a truncated-power basis of degree p with k knots $\tau_1, \dots, \tau_k$. The polynomial terms describe the global age trend, whereas each truncated term $(t - \tau_r)_+^p$ contributes only when t exceeds the corresponding knot. To ensure model identifiability, only one intercept is included in the combined design. In the notation used here, the intercept is retained in the parametric component and the spline intercept is constrained to

zero; equivalently, the nonparametric polynomial sum begins at degree one. This restriction removes the duplicate columns of ones that would otherwise make $C'W^{-1}C$ singular. The number of estimable coefficients is therefore $q = 2 + p + k$: two coefficients for the parametric component, p polynomial-age coefficients, and k truncated-spline coefficients.

2) Equations (1) and (2) can be expressed in matrix form as:

$$\underline{y} = \underline{C}\underline{\omega} + \underline{\varepsilon}; \underline{\varepsilon} \sim N(0, \sigma^2 \underline{W})$$

$$\text{where } \underline{C} = [\underline{X} : \underline{T}] \text{ and } \underline{\omega} = [\underline{\beta} \quad \underline{\alpha}]^T$$

After stacking all subject-level vectors, the model can be expressed as $y = C \omega + \epsilon$, where $C = [X:T]$ is the combined design matrix and ω contains all parametric and spline coefficients. The error vector was assumed to follow a multivariate normal distribution with mean zero and covariance $\sigma^2 W$. The working matrix W is block diagonal across infants. In the present application, each within-infant block is diagonal, which permits subject-specific variances but imposes working independence among repeated measurements after conditioning on the fitted mean. Thus, the weighting structure addresses heteroscedasticity but does not estimate residual serial correlation. This assumption is stated explicitly because repeated observations from the same infant are not intrinsically independent; if a non-diagonal covariance structure such as compound symmetry, autoregressive correlation, or an unstructured block is available, the same generalized least-squares form can be used by replacing the corresponding blocks of W .

The positive-definite weight matrix was treated as known up to the common scale parameter in the theoretical derivation. In an empirical implementation, W must either be specified from substantive assumptions or estimated in a feasible weighted-least-squares step. The reported inference is conditional on that working matrix. Misspecification of W may affect standard errors and confidence limits even when the fitted mean curve remains descriptively useful. For this reason, the interpretation of the application is restricted to the assumed diagonal working covariance, and no claim is made that all within-infant dependence has been eliminated.

3) The parameter ω is estimated using the weighted least squares method by minimizing the following function:

$$Q = (\underline{y} - \underline{C}\underline{\omega})^T \underline{W}^{-1}(\underline{y} - \underline{C}\underline{\omega})$$

For fixed p , knots, and W , minimizing the weighted residual sum of squares gives $\hat{\omega} = (C'W^{-1}C)^{-1}C'W^{-1}y$, provided that C has full column rank. The fitted mean vector can then be written as $\hat{f} = A y$, with $A = C(C'W^{-1}C)^{-1}C'W^{-1}$. Candidate spline specifications were compared using generalized cross-validation. In weighted form, the criterion was evaluated as $GCV = \{M^{-1}(y - \hat{f})'W^{-1}(y - \hat{f})\} / \{1 - \text{tr}(A)/M\}^2$. The model with the smallest GCV among the candidate orders and knot configurations was selected. GCV was used only as a relative selection criterion; the minimum value does not by itself verify the covariance assumption or the distributional adequacy of the selected model.

4) Derive the sampling distribution of $\hat{\omega}$.

5) Equation (2) is rewritten as:

Under the normal-error assumption and conditional on the selected design and working covariance, $\hat{\omega}$ is normally distributed with expectation ω and covariance $\sigma^2(C'W^{-1}C)^{-1}$. At a specified predictor combination (x_{ij}, t_{ij}) , the fitted mean is a linear combination of $\hat{\omega}$ and is therefore normally distributed. Its variance is obtained from the corresponding diagonal element of $C(C'W^{-1}C)^{-1}C'$. When σ^2 is unknown, it is estimated from the weighted residual sum of squares divided by $M - q$, where $q = 2 + p + k$.

The resulting standardized statistic follows a Student's t distribution with $M - q$ degrees of freedom under the stated assumptions.

The resulting $(1 - \alpha)100\%$ interval is a pointwise confidence interval for the mean regression curve at a particular predictor value. It is not a prediction interval for an individual infant because it does not add future-observation variability, and it is not a simultaneous confidence band because the nominal coverage applies separately at each evaluation point. The theoretical validity of the t -based interval requires a fixed full-rank design, normally distributed errors, a correctly specified covariance matrix up to scale, and an error-variance estimator independent of the fitted mean. Because the spline order and knots were selected using the same observations, the empirical interval should be interpreted as conditional on the selected model.

$$y_{ij} = f(x_{ij}, t_{ij}) + \varepsilon_{ij}$$

6) Derive the sampling distribution of the regression curve $\hat{f}(x, t)$.

7) Construct the pivotal quantity for the regression curve $\hat{f}(x_{ij}, t_{ij})$.

Let the pivotal quantity $T_{ij}(x, t)$ for the regression curve be $\hat{f}(x_{ij}, t_{ij})$.

8) Solve the pivotal equation in probability form.

$$P(a_{ij} \leq T_{ij} \leq b_{ij}) = 1 - \alpha$$

The computational implementation was developed in R. For every candidate specification, the program constructed the parametric and truncated-spline design matrices, evaluated the weighted estimator, generated fitted values, calculated GCV, and retained the specification with the minimum criterion value. The selected model was then used to estimate σ^2 , the covariance of the fitted mean, and the lower and upper 95% pointwise confidence limits. Calculations were performed using matrix operations consistent with the derivation. Numerical implementation should verify the rank and condition number of $C'W^{-1}C$ before inversion; a singular matrix indicates duplicated basis functions or an unidentified parameterization.

b. Develop an algorithm and R program to estimate confidence intervals for semiparametric regression curves based on the truncated spline estimator for longitudinal data.

Apply the developed algorithm and program to the infant growth data.

For the application, the candidate models comprised the spline orders and knot configurations reported in Table 1. Model adequacy was evaluated using the GCV criterion, the coefficient of determination, mean squared error, and a residual normality assessment. The Kolmogorov-Smirnov result was interpreted as evidence for failing to reject the null hypothesis of normality rather than proof that the residuals were normal. The fitted sex-specific trajectories, their pointwise confidence limits, and the observations were subsequently displayed to support substantive interpretation. Because the diagonal working covariance does not model residual within-infant correlation and only 16 infants were observed, the empirical findings are considered illustrative and should not be generalized beyond the analyzed sample without external validation.

Results

Construction of Confidence Intervals for Semiparametric Regression Curves Based on the Truncated Spline Estimator for Longitudinal Data

In the semiparametric regression model for longitudinal data, paired observations are given by (x_{ij}, t_{ij}, y_{ij}) , $i = 1, 2, \dots, n; j = 1, 2, \dots, m_i$ where y_{ij} represents the response of subject i at observation j . The model is written as follows:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + g(t_{ij}) + \varepsilon_{ij}$$

Here β_0 and β_1 are the parametric regression coefficients, $g(t_{ij})$ is an unknown function that is approximated using a nonparametric regression approach, and ε_{ij} is normally distributed with mean zero and variance σ_i^2 . The function $g(t_{ij})$ in equation (6) can be expressed as

$$g(t_{ij}) = \sum_{s=0}^p \alpha_s t_{ij}^s + \sum_{r=1}^k \alpha_{(p+r)} (t_{ij} - \tau_r)_+^p$$

where $(t_{ij} - \tau_r)_+^p$ is defined as follows:

$$(t_{ij} - \tau_r)_+^p = \begin{cases} (t_{ij} - \tau_r)^p, & t_{ij} \geq \tau_r \\ 0, & t_{ij} < \tau_r \end{cases}$$

Equations (6)-(8) contain an intercept in the parametric component and a constant term in the truncated-power basis. These two columns are identical if they are estimated simultaneously. Therefore, the spline constant was constrained to zero and its corresponding design-matrix column was omitted before estimation. Under this identifiable parameterization, the coefficient vector contains $q = 2 + p + k$ estimable parameters. This clarification is essential because the inverse in the weighted least-squares estimator exists only when the combined design matrix has full column rank.

Equation (6) can be expressed in matrix form as

$$\underline{y} = \underline{X}\underline{\beta} + \underline{T}\underline{\alpha} + \underline{\varepsilon} \quad (9)$$

with

$$\underline{y}_i = [y_{i1} \quad y_{i2} \quad \cdots \quad y_{im_i}]^T, i = 1, 2, \dots, n$$

$$\underline{X}_i = \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{im_i} \end{bmatrix},$$

$$\underline{T}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & \cdots & t_{i1}^p & (t_{i1} - \tau_1)_+^p & \cdots & (t_{i1} - \tau_k)_+^p \\ 1 & t_{i2} & t_{i2}^2 & \cdots & t_{i2}^p & (t_{i2} - \tau_1)_+^p & \cdots & (t_{i2} - \tau_k)_+^p \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & t_{im_i}^2 & \cdots & t_{im_i}^p & (t_{im_i} - \tau_1)_+^p & \cdots & (t_{im_i} - \tau_k)_+^p \end{bmatrix},$$

$$\underline{\beta} = [\beta_0 \quad \beta_1]^T, \underline{\alpha} = [\alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_p \quad \alpha_{p+1} \quad \cdots \quad \alpha_{p+k}]^T$$

and

$$\underline{\varepsilon}_i = [\varepsilon_{i1} \quad \varepsilon_{i2} \quad \cdots \quad \varepsilon_{im_i}]^T$$

Equation (9) can be written as:

$$\underline{y} = \underline{C}\underline{\omega} + \underline{\varepsilon}, \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{W})$$

where $\underline{C} = [\underline{X} \quad \underline{T}]$ and $\underline{\omega} = [\underline{\beta} \quad \underline{\alpha}]^T$. The parameter $\underline{\omega}$ is estimated using weighted least squares (WLS) to handle heteroscedasticity. Therefore, a weighting matrix $\underline{W}$ (variance-covariance matrix) with size $M \times M$ is required. The estimator of $\underline{\omega}$ in equation (10) can be obtained by solving the following WLS optimization problem:

$$\min_{\underline{\omega}} \{(\underline{y} - \underline{C}\underline{\omega})^T \underline{W}^{-1}(\underline{y} - \underline{C}\underline{\omega})\}$$

with

$$\mathbf{W} = \text{cov}(\mathbf{y}) = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{\Sigma}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{\Sigma}_n \end{bmatrix}$$

The block-diagonal form of $\mathbf{W}$ separates different infants. In the empirical model, each block $\mathbf{\Sigma}_i$ is diagonal; therefore, the procedure allows heteroscedastic variances but uses working independence for repeated measurements within an infant. The estimates and confidence limits reported below are conditional on this covariance choice. A non-diagonal block could be used in the same generalized least-squares estimator, but doing so would require re-estimation of the model and its standard errors.

and

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_i^2 & 0 & \cdots & 0 \\ 0 & \sigma_i^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_i^2 \end{bmatrix}$$

for $i = 1, 2, \dots, n$ and $\text{cov}(y_{ik}, y_{ir}) = 0, k \neq r$. The optimization problem in equation (12) is solved as follows.

$$\{(\mathbf{y} - \mathbf{C}\boldsymbol{\omega})^T \mathbf{W}^{-1}(\mathbf{y} - \mathbf{C}\boldsymbol{\omega})\} = \mathbf{y}^T \mathbf{W}^{-1} \mathbf{y} - 2\boldsymbol{\omega}^T \mathbf{C}^T \mathbf{W}^{-1} \mathbf{y} + \boldsymbol{\omega}^T \mathbf{C}^T \mathbf{W}^{-1} \mathbf{C} \boldsymbol{\omega}$$

Let $Q = \mathbf{y}^T \mathbf{W}^{-1} \mathbf{y} - 2\boldsymbol{\omega}^T \mathbf{C}^T \mathbf{W}^{-1} \mathbf{y} + \boldsymbol{\omega}^T \mathbf{C}^T \mathbf{W}^{-1} \mathbf{C} \boldsymbol{\omega}$, to obtain the estimator of $\boldsymbol{\omega}$, the partial derivative of Q with respect to $\boldsymbol{\omega}$ is taken as follows:

$$\frac{\partial Q}{\partial \boldsymbol{\omega}} = -2\mathbf{C}^T \mathbf{W}^{-1} \mathbf{y} + 2\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C} \boldsymbol{\omega}$$

The minimum value Q is obtained from $\frac{\partial Q}{\partial \boldsymbol{\omega}} = 0$

Thus, we obtain

$$\begin{aligned} -2\mathbf{C}^T \mathbf{W}^{-1} \mathbf{y} + 2\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C} \boldsymbol{\omega} &= 0 \\ \mathbf{C}^T \mathbf{W}^{-1} \mathbf{C} \boldsymbol{\omega} &= \mathbf{C}^T \mathbf{W}^{-1} \mathbf{y} \\ \hat{\boldsymbol{\omega}} &= (\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{W}^{-1} \mathbf{y} \end{aligned}$$

Equation (13) shows that the estimator is a linear function of the response vector. Its unbiasedness follows from $E(\epsilon) = 0$ together with a correctly specified full-rank mean structure, whereas normality follows from the assumed multivariate normal error distribution. These properties are model-based rather than distribution-free. They do not remain exact if the covariance matrix is misspecified, the design is rank deficient, or spline-selection uncertainty is incorporated after the model has been selected.

Several properties of the estimator $\boldsymbol{\omega}$ can be derived as follows:

1. $\hat{\boldsymbol{\omega}}$ is linear in $\mathbf{y}$
2. $\hat{\boldsymbol{\omega}}$ is unbiased

In equation (10), the random error is known as $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{W})$ and $\mathbf{y}$ a linear combination of $\boldsymbol{\epsilon}$, hence

$$\underline{y} \sim N(\underline{C}\omega, \sigma^2 \underline{W})$$

Based on equation (14), the expectation of $\hat{\omega}$ can be obtained as follows.

$$\begin{aligned} E(\hat{\omega}) &= E[(\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1} \underline{y}] \\ &= [(\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1} E(\underline{y})] \\ &= (\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1} \underline{C}\omega \\ &= \omega \end{aligned}$$

Based on equation (15), it can be concluded that $\hat{\omega}$ is an unbiased estimator because $E(\hat{\omega}) - \omega = 0$.

3. $\hat{\omega}$ is normally distributed

Since $\hat{\omega}$ is linear in $\underline{y}$, it can be concluded that $\hat{\omega}$ is normally distributed and can be written as

$$\hat{\omega} \sim N(E(\hat{\omega}), Var(\hat{\omega}))$$

$$\begin{aligned} \text{Where } Var(\hat{\omega}) &= (\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1} Var(\underline{y}) (\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1})^T \\ &= \sigma^2 (\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \end{aligned}$$

Based on equations (15) and (16), the distribution of the parameter estimator $\hat{\omega}$ is therefore given by:

$$\hat{\omega} \sim N(\omega, \sigma^2 (\underline{C}^T \underline{W}^{-1} \underline{C})^{-1})$$

Next, the point estimate of the semiparametric truncated spline regression curve for longitudinal data is obtained by substituting equation (13) into equation (10), which gives:

$$\begin{aligned} \hat{\underline{y}} &= \hat{\underline{f}}(x, t) = \underline{C}\hat{\omega} \\ &= \underline{C}(\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1} \underline{y} = \underline{A}_{semipar} \underline{y} \end{aligned}$$

$$\text{Where } \underline{A}_{semipar} = \underline{C}(\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \underline{W}^{-1}$$

Several properties of the point estimator $\hat{\underline{f}}(x, t)$ for the semiparametric truncated spline regression curve in longitudinal data can be derived as follows:

1. $\hat{\underline{f}}(x, t)$ is linear in $\underline{y}$.
2. $\hat{\underline{f}}(x, t)$ is unbiased.

The estimator $\hat{\underline{f}}(x, t)$ is unbiased if $E(\hat{\underline{f}}(x, t)) - \underline{f}(x, t) = 0$, that can be shown as

$$E(\hat{\underline{f}}(x, t)) = E(\underline{C}\hat{\omega}) = \underline{C}E(\hat{\omega}) = \underline{C}\omega = \underline{f}(x, t)$$

3. $\hat{\underline{f}}(x, t)$ is normally distributed

Since $\hat{\underline{f}}(x, t)$ is linear in $\underline{y}$, it follows that $\hat{\underline{f}}(x, t)$ is normally distributed and can be written as

$$\begin{aligned} \hat{\underline{f}}(x, t) &\sim N\left(E(\hat{\underline{f}}(x, t)), Var(\hat{\underline{f}}(x, t))\right) \\ Var(\hat{\underline{f}}(x, t)) &= Var(\underline{C}\hat{\omega}) \\ &= \underline{C}Var(\hat{\omega})\underline{C}^T \\ &= \sigma^2 \underline{C}(\underline{C}^T \underline{W}^{-1} \underline{C})^{-1} \underline{C}^T \end{aligned}$$

Based on equations (19) and (20), the distribution of $\hat{f}(x, t)$ is obtained as

$$\hat{f}(x, t) \sim N(\hat{f}(x, t), \sigma^2 \mathbf{C}(\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C})^{-1} \mathbf{C}^T)$$

The distribution in equation (21) concerns the estimated conditional mean curve. Its variance quantifies uncertainty in estimating the mean at a specified value of sex and age. It does not include the additional residual variance required to predict a new individual height measurement. Consequently, the inferential result developed in the following equations is a confidence interval for the mean curve rather than a prediction interval.

To construct confidence interval estimates for the semiparametric truncated spline regression curve for longitudinal data, the next step after deriving the distribution of the point estimator is to obtain the pivotal quantity for $\hat{f}_{ij}(x, t)$ where $i = 1, 2, \dots, n; j = 1, 2, \dots, m_i$. The pivotal quantity is obtained using the following transformation:

$$T_{ij} = \frac{\hat{f}(x_{ij}, t_{ij}) - E(\hat{f}(x_{ij}, t_{ij}))}{\sqrt{\left(Var(\hat{f}(x, t))\right)_{ij}}}$$

By substituting equation (21) into equation (22), we obtain

$$T_{ij}(x, t) = \frac{\hat{f}(x_{ij}, t_{ij}) - f(x_{ij}, t_{ij})}{\sqrt{\sigma^2 (\mathbf{C}(\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C})^{-1} \mathbf{C}^T)_{ij}}}$$

Here, $T_{ij}(x, t)$ denotes the pivotal quantity for the regression curve of the subject i observation j . The term $(\mathbf{C}(\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C})^{-1} \mathbf{C}^T)_{ij}$ denotes the diagonal element corresponding to subject i at observation j for $i = 1, 2, \dots, n; j = 1, 2, \dots, m_i$. In practice, the population variance is unknown. Therefore, σ^2 is estimated by $\hat{\sigma}^2$ as follows.

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T (\mathbf{I} - \mathbf{A}_{\text{semipar}}) \mathbf{y}}{M - (2 + p + k)}$$

For consistency with weighted least squares, the residual-variance estimator is evaluated in weighted form as $\sigma\text{-hat squared} = (\mathbf{y} - \mathbf{y}\text{-hat})^T \mathbf{W}^{-1} (\mathbf{y} - \mathbf{y}\text{-hat}) / [M - (2 + p + k)]$. The unweighted residual expression is recovered only when $\mathbf{W}$ is the identity matrix. The denominator corresponds to the total number of observations minus the $q = 2 + p + k$ estimable coefficients.

Based on equation (24), the pivotal quantity in equation (23) can therefore be written as follows:

$$T_{ij}(x, t) = \frac{\hat{f}(x_{ij}, t_{ij}) - f(x_{ij}, t_{ij})}{\sqrt{\frac{\mathbf{y}^T (\mathbf{I} - \mathbf{A}_{\text{semipar}}) \mathbf{y}}{M - (2 + p + k)} (\mathbf{C}(\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C})^{-1} \mathbf{C}^T)_{ij}}}$$

After obtaining the pivotal quantity in equation (25), the next step is to determine its distribution.

$$T_{ij}(x, t) = \frac{\hat{f}(x_{ij}, t_{ij}) - f(x_{ij}, t_{ij})}{\sqrt{\frac{\mathbf{y}^T (\mathbf{I} - \mathbf{A}_{\text{semipar}}) \mathbf{y}}{M - (2 + p + k)} (\mathbf{C}(\mathbf{C}^T \mathbf{W}^{-1} \mathbf{C})^{-1} \mathbf{C}^T)_{ij}}} \sim t_{(M - (2 + p + k))}$$

Equation (26) shows that the pivotal quantity follows a Student's t -distribution with $M - (2 + p + k)$ degrees of freedom. Therefore, to construct the confidence interval for the semiparametric

truncated spline regression curve for longitudinal data, it is necessary to solve a probability equation of the specified size $1 - \alpha$ where $0 < \alpha < 1$. Let $a_{ij}(x, t) = a_{ij}$, $b_{ij}(x, t) = b_{ij}$, and let $T_{ij}(x, t)$ be pivotal quantity. Then, the probability equation can be written as follows.

$$P(a_{ij} \leq T_{ij}(x, t) \leq b_{ij}) = 1 - \alpha, 0 < \alpha < 1, i = 1, 2, \dots, n; j = 1, 2, \dots, m_i$$

Based on equation (26), equation (27) can be rewritten as:

$$P(a_{ij} \leq \frac{\hat{f}(x_{ij}, t_{ij}) - f(x_{ij}, t_{ij})}{\sqrt{\frac{y^T(I - A_{\text{semipar}})y}{M - (2 + p + k)}(C(C^T W^{-1} C)^{-1} C^T)_{ij}}} \leq b_{ij}) = 1 - \alpha$$

Equation (28) can be rewritten as:

$$f(x_{ij}, t_{ij}) \leq \hat{f}(x_{ij}, t_{ij}) - a_{ij}\sqrt{s_{ij}} \text{ dan } f(x_{ij}, t_{ij}) \geq \hat{f}(x_{ij}, t_{ij}) - b_{ij}\sqrt{s_{ij}} \quad (29)$$

Where $s_{ij} = \frac{y^T(I - A_{\text{semipar}})y}{M - (2 + p + k)}(C(C^T W^{-1} C)^{-1} C^T)_{ij}$ is the main diagonal element of matrix s corresponding to subject i at observation j .

Therefore, equation (29) can be rewritten as:

$$P(\hat{f}(x_{ij}, t_{ij}) - b_{ij}\sqrt{s_{ij}} \leq f(x_{ij}, t_{ij}) \leq \hat{f}(x_{ij}, t_{ij}) - a_{ij}\sqrt{s_{ij}}) = 1 - \alpha$$

Based on equations (26) and (30), the $(1 - \alpha)100\%$ confidence interval

for the semiparametric truncated spline regression curve for longitudinal data is obtained as follows.

$$P\left(\hat{f}(x_{ij}, t_{ij}) - t_{\left(\frac{\alpha}{2}, M - (2 + p + k)\right)}\sqrt{s_{ij}} \leq f(x_{ij}, t_{ij}) \leq \hat{f}(x_{ij}, t_{ij}) + t_{\left(\frac{\alpha}{2}, M - (2 + p + k)\right)}\sqrt{s_{ij}}\right)$$

Interpretation of the Theoretical Result

Equation (31) is the principal theoretical result of the study. It provides a $(1 - \alpha)100\%$ pointwise confidence interval for the semiparametric mean curve at one predictor combination. The Student's t quantile reflects estimation of the common variance scale, while the relevant diagonal element of the fitted-curve covariance matrix determines the local standard error. The interval is conditional on the selected spline order, knot locations, working covariance matrix, and normal-error assumption. It should not be described as a simultaneous band covering the complete curve with probability $1 - \alpha$.

The derivation extends the usual linear-model confidence interval to a combined parametric and truncated-spline design. Its practical value is the transparent matrix representation, which allows the same calculations to be applied to different ages and parametric predictor values after the design matrix has been constructed. This result is consistent with the broader spline literature, in which the fitted curve is a linear smoother conditional on its basis and smoothing specification (Eilers & Marx, 1996; Ruppert et al., 2003).

Algorithm for Constructing Confidence Intervals for Semiparametric Regression Curves Based on the Truncated Spline Estimator for Longitudinal Data

- 1) Input paired data (y_{ij}, x_{ij}, t_{ij}) with $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m_i$.
- 2) Model the relationship between x , t , and y using semiparametric regression based on the truncated spline estimator.
- 3) Determine the optimal knot points based on the minimum GCV criterion.
- 4) Calculate the upper and lower confidence limits based on the best model.

Application of Confidence Intervals for Semiparametric Regression Curves Based on the Truncated Spline Estimator to Height-Growth Data of Infants Aged 0-24 Months in Surabaya in 2015

$t_{ij}xyt_{ij}$ The empirical application used longitudinal height measurements of infants aged 0-24 months in Surabaya in 2015. The dataset contained 16 infants, comprising eight male and eight female infants, with 252 repeated measurements in total: 119 male observations and 133 female observations. Height in centimeters was the response, age in months was modeled through the truncated-spline component, and sex was represented by the parametric component using $x = 1$ for male infants and $x = 0$ for female infants. The records had previously been screened according to the WHO criteria stated by the data provider.

The age-height relationship was modeled using candidate truncated splines because infant growth is not expected to follow one constant linear trend throughout the first 24 months. Candidate specifications differed in polynomial order, number of knots, and knot positions. Each model was fitted to the same response and predictor structure, and the optimal specification was selected using the minimum GCV criterion. The comparative results are presented in Table 1.

Table 1. Selection of the Optimal Order Based on the Optimal Number of Knots for Infant Height

Order	Number of Knots	Optimal Knot Points	GCV	R^2
1	1	4	9.4135	0.9016
	2	4 ; 8	9.0084	0.9066
	3	4 ; 8 ; 20	9.0020	0.9074
	4	4 ; 8 ; 12 ; 20	9.0501	0.9077
	5	4 ; 8 ; 12 ; 16 ; 20	9.1216	0.9077
2	1	8	8.9343	0.9074
	2	4 ; 8	8.8820	0.9087
	3	4 ; 12 ; 20	8.8842	0.9094
	4	4 ; 8 ; 12 ; 16	8.8505	0.9105
	5	4 ; 8 ; 12 ; 16 ; 20	8.9075	0.9106

Table 1 shows that the smallest GCV value was 8.8505, obtained for a second-order truncated spline with four knots at 4, 8, 12, and 16 months. Its coefficient of determination was 0.9105, indicating that approximately 91.05% of the observed variation in height was accounted for by the fitted mean structure. Although the five-knot second-order model produced a slightly larger R-squared value of 0.9106, its GCV increased to 8.9075. The four-knot model was therefore retained because it achieved the best balance among the candidate specifications according to GCV. This conclusion is relative to the models evaluated and does not establish that the selected knot configuration is uniquely optimal outside the candidate set.

$$\hat{y} = 50,8965 + 0,4926x + 5,0768t - 0,4457t^2 + 0,4543(t-4)_+^2 - 0,1650(t-8)_+^2 + 0,2815(t-12)_+^2 - 0,2052(t-16)_+^2$$

Equation (32) can be interpreted as the following piecewise form.

$$\hat{y} = \begin{cases} 50,8965 + 0,4926x + 5,0768t - 0,4457t^2 & ; (0 \leq t < 4) \\ 58,1653 + 0,4926x + 1,4424t + 0,0086t^2 & ; (4 \leq t < 8) \\ 47,6053 + 0,4926x + 4,0824t - 0,1564t^2 & ; (8 \leq t < 12) \\ 88,1413 + 0,4926x - 2,6736t + 0,1251t^2 & ; (12 \leq t < 16) \\ 35,6101 + 0,4926x + 3,8928t - 0,0801t^2 & ; (t \geq 16) \end{cases}$$

For the selected model, the reported MSE was 8.2975 and the coefficient of determination was 91.0469%. These values indicate a close in-sample fit, but they should be interpreted together with the small number of independent subjects and the working covariance assumption.

Equation (32) gives the estimated population-average height curve conditional on sex and age. Substituting the binary sex code into equation (32) produces the male and female equations reported below.

a. Height growth of male infants

The estimated height-growth model for male infants is given by:

$$\hat{y} = 51,3891 + 5,0768t - 0,4457t^2 + 0,4543(t-4)_+^2 - 0,1650(t-8)_+^2 + 0,2815(t-12)_+^2 - 0,2052(t-16)_+^2$$

Equation (34) can be rewritten in the following piecewise form.

$$\hat{y} = \begin{cases} 51,3891 + 5,0768t - 0,4457t^2 & ; (0 \leq t < 4) \\ 58,6579 + 1,4424t + 0,0086t^2 & ; (4 \leq t < 8) \\ 48,0979 + 4,0824t - 0,1564t^2 & ; (8 \leq t < 12) \\ 88,6339 - 2,6736t + 0,1251t^2 & ; (12 \leq t < 16) \\ 36,1027 + 3,8928t - 0,0801t^2 & ; (t \geq 16) \end{cases}$$

Using the relevant polynomial segment in equation (35), the estimated mean height of a male infant at four months was 64.5651 cm. This is a fitted mean from the selected model, not an individual prediction. Variation among infants remains around the mean trajectory and is represented separately by the model error and the reported confidence limits.

b. Height growth of female infants

The estimated height-growth model for female infants is given by

$$\hat{y} = 50,8965 + 5,0768t - 0,4457t^2 + 0,4543(t-4)_+^2 - 0,1650(t-8)_+^2 + 0,2815(t-12)_+^2 - 0,2052(t-16)_+^2$$

Equation (36) can be rewritten in the following piecewise form.

$$\hat{y} = \begin{cases} 50,8965 + 5,0768t - 0,4457t^2 & ; (0 \leq t < 4) \\ 58,1653 + 1,4424t + 0,0086t^2 & ; (4 \leq t < 8) \\ 47,6053 + 4,0824t - 0,1564t^2 & ; (8 \leq t < 12) \\ 88,1413 - 2,6736t + 0,1251t^2 & ; (12 \leq t < 16) \\ 35,6101 + 3,8928t - 0,0801t^2 & ; (t \geq 16) \end{cases}$$

Using the relevant polynomial segment in equation (37), the estimated mean height of a female infant at 12 months was 74.0725 cm. Equations (34) and (36) differ by the constant sex coefficient of 0.4926 cm. Under the model coding, male infants therefore had an estimated mean height 0.4926 cm greater than female infants at the same modeled age. This coefficient represents a constant difference in expected height, not a difference in growth velocity. A sex-specific growth-rate conclusion would require an interaction between sex and the age-spline basis. Because no standard error or hypothesis test for the sex coefficient is reported, the estimated difference should be interpreted descriptively rather than as evidence of a statistically significant biological difference. Figure 1 compares the two fitted mean curves.

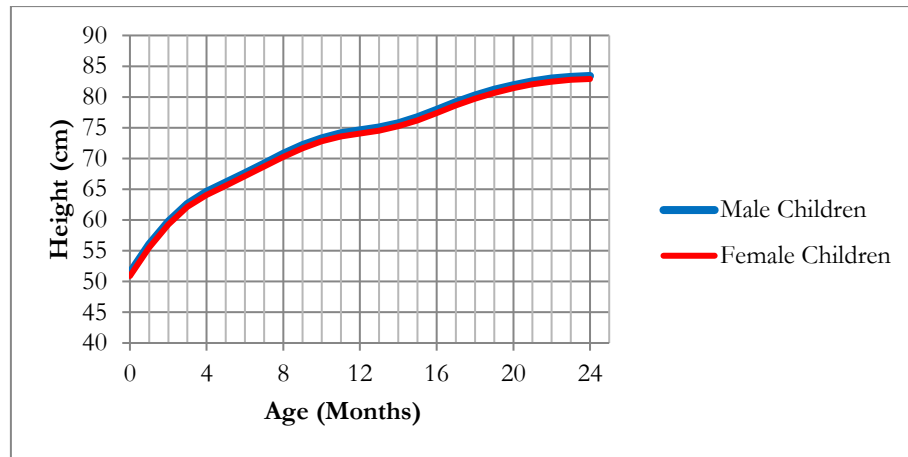


Figure 1. Comparison of the estimated average height-growth curves for male and female infants

Equations (34) and (36) represent the selected second-order spline model with knots at 4, 8, 12, and 16 months. The 95% pointwise confidence limits for the male and female mean curves were calculated from equation (31). Their validity depends on the normal-error and working-covariance assumptions described in the Method section. Residual normality was evaluated using the hypotheses stated below; this test was treated as one diagnostic component rather than a complete assessment of longitudinal model adequacy.

H_0 : The residuals are normally distributed

H_1 : The residuals are not normally distributed

The Kolmogorov-Smirnov test produced $p = 0.0645$, which exceeded $\alpha = 0.05$. Therefore, the null hypothesis was not rejected. This result indicates that the test did not identify a statistically significant departure from normality at the 5% level; it does not prove that the residuals were normally distributed. Moreover, a marginal residual-normality test does not assess within-infant correlation or heteroscedasticity. The pointwise lower and upper confidence limits are presented in Table 2.

Table 2. Lower and Upper Bounds of the Confidence Intervals for the Infant Height-Growth Curves

Male Infants			Female Infants		
Age	Lower Bound	Upper Bound	Age	Lower Bound	Upper Bound
0	37.9862	64.7920	0	37.6890	64.1040
1	46.4023	65.6381	1	46.2817	64.7735
2	48.8160	70.7039	2	48.6454	69.8892
3	51.7406	73.3132	3	51.4798	72.7515
4	55.8171	73.4759	4	55.3896	72.7556
5	57.1441	75.0259	5	56.5345	74.6503
6	57.7523	77.4920	6	57.1472	77.1118
7	59.7631	78.5900	7	59.3105	78.0573
8	62.1824	79.3140	8	62.0994	78.4118
9	62.3693	81.9749	9	62.5670	80.7920
10	62.5649	83.2106	10	62.8654	81.8446
11	64.0194	84.0019	11	64.3993	82.7162
12	65.9254	84.0166	12	65.9506	82.7795
13	66.0290	84.1447	13	66.1480	82.9123
14	66.0564	85.3995	14	66.3061	84.5201
15	67.4293	85.9017	15	67.1800	85.2017
16	69.8769	85.9376	16	69.5079	85.2855
17	70.8114	87.0450	17	70.3590	86.5549
18	71.2356	88.8885	18	70.7405	88.4824
19	71.5739	90.3047	19	70.9948	89.9442
20	72.0192	90.7859	20	71.3946	90.4195
21	72.9502	90.8367	21	72.3020	90.5036
22	74.2496	90.9150	22	73.5975	90.5781

Male Infants			Female Infants		
Age	Lower Bound	Upper Bound	Age	Lower Bound	Upper Bound
23	74.3617	92.2061	23	73.8269	91.7558
24	75.2013	96.0168	24	74.5825	95.4841

Table 2 reports the 95% pointwise confidence limits for both sex-specific mean curves. Figure 2 provides the available graphical illustration for the female observations, fitted mean trajectory, and pointwise confidence limits. These limits describe uncertainty in the estimated mean height at each age and should not be interpreted as reference ranges within which 95% of individual measurements are expected to fall.

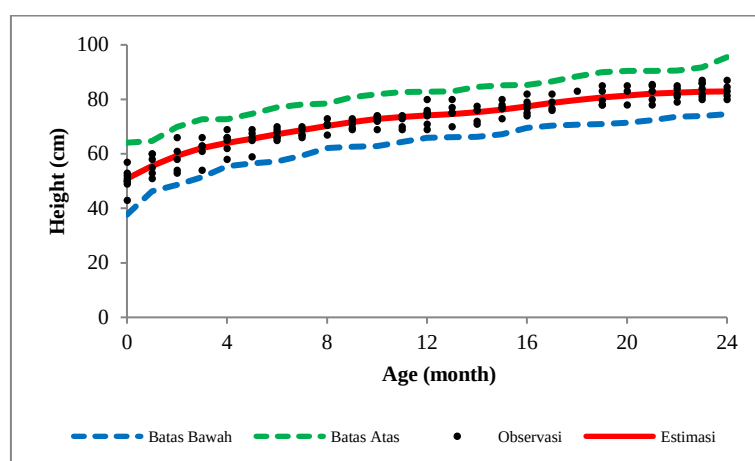


Figure 2. Height-Growth Curve of Female Infants

Interpretation and Discussion of the Empirical Findings

Descriptively, 117 of the 119 male observations (98.3193%) and all 133 female observations fell between the reported lower and upper limits. These proportions summarize the relationship between the observed sample and the fitted limits, but they are not estimates of confidence-interval coverage. A pointwise confidence interval targets the unknown mean curve, whereas the observed heights contain individual residual variability. Formal coverage evaluation would require repeated samples or a Monte Carlo simulation under known data-generating conditions. Therefore, the observed proportions cannot be used alone to conclude that the interval procedure is adequate or that it achieves nominal 95% coverage.

The piecewise coefficients in equation (32) indicate that the steepest estimated increase occurred from birth to four months. Over this interval, the fitted mean height increased by approximately 3.2940 cm per month. The smallest estimated increase occurred from 16 to 24 months, with an average increase of approximately 0.6906 cm per month. This declining velocity is broadly consistent with international infant-growth patterns, which show rapid linear growth during early infancy followed by gradual deceleration (WHO Multicentre Growth Reference Study Group, 2006). These quantities describe average slopes of the selected piecewise curve; they should not be interpreted as subject-specific growth rates or causal effects of age.

The empirical findings illustrate the principal advantage of the truncated-spline representation: one model can accommodate different local age trends while preserving a continuous fitted trajectory. The selected knots identify locations at which the polynomial expression changes, but they should not automatically be treated as biological change points. Their positions were selected to improve the statistical fit among the candidate models and may vary in another sample. This conditional interpretation is consistent with spline research emphasizing that basis and knot selection influence both curve shape and uncertainty (Eilers & Marx, 1996; Welsh et al., 2002).

Several limitations qualify the application. First, only 16 infants were observed, so the effective independent sample size is much smaller than the total of 252 repeated measurements. Second, the diagonal within-infant covariance assumes working independence after adjustment for the mean trajectory; residual serial correlation could change the estimated standard errors. Third, the reported intervals condition on the selected order and knots and do not incorporate model-selection uncertainty. Fourth, the sex effect is constrained to be constant across age because no sex-by-age interaction is included. Finally, the local Surabaya sample does not constitute a population growth standard. These limitations do not invalidate the fitted illustration, but they restrict inferential and external generalization.

Future work should compare diagonal, compound-symmetry, autoregressive, and unstructured working covariance matrices; evaluate interval coverage through simulation; and consider bootstrap or simultaneous-band procedures that account for knot selection. A sex-by-age spline interaction could also be assessed if the sample size is sufficient. These extensions would determine whether the apparent trajectory differences persist when within-infant dependence and model-selection uncertainty are handled more fully.

Conclusion

This study derived a pointwise confidence interval for the conditional mean curve in a semiparametric longitudinal regression model based on a truncated spline estimator. The derivation combines the parametric and spline components in one identifiable design matrix, estimates the coefficients through weighted least squares, and uses a Student's t pivotal quantity when the common variance scale is unknown. Under the stated assumptions, the interval is determined by the fitted mean, the appropriate diagonal element of its covariance matrix, and $M - (2 + p + k)$ degrees of freedom. The result provides a transparent matrix-based procedure that can be implemented after the spline order, knots, and working covariance have been specified. It is a pointwise confidence interval for the mean curve and should not be interpreted as either a prediction interval for an individual observation or a simultaneous confidence band for the entire trajectory.

$$P\left(\hat{f}(x_{ij}, t_{ij}) - t_{\left(\frac{\alpha}{2}, M-(2+p+k)\right)} \sqrt{s_{ij}} \leq f(x_{ij}, t_{ij}) \leq \hat{f}(x_{ij}, t_{ij}) + t_{\left(\frac{\alpha}{2}, M-(2+p+k)\right)} \sqrt{s_{ij}}\right)$$

In this expression, the fitted value is the estimated conditional mean at subject i and occasion j , s_{ij} is the corresponding diagonal element of the fitted-curve covariance factor, and the Student's t quantile uses $M - (2 + p + k)$ degrees of freedom. The formula is conditional on the selected spline specification and working covariance matrix.

In the infant-growth application, the minimum GCV criterion selected a second-order truncated spline with knots at 4, 8, 12, and 16 months. The selected model yielded an MSE of 8.2975 and explained 91.0469% of the observed variation in height. The fitted mean curve showed the largest average increase during the first four months and the smallest average increase from 16 to 24 months, which is consistent with the general deceleration of linear growth during infancy. The estimated sex coefficient of 0.4926 cm represents a constant difference in expected height between male and female infants at the same modeled age; it does not demonstrate a difference in growth velocity or a statistically significant biological effect. Similarly, the proportions of observations located between the reported limits are descriptive sample summaries and do not establish nominal confidence-interval coverage.

The findings are subject to several limitations. Only 16 infants were observed, the within-infant covariance was represented by a diagonal working structure, and the confidence intervals were calculated conditionally on the spline order and knots selected from the same data. Residual

serial correlation, model-selection uncertainty, and sex-specific age trajectories were therefore not fully addressed. In addition, the local sample from Surabaya cannot be generalized as a population growth standard. Future research should compare alternative within-subject covariance structures, evaluate interval coverage through Monte Carlo simulation or bootstrap procedures, develop simultaneous confidence bands, and assess sex-by-age spline interactions using a larger longitudinal sample. These extensions would provide stronger evidence regarding the inferential performance and external applicability of the proposed method.

Acknowledgment

-

References

- Brumback, B. A., & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93(443), 961-976. <https://doi.org/10.1080/01621459.1998.10473755>
- Budiantara, I. N. (2014). Nonparametric and semiparametric spline regression modeling. In *Proceedings of the 2014 National Seminar on Mathematics*, Universitas Udayana, Denpasar, Indonesia.
- Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3), 529-544. <https://doi.org/10.1093/biomet/asp035>
- Crainiceanu, C. M., Ruppert, D., & Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14(14), 1-24. <https://doi.org/10.18637/jss.v014.i14>
- Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford University Press.
- Durban, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24(8), 1153-1167. <https://doi.org/10.1002/sim.1991>
- Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89-121. <https://doi.org/10.1214/ss/1038425655>
- Fadhilah, K. N., Suparti, & Tarno. (2016). Truncated spline regression modeling for longitudinal data: A case study of stock prices in the banking stock group, January 2009-December 2015. *Jurnal Gaussian*, 5(3), 447-454.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Wiley.
- Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Chapman & Hall. <https://doi.org/10.1201/b15710>
- Gu, C., & Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *The Annals of Statistics*, 33(3), 1357-1379. <https://doi.org/10.1214/009053605000000110>
- Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1), 121-128. <https://doi.org/10.1111/j.0006-341X.2002.00121.x>
- Hoover, D. R., Rice, J. A., Wu, C. O., & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4), 809-822. <https://doi.org/10.1093/biomet/85.4.809>

- Huang, J. Z., Wu, C. O., & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1), 111-128. <https://doi.org/10.1093/biomet/89.1.111>
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22. <https://doi.org/10.1093/biomet/73.1.13>
- Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2), 381-400. <https://doi.org/10.1111/1467-9868.00183>
- Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2(1), 321-359. <https://doi.org/10.1146/annurev-statistics-010814-020413>
- Rice, J. A., & Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1), 253-259. <https://doi.org/10.1111/j.0006-341X.2001.00253.x>
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511755453>
- Verbyla, A. P., Cullis, B. R., Kenward, M. G., & Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3), 269-311. <https://doi.org/10.1111/1467-9876.00154>
- Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611970128>
- Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90(1), 43-52. <https://doi.org/10.1093/biomet/90.1.43>
- Welsh, A. H., Lin, X., & Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*, 97(458), 482-493. <https://doi.org/10.1198/016214502760047014>
- WHO Multicentre Growth Reference Study Group. (2006). WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatrica*, 95(S450), 76-85. <https://doi.org/10.1111/j.1651-2227.2006.tb02378.x>
- Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673-686. <https://doi.org/10.1198/016214504000000980>
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3-36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>
- Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press. <https://doi.org/10.1201/9781315370279>
- Wu, H., & Zhang, J.-T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97(459), 883-897. <https://doi.org/10.1198/016214502388618672>
- Zeger, S. L., & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50(3), 689-699. <https://doi.org/10.2307/2532782>

Zhang, D., Lin, X., Raz, J., & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93(442), 710-719.
<https://doi.org/10.1080/01621459.1998.10473723>