Paper 4739-2020

Estimating Unknown Change Points and Variation Using SAS®

Depeng Jiang, University of Manitoba; Lin Yan, University of Manitoba;
Zhongyuan Zhang, University of Toronto; Shaoping Jiang, Yunnan Minzi University

ABSTRACT

Change point (or knot, joint, turning point) was defined as "the time when development switches from one phase to another". Piecewise growth curve model (PGCM) is often used to estimate the underlying growth process. When fitting a PGCM, the conventional practice is to specify the change points a priori. However, the change points were often unknown and misspecifications of turning points could lead to bias of growth trait estimation. Also, there was individual variation in the change points. To estimate the individual specific change point, several different estimation methods (e.g., the profile likelihood, the first-order Taylor expansion and the Bayesian estimation methods) were proposed. Some R packages were developed to estimate the unknown change point. In SAS, the NLMIXED procedure was used to fit the nonlinear random effects models and could potentially be used to estimate the change points. We present the PGCMs to allow individual specific change points as a function of time-varying predictor. We illustrate these respective models with an empirical example to demonstrate the use of SAS in estimation of unknown change points and non-linear growth curve model. The implication and challenges in fitting these models are discussed.

INTRODUCTION

Longitudinal studies are designed to measure intra-individual change over time and inter-individual differences in these changes. In analyzing longitudinal data, individual differences in growth trajectories over time, are typically captured by random effects using mixed-effects models. These random effects or latent traits describe each person's trend across time, and explain the correlational structure of the longitudinal data.

The intra-individual growth trajectory in health and behavioral outcomes often consists of distinct segments (phases) of growth (e.g., Kreisman, 2003; Paris, 2005; Silverman, Speece, Harring, & Ritchey, 2012). For example, in studies of interventions, the individual trajectories before the intervention and after the intervention is different. Thus the overall growth trajectory in this scenario include two distinct segments of change, one segment that describe the trend before the intervention and one segment that characterizes the trend after the intervention. The change rate might also be quite different between post-intervention (short term) and follow-up assessments (long term). Because the rate of change is different in different segments, piecewise growth curve models or piece-wise random-effect models are often used to analyze this kind of segmented data set. The specification of separate growth profiles (functional forms) corresponds to multiple developmental phases of the overall change process and random effects describe the inter-individual differences in theses phases.

The major challenge in applying the piece-wise growth curve model (PGCM) is to specify the change points. Change point (or knot, joint, turning point) was defined as "the time when development switches from one phase to another". When fitting a PGCM, the conventional practice is to specify the change points a priori according to theories or designs (e.g., the start point of an intervention). Yet, such consideration may not always be possible or reasonable. In the intervention, even we know the start point of an intervention, the change or turning point may occur after the intervention due to delay in response to

intervention (Ning & Luo, 2017). The misspecifications of turning points could lead to bias of growth trait (e.g., growth rate) estimation, and may attenuate the relationship between the predictors and growth rates, leading to misleading inferential conclusions (Ning & Luo, 2017). In this situation, the unknown change point needs to be estimated based on data. With free estimation of change points, we can discover more optimal function form of each growth phases and give a more adequate description of the growth patterns in the data (Kwok, Luo, & West, 2010; Wood, & Jackson, 2013).

A piece-wise growth curve model (PGCM) with unknown change points is a type of nonlinear random-effects model (Cudeck & Klebe, 2002; Du Toit & Cudeck, 2009; Wang & McArdle, 2008). The change point as an unknown parameter in the PGCM is often treated as a fixed effect, in which each individual's change point is assumed to be the same. It might be beneficial to treat the change points as randomly varying across individuals to more accurately mirror individual differences in the timing of development switch. The estimation of the PGCM with individual specific change points is computationally very challenging because of the nonlinear random parameter (i.e., change point) where we estimate both its mean and its variance as well as its covariance with other random effects (Kohli, Harring, & Zopluoglu, 2016).

Individual-specific change points are relatively novel, and there are very few empirical studies to successfully model the random change points using traditional and Bayesian mixed effects models and growth mixture models (e.g., Cudeck, 1996; Dominicus, Ripatti, Pedersen, & Palmgren, 2008; McArdle & Wang, 2008; Muniz Terrera, van den Hout, & Matthews, 2011; Kohli, 2011; Li, Duncan, Duncan, & Hops, 2001). If the change point is not the same for all individuals, one of the most interesting features in PGCM is how to predict the change points using other time-invariant (e.g., individual characters) and time-varying predictors (e.g. individual time-specific intervention). Preacher and Hancock (2012, 2015) proposed and illustrated the a four-step strategy to use the structural equation modeling (SEM) based structured latent curve modeling (LCM) approach to estimate the random change points in PGCM. They not only show that how the change point may be treated as a random coefficient within the SEM/LCM framework but also demonstrate how to predict individual differences in the change point using time-invariant predictor. However, this SEM/LCM-based approach for PGCM requires some degree of balance in measurement schedules and it cannot predict the change point using time-varying predictors.

To best of our knowledge, there is no empirical study to include the time-varying predictors to predict individual differences in the change point. The goal of this paper is to present the statistical model for estimating individual specific change points as a function of other predictors. We illustrate these respective models with an empirical example to demonstrate the use of SAS in handling the PGCM with individual specific change point. We reexamine data published by Murnane, Willett, & Boudett (1999) related to the benefit of obtaining the General Educational Development certificate, or GED for male dropouts. Additionally, emphasis is placed on the interpretation of the empirical results and how it differs from conventional practice to specify the change points a priori.

EMPRICAL EXAMPLE

Murnane and colleagues (1999) used longitudinal data from the National Longitudinal Survey of Youth (NLSY) to examine whether the wage trajectories of male high school dropouts are affected by the acquisition of the GED credential. Approximately 500,000 school dropouts acquire this credential each year by passing the GED examinations, which test knowledge and/or skills in writing, social studies, science, mathematics, and interpreting literature and the arts. Their analytic sample included 901 males who left high school before graduation. Approximately two in five dropouts obtained a GED and did so at an average age of 20. They conducted the random intercept mixed-effect models and specified the log wage as a quadratic function of potential labor market experience. To

examine the impact of the acquisition of the GED, they also included a quadratic function of the potential labor market experience of a dropout from his receipt of GED. In addition, to examine whether the impact of GED acquisition on the shape of the wage profile was moderated by other important characteristics of the dropouts, they tested for statistical interactions between the Years Since GED predictor (and its square) and selected time-invariant characteristics of dropouts (e.g., indicator of low score on the Armed Forces Qualifying Test (AFQT) - a test of reading and mathematics skills). They show that acquisition of the GED results in wage increases for dropouts who left school with weak skills, but not for dropouts who left high school with stronger skills.

Singer and Willett (2003) reanalyzed the high school dropout wage data with a sample of 888 male dropouts to illustrate how to model discontinuity in longitudinal data analyses. They hypothesized that dropouts who obtain a GED might command higher salaries. If so, their (log) wage trajectories could exhibit a discontinuity upon GED receipt. They specified the log wage is a linear function of potential labor market experience and compared three different discontinuities upon GED receipt: an immediate shift in elevation but no shift in slope, an immediate shift in slope but no shift in elevation, and immediate shifts in both elevation and slope. They also examined whether these discontinuities vary by subjects by including both random intercept and random slopes in the mixed models. They found that on labor-force entry, a White male who dropped in 9th grade and who live in a community with an unemployment rate of 7% is expected to earn an hourly log wage of 1.7386 ($5.69 in constant 1990 dollars). Before GED attainment, log wages rise annually by 0.0415 (4.2% in raw wages). Upon GED receipt, log wages rise immediately by 0.0409 (4.2%) and then annually by 0.0509 (5.2%). Although the average effects of shifts in elevation and slope upon GED receipt were not statistically significant at conventional level (alpha = 0.05), there were significant variations of these effects. The effects of the three substantive predictors − local area unemployment rate, race, and highest grade completed- remain similar to those found by Murnane and colleagues (1999).

Both Murnane and colleagues (1999) and Singer and Willett (2003) fitted the random effect regression models and specify the change point a prior − upon the time in which the dropout obtained his GED. In this paper, we postulate that a discontinuity in individual wage trajectories might not occur right at the time of GED receipt. We fit the PGCM and treat the change points as an unknown function of the time of GED receipt.

METHODS

A two-phase linear-linear piecewise growth curve model with one unknown change points was used as an illustrate example. It can be specified in the form of two-level models. The Level 1 (repeated measures) model is specified as

$$y_{ij} = \begin{cases} \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij} & t_{ij} < \gamma_i \\ \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i}(t_{ij} - \gamma_i) + \varepsilon_{ij} & t_{ij} \geq \gamma_i \end{cases} \quad (1)$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad (2)$$

where y_{ij} is the response at the j^{th} measurement for the i^{th} individual, β_{0i} and β_{1i} are the subject-specific intercept and the slope growth factor before the change point, and β_{2i} denotes the subject-specific difference between slopes after the change point and slopes before the change point. γ_i is the location of the change point marking the shift from one growth phase to the other. ε_{ij} is the Level-1 residual for individual i at measurement j.

We assume that the individual-specific change point is a function of another time-varying covariate (e.g., indicator whether a dropout attained a GED at a specific year).

$$\gamma_i = f(x_{ij}) \tag{3}$$

The Level-2 (between-subject) model is specified as

$$\begin{cases} \beta_{0i} = u_{00} + u_{01} w_i + \varsigma_{0i} \\ \beta_{1i} = u_{10} + u_{11} w_i + \varsigma_{1i} \\ \beta_{2i} = u_{20} + u_{21} w_i + \varsigma_{2i} \end{cases} \tag{4}$$

with

$$\begin{bmatrix} \varsigma_{0i} \\ \varsigma_{1i} \\ \varsigma_{2i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Phi_{11} & & \\ \Phi_{21} & \Phi_{22} & \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \right), \tag{5}$$

where w is a predictor for Level-1 latent growth factors (β_{0i}, β_{1i} and β_{2i}). u_{k0} and u_{k1} (k=1,2,3) are intercept and slope of w for regression of β_{ki}; ζ_k is random disturbance (Level-2 residual) for the regression of β_{ki}. The Level-1 residuals and the Level-2 residuals are also assumed to be uncorrelated with each other and with the latent growth factors. The Level-2 residuals are assumed to follow multivariate normal distributions with means, variances, and covariance as in (5).

The estimation of above PGCM is rather complicated and time intensive, because the estimation of the location of the change point(s) (an intrinsically nonlinear parameter), along with the variance of other growth factors makes the overall computation of the model very challenging. The above model also cannot be estimated using the structural equation modeling (SEM) based structured latent curve modeling (LCM) approach proposed by Preacher and Hancock (2012, 2015). Most of the standard commercial software cannot easily fit above models, with the exception of SAS. The above PGCM can be fit in SAS using the NLMIXED procedure. PROC NLMIXED is a very flexible program that can be used to fit many nonlinear growth model that are linear or nonlinear in their parameters.

DATA

We reanalyze the dropout wage data used in Singer and Willett (2003). To track wages on a common temporal scale, Murnane and colleagues (1999) decided to clock time from each respondent's first day of work. This allow each hourly wage to be associated with a temporally appropriate point in the respondent's labor force history. The original data set has an unusual temporal schedule, varying not only in spacing but length. Some dropouts had more than one interviews within same year. We reorganized the dataset to create a year-based wage data set. If a respondent has two or more responses within one year, we take average of wages (log of wages) for those responses within that year. This data manipulation makes a sample of 5206 from 888 individuals available for our analyses. Across the full sample of 888 dropouts, 134 men have 1 or 2 waves (years) of data, 171 have 3 or 4, 187 have 5 or 6, 227 have 7 or 8, and 169 have more than 9.

In our sample, 27.7% of the dropouts are Black, 23.0% are Hispanic, and 49.3% are Non-Hispanic White (see Table 1). More than 50% of them drop out of school at Grade 9 or 10. In this sample, 581 dropouts did not receive a GED; among the remaining 307, the timing of GED attainment varies, with mean of 2.2 years and range from 0 to 12 years.

	Permanent Dropouts (N=581)	Eventual GED Holders (N=307)	p-value
Black	26.2%	30.6%	.16

Hispanic	23.2%	22.5%	.80
Non-Hispanic White	50.6%	46.9%	.29
Grade in high school at dropout	8.8(1.4)	9.1(1.3)	.003
Number of years between labor force entry and GED	--	2.2(1.7)	--

Table 1. Sample Characteristics of Dropouts by GED Status

MODELS

We specify five PGCMs that examine the impact of GED acquisition and whether the impact of GED acquisition on the shape of the wage profile was moderated by race of the dropouts. In Model A, we follow the similar specification as in Murnane and colleagues (1999): log wage is a quadratic function of potential labor market experience. The specification of Model A is as below

$$\begin{aligned}\ln(w_{ij}) = {} & \beta_{00} + \beta_{10}\text{Year\_1}_{ij} + \beta_{20}\text{Year\_1}_{ij}^2 + \beta_{30}\text{Year\_2}_{ij} + \beta_{40}\text{Year\_2}_{ij}^2 \\ & + \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} + \beta_{31}\text{Black}_i * \text{Year\_2}_{ij} \\ & + \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varepsilon_{ij}\end{aligned} \qquad \text{(Model A)}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad \varsigma_{0i} \sim N(0, \Phi_{11})$$

where ln(w_{ij}) represents the natural logarithm of the hourly wage earned by person *i* (*i* =1, 2, …, 888) on the j^{th} (*j* =1, 2, …, 12) year since work force entrance. *Year_1* denotes the number of years of potential labor market experience. In this model, we specified the log wages is a quadratic function of potential market experience by including both *Year_1* and its square. *Year_2* denotes the number of years of potential labor market experience of a dropout from his receipt of GED, and was set to zero on all occasions prior to GED receipt. The slopes of *Year_2* and its square ensured that the nonlinear shape of wage-experience trajectory could differ before and after receipt of the GED. Similarly, as in Murnane and colleagues (1999), we included the interactions of Black indicator and its interactions with *Year_1* and *Year_2* to examine whether wage trajectory and impact of GED receipt were different between African American and White (Hispanic and Non-Hispanic White). In Model A, we also include two additional predictors: highest grade completed (HGC_9, centered around Grade 9) and local area unemployment rate (UERate_C, grand mean centered).

In Model A, only the intercept was specified as a random effect while slopes of all predictors were specified as fixed effects. In Model B, we assume that linear slopes of *Year_1* and *Year_2* could vary across dropouts.

$$\begin{aligned}\ln(w_{ij}) = {} & \beta_{00} + \beta_{10}\text{Year\_1}_{ij} + \beta_{20}\text{Year\_1}_{ij}^2 + \beta_{30}\text{Year\_2}_{ij} + \beta_{40}\text{Year\_2}_{ij}^2 \\ & + \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} + \beta_{31}\text{Black}_i * \text{Year\_2}_{ij} \\ & + \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varsigma_{1i}\text{Year\_1}_{ij} + \varsigma_{2i}\text{Year\_2}_{ij} + \varepsilon_{ij}\end{aligned} \qquad \text{(Model B)}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad \begin{bmatrix} \varsigma_{0i} \\ \varsigma_{1i} \\ \varsigma_{2i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Phi_{11} & & \\ \Phi_{21} & \Phi_{22} & \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \right),$$

In both Model A and B, we assume that the receipt of GED will change the wage-experience of a dropout immediately from the time of his receipt of GED. This might not occur in reality. In Model C, we extend the Model B by including the estimation of individual change point of his wage trajectory as a function of the time of his receipt of GED (some delay).

$$\text{If } \text{Year\_2} \le \gamma$$

$$\begin{aligned}\ln(w_{ij}) &= \beta_{00} + \beta_{10}\text{Year\_1}_{ij} + \beta_{20}\text{Year\_1}_{ij}^2 \\ &+ \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} \\ &+ \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varsigma_{1i}\text{Year\_1}_{ij} + \varepsilon_{ij}\end{aligned} \qquad \text{(Model C)}$$

$$\text{If } \text{Year\_2} > \gamma$$

$$\begin{aligned}\ln(w_{ij}) &= \beta_{00} + \beta_{10}\text{Year\_1}_{ij} + \beta_{20}\text{Year\_1}_{ij}^2 + \beta_{30}(\text{Year\_2}_{ij} - \gamma) + \beta_{40}(\text{Year\_2}_{ij} - \gamma)^2 \\ &+ \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} + \beta_{31}\text{Black}_i * (\text{Year\_2}_{ij} - \gamma) \\ &+ \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varsigma_{1i}\text{Year\_1}_{ij} + \varsigma_{2i}\text{Year\_2}_{ij} + \varepsilon_{ij}\end{aligned}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad \begin{bmatrix}\varsigma_{0i} \\ \varsigma_{1i} \\ \varsigma_{2i}\end{bmatrix} \sim MVN\left(\begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}, \begin{bmatrix}\Phi_{11} & & \\ \Phi_{21} & \Phi_{22} & \\ \Phi_{31} & \Phi_{32} & \Phi_{33}\end{bmatrix}\right),$$

Both Model A and B can be easily fit in SAS using the MIXED procedure. An MIXED script for Model A and B can be written as

```
*  Model A: Similar specification as in Murnane and colleagues (1999);
PROC MIXED Data= wages_Y2  Method=ml noclprint noinfo COVTEST;
   class id;
   model LNW_Y =  Year_1 Year_1*Year_1 year_2 year_2*year_2
         black  black*Year_1
         black* Year_2  HGC_9 UERATE_C / solution notest outpm = mc;
   random intercept   / subject=id type=un;
RUN;

 *  Model B: Model A + Random effects of linear slopes before and after GED;
PROC MIXED Data= wages_Y2  Method=ml noclprint noinfo COVTEST;
  class id;
  model LNW_Y =  Year_1 Year_1*Year_1 year_2 year_2*year_2
        black  black*Year_1
        black* Year_2  HGC_9 UERATE_C / solution notest outpm = mc;
  random intercept Year_1 year_2 / subject=id type=un;
RUN;
```

The scripts begin by calling the MIXED procedure and wages_Y2 dataset, followed by specifying the fixed effect part of PGCM using MODEL statement and random effect part using RANDOM statement.

We then use the NLMIXED procedure to fit Model B again to make sure we can get the same results with two different SAS procedures. In PROC NLMIXED, we begin by specify the starting values for the estimation of all unknown parameters using the PARMS statement. Starting values can be difficult to generate, but are important for obtaining convergence within a reasonable time. We obtained the starting values by using results from fitting Model B with the MIXED procedure. We also tried different number of quadrature points used in the estimation (we chose QPOINTS = **20**). The next few lines of the script define the

outcome variable, its distribution, and how the random effects should be included in the model.

For an unstructured covariance matrix for random effects, we use Cholesky-root reparameterization. That is, if Φ is a *p x p* positive definite matrix, you can find an upper triangular matrix T such that Φ =T′T, so that is a type of square root of Φ. For a 3 x 3 matrix, it can be shown the following:

$$T' = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix},$$

$$T'T = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & t_{31} \\ 0 & t_{22} & t_{32} \\ 0 & 0 & t_{33} \end{bmatrix} = \begin{bmatrix} t_{11}^2 & t_{11}t_{21} & t_{11}t_{31} \\ t_{11}t_{21} & t_{21}^2 + t_{22}^2 & t_{21}t_{31} + t_{22}t_{32} \\ t_{11}t_{31} & t_{21}t_{31} + t_{22}t_{32} 0 & t_{31}^2 + t_{32}^2 + t_{33}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \Phi_{11} & & \\ \Phi_{21} & \Phi_{22} & \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} = \Phi$$

We re-parameterized the unstructured matrix with the above Cholesky root. Based on the mathematical relationship between the UN and CHOL structures, we then compute the variances and the covariance in the UN structure accordingly. Cholesky-root re-parameterizations generally have better numerical behaviors than the UN structure, and are useful when the program failed to converge or was experiencing a long run time.

```
* Model B: Duplicate above analyses, but using NLMIXED ;
PROC NLMIXED DATA = Wages_Y2 MAXITER = 2000 NOAD QPOINTS = 20 GCONV = 0;
   * Initial values;
   PARMS G00 = 1.70 G10 = 0.06 G20 = -0.003  G30 = 0.03 G40 = -0.003
             G01 = 0  G11 =  -0.02   G31 = 0
             G50 = 0.04 G60 = -0.01
             s2e= 0.08
             t11 = 0.7
             t21 = 0 t22 = 0.01
             t31 = 0 t32 = 0 t33 = 0.01
              ;
   * Define the piecewise mean trajectory;
         y_pred =   (G00+ G01*black + u0) + (G10 +G11*black + u1)*year_1
                     + G20 * year_1* year_1
                     + G50* HGC_9 + G60* UERATE_C
                     + (G30 + G31*black + u2)* year_2
                     + G40*year_2*year_2    ;
         MODEL LNW_Y ~ normal(y_pred,s2e);
   * Cholesky decomposition of the covariance matrix of random effects;
   phi11 = t11*t11;
   phi21 = t21*t11;
   phi22 = t21*t21 + t22*t22;
   phi31 = t31*t11;
   phi32 = t31*t21 + t32*t22;
   phi33 = t31*t31 + t32*t32 + t33*t33;

   RANDOM u0 u1 u2 ~ normal([0, 0, 0],[phi11,
                                       phi21, phi22,
                                       phi31, phi32, phi33] ) subject=id;
```

```
*Recover parameters of the covariance matrix of random effects;
            ESTIMATE 'phi11' t11*t11;
            ESTIMATE 'phi21' t21*t11;
            ESTIMATE 'phi22' t21*t21 + t22*t22;
            ESTIMATE 'phi31' t31*t11;
            ESTIMATE 'phi32' t31*t21 + t32*t22;
            ESTIMATE 'phi33' t31*t31 + t32*t32 + t33*t33;

   RUN;
```

Adjusting the above script for Model B with estimation of change point as a function of another time-varying predictor (Model C) is straight. We assume the simple function form and the function form is known. We added a parameter (gamma) to denote the individual change point as a function of the time the dropout obtained the GED. This non-linear model cannot be specified with the MIXED procedure, but can be specified with the NLMIXED.

```
   *  Model C: Model B + Estimation of unknown change points as a function of time at
   which GED was obtained;
   PROC NLMIXED DATA = Wages_Y2 MAXITER = 2000 NOAD QPOINTS = 20 GCONV = 0;
      * Initial values;
      PARMS G00 = 1.70 G10 = 0.06 G20 = -0.003  G30 = 0.03 G40 = -0.003
                  G01 = 0  G11 =  -0.02   G31 = 0
                  G50 = 0.04 G60 = -0.01
                  s2e= 0.08
                  t11 = 0.22
                  t21 = 0 t22 = 0.04
                  t31 = -0.01 t32 = -0.01 t33 = 0.04
                  gamma = 0.8;
      * Gammai denotes the random knot;
                  Gammai = Gamma ;
      * Define the piecewise mean trajectory;
                  y_pred =   (G00+ G01*black + u0)
                              + (G10 +G11*black + u1)*year_1
                              + G20 * year_1* year_1
                              + G50* HGC_9 + G60* UERATE_C;

                  IF (year_2 > Gammai)   THEN DO;
                   y_pred =   (G00+ G01*black + u0)
                              + (G10 +G11*black + u1)*year_1
                              + G20 * year_1* year_1
                              + G50* HGC_9 + G60* UERATE_C
                              + (G30 + G31*black + u2)* (year_2 - gammai)
                              + G40 * (year_2 - gammai) * (year_2 - gammai)
                        ;
                                                  END;

            MODEL LNW_Y ~ normal(y_pred,s2e);
      * Cholesky decomposition of the covariance matrix of random effects;
      phi11 = t11*t11;
      phi21 = t21*t11;
      phi22 = t21*t21 + t22*t22;
      phi31 = t31*t11;
      phi32 = t31*t21 + t32*t22;
      phi33 = t31*t31 + t32*t32 + t33*t33;

      RANDOM u0 u1 u2 ~ normal([0, 0, 0],[phi11,phi21, phi22, phi31, phi32,
phi33] ) subject=id;
```

```
*Recover parameters of the covariance matrix of random effects;
            ESTIMATE 'phi11' t11*t11;
            ESTIMATE 'phi21' t21*t11;
            ESTIMATE 'phi22' t21*t21 + t22*t22;
            ESTIMATE 'phi31' t31*t11;
            ESTIMATE 'phi32' t31*t21 + t32*t22;
            ESTIMATE 'phi33' t31*t31 + t32*t32 + t33*t33;

    RUN;
```

In Model D, we follow the similar specification as in Singer and Willett (2003): log wage is a linear function of potential labor market experience. The specification of Model D is as below

$$\begin{aligned}\ln(w_{ij}) &= \beta_{00} + \beta_{10}\text{Year\_1}_{ij} + \beta_{20}\text{Year\_2}_{ij} \\ &+ \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} + \beta_{31}\text{Black}_i * \text{Year\_2}_{ij} \\ &+ \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varsigma_{1i}\text{Year\_1}_{ij} + \varsigma_{2i}\text{Year\_2}_{ij} + \varepsilon_{ij}\end{aligned} \qquad \text{(Model D)}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad \begin{bmatrix}\varsigma_{0i} \\ \varsigma_{1i} \\ \varsigma_{2i}\end{bmatrix} \sim MVN\left(\begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}, \begin{bmatrix}\Phi_{11} & & \\ \Phi_{21} & \Phi_{22} & \\ \Phi_{31} & \Phi_{32} & \Phi_{33}\end{bmatrix}\right),$$

Similarly, in Model E, we extend the Model D by including the estimation of individual change point of his wage trajectory as a function of the time of his receipt of GED (some delay).

$$\begin{aligned}&\text{If } \text{Year\_2} \le \gamma \\ &\ln(w_{ij}) = \beta_{00} + \beta_{10}\text{Year\_1}_{ij} \\ &\quad + \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} \\ &\quad + \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varsigma_{1i}\text{Year\_1}_{ij} + \varepsilon_{ij} \qquad \text{(Model E)} \\ &\text{If } \text{Year\_2} > \gamma \\ &\ln(w_{ij}) = \beta_{00} + \beta_{10}\text{Year\_1}_{ij} + \beta_{30}(\text{Year\_2}_{ij} - \gamma) \\ &\quad + \beta_{01}\text{Black}_i + \beta_{11}\text{Black}_i * \text{Year\_1}_{ij} + \beta_{31}\text{Black}_i * (\text{Year\_2}_{ij} - \gamma) \\ &\quad + \beta_{50}\text{HGC\_9}_i + \beta_{60}\text{UERate\_C}_{ij} + \varsigma_{0i} + \varsigma_{1i}\text{Year\_1}_{ij} + \varsigma_{2i}\text{Year\_2}_{ij} + \varepsilon_{ij}\end{aligned}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad \begin{bmatrix}\varsigma_{0i} \\ \varsigma_{1i} \\ \varsigma_{2i}\end{bmatrix} \sim MVN\left(\begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}, \begin{bmatrix}\Phi_{11} & & \\ \Phi_{21} & \Phi_{22} & \\ \Phi_{31} & \Phi_{32} & \Phi_{33}\end{bmatrix}\right),$$

The SAS script for the Model D and E are similar as Model B and C, which is contained in Appendix A.

RESULTS

Table 2 displays estimated parameters (and standard errors) and goodness of fit from Model A to C and Table 3 for Model D and E.

Results of Model A indicate that a White male who dropped out school in 9th Grade and who lives in a community with an unemployment rate at average level (about 7%) is predicted to earn $5.67 (exp(1.736)) per hour (in 1990 dollars) upon entry into the labor market after dropping out of school. Before GED attainment, he experiences a real wage increase of almost 6.2% during the first year after leaving school and the increase rate will decrease slightly over years (negative slope of quadratic term). The linear rate of subsequent wage growth of GED White recipients increased by approximately 2.3% over the predicted rate in the absence of the credential. At work force entry, there was no racial differences in wages, but racial disparities increase over time because wages for Black dropouts increase at a slower rate. There was no race difference in the wage benefit from GED attainment.

Results from Model B in Table 2 indicate the impact of the GED and race differences in the impact of the GED tell basically the same story as those from Model A. From Model B, we also found there were significant individual variations among dropouts in the initial wages and annual increase before GED receipt. The variation in the impact GED receipt is not statistically significant or only marginally significant.

Results from Model C in Table 2 indicate the impact of the GED on wage trajectories is not taking effect right at the time when dropout boys obtained their GEDs. On average, about one year after the GED receipt, the linear rate of wage growth increased by approximately 3.9% over the predicted rate in the absence of the credential.

	Model A	Model B	Model C
Fixed Effects			
Intercept	1.736(0.014)***	1.734(0.014)***	1.737(0.014)***
Years Since Dropout	0.062(0.005)***	0.064(0.006)***	0.064(0.006)***
(Years Since Dropout)*(Years Since Dropout)	-0.002(0.0005)***	-0.003(0.0006)***	-0.003(0.0006)***
Knot - GED			1.000(0.002)***
Years Since GED or Knot	0.023(0.009)**	0.033(0.009)***	0.038(0.001)***
(Years Since GED or Knot) * (Years Since GED or Knot)	-0.002(0.001)+	-0.003(0.001)**	-0.004(0.002)**
Black	0.002(0.024)	0.000(0.023)	0.002(0.023)
Black* (Years Since Dropout)	-0.025(0.005)***	-0.024(0.006)***	-0.023(0.006)***
Black*(Years Since GED or Knot)	0.011(0.007)	0.008(0.011)	0.005(0.013)
Grade at Dropout	0.042(0.007)***	0.038(0.006)***	0.038(0.006)***
Unemployment Rate	-0.012(0.002)***	-0.012(0.001)***	-0.012(0.002)***
Variance Components			
Level-1	0.087(0.002)***	0.075(0.002)***	0.075(0.002)***
Level-2: Intercept	0.053(0.003)***	0.049(0.004)***	0.048(0.004)***
Years Since Dropout		0.001(0.0002)***	0.001(0.0002)**
Years Since GED or Knot		0.0016(0.002)	0.003(0.002+
Intercept & (Years Since Dropout)		-0.001(0.001)	-0.001(0.001)
Intercept & (Years Since GED or Knot)		-0.003(0.0014)+	-0.004(0.002)*

(Years Since Dropout) & (Years Since GED or Knot)		-0.000(0.001)	-0.001(0.001)
Goodness-of-fit			
Deviance statistics	3331.2	3100.7	3096.9
AIC	3355.2	3134.7	3132.9
BIC	3412.7	3216.1	3219.1

Note: GED = General Educational Development. + p <.10, * p <.05, ** p <.01, *** p <.001.

Table 2. Parameter Estimates (standard errors) from Fitted PGCV Predicting Log of Hourly Wages (N=888): Murnane et al. (1999) vs. Our Model

Table 3 reports the results from Model D and E. The impact of the GED and race differences in the impact of the GED tell basically the same story as in Table 2. Similarly, Table 3 also indicates the impact of the GED on wage trajectories is not taking effect right at the GED receipt. On average, about one year after the GED receipt, the linear rate of log wage trajectory increased by approximately 3.9% over the predicted rate in the absence of the credential.

	Model D	Model E
Fixed Effects		
Intercept	1.769(0.012)***	1.770(0.012)***
Years Since Dropout	0.041(0.003)***	0.041(0.003)***
Knot - GED		1.000(0.020)***
(Years Since GED or Knot)	0.013(0.006)*	0.014(0.007)*
Black	-0.009(0.023)	0.008(0.023)
Black* (Years Since Dropout)	-0.022(0.006)***	-0.021(0.006)**
Black* (Years Since GED or Knot)	0.010(0.011)	0.006(0.013)
Grade at Dropout	0.039(0.006)***	0.039(0.007)***
Unemployment Rate	-0.013(0.002)***	-0.013(0.002)***
Variance Components		
Level-1	0.076(0.002)***	0.076(0.002)***
Level-2: Intercept	0.050(0.005)***	0.049(0.004)***
Years Since Dropout	0.0014(0.0002)***	0.001(0.0002)***
Years Since GED or Knot	0.0014(0.002)	0.003(0.002)+
Intercept & (Years Since Dropout)	-0.002(0.001)+	-0.001(0.001)
Intercept & (Years Since GED or Knot)	-0.003(0.0014)*	-0.004(0.002)*
(Years Since Dropout) & (Years Since GED or Knot)	-0.000(0.001)	-0.001(0.001)
Goodness-of-fit		
Deviance statistics	3142.2	3139.8
AIC	3172.2	3171.8
BIC	3244.0	3248.4

Note: GED = General Educational Development. + p <.10, * p <.05, ** p <.01, *** p <.001.

Table 3. Parameter Estimates (standard errors) from Fitted PGCV Predicting Log of Hourly Wages (N=888): Singer et al. (2005) vs. Our Model

Table 2 and 3 also indicate that the effects of two other substantive predictors − highest grade completed and local area unemployment rate − remains similar across all five models. Dropouts who stay in school longer earn higher wages on labor force entry and each more year longer is associated with a 4.2% higher in wage. Results from these models also indicate that a 1% increase in the local unemployment rate is associated with a 1.2% decrease in wage.

Goodness of fit index reported in Table 2 shows that PGCM with multiple random effects can substantially improve fit (Model B vs. A). The deviance statistics test suggested that our proposed model (Model C) is significantly better than Murnane and colleagues's model (Model A) and Model B. Results in Table 3 shows that our proposed Model E is better than Singer and Willett's Model D. The comparison of AIC statistics also suggest that our models are better than their models. Though BIC giving the nod to Model D over E, they are roughly comparable

CONCLUSION AND DISCUSSION

In both health and behavioral sciences, researchers are quite often interested in not only the intra-individual change over time but also the inter-individual differences in these changes. The models of nonlinear change such as piece-wise growth curve model (PGCM) might be required to provide more accurate, complete, and easily interpretable description of how individual change over time and inter-individual differences in such change.

The major challenge in applying the PGCM is to specify the change points. This paper expanded on previous literature on change points in PGCM. We provide models to specify the individual change point as a function of a time-varying predictor and illustrate these models using an empirical example to demonstrate how to estimate these models in SAS and its impact on other features that describe the growth profiles.

We reanalyze data published by Murnane et al. (1999) related to the benefit of obtaining GED for male dropouts and hypothesize that individual change point of his wage trajectory is a function of the time of his receipt of GED. Our results demonstrate that about one year after the GED receipt, male dropouts' wage profile shifted from first phase linear or non-linear growth to second phase growth with significantly higher growth rate. The delay in the increased wages resulting from GED acquisition is consistent with Murnane et al.'s justification of economic benefits from obtaining a GED. The GED recipients use the credential to gain access to a significant amount of postsecondary education or training earn wages that are considerably higher than those GED recipients who do not. A large percentage of GED recipients avail themselves of improved access to postsecondary education and training that the indirect effects of the credential on subsequent wages are a substantial part of the total effect for the average GED recipient.

Though the specification of PGCM with individual change points as a function of the time of GED acquisition has minimal impact on the growth rates before the change points in our empirical example, the magnitude of growth rates after the change points and its variation across the subjects are different from the conventional PGCM with change point specified a priori. This lead to a more optimal descriptions of the growth pattern in the wage data and provide the further empirical support for Cameron and Heckman (1993) argument about that to the extent school dropouts derive labor market economic benefits from obtaining a GED, the benefits come primarily through the mechanism of improving access to postsecondary training and education.

Currently, there are only a few program, in addition to NLMIXED, that can be used to fit PGCM with individual specific change point and with multiple random effects. Though some R packages (e.g., FitMM and BayesianPGMM) were developed to estimate the unknown change point and some SEM-based software can be used to fit PGCM with random change points, most of them requires some degree of balance in measurement schedules and it

cannot predict the change point using time-varying predictors. As we have illustrated, PROC NLMIXED is a useful procedure to fit this kind of PGCM. It allows us to fit different shape of non-linear growth and models with more than one random effect or even with multiplicative random effects. Additionally, NLMIXED allows for flexibility in timing basis and allow individual have a unique time values at each assessment.

The estimation of PGCM with unknown change points, in general, to be more challenging with respect to model convergence as well as the accuracy and precision of estimated model parameters, irrespective of the estimation approach used. It is recommended that careful consideration of practical and substantive implication should be made before incorporating any complexity into these models.

PGCM with individual change point as a function of other factors provide a flexible framework for researchers and practitioners that allow them to characterize individual pathways that exhibits distinct phases of development. This is very useful for practitioners who are seeking to know when the treatment takes effect and measure the effectiveness of treatment or intervention. It can also be used for healthcare professional and police makers who want to know when individual may need to seek professional services for mental or health disability.

REFERENCES

Cameron, S. V., and J. J. Heckman. 1993. "The nonequivalence of high school equivalents". *Journal of Labor Economics* 11:1-47.

Cudeck, R. (1996). "Mixed-effects models in the study of individual differences with repeated measures data". *Multivariate Behavioral Research, 31,* 371– 403.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. Psychological Methods, 7, 41–63.

Dominicus, A., Ripatti, S., Pedersen, N. L., & Palmgren, J. (2008). "A random change point model for assessing variability in repeated measures of cognitive function". *Statistics in Medicine, 27,* 5786 –5798.

Kohli, N. (2011). "*Estimating unknown knots in piecewise linear-linear latent growth mixture models"* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Kohli, N., Harring, J. R., & Zopluoglu, C. (2016). "A finite mixture of nonlinear random coefficient models for continuous repeated measures data". Psychometrika, 81, 851–880.

Kreisman, M. B. (2003). "Evaluating academic outcomes of Head Start: An application of general growth mixture modeling". *Early Childhood Research Quarterly, 18*, 238–254.

Kwok O, Luo W, West SG. (2010). "Using modification indexes to detect turning points in longitudinal data: a monte carlo study". *Structure Equation Modeling.* (2010). **17**:216–40.

Li, F., Duncan, T. E., Duncan, S. C., & Hops, H. (2001). "Piecewise growth mixture modeling of adolescent alcohol use data". *Structural Equation Modeling, 8,* 175–204.

McArdle, J. J., & Wang, L. (2008). "Modeling age-based turning points in longitudinal life-span growth curves of cognition". In P. Cohen (Ed.), *Applied data analytic techniques for turning points research* (pp. 105– 127). New York, NY: Routledge.

Muniz Terrera, G., van den Hout, A., & Matthews, F. E. (2011). "Random change point models: Investigating cognitive decline in the presence of missing data". *Journal of Applied Statistics, 38,* 705–716.

Murnane, R.J. , Willett, J.B. , & Boudett, K.P. (1999). Do male dropouts benefit from obtaining a GED, postsecondary education, and training? Education Review, 22, 475-502.

Ning L., & Liu, W. (2017). Specifying turing point in piecewise growth curve models: Challenges and solutions. Frontiers in Applied Mathematics and Statistics, 3, 1-15.

Paris, S. G. (2005). "Reinterpreting the development of reading skills". *Reading Research Quarterly, 40*, 184–202.

Preacher, K. J., & Hancock, G. R. (2012). On interpretable reparameterizations of linear and nonlinear latent growth curve models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 25–58). Charlotte, NC: Information Age Publishing.

Preacher, K. J., & Hancock, G. R. (2015). Meaningful aspects of change as novel random coefficients: A general method for reparameterizing longitudinal models. *Psychological Methods, 20(1),* 84-101.

Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2012). "Fluency has a role in the simple view of reading". *Scientific Studies of Reading, 16*, 1–26.

Singer J.D. & Willett J.B. (2003). "Applied Longitudinal Data Analysis". New York: Oxford University Press.

Wang, L., & McArdle, J. J. (2008). A simulation study comparison of Bayesian estimation with conventional methods for estimating unknown change points. Structural Equation Modeling, 15, 52–74.

Wood PK, Jackson KM. (2013). "Escaping the snare of chronological growth and launching a free curve alternative: general deviance as latent growth model". *Developmental Psychopathology. **25***:739–54.

ACKNOWLEDGMENTS

Preparation of this paper was supported in part by a grant from the Research Manitoba Applied Health Services Program, Canadian Institute for Health Research (CIHR) Project Grant, and Children's Hospital Research Institute of Manitoba (CHRIM) Foundation Operating Grant.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Depeng Jiang
University of Manitoba, Winnipeg, Canada
Depeng.Jiang@umanitoba.ca