Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/090312

Piecewise Baseline Hazard Model with Gamma Frailty: Analysing the Transitions into The Labour Force Entry by The Youths in India

Tapan Kumar Chakrabarty and Jayanta Deb[*]

Department of Statistics, North-Eastern Hill University, Shillong, Meghalaya, India

Received: 29 Jun. 2019, Revised: 27 Nov. 2019, Accepted: 29 Nov. 2019
Published online: 1 Nov. 2020

Abstract: The present paper demonstrates piecewise constant baseline hazard model with shared frailty for analysing the timing of entry into workforce after schooling that are clustered into geographical domain. Observations from the same cluster are usually correlated because, unknowingly, they share certain unobserved characteristics. Including these within cluster correlations in the model allows correctly measuring the covariate effects and avoiding underestimation or overestimation of the parameters of interest. Besides analysing the effect of substantial demographic and socio-economic circumstances on the time to entry into the workforce for geographically clustered event history data, comparison of estimates obtained from the Cox regression model, the Cox regression model with shared frailty, and the piecewise constant baseline hazard with shared frailty are also outlined here.

Keywords: Cox regression, baseline hazard, frailty, gamma distribution, log-likelihood

1 Introduction

Entry into workforce after schooling is an important social transition, where the timing of transition plays a crucial role of an individual's life. Both social activities are closely connected to the family background and are interdependent, in such a way that, under normal conditions, economic independence follows job stability which is linked to leaving education and marriage [1]. There are several observable [2] and unobservable [3] predictors which influence the transition. Examination of those predictors that influence the distribution of timing of transition to work with respect to specific age intervals has an important issue to study. In Indian scenario, age at marriage of youth usually get affected by the timing of entry into workforce, hence a longer timing of entry into work imply a higher age at marriage as well as at parenthood. Moreover, average age at entry into work can affect a country's employment structure, socio-economic development as well as the quality of life. The socio-economic and demographic factors that are influencing the timing of entry into work can provide comprehensive knowledge about the employment rate of the country. Analysis of those factors affecting the span of entry into work has proven useful since in many cases they appear to vary substantially across the states in India.

Virtually in all fields of research, regression modelling has been used long since to estimate the relationship between an outcome variable and independent predictor variables. It is a well-known approach because of the fact that, biologically plausible models can be fitted easily, as well as can be easily evaluated and interpreted. Statistically, a systematic and an error component must be specified in a model to estimate the effect of various systematic factors on the response or outcome variable. When the response is a dichotomous variable, a logistic regression is well defined through an odds ratio. Most of the methods in statistical analysis do not include the length of time as a variable. After completion of the study period, studies are made only based on the trail outcomes. The method which accounts for the length of time in the response variable is known as cox proportional hazard model. It emphasizes on yielding different interpretations at different points in time. When individuals are followed over a certain time period, the inherent aging process distinguishes survival time from the other dependent variables. The instantaneous risk of an event occurrence is known as the hazard function and those hazard functions are multiplicatively related. Cox model may be defined as a multiple linear regression of the logarithm of the hazard on the variables with the baseline hazard as intercept term. The effect of covariates is to

[*] Corresponding author e-mail: jdeb888@gmail.com

multiply the hazard function by a function of the explanatory variables. Through the exponentiated coefficient, this model interprets the effects of those covariates on the outcome variable, and is broadly known as hazard ratio. In this way, Cox regression model takes care about the observable covariates.

Most demographic surveys in India collect data that are clustered according to geographical regions due to the sampling design. Individuals in the same cluster (state) usually share certain unobservable characteristics and as a result, duration of transition to work of the same cluster tend to be correlated.

Timing of transition to work can be considered as an event history data as it gives us a longitudinal record of when transition to work occurred for an individual or a group of individuals. Timing is defined as the first age at entry into work after leaving education. Considering exit from study but not entered into workforce as censored and entry into workforce as failure or an event, Cox proportional hazards model can be used to identify important observable factors for timing of transition to work, provided the timings are independent. The random effect models for event history data which are known as frailty models can be used to analyse correlated durations of entry into work for an individual level [4]. A shared frailty model which is known as conditional independence model can be used where the random effect is common to all subjects in a cluster (state) [5]. But only the shared frailty model is not a good idea in some situations where the effects of selected covariates on subject frailty may not be the same for the whole period of time [6, 7, 8]. In that situation a partition in the time axis with adequate intervals having accurate cut points are needed. In our case, we expect the rate of transition to workforce is higher at the median age of entry into work for male and female respectively, where beyond that age the expected rate should show a lower value. In order to accommodate for this expected pattern in the transition rate, time is split into two intervals for both male & female respectively. And here comes the scenario for applying the piecewise constant baseline hazard with random effect shared by the cluster (state).

There are numerous theoretical developments on frailty models that have appeared in demographic literature. Frailty models are the extension of Cox PH models [9], the aim of which is to account for unobserved heterogeneity in survival or time to event data [10], where random frailty have a multiplicative effect on baseline hazard function [11]. The univariate frailty models suggested by [12] was significantly adopted by its application to multivariate survival data in a paper by [13] on Chronic disease incidence. An application of random effect model for analysing birth interval in Bangladeshi women was used by Mahmood, Zainab and Latif [14]. Moreover, piecewise exponential models have been used to assess the influence of job-specific experience on the hazard of acute injury for hourly factory workers [8]. Bayesian estimation was used to develop several machine learning models [16, 17]. The paper of Sari, Thamrin and Lawi [7] examines Bayesian estimation of piecewise exponential frailty models for multivariate survival data. Estimation of the parameters was done using the piecewise exponential frailty model with changing prior for baseline hazard function [6]. Recently, Singh, Singh, Bharti and Singh [15] used frailty model approach to assess the survival of children below 5 years of age in EAG states & Assam. As long as the days going on, the use of this model is constantly increasing.

In this paper, the main objective is to select an appropriate model for analysing duration of transition to work, which will be able to identify important factors associated with the timings of entry into work for the youth in India. For simplicity, we have considered gamma distribution for frailty. The log-likelihood is used to compare the performance of Cox proportional hazard, Cox proportional hazard with shared frailty and piecewise constant baseline hazard with shared frailty model. In section 2 we give an introduction to the dataset and methods used, section 3 provides the descriptive analysis of the dataset to check the unobserved heterogeneity, and section 4, section 5 & section 6 develop the covariates selection procedure results from the analysis and concluding remarks.

2 Materials and Methods

2.1 Data

In this article, we use a dataset from a survey entitled " The Youth in India: Situation and Needs 2006-2007" conducted by the International Institute for Population Sciences, Mumbai and the Population Council, New Delhi [18] is the first-ever sub nationally representative study that provides data on young people's transition to various adulthood events. Research has been conducted in a total of six states of India namely: Andhra Pradesh, Bihar, Jharkhand, Maharashtra, Rajasthan and Tamil Nadu and these six states are purposively selected to represent the different geographic and sociocultural regions within the country and together they represent two-fifths of the country's population. It provides a wealth of evidence on married and unmarried young women and men (aged 15-24 & 15-29) from both rural and urban settings of each state. The surveys are undertaken in a phased manner and took place between January 2006 and April 2008. In all, 58,728 young people are contacted, of which a total of 50,848 married and unmarried young women and men were successfully interviewed. Using the information on time to entry into work data from Youth Study, the present article has made an attempt to identify potential covariates accounting for such transition.

Since the data have been collected from different geographical regions or states in India, there is a high probability of unobserved heterogeneity to be present in the data. The people staying in the state of Rajasthan will share similar kind of environment, culture, verbal & visual thinking, style of living, etc., which are different from the people staying in the state of Tamil Nadu. This kind of heterogeneity influences the people's social transitions. Evidence of the existence of unobserved heterogeneity is strongly visible from the dataset, and has been analysed briefly through density plot, Theil?s Index, Kaplan-Meier failure estimates, and proportions, respectively in subsection 3.1, subsection 3.3, subsection 3.4 & subsection 3.5, which gives an indication of adding a frailty model during the analysis. Moreover, it is well-known that, the baseline hazard function is not constant over the whole period of time as the probability of entering into work force is lower at the beginning ages (ex: 12-15), highest at a certain year of age (ex: 16), and then gradually decreases as the time moves on. This pattern is evident from density plots, and median ages, respectively in subsection 3.1 and subsection 3.2. Overall median survival is calculated for men and women separately and found that 50% of the total males have experienced the entry into labour force at the age of 16, while 50% of the total females have experienced the entry into labour force at the age of 23 (briefly discussed in subsection 3.2). These changing patterns suggest us to consider a piecewise baseline hazard model for capturing the effects of covariates more efficiently.

2.2 Shared Frailty Model

The generalization of the Cox proportional hazards model [9] is the best to assess the observable covariate effects on the hazard function and widely applied model that allows for the random effect by multiplicatively adjusting the baseline hazard function. Including frailties in the model allows correctly measuring the covariate effects and avoiding underestimation or overestimation of the parameters.

Let t represent the duration of transition to work with respect to j^{th} individual and Z_j be a covariate or a covariate vector with respect to j^{th} individual. Then the hazard function conditional on both covariates and individual random effect can be written as

$$h_j(t) = h_0(t) u_j e^{\beta^t Z_j}; j = 1, 2 \ldots, n \tag{1}$$

Where $h_0(t)$ is an arbitrary baseline hazard rate, β is a parameter, and Z is a covariate with respect to j^{th} observation.

In this study, there are several covariates, namely place of residence, caste, total number of brothers & sisters, work status of father and mother, class last attended, father's education and type of school last attended which are discussed briefly in section 6. Moreover, the durations of transition to work, which are obtained from a sub-nationally conducted survey, are assumed to be correlated because duration of transition for individuals from the same cluster (state) are assumed to be more alike compared to that from different clusters (states). It is assumed that the correlations are due to unobservable cluster-specific covariates. To adjust state-level heterogeneity, clusters (states) are considered as random for the following frailty model which is again conditional on both covariates and random cluster effect (frailty). Suppose we have j individuals and i clusters (states). Each cluster consists of n_i individuals and $\sum_{i=1}^{G} n_i = n$, where n is the total sample size. Then the hazard rate for the j^{th} individual in the i^{th} cluster (state) conditional on both the covariates and random cluster (state) effect is given by

$$h_{ij}(t) = h_0(t) u_i e^{\beta^t Z_{ij}}; i = 1, 2 \ldots, G; j = 1, 2 \ldots, n_i \tag{2}$$

It assumes that, given the random effect or frailty, all event times in a cluster are independent. This model (2) is known as shared frailty model because all the individuals in a specific cluster (state) share the same frailty, i.e. each cluster is represented by one frailty and frailty is common to all cluster members. Shared frailty model was introduced by Clayton [13] without using the notion frailty and extensively studied in Hougaard [19], Therneau and Grambsch [20], Duchateau, Janssen, Lindsey, Legrand, Nguti and Sylvester [21], Duchateau, Janssen, Kezic and Fortpied [22], and Duchateau and Janssen [23].

2.3 Piecewise Constant Baseline Hazard with Shared Frailty

The first and common approach to define the hazard function as

$$h_{ij}(t|u_i) = h_0(t) u_i e^{\beta^t Z_{ij}}; i = 1, 2 \ldots, G; j = 1, 2 \ldots, n_i$$

which is the hazard function of the j^{th} individual of cluster (state) i, given the frailty of cluster (state) i (u_i), where $h_0(t)$ is an arbitrary baseline hazard rate and Z_{ij} is the corresponding covariate vector. The frailty U is supposed to follow a gamma distribution g(u; θ, θ). To account for piecewise structure, we divide the time (timing of transition) into some pre-specified intervals $I_k(y_{k-1}, y_k)$ for $k = 1, 2 \ldots, g$, where $0 = y_0 < y_1 < y_2 < \ldots\ldots < y_g < \infty$, y_g being the last survival or censored time, and assume the baseline hazard to be constant within time intervals. The model was first introduced by Breslow [24] who used distinct failure times as end points of intervals.

Kalbfleisch and Prentice [25] suggested that the selection of the grid { y_0, y_1 ,y_2 ,.....y_g } should be made independent of the data. Since the baseline hazard is not constant, a modified model is written as:

$$\begin{aligned} h_{ij}(t) &= (h_1 I(y_0 < t \leqslant y_1) + h_2 I(y_1 < t \leqslant y_2) + \ldots\ldots + h_g I(y_{g-1} < t \leqslant y_g)) u_i e^{\beta' Z_{ij}} \\ &= \sum_{k=1}^{g} h_k I(y_{k-1} < t \leqslant y_k) u_i e^{\beta' Z_{ij}} \\ &; i = 1,2\ldots,G; j = 1,2\ldots,n_i; k = 1,2\ldots,g \end{aligned} \tag{3}$$

2.4 Likelihood Specification

The likelihood function for survival data is given by

$$L = \prod_{j=1}^{n} [(1-G_j(t)) f_j(t)]^{\delta_j} \; [(1-F_j(t)) g_j(t)]^{1-\delta_j}$$

where δ_j is the censoring indicator, g and G are the density function and the cumulative distribution function of the censoring time, f and F are the density function and the cumulative distribution function of the event time, respectively. Ignoring the distribution of censoring times since it doesn't depend on the parameter of interest, the likelihood function [26] for ith cluster (state) is of the form

$$\begin{aligned} Ł_i &= \prod_{j=1}^{n_i} \left((f_{ij}(t))^{\delta_{ij}} (S_{ij}(t))^{1-\delta_{ij}} \right) \\ &= \prod_{j=1}^{n_i} (h_{ij}(t))^{\delta_{ij}} (S_{ij}(t)) \qquad \text{Since, } h_{ij}(t) = \frac{f_{ij}(t)}{S_{ij}(t)} \end{aligned} \tag{4}$$

Cox proportional hazards model for frailties is given by

$$h_{ij}(t) = h_0(t) u_i e^{\beta' Z_{ij}} \implies \frac{f_{ij}(t)}{S_{ij}(t)} = h_0(t) u_i e^{\beta' Z_{ij}} \tag{5}$$

where u_i's are independent and identically distributed random sample from a distribution with mean of 1 and some unknown variance of θ.

Integrating both sides of (5), we can get the expression for the survival function.

$$\int_{0}^{\infty} \frac{f_{ij}(t)}{S_{ij}(t)} dt = \int_{0}^{\infty} h_0(t) u_i e^{\beta' Z_{ij}} dt$$

$$\implies \quad S_{ij}(t) = exp(-H_0(t) u_i e^{\beta' Z_{ij}}) \tag{6}$$

Now, for i^{th} subgroup, the conditional and marginal likelihood are given by

$$L_i(\beta | u_i) = \prod_{j=1}^{n_i} (h_0(t) u_i e^{\beta' Z_{ij}})^{\delta_{ij}} exp^{-H_0(t) u_i e^{\beta' Z_{ij}}} \tag{7}$$

$$L_i(\theta,\beta) = \prod_{j=1}^{n_i} \int_{0}^{\infty} (h_0(t) u e^{\beta' Z_{ij}})^{\delta_{ij}} exp^{-H_0(t) u e^{\beta' Z_{ij}}} g(u) du \tag{8}$$

where g(u) is the probability distribution function of frailties u_1, ..., u_G.

To obtain the marginal log-likelihood for the gamma frailty model, let u_i be independent and identically distributed sample of gamma random variables with density function

$$g(u) = \frac{u^{\frac{1}{\theta}-1} e^{-\frac{u}{\theta}}}{\Gamma \frac{1}{\theta} \theta^{\frac{1}{\theta}}} \qquad u > 0, \theta > 0$$

with E(U) = 1 and Var(U) = θ. Larger values of θ indicate that there is a higher degree of heterogeneity among states and strong association within states. The marginal likelihood function for the ith group is given by

$$L_i(\theta,\beta) = \prod_{j=1}^{n_i} \int_0^\infty (h_0(t)ue^{\beta^t Z_{ij}})^{\delta_{ij}} exp^{-H_0(t)ue^{\beta^t Z_{ij}}} \frac{u^{\frac{1}{\theta}-1}e^{-\frac{u}{\theta}}}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}}du \tag{9}$$

Rearranging the terms in Equation (9), we obtain the following expression:

$$L_i(\theta,\beta) = \prod_{j=1}^{n_i} h_0(t)^{\delta_{ij}} e^{\beta^t Z_{ij}\delta_{ij}} \frac{\Gamma(\frac{1}{\theta}+d_i)\theta^{\frac{1}{\theta}+d_i}}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}} \int_0^\infty exp^{-u(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}})} \frac{u^{\frac{1}{\theta}+d_i-1}}{\Gamma(\frac{1}{\theta}+d_i)\theta^{\frac{1}{\theta}+d_i}}du$$

where $d_i = \sum_{i=1}^{n_i} \delta_{ij}$.

To make our problem tractable, we integrate out the frailty term u. The term under the integral is the moment generating function of a gamma distribution with a pdf $\Gamma(\frac{1}{\theta}+d_i, \frac{1}{\theta})$. Using this fact, we can derive the expression for marginal likelihood function as

$$L_i(\theta,\beta) = \prod_{j=1}^{n_i} h_0(t)^{\delta_{ij}} e^{\beta^t Z_{ij}\delta_{ij}}$$

$$\frac{\Gamma\left(\frac{1}{\theta}+d_i\right)\theta^{\frac{1}{\theta}+d_i}}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}\theta^{\frac{1}{\theta}+d_i}\left(\left[\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right]^{\frac{1}{\theta}+d_i}\right)} \int_0^\infty exp^{-u\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)}$$

$$\frac{\left[\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right]^{\frac{1}{\theta}+d_i}}{\Gamma(\frac{1}{\theta}+d_i)}du$$

$$L_i(\theta,\beta) = \prod_{j=1}^{n_i} h_0(t)^{\delta_{ij}} e^{\beta^t Z_{ij}\delta_{ij}}$$

$$\frac{\Gamma\left(\frac{1}{\theta}+d_i\right)}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}\left(\left[\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right]^{\frac{1}{\theta}+d_i}\right)} \int_0^\infty exp^{-u\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)}$$

$$\frac{\left[\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right]^{\frac{1}{\theta}+d_i}}{\Gamma(\frac{1}{\theta}+d_i)}du$$

It is easy to see that the term under the integral is the pdf of $\Gamma(\frac{1}{\theta}+d_i, \frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}})$, which integrates to 1. Therefore, the obtained marginal likelihood function is

$$L_i(\theta,\beta) = \frac{\Gamma\left(\frac{1}{\theta}+d_i\right)\prod_{j=1}^{n_i} h_0(t)^{\delta_{ij}} e^{\beta^t Z_{ij}\delta_{ij}}}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}\left(\left[\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right]^{\frac{1}{\theta}+d_i}\right)} \tag{10}$$

Taking the logarithm of this expression and summing over the i clusters, we obtain the marginal log-likelihood function, $l(\theta,\beta)$.

$$l(\theta,\beta) = \sum_{i=1}^{G}[d_i \log(\theta) - \log\left(\Gamma\frac{1}{\theta}\right) + \log\left(\Gamma\frac{1}{\theta}+d_i\right) - \left(\frac{1}{\theta}+d_i\right)\log\left(1+\theta\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right) + \sum_{j=1}^{n_i}\delta_{ij}(\beta^t Z_{ij}+\log(h_0(t)))] \tag{11}$$

By maximizing this log-likelihood function, we can obtain maximum likelihood estimators for θ and β. Incorporating the hazard and cumulative hazard for exponential model, the marginal log-likelihood function for gamma frailty with exponential baseline hazard rate is given by

$$l(\theta,\beta,\lambda)=\sum_{i=1}^{G}[d_i log(\theta)-log\left(\Gamma\frac{1}{\theta}\right)+log\left(\Gamma\frac{1}{\theta}+d_i\right)$$
$$-\left(\frac{1}{\theta}+d_i\right)log\left(1+\theta\sum_{j=1}^{n_i}\lambda t e^{\beta^t Z_{ij}}\right)+\sum_{j=1}^{n_i}\delta_{ij}(\beta^t Z_{ij}+log(\lambda))] \quad (12)$$

Let us consider the piecewise constant baseline hazard model for frailty as

$$h_{ij}(t)=\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)u_i e^{\beta^t Z_{ij}}$$
$$\Longrightarrow \frac{f_{ij}(t)}{S_{ij}(t)}=\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)u_i e^{\beta^t Z_{ij}} \quad (13)$$

Thus, integrating both sides of Equation (13), we can get the survival function as

$$\int_0^\infty \frac{f_{ij}(t)}{S_{ij}(t)}dt=\int_0^\infty\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)u_i e^{\beta^t Z_{ij}}dt$$
$$\Longrightarrow S_{ij}(t)=e^{-\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)u_i e^{\beta^t Z_{ij}}} \quad (14)$$

Then the conditional and marginal likelihood is given by

$$L_i(\beta|u_i)=\prod_{j=1}^{n_i}(\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)u_i e^{\beta^t Z_{ij}})^{\delta_{ij}} e^{-\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)u_i e^{\beta^t Z_{ij}}} \quad (15)$$

$$L_i(\theta,\beta)=\prod_{j=1}^{n_i}\int_0^\infty(\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)u e^{\beta^t Z_{ij}})^{\delta_{ij}} e^{-\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)u e^{\beta^t Z_{ij}}}g(u)du \quad (16)$$

Now incorporating the gamma distribution for frailty, we can write the expression as

$$L_i(\theta,\beta)=\prod_{j=1}^{n_i}\int_0^\infty[\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)u e^{\beta^t Z_{ij}}]^{\delta_{ij}} e^{-\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)u e^{\beta^t Z_{ij}}}\frac{u^{\frac{1}{\theta}-1}e^{-\frac{u}{\theta}}}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}}du$$

$$=\prod_{j=1}^{n_i}[\sum_{k=1}^{g}h_k I(y_{k-1}<t\leqslant y_k)]^{\delta_{ij}}[e^{\beta^t Z_{ij}}]^{\delta_{ij}}\frac{\Gamma(\frac{1}{\theta}+d_i)\theta^{\frac{1}{\theta}+d_i}}{\Gamma\frac{1}{\theta}\theta^{\frac{1}{\theta}}}$$
$$\int_0^\infty e^{-u[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)e^{\beta^t Z_{ij}}]}\frac{u^{\frac{1}{\theta}+d_i-1}}{\Gamma(\frac{1}{\theta}+d_i)\theta^{\frac{1}{\theta}+d_i}}du \quad (17)$$

where $d_i=\sum_{j=1}^{n_i}\delta_{ij}$. Then under the integral is the moment generating function of a gamma distribution with a pdf of $\Gamma(\frac{1}{\theta}+d_i,\frac{1}{\theta})$. Using this, we can derive the likelihood function for a piecewise constant baseline hazard as

$$L_i(\theta,\beta)=\prod_{j=1}^{n_i}[\sum_{k=1}^{g}h_k I(y_{k-1}<t<y_k)]^{\delta_{ij}}e^{\beta^t Z_{ij}\delta_{ij}}$$
$$\frac{\Gamma(\frac{1}{\theta}+d_i)\theta^{\frac{1}{\theta}+d_i}}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}\theta^{\frac{1}{\theta}+d_i}[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)e^{\beta^t Z_{ij}}]^{\frac{1}{\theta}+d_i}}$$
$$\int_0^\infty e^{-u[\frac{1}{\theta}+\sum_{j-1}^{n_i}\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)e^{\beta^t Z_{ij}}]}\frac{[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g}H_k I(y_{k-1}<t\leqslant y_k)e^{\beta^t Z_{ij}}]^{\frac{1}{\theta}+d_i}}{\Gamma(\frac{1}{\theta}+d_i)}du$$

$$= \prod_{j=1}^{n_i}[\sum_{k=1}^{g} h_k I(y_{k-1} < t \leqslant y_k)]^{\delta_{ij}} e^{\beta^t Z_{ij}\delta_{ij}}$$

$$\frac{\Gamma(\frac{1}{\theta}+d_i)}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g} H_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}]^{\frac{1}{\theta}+d_i}}$$

$$\int_0^{\infty} e^{-u[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g} H_k I(y_{k-1}<t\leqslant y_k)e^{\beta^t Z_{ij}}]} \frac{[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g} H_k I(Y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}]^{\frac{1}{\theta}+d_i}}{\Gamma(\frac{1}{\theta}+d_i)} du \quad (18)$$

The term under the integral is the pdf of $\Gamma(\frac{1}{\theta}+d_i, \frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g} H_k I(y_{k-1} < t \leqslant y_k)e^{\beta_t Z_{ij}})$ which integrates to 1. Therefore we can obtain the marginal likelihood function as

$$L_i(\theta,\beta) = \frac{\Gamma(\frac{1}{\theta}+d_i)\prod_{j=1}^{n_i}[\sum_{k=1}^{g} h_k I(y_{k-1} < t \leqslant y_k)]^{\delta_{ij}} e^{\beta^t Z_{ij}\delta_{ij}}}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}[\frac{1}{\theta}+\sum_{j=1}^{n_i}\sum_{k=1}^{g} H_k I(y_{k-1} < t \leqslant y_k)e^{\beta_t Z_{ij}}]^{\frac{1}{\theta}+d_i}} \quad (19)$$

Taking the logarithm and summing over all the clusters,

$$l(\theta,\beta) = \sum_{i=1}^{G}[d_i log(\theta) - log\Gamma(\frac{1}{\theta}) + log\Gamma(\frac{1}{\theta}+d_i) - (\frac{1}{\theta}+d_i)log[1+\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g} H_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}] + \sum_{j=1}^{n_i}\delta_{ij}[\beta^t Z_{ij} + log(\sum_{k=1}^{g} h_k I(y_{k-1} < t \leqslant y_k))]] \quad (20)$$

Considering the baseline hazard follows an exponential distribution, we get

$$l(\theta,\beta,\lambda_k) = \sum_{i=1}^{G}[d_i log(\theta) - log\Gamma(\frac{1}{\theta}) + log\Gamma(\frac{1}{\theta}+d_i) - (\frac{1}{\theta}+d_i)log[1+\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}] + \sum_{j=1}^{n_i}\delta_{ij}[\beta^t Z_{ij} + log(\sum_{k=1}^{g}\lambda_k)]] \quad (21)$$

By maximizing the likelihood function, the maximum likelihood estimators for the parameters can be obtained. We derive the first derivatives for the gamma frailty model and exponential baseline hazard as

$$\frac{\partial l(\theta,\beta,\lambda_k)}{\partial\theta} = \sum_{i=1}^{G}[\frac{d_i}{\theta} + \Gamma'(\frac{1}{\theta}) - \frac{\Gamma'(\frac{1}{\theta}+d_i)}{\theta^2(\frac{1}{\theta}+d_i)} - \frac{(\frac{1}{\theta}+d_i)\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}}{1+\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}} - \frac{log(1+\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}})}{\theta^2}] \quad (22)$$

$$\frac{\partial l(\theta,\beta,\lambda_k)}{\partial\beta} = \sum_{i=1}^{G}[\sum_{j=1}^{n_i}\delta_{ij}Z_{ij} - (\frac{1}{\theta}+d_i)\frac{\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}Z_{ij}}{1+\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}Z_{ij}}] \quad (23)$$

$$\frac{\partial l(\theta,\beta,\lambda_k)}{\partial\lambda_k} = \sum_{i=1}^{G}(\sum_{j=1}^{n_i}(\frac{1}{\sum_{k=1}^{g}\lambda_k})\delta_{ij} - (\frac{1}{\theta}+d_i)\frac{\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g} I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}}}{log(1+\theta\sum_{j=1}^{n_i}\sum_{k=1}^{g}\lambda_k I(y_{k-1} < t \leqslant y_k)e^{\beta^t Z_{ij}})}) \quad (24)$$

The maximum likelihood estimates can be obtained by setting each of the first-order derivatives to 0 and solving for the parameters of interest.

2.5 Asymptotic Variance Co-variance Matrix

Asymptotic variance Co-variance matrix can be obtained from the log-likelihood expression. Let us consider H to be the Hessian matrix of the second partial derivatives of the marginal log-likelihood. Then the negative expected value of the Hessian matrix is known as Fischer information. Consider $\mathscr{I}$ to denote the Fischer information matrix, then

$$\mathscr{I}(\theta,\beta,\lambda) = -E(H(\theta,\beta,\lambda)) \tag{25}$$

Then the observed information matrix I is the negative of Hessian matrix and is written as

$$I(\theta,\beta,\lambda) = -H(\theta,\beta,\lambda) \tag{26}$$

Taking the inverse of Fischer information matrix, we can obtain the asymptotic variance co-variance matrix of the estimates. Taking the inverse of the observed information matrix, one can get the estimated variance co-variance matrix.

3 Pre-test for Visualizing The Existence of Heterogeneity

3.1 Density plots

While the period of transition to adulthood is marked by discontinuation of schooling and entry into the labour market for many young people, some combine schooling and work and others are neither in school nor working. Data collected through the Life Event Calendar component of the Youth Study [18] provided an opportunity to explore the pattern of these events (that is, studying, working, both studying and working, and neither studying nor working) in young people's lives through density plots [27] and these are presented in Figure 1.

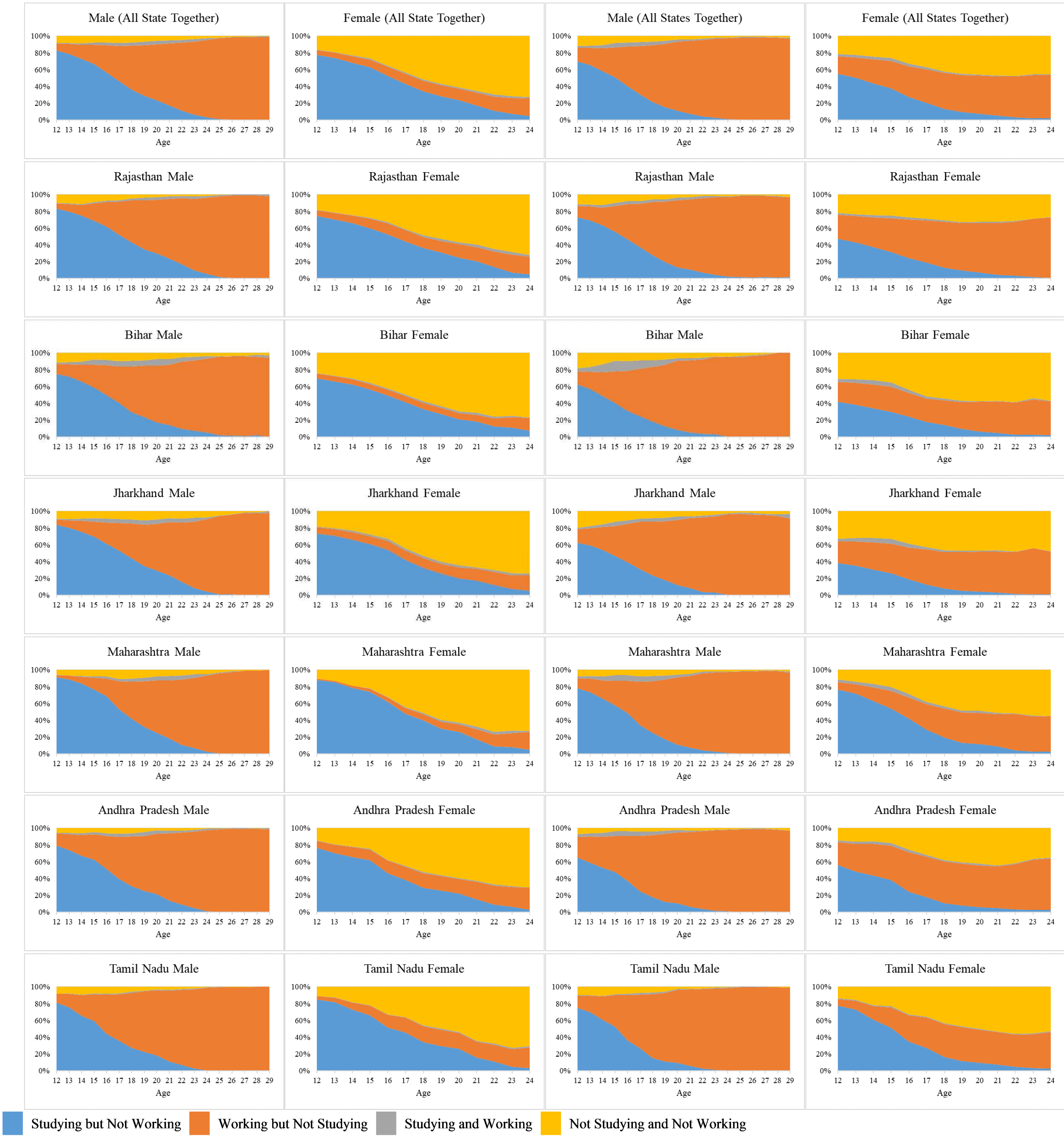
Pattern is varied widely by state, sex and place of residence. A comparison of the 24 panels of Figure 1 shows, first, that the proportion of youth reporting school attendance declined across all groups as young people transitioned out of childhood or early adolescence into late adolescence and young adulthood. While 83% urban young men, 72% rural young men, 78% urban young women and 57% rural young women were in school (a small minority of them were also working) at age 12, the percentage who remained in school at age 15 fell to 69% for urban men, 56% for rural men, 64% for urban young women and 42% for rural young women.

Second, very few young people i.e., 4% or fewer urban young men, 5% or fewer rural young men, 2% or fewer urban young women and 4% or fewer rural young women reported having combined studying and working at any age. Third, exit from school was accompanied by a steady rise in work participation over the ages among young men, and a much more gradual rise among young women. Moreover, while more young rural women than men were working at early ages (12-13), a reverse pattern was evident by age 15, and the gender gap widened with age thereafter. Fourth, age at which more youth were working than attending school occurred at age 18 for urban men, 16 for rural men, 22 for urban female and 16 for rural female. Finally, significant proportion of young women, but not young men, were neither in school nor working from age 12 onwards. Among young men, small proportions (9% or fewer for urban men and 12% or fewer for rural men) were neither working nor in school at any age. Among young women, there was a steady increase by age. At age 12, 16% of urban young women and 21% of rural young women were neither studying nor working; percentages increased to 73% at age 24 for urban young women and 48% for rural women at age 22.

The dataset shows a huge differences in percentages among the states by rural and urban area at different ages. 77% (lowest) urban men in Bihar & 91% (highest) urban men in the state Maharashtra were in study at age 12 whereas, 64% (lowest) rural men in Jharkhand and 80% (highest) rural men in Maharashtra were in study at age 12. For urban female, 71% from Bihar and 89% (highest) from Maharashtra were in study at age 12 whereas, 42%(lowest) Jharkhand rural female and 80%(highest) Maharashtra rural female were in study at age of 12. Moreover, 98-100% young men were working between age 25-29. There exists a significantly huge differences in rural (43% (Bihar)-73% (Rajasthan)) and urban (16-26%) young women.

3.2 Median Age of Leaving Study and Entry into Work

In order to measure the homogeneity or heterogeneity, the second approach was to estimate the "median age" for both the events as shown in Figure 2. Figure 2 shows the median ages at experiencing each event that urban young men, urban young women, rural young men and rural young women achieved. In general, median ages showed the delays that urban young men and women experienced both the events compared with their rural counterparts. Integrating the results, it can be seen that the urban young men and women showed higher median ages for leaving education & entry into work.



Place of residence: Urban Place of Residence: Rural

Fig. 1: Respondent-type wise density plots for various states

Moreover, we can observe three different types of median ages for men in both the events among all the six states. Similar duration for both events are observed among urban area of Rajasthan, Bihar, Jharkhand & Andhra Pradesh, but Rajasthan, Bihar, Maharashtra & Tamil Nadu show a similar duration in the rural areas. Men, living in the rural areas of Jharkhand are leaving education at a very early age, where urban men of Maharashtra & Rural men of Andhra Pradesh experienced both events in the same year. Urban women are entering the work after 7 years of leaving education where the rural

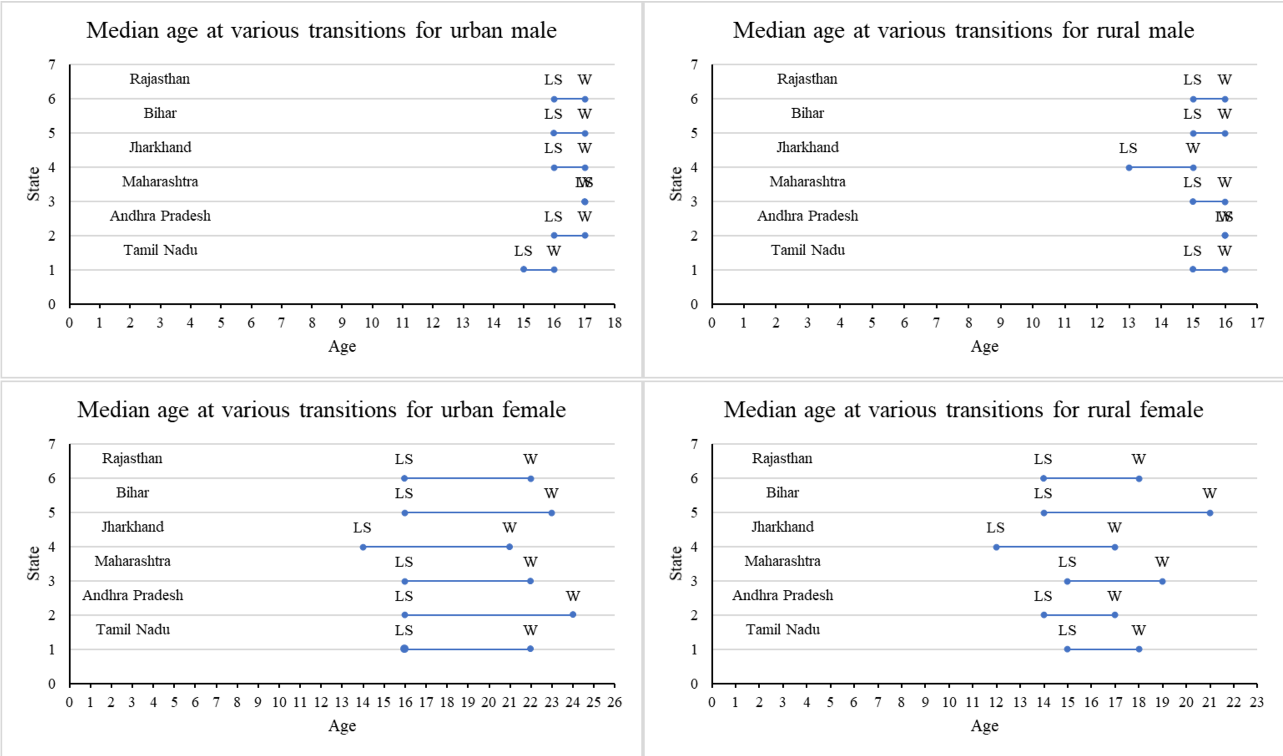


Fig. 2: Respondent-type and place of residence wise median ages for various states

women show the trend of entering the work after 4 years. Women of Jharkhand are already realising these two events before women in other states. In all, it turns out that women are delaying their entry into work by minimum of 4 years, where men leave their education and enter the work in the same year. For all the states, men and women staying in rural & urban area show different median ages for both the events which signifies the existence of an unobserved heterogeneity.

3.3 Theil's Index for Comparing State Level Heterogeneity

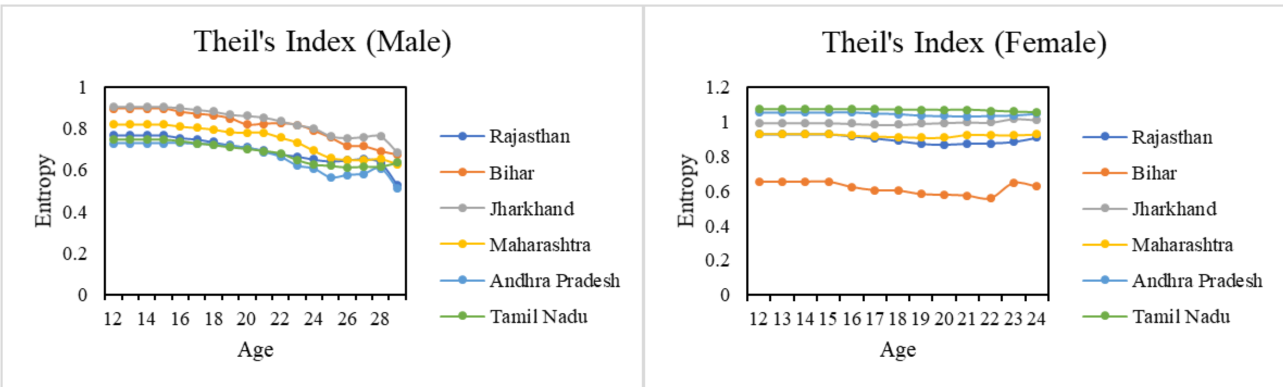


Fig. 3: Respondents age & state-wise Theil's inequality index

To figure out the existence of unobserved heterogeneity more specifically, one can use Theil's Index [28] used by Billari [29] and is written as $E = \sum_{s=1}^{S} p_s log(p_s)$ where S is the number of the type of transition and p_s is the relative frequency of the transition S. One might simply examine the heterogeneity of three transitions, namely leaving study to work at same age (LSW), entry into work at least after one year of leaving study (LSW) and left study but not entered into work (LS), at each age and compare the distribution of their values for different states. Since all the individuals in different states are associated with different social and economic features, then a great discrimination can be predicted among them. Thus in this third approach we have considered Theil's index and calculated the values for male & female separately at various ages for the six states and are presented in the following Table 1 & Table 2, respectively.

Table 1: Theil's Index for interpreting heterogeneity for male

Age	State					
	Rajasthan	Bihar	Jharkhand	Maharashtra	Andhra Pradesh	Tamil Nadu
Age 12	0.769	0.896	0.906	0.821	0.729	0.746
Age 13	0.769	0.896	0.906	0.821	0.729	0.746
Age 14	0.769	0.896	0.906	0.821	0.729	0.746
Age 15	0.769	0.896	0.906	0.821	0.729	0.746
Age 16	0.756	0.881	0.900	0.810	0.730	0.740
Age 17	0.749	0.871	0.891	0.804	0.725	0.729
Age 18	0.735	0.863	0.882	0.796	0.724	0.721
Age 19	0.720	0.849	0.867	0.783	0.715	0.710
Age 20	0.710	0.819	0.861	0.781	0.708	0.702
Age 21	0.694	0.823	0.854	0.780	0.687	0.692
Age 22	0.676	0.825	0.836	0.756	0.664	0.681
Age 23	0.666	0.818	0.816	0.734	0.621	0.649
Age 24	0.651	0.790	0.801	0.693	0.606	0.628
Age 25	0.643	0.756	0.763	0.659	0.564	0.622
Age 26	0.647	0.718	0.753	0.650	0.575	0.614
Age 27	0.653	0.714	0.759	0.648	0.582	0.618
Age 28	0.635	0.691	0.762	0.654	0.607	0.618
Age 29	0.528	0.674	0.684	0.628	0.511	0.637

Table 2: Theil's Index for interpreting heterogeneity for female

Age	State					
	Rajasthan	Bihar	Jharkhand	Maharashtra	Andhra Pradesh	Tamil Nadu
Age 12	0.931	0.656	0.996	0.930	1.056	1.073
Age 13	0.931	0.656	0.996	0.930	1.056	1.073
Age 14	0.931	0.656	0.996	0.930	1.056	1.073
Age 15	0.931	0.656	0.996	0.930	1.056	1.073
Age 16	0.919	0.625	0.990	0.921	1.057	1.073
Age 17	0.906	0.606	0.986	0.916	1.050	1.073
Age 18	0.891	0.604	0.985	0.915	1.047	1.071
Age 19	0.874	0.583	0.991	0.910	1.038	1.071
Age 20	0.868	0.579	0.996	0.910	1.035	1.069
Age 21	0.875	0.574	1.000	0.925	1.032	1.070
Age 22	0.877	0.558	0.998	0.925	1.036	1.066
Age 23	0.885	0.648	1.019	0.923	1.038	1.061
Age 24	0.910	0.629	1.012	0.928	1.049	1.058

Figure 3 shows the differences among the indices clearly. Among all states, Jharkhand men have shown the highest level of diversity among these three transitions, whereas the people of Andhra Pradesh have shown the lowest level of heterogeneity. Simillarly, Tamil Nadu has shown the highest heterogeneity in the case of women, where Bihar state has got the lowest level of heterogeneity. By all means, we can say that every state has shown different levels of unobserved heterogeneity among all the three transitions.

3.4 Kaplan-Meier Failure Estimates of Leaving Education and Entry into Workforce

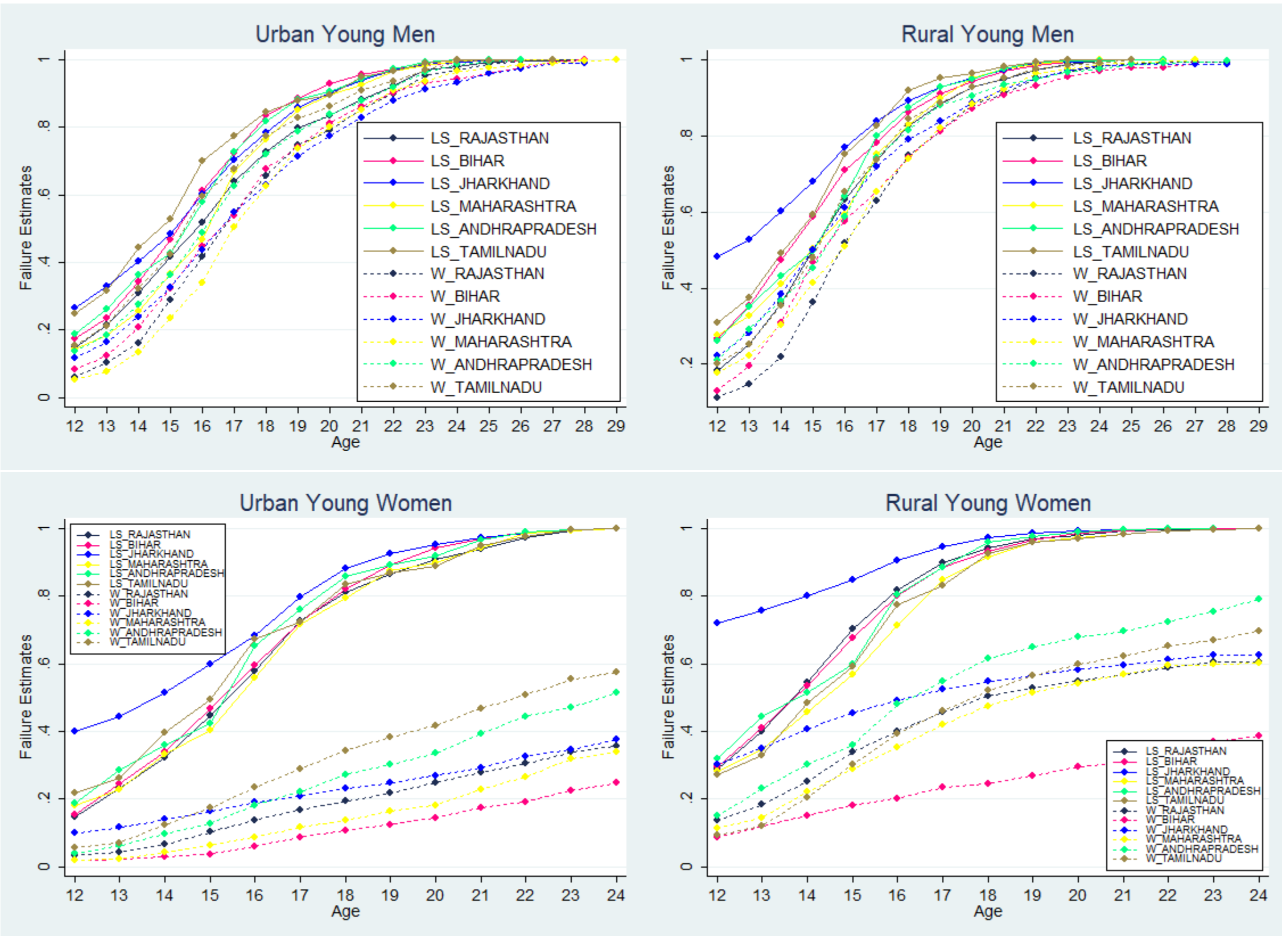


Fig. 4: Kaplan-Meier failure estimates for six states by sex & place of residence-wise

Kaplan-Meier [30] failure estimates were used to visualize the patterns of both the events, namely leaving study and entry into work. Figure 4 shows that patterns of leaving study and entry into work are not only differed by areas of residence, but also considerably by gender and States. The timing at experiencing both events generated different patterns specific to different groups of population [31]. People in Jharkhand state are showing an early trend in discontinuation of schooling rather than that of other states. Until 18 years of age, Rajasthan's rural young men and urban young men in Maharashtra are delaying the entry into the workforce than that of other states. While, Andhra Pradesh and Jharkhand rural young men entered the work earlier until age 15, but, after 16 years of age onwards, men in Tamil Nadu entered into work earlier than the men of other states. In case of females, Jharkhand women are showing remarkable early entry into workforce than the women in other states, whereas, Bihar women are delaying entry into workforce than the women in other states. After 15 years of age, urban women in Andhra Pradesh & Tamil Nadu entered into work earlier than that of any other states. In the rural areas, after 16 years of age, women in Andhra Pradesh, and after 19 years of age, women in Tamil Nadu showed earlier trend in entry into work in comparison to others. There are significant delay in entry into work for women rather than men. Initially for men, the differences are clear for all the states but as the time moves, failure estimates are also getting closer & closer.

Table 3: Proportions of young men and women having followed different social trajectories by gender and place of residence

State	Respondent type Place of residence	Male Urban	 Rural	Female Urban	 Rural
Rajasthan	E→W	15%	15%	7%	7%
	EW(Simul)	35%	32%	6%	12%
	E	2%	2%	35%	18%
	W	13%	21%	13%	36%
	Initial State(S)	35%	30%	40%	27%
	N	1442	2101	2338	3258
	Total	100%	100%	100%	100%
Bihar	E→W	17%	15%	4%	3%
	EW(Simul)	23%	26%	2%	5%
	E	4%	3%	35%	21%
	W	28%	37%	12%	36%
	Initial State(S)	28%	18%	46%	35%
	N	1168	1023	2487	2820
	Total	100%	100%	100%	100%
Jharkhand	E→W	19%	27%	9%	18%
	EW(Simul)	31%	42%	9%	28%
	E	6%	4%	46%	35%
	W	10%	9%	6%	7%
	Initial State(S)	34%	18%	31%	13%
	N	1802	1145	2279	2765
	Total	100%	100%	100%	100%
Maharashtra	E→W	22%	18%	8%	16%
	EW(Simul)	34%	50%	7%	21%
	E	4%	3%	49%	34%
	W	9%	12%	4%	7%
	Initial State(S)	31%	17%	32%	21%
	N	1574	1090	2123	2030
	Total	100%	100%	100%	100%
Andhra Pradesh	E→W	16%	9%	12%	15%
	EW(Simul)	42%	48%	12%	24%
	E	5%	3%	39%	20%
	W	15%	25%	10%	26%
	Initial State(S)	23%	14%	28%	15%
	N	1539	1423	2023	2530
	Total	100%	100%	100%	100%
Tamil Nadu	E→W	23%	20%	18%	23%
	EW(Simul)	51%	55%	14%	26%
	E	3%	4%	35%	31%
	W	3%	4%	2%	3%
	Initial State(S)	21%	17%	30%	18%
	N	1376	1458	2075	2680
	Total	100%	100%	100%	100%

3.5 Proportion of Young Men & Women Having Followed Different Trajectories

Since the Kaplan-Meier failure estimates produced cumulative proportions of transitions at a given age, so the estimates provide patterns that did not consider individual trajectories between the transition from leaving education to entry into work. Table 3 displays the different trajectories achieved by age 12-29 from leaving education to entry into work by states, respondent type & place of residence. The first trajectory includes respondents that left education and subsequently entered the work (E $\rightarrow$ W); the second trajectory is that in which both transitions occurred during the same year of age (EW simultaneously). The next three sequences correspond to respondents that after leaving education did not enter into the work (E); those who entered the work without leaving education (W); and finally, those who did experience neither of these two social transitions, and were in education (Student). The analysis considers two genders, two areas, together with five possible outcomes. This means that there are up to 120 different results to look at. Therefore the main patterns that come out on this analysis are summarized as follows.

Table 4: Estimated median age for male by different variables and examining the differences using four statistical tests

Variable	Category	Median Duration			P-values			
		Median	LCL	UCL	LR	GW	TW	PP
State	Rajasthan	17	17	17	0	0	0	0
	Bihar	16	16	16				
	Jharkhand	16	16	16				
	Maharashtra	17	17	17				
	Andhra Pradesh	16	16	16				
	Tamil Nadu	16	16	16				
Place of residence	Urban	17	17	17	0	0	0	0
	Rural	16	16	16				
Religion	Hindu	16	16	16	0	0	0	0
	Muslim	16	16	16				
	Others	16	16	16				
Caste	ST/SC	16	16	16	0	0	0	0
	OBC	16	16	16				
	General	17	17	17				
	Others	17	16	18				
Work status of father	Yes	16	16	16	0	0	0	0
	No	17	17	17				
Work status of mother	Yes	16	16	16	0	0	0	0
	No	17	17	17				
Father's education	Illiterate	15	15	15	0	0	0	0
	Literate	17	17	17				
Mother's education	Illiterate	16	16	16	0	0	0	0
	Literate	18	18	18				
Result of last exam	Pass	16	16	16	0	0	0	0
	Fail	17	17	17				
Type of school attended	Private	18	18	18	0	0	0	0
	Government	16	16	16				
	Others	18	18	18				

Table 5: Estimated median age for female by different variables and examining the differences using four statistical tests

Variable	Category	Median Duration			P-values			
		Median	LCL	UCL	LR	GW	TW	PP
State	Rajasthan	20*	20	21	0	0	0	0
	Bihar	22*	22	22				
	Jharkhand	19*	19	19				
	Maharashtra	21*	21	21				
	Andhra Pradesh	20	19	21				
	Tamil Nadu	20	20	20				
Place of residence	Urban	22*	21	22	0	0	0	0
	Rural	18	18	18				
Religion	Hindu	22	22	22	0	0	0	0
	Muslim	21*	21	21				
	Others	18	17	19				
Caste	ST/SC	18	18	18	0	0	0	0
	OBC	20*	20	20				
	General	22*	22	22				
	Others	21*	20	21				
Work status of father	Yes	20*	20	20	0	0	0	0
	No	21*	21	21				
Work status of mother	Yes	17	17	17	0	0	0	0
	No	22*	22	22				
Father's education	Illiterate	18	18	18	0	0	0	0
	Literate	21*	21	21				
Mother's education	Illiterate	20	20	20	0	0	0	0
	Literate	22*	21	22				
Result of last exam	Pass	20*	20	21	0	0	0	0
	Fail	20*	20	21				
Type of school attended	Private	22*	21	22	0	0	0	0
	Government	20*	20	20				
	Others	21*	20	21				

Table 3 shows important differences between urban and rural young men as well as urban and rural young women in the experience of social trajectories. Results show the delay in the experience of transitions by urban young men compared with rural male respondents by age 12-29. There was a tendency to experience both the transitions simultaneously in case of young men. Results show the delay in the experience of transitions by urban young women compared with rural female respondents. 28% of the Bihar young men had not left study during the work whereas, 36% of Rajasthan & Bihar women & 26% of Andhra Pradesh young women had not left education during work.

4 Selected Covariates

Researchers hope to get the best results for any kind of analysis. Thus, a proper selection of a set of covariates is required. In event history analysis, one such way is to compare the survivor functions between the groups within each variable. To compare the differences among the groups within each variable, P-values for four different tests are calculated. P-values for Log Rank test, Gehan Wilcoxon test, Tarone-Ware test and Peto-Peto-Prentice test with survival median are shown in the Table 4 and Table 5 for male and female respectively.

There are several variables in the Youth study dataset. Among these, some of the selected variables are considered as well as tested with the survivor functions for those variables using the four tests for this analysis. State is one of the most important variables to be included. The P-values for State variable is 0 and thus the data shows a significant difference among the survivor functions of these six states. Secondly, significant differences are also being observed in places of

Table 6: Hazard ratios for covariates of different types of model with p-values for male (without time varying covariate)

Covariate	Cox regression model		Shared frailty model		Piecewise constant baseline hazard with shared frailty model	
	exp(coef)	p-value	exp(coef)	p-value	exp(coef)	p-value
State						
Rajasthan						
Bihar	1.29853	0.000*				
Jharkhand	0.97241	0.376				
Maharashtra	1.19173	0.000*				
Andhra Pradesh	1.42817	0.000*				
Tamil Nadu	1.64164	0.000*				
Place of residence						
Urban						
Rural	1.06592	0.001*	1.06592	0.001*	1.07653	0.000*
Religion						
Hindu						
Muslim	1.28185	0.000*	1.28184	0.000*	1.23704	0.000*
Others	1.12822	0.005*	1.12818	0.005*	1.02024	0.636
Caste						
SC/ST						
OBC	0.96227	0.085	0.96228	0.085	0.98325	0.447
GEN	0.87285	0.000*	0.87283	0.000*	0.88144	0.000*
Others	1.02632	0.741*	1.02630	0.740	0.97088	0.697
Work status of father						
Yes						
No	1.21927	0.000*	1.21928	0.000*	1.15239	0.000*
Work status of mother						
Yes						
No	0.78119	0.000*	0.78117	0.000*	0.83106	0.000*
Father's education						
Illiterate						
Literate	0.67415	0.000*	0.67414	0.000*	0.73518	0.000*
Mother's education						
Illiterate						
Literate	0.63094	0.000*	0.63096	0.000*	0.69173	0.000*
Result of last exam						
Pass						
Fail	1.45406	0.000*	1.45406	0.000*	1.27470	0.000*
Type of school last attended						
Government						
Private	1.45713	0.000*	1.45710	0.000*	1.37501	0.000*
Others	2.29170	0.000*	2.29169	0.000*	1.83970	0.000*
Total no of siblings	1.03900	0.000*	1.03899	0.000*	1.03191	0.000*
Frailty				0.000*	1.07059	0.000*
Log-likelihood	-106842.5		-106842.5		-42887.65	

residence which are classified as rural and urban. Thirdly, religion which is classified as Hindus, Muslims and others in three different types, Caste which is classified as Scheduled Caste/ Scheduled Tribe (SC/ST), other backward caste (OBC), General and others, are also showing significant differences within them. Father's and Mother's working status, in many cases, do affect the educational status of adults and have a lot of influence on entering the job. Its reliability is significantly proven from the results of these tests. There is also a significant difference between the survivor function of the adults who pass the last exam and the survivor function of the adults who fail in it. Data also shows significant P-values of these four tests for the variable Type of school last attended. Summarily, the results of these tests give us an indication to choose these variables for further analysis.

Table 7: Hazard ratios for covariates of different types of model with p-values for female (without time varying covariate)

Covariate	Cox regression model		Shared frailty model		Piecewise constant baseline hazard with shared frailty model	
	exp(coef)	p-value	exp(coef)	p-value	exp(coef)	p-value
State						
Rajasthan						
Bihar	0.71772	0.000*				
Jharkhand	0.85483	0.000*				
Maharashtra	0.78631	0.000*				
Andhra Pradesh	1.13570	0.000*				
Tamil Nadu	0.99595	0.901*				
Place of residence						
Urban						
Rural	1.88562	0.000*	1.88562	0.000*	1.55567	0.000*
Religion						
Hindu						
Muslim	0.87595	0.000*	0.87594	0.000*	0.88185	0.000*
Others	1.28402	0.000*	1.28403	0.000*	1.20443	0.000*
Caste						
SC/ST						
OBC	0.89477	0.000*	0.89477	0.000*	0.93423	0.002*
GEN	0.66562	0.000*	0.66561	0.000*	0.74841	0.000*
Others	0.75825	0.003*	0.75819	0.003*	0.74891	0.001*
Work status of father						
Yes						
No	1.00839	0.714	1.00839	0.710	1.06916	0.003*
Work status of mother						
Yes						
No	0.42284	0.000*	0.42283	0.000*	0.52196	0.000*
Father's education						
Illiterate						
Literate	0.76618	0.000*	0.76616	0.000*	0.81587	0.000*
Mother's education						
Illiterate						
Literate	0.73716	0.000*	0.73716	0.000*	0.75476	0.000*
Result of last exam						
Pass						
Fail	1.10615	0.000*	1.10615	0.000*	1.13503	0.000*
Type of school last attended						
Government						
Private	1.26976	0.000*	1.26977	0.000*	1.24592	0.000*
Others	2.32635	0.000*	2.32628	0.000*	1.73976	0.000*
Total no of siblings	1.03020	0.000*	1.03019	0.000*	1.02032	0.000*
Frailty				0.000*	1.05989	0.000*
Log-likelihood	-111075.3		-111075.3		-53438.29	

5 Results and Discussion

Time to entry into work is defined as the current age of respondent in this analysis. A total number of 46549 respondents were selected for the analysis, out of which 17141 were men and 29408 were women. 2720 of them experienced the event leaving education and rest of them were used as the reference category in the regression analysis. In this analysis a small part of Youth Study Dataset considered as an event history data with 'not having entered into work' as censored observation. In addition to the commonly used Cox's regression model for analysing survival data, two types of frailty models are also considered for examining the effects of different demographic and socio-economic factors on the time to entry into work. These two frailty models (shared gamma frailty & piecewise baseline hazard with shared gamma frailty)

Table 8: Hazard ratios for covariates of different types of model with p-values for male (with time varying covariate)

Covariate	Cox regression model		Shared frailty model		Piecewise constant baseline hazard with shared frailty model	
	exp(coef)	p-value	exp(coef)	p-value	exp(coef)	p-value
TVC						
ref.						
0 Year	2.82474	0.000*	2.82471	0.000*	2.15211	0.000*
1 Year	2.03145	0.000*	2.03144	0.000*	1.77461	0.000*
2 Year	1.87165	0.000*	1.87163	0.000*	1.67963	0.000*
3-4 Years	1.33752	0.000*	1.33751	0.000*	1.32875	0.000*
5-6 Years	0.72286	0.000*	0.72284	0.000*	0.77159	0.000*
7+ Years	0.41771	0.000*	0.41770	0.000*	0.51162	0.000*
State						
Rajasthan						
Bihar	1.46722	0.000*				
Jharkhand	1.05216	0.112				
Maharashtra	1.11345	0.001*				
Andhra Pradesh	1.22112	0.000*				
Tamil Nadu	1.42572	0.000*				
Place of residence						
Urban						
Rural	1.05745	0.005*	1.05745	0.005*	1.05948	0.003*
Religion						
Hindu						
Muslim	1.38600	0.000*	1.38600	0.000*	1.28500	0.000*
Others	1.10742	0.018*	1.10739	0.019*	1.00464	0.913
Caste						
SC/ST						
OBC	0.92988	0.001*	0.92989	0.001*	0.96973	0.167
GEN	0.82648	0.000*	0.82647	0.000*	0.85679	0.000*
Others	1.00631	0.936*	1.00627	0.940	0.94166	0.428
Work status of father						
Yes						
No	1.13967	0.000*	1.13967	0.000*	1.09580	0.000*
Work status of mother						
Yes						
No	0.77235	0.000*	0.77235	0.000*	0.83788	0.000*
Father's education						
Illiterate						
Literate	0.65876	0.000*	0.65875	0.000*	0.73783	0.000*
Mother's education						
Illiterate						
Literate	0.65868	0.000*	0.65869	0.000*	0.72112	0.000*
Result of last exam						
Pass						
Fail	1.37034	0.000*	1.37033	0.000*	1.20825	0.000*
Type of school last attended						
Government						
Private	1.35100	0.000*	1.35100	0.000*	1.29265	0.000*
Others	3.14778	0.000*	3.14782	0.000*	2.35585	0.000*
Total no of siblings	1.02975	0.000*	1.02975	0.000*	1.03044	0.000*
Frailty				0.000*	1.03614	0.000*
Log-likelihood	-105464.7		-105464.7		-42128.62	

are also considered for adjusting the heterogeneity in time to entry into work due to various geographical regions i.e. states

Table 9: Hazard ratios for covariates of different types of model with p-values for female (with time varying covariate)

Covariate	Cox regression model		Shared frailty model		Piecewise constant baseline hazard with shared frailty model	
	exp(coef)	p-value	exp(coef)	p-value	exp(coef)	p-value
TVC						
ref.						
0 Year	3.72919	0.000*	3.72948	0.000*	2.25032	0.000*
1 Year	1.72636	0.000*	1.72647	0.000*	1.59553	0.000*
2 Year	1.35794	0.000*	1.35805	0.000*	1.35136	0.000*
3-4 Years	0.82069	0.000*	0.82076	0.000*	0.97431	0.476*
5-6 Years	0.37951	0.000*	0.37954	0.000*	0.57217	0.000*
7+ Years	0.11364	0.000*	0.11366	0.000*	0.22108	0.000*
State						
Rajasthan						
Bihar	0.73960	0.000*				
Jharkhand	1.10999	0.002*				
Maharashtra	0.91646	0.023*				
Andhra Pradesh	1.11406	0.001*				
Tamil Nadu	1.09761	0.006*				
Place of residence						
Urban						
Rural	1.73403	0.000*	1.73403	0.000*	1.38217	0.000*
Religion						
Hindu						
Muslim	0.91420	0.009*	0.91420	0.009*	0.90469	0.003*
Others	1.11389	0.002*	1.11391	0.003*	1.06558	0.070
Caste						
SC/ST						
OBC	0.92104	0.000*	0.92102	0.000*	0.96694	0.125
GEN	0.69054	0.000*	0.69051	0.000*	0.78555	0.000*
Others	0.81913	0.034*	0.81908	0.035*	0.80953	0.020*
Work status of father						
Yes						
No	1.07221	0.002*	1.07222	0.002*	1.10261	0.000*
Work status of mother						
Yes						
No	0.48762	0.000*	0.48762	0.000*	0.60037	0.000*
Father's education						
Illiterate						
Literate	0.69881	0.000*	0.69880	0.000*	0.80844	0.000*
Mother's education						
Illiterate						
Literate	0.64943	0.000*	0.64944	0.000*	0.74373	0.000*
Result of last exam						
Pass						
Fail	1.08462	0.001*	1.08461	0.001*	1.10442	0.000*
Type of school last attended						
Government						
Private	1.41815	0.000*	1.41813	0.000*	1.27242	0.000*
Others	3.27121	0.000*	3.27121	0.000*	2.01899	0.000*
Total no of siblings	1.04918	0.000*	1.04916	0.000*	1.02875	0.000*
Frailty				0.000*	1.05134	0.000*
Log-likelihood	-107423.4		-107423.4		-51856.01	

and the gamma distribution is used as the frailty distribution for the states. In all the models, the same set of covariates are

Table 10: log-likelihood, coefficient of bases and variance of random effect for the three models

Respondent type		Male			
	Model	**Cox regression model**	**Shared frailty model**	**Piecewise constant baseline hazard with shared frailty**	
Without TVC	Variance of random effect		1.14803		
	Coefficient of baseline			0.01396	0.16570
	exp(Coef)			1.01406	1.18022
	se(Coef)			0.00060	0.00760
	log-likelihood	-106842.5	-106842.5	-42887.65	
With TVC	Variance of random effect		1.09135		
	Coefficient of baseline			0.01031	0.12685
	exp(Coef)			1.01037	1.13525
	se(Coef)			0.00050	0.00590
	log-likelihood	-105464.7	-105464.7	-42128.62	
Respondent type		**Female**			
Without TVC	Variance of random effect		1.10320		
	Coefficient of baseline			0.01593	0.12641
	exp(Coef)			1.01606	1.13475
	se(Coef)			0.00080	0.01110
	log-likelihood	-111075.3	-111075.3	-53438.29	
With TVC	Variance of random effect		1.06087		
	Coefficient of baseline			0.01320	0.18804
	exp(Coef)			1.01328	1.20688
	se(Coef)			0.00060	0.01630
	log-likelihood	-107423.4	-107423.4	-51856.01	

used which are, place of residence, religion, caste, total number of siblings, work status of father, work status of mother, father's education, mother's education, result of last exam & type of school last attended. Another important point in this analysis is that when the Cox's model is used, only then, the state has been kept as a covariate and when the frailty model is taken, the 'state' has been considered as a frailty variable for unobserved heterogeneity. For analytical convenience, mostly used statistical software like R, SPSS & STATA has been used.

5.1 Analysis of Survival Models with & without Time Varying Covariate

Table 6 and Table 7 show the estimated hazard ratio corresponding to the parameters and the corresponding P-values for all the three models considered in the analysis. The fit of the Cox's regression model shows that all the covariates have significant effects on the likelihood of entry into work. Examining the total number of siblings in the model, we see that with a hazard ratio of 1.03020 for male & 1.02975 for female, each sibling increases the hazard of entry into work by a factor almost 3 percent (i.e. (1.03020-1)100 $\sim$ 3.0 & (1.02975-1)100 $\sim$ 3.0).

State found to be an important covariate for time to entry into work as the analysis shows that all the states has significantly higher likelihood of entry into work compared to Rajasthan for men, but for women there are significantly lower likelihood of entry into work compared to the state of Rajasthan.

The result of the analysis shows that place of residence is another important factor for time to entry into work, where men & women from a rural area have significantly higher likelihood of entry into work compared to that of urban regions of the states. Religion shows a significant effect on time to entry into work where Muslim men and men & women in other communities have a significantly higher likelihood of entry into work than that of Hindus. We can also see that Muslim women are associated with a decreased hazard of entry into work after leaving education. Compared to SC/ST categories, OBC & GEN show a lower likelihood of entry into work for both types of respondents but other categories have a 1.02 greater chance of entry into work for male respondents. In case of male respondents, whose father doesn't work have a significantly higher likelihood of entry into work than those whose father does work. In case of female, the value is not at all significant. which is vice versa in case of mother.

For literate parents, the likelihood of entry into work in case of all the respondents is significantly lower than that of illiterate parents which is naturally common. For both men & women, those who have failed in their last exam have a significantly higher likelihood of entry into work than those who have passed in their last exam. Moreover, those who have studied in private & other institutions, for them there is a much higher likelihood to go to work as compared to those who studied in government institutions.

In addition to the Cox's proportional hazard model, two types of shared gamma frailty models are considered to examine the effects of states (cluster/frailty term) on the time to entry into work in the Indian context. The estimated hazard ratios corresponding to the parameters & the corresponding P-values of the shared frailty models are also reported in Table 6 and Table 7 for male & female respectively. The corresponding Log-likelihood is used for testing the significance of frailty parameter and the third model is found to be the best model among the competing models for analysing time to entry into work in Indian scenario. The piecewise constant baseline hazard with shared frailty model fit the time to entry into work data significantly better than that of Cox's proportional hazard & shared frailty model, i.e., frailty term state have a significantly higher effects on the time to entry into work for the respondent.

Except for the covariate Religion (Others) & Caste (OBC), the sign of estimates are found to be the same for all the three models, but the size of the estimates & P-values are found to be different for some covariates considered in the models. After adjusting the heterogeneity due to states, the analysis of the piecewise constant baseline hazard with shared frailty shows that there is a significant difference ($p<0.05$) in the likelihood of entry into work among the fathers who are not working & fathers who are working in case of female respondents which is insignificant for the other two models. This may be due to the strong correlation between work status of father and time to entry into work. It is also found that, other categories in caste have a lower likelihood of entry into work compared to that of SC/ST category.

For all the models, the sign of time varying covariate are similar except the 3-4 years for female in the baseline models. Just after leaving education there are higher likelihood for both men & women to go to work as compared to their counterparts.

Since the size of the variance parameter (Table 10) is large for shared frailty, it indicates that clusters are heterogeneous with respect to the risk of entry into work. In our analysis, a substantial estimate of variance parameter of size 1.09135, 1.06087, 1.14803 & 1.10320 are obtained for the frailty term 'state'. For piecewise baseline hazard, the baseline is divided into two parts according to the median ages of entry into work for male & female, which are 16 & 23, respectively. The coefficient of the baseline estimate with their standard errors are also mentioned in the tables. These coefficients are also showing higher differences between the two bases.

6 Conclusion

Time to entry into work is considered as one of the important indicators for describing the overall employment pattern of a country. Cox's proportional hazard model is often used to examine the effects of different covariates on time to various events but geographical regions can produce unobserved heterogeneity in the data which can't be captured using Cox's model. Moreover, combining the age interval 12-29 in a single baseline hazard model may not be appropriate for the analysis as the results may provide inefficient estimates. The states were defined as clusters and it is assumed that time to entry into work for male & female residing in the same states are correlated because they share the same environment. In this analysis, all the covariates are found to have significant effect on time to entry into work for both the respondent types. The observations within a specific state are found to be correlated and the analysis with baseline hazard with shared frailty model shows that working status of father is also a significant covariate for females where the Cox's & shared frailty model couldn't observe this.

Acknowledgement

We are grateful to Government of India, as the government has served with non-net fellowship to the second author for our research work. We are also thankful to International Institute for Population Sciences, Mumbai, as they provided the valuable dataset for this research work. We are also immensely grateful to the editors and reviewers who moderated this paper with their valuable suggestions and in that line improved the manuscript significantly.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] J.A. Cañada-Vicinay, Growing up in an intact vs. non-intact family and the transition from school to permanent work: A gender approach for spain in 2000, *Economics of Education Review*, **24**, 691-704 (2005).
[2] M. Manacorda, F.C. Rosati, M. Ranzani and G. Dachille, Pathways from school to work in the developing world, *IZA Journal of Labor & Development*, **6**, 1 (2017).
[3] M. Salas-Velasco, The transition from higher education to employment in europe: the analysis of the time to obtain the first job, *Higher Education*, **54**, 333-360 (2007).
[4] B. Nilsson, The school-to-work transition in developing countries, *The Journal of Development Studies*, 1-20 (2018).
[5] G.J. Van den Berg and B. Drepper, Inference for shared-frailty survival models with left-truncated data, *Econometric Reviews*, **35**, 1075-1098 (2016).
[6] S.A. Thamrin and A. Lawi, *Estimating piecewise exponential frailty model with changing prior for baseline hazard function*, In AIP Conference Proceedings, **1707**, (2016).
[7] I.M. Sari, A. Thamrin and A. Lawi. *Piecewise exponential frailty models on survival data using bayesian approach*, In Conference paper: international conference on science (ICOS-1), (2014).
[8] J. Kubo, M.R. Cullen, L. Cantley, M. Slade, B. Tessier-Sherman, O. Taiwo and M. Desai, Piecewise exponential models to assess the influence of job-specific experience on the hazard of acute injury for hourly factory workers, *BMC medical research methodology*, **13**, 89 (2013).
[9] D.R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187-220 (1972).
[10] W. Mietlowski, A review of: "frailty models in survival analysis, by A. Wienke", *Journal of Biopharmaceutical Statistics*, 860-862 (2011).
[11] L. Duchateau and P. Janssen, *The frailty model*, Springer Science & Business Media, (2007).
[12] J.W. Vaupel, K.G. Manton and E. Stallard, The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**, 439-454 (1979).
[13] D.G. Clayton, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**, 141-151 (1978).
[14] S. Mahmood, B. Zainab and A.M. Latif, Frailty modeling for clustered survival data: an application to birth interval in bangladesh, *Journal of Applied Statistics*, **40**, 2670-2680 (2013).
[15] M. Singh, N.K. Singh, A. Bharti and R. Singh, Frailty model approach for survival of children below age 5 in eag states and assam, *International Journal of Current Advanced Research*, **7**, (2018).
[16] M. Zidan, A. Sagheer and N. Metwally, An Autonomous Competitive Learning Algorithm using Quantum Hamming Neural Networks, Proceedings of the International Joint Conference on Neural Networks (IJCNN, Killarney, Ireland, 2015), IEEE, 1-7 (2015)
[17] A. Sagheer, M. Zidan, M. M. Abdelsamea, A novel autonomous perceptron model for pattern classification applications, Entropy 21 (8), 763 (2019).
[18] I.I. for Population Sciences (IIPS) and P. Council, Youth in India: Situation and needs 2006-2007, (2010).
[19] P. Hougaard, *Analysis of multivariate survival data*, Springer Science & Business Media, (2002).
[20] T.M. Therneau and P.M. Grambsch, *Modeling survival data: extending the Cox model*, Springer Science & Business Media, (2013).
[21] L. Duchateau, P. Janssen, P. Lindsey, C. Legrand, R. Nguti and R. Sylvester, The shared frailty model and the power for heterogeneity tests in multicenter trials, *Computational Statistics & Data Analysis*, **40**, 603-620 (2002).
[22] L. Duchateau, P. Janssen, I. Kezic and C. Fortpied, Evolution of recurrent asthma event rate over time in frailty models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 355-363 (2003).
[23] L. Duchateau and P. Janssen, Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows, *Biometrics*, **60**, 608-614 (2004).
[24] N. Breslow, Covariance analysis of censored survival data, *Biometrics*, 89-99 (1974).
[25] J.D. Kalbfleisch and R.L. Prentice, Marginal likelihoods based on Cox?s regression and life model, *Biometrika*, **60**, 267-278 (1973).
[26] U. Abdulkarimova. *Frailty Models for Modelling Heterogeneity*, PhD thesis, (2013).
[27] U. Ram, S. Mohanty and A. Singh, *Youth in india: Situation and needs*, (2010).
[28] H. Theil, Statistical decomposition analysis; with applications in the social and administrative sciences, Technical report, (1972).
[29] F.C. Billari, The analysis of early life courses: Complex descriptions of the transition to adulthood *Journal of population research*, **18**, 119-142 (2001).
[30] E.L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, *Journal of the American statistical association*, **53**, 457-481 (1958).

[31] G. Mejia Pailles. A life course perspective on social and family formation transitions to adulthood of young men and women in Mexico. PhD thesis, The London School of Economics and Political Science (LSE), (2012).

---

7 Stata Code with respect to Fig. 4

Urban

```
keep if LS_STA == 1
stset AGELS2, failure(LS_STA == 1)
sts graph, failure tmax(29) tmin(12) noorigin
sts gen RS=S if VTYPE==1 & HSTATE==8
sts gen BS=S if VTYPE==1 & HSTATE==10
sts gen JS=S if VTYPE==1 & HSTATE==20
sts gen MS=S if VTYPE==1 & HSTATE==27
sts gen APS=S if VTYPE==1 & HSTATE==28
sts gen TNS=S if VTYPE==1 & HSTATE==33

generate LS_RAJASTHAN=1-RS if VTYPE==1 & HSTATE==8
generate LS_BIHAR=1-BS if VTYPE==1 & HSTATE==10
generate LS_JHARKHAND=1-JS if VTYPE==1 & HSTATE==20
generate LS_MAHARASHTRA=1-MS if VTYPE==1 & HSTATE==27
generate LS_ANDHRAPRADESH=1-APS if VTYPE==1 & HSTATE==28
generate LS_TAMILNADU=1-TNS if VTYPE==1 & HSTATE==33

stset AGEW2, failure(W_STA == 1)
sts graph, failure tmax(29) tmin(12) noorigin
sts gen RW=S if VTYPE==1 & HSTATE==8
sts gen BW=S if VTYPE==1 & HSTATE==10
sts gen JW=S if VTYPE==1 & HSTATE==20
sts gen MW=S if VTYPE==1 & HSTATE==27
sts gen APW=S if VTYPE==1 & HSTATE==28
sts gen TNW=S if VTYPE==1 & HSTATE==33

generate W_RAJASTHAN=1-RW if VTYPE==1 & HSTATE==8
generate W_BIHAR=1-BW if VTYPE==1 & HSTATE==10
generate W_JHARKHAND=1-JW if VTYPE==1 & HSTATE==20
generate W_MAHARASHTRA=1-MW if VTYPE==1 & HSTATE==27
generate W_ANDHRAPRADESH=1-APW if VTYPE==1 & HSTATE==28
generate W_TAMILNADU=1-TNW if VTYPE==1 & HSTATE==33

twoway (connected LS_RAJASTHAN AGELS2, sort) (connected LS_BIHAR AGELS2, sort)(connected LS_JHARKHAND AGELS2, sort)
(connected LS_MAHARASHTRA AGELS2, sort)(connected LS_ANDHRAPRADESH AGELS2, sort)(connected LS_TAMILNADU AGELS2, sort)
(connected W_RAJASTHAN AGEW2, sort) (connected W_BIHAR AGEW2, sort)(connected W_JHARKHAND AGEW2, sort)
(connected W_MAHARASHTRA AGEW2, sort)(connected W_ANDHRAPRADESH AGEW2, sort)(connected W_TAMILNADU AGEW2, sort) , legend(pos(5)ring(0)col(1))
```

Rural

```
keep if LS_STA == 1
stset AGELS2, failure(LS_STA == 1)
sts graph, failure tmax(29) tmin(12) noorigin
sts gen RS=S if VTYPE==2 & HSTATE==8
```

sts gen BS=S if VTYPE==2 & HSTATE==10
sts gen JS=S if VTYPE==2 & HSTATE==20
sts gen MS=S if VTYPE==2 & HSTATE==27
sts gen APS=S if VTYPE==2 & HSTATE==28
sts gen TNS=S if VTYPE==2 & HSTATE==33

generate LS_RAJASTHAN=1-RS if VTYPE==2 & HSTATE==8
generate LS_BIHAR=1-BS if VTYPE==2 & HSTATE==10
generate LS_JHARKHAND=1-JS if VTYPE==2 & HSTATE==20
generate LS_MAHARASHTRA=1-MS if VTYPE==2 & HSTATE==27
generate LS_ANDHRAPRADESH=1-APS if VTYPE==2 & HSTATE==28
generate LS_TAMILNADU=1-TNS if VTYPE==2 & HSTATE==33

stset AGEW2, failure$(W_STA == 1)$
sts graph, failure tmax(29) tmin(12) noorigin
sts gen RW=S if VTYPE==2 & HSTATE==8
sts gen BW=S if VTYPE==2 & HSTATE==10
sts gen JW=S if VTYPE==2 & HSTATE==20
sts gen MW=S if VTYPE==2 & HSTATE==27
sts gen APW=S if VTYPE==2 & HSTATE==28
sts gen TNW=S if VTYPE==2 & HSTATE==33

generate W_RAJASTHAN=1-RW if VTYPE==2 & HSTATE==8
generate W_BIHAR=1-BW if VTYPE==2 & HSTATE==10
generate W_JHARKHAND=1-JW if VTYPE==2 & HSTATE==20
generate W_MAHARASHTRA=1-MW if VTYPE==2 & HSTATE==27
generate W_ANDHRAPRADESH=1-APW if VTYPE==2 & HSTATE==28
generate W_TAMILNADU=1-TNW if VTYPE==2 & HSTATE==33

twoway (connected LS_RAJASTHAN AGELS2, sort) (connected LS_BIHAR AGELS2, sort)(connected LS_JHARKHAND AGELS2, sort)
(connected LS_MAHARASHTRA AGELS2, sort)(connected LS_ANDHRAPRADESH AGELS2, sort)(connected LS_TAMILNADU AGELS2, sort)
(connected W_RAJASTHAN AGEW2, sort) (connected W_BIHAR AGEW2, sort)(connected W_JHARKHAND AGEW2, sort)
(connected W_MAHARASHTRA AGEW2, sort)(connected W_ANDHRAPRADESH AGEW2, sort)(connected W_TAMILNADU AGEW2, sort), legend(pos(5)ring(0)col(1))

Tapan Kumar Chakrabarty is currently a Professor of Statistics in North-Eastern Hill University, Shillong, India. He has been teaching Statistics during the last three decades at several Universities and Institutes in India including Manipur University, Nagaland University, and Indian Statistical Institute, North-East Centre, Tezpur, Assam. He pursued collaborative research with Indian Statistical Institute, Kolkata, and University of Calcutta among others. He delivered invited talks at European Population Conference 2008, Barcelona, Spain and in Third and Fourth Asian Population Conferences held in Kuala Lumpur, Malaysia and in Shanghai University, Shanghai, China in 2015 and 2018, respectively.

Jayanta Deb is currently a bona-fide research scholar in the Department of Statistics, NEHU, Shillong, India. He is graduated from Gurucharan college, Silchar, Assam and completed his master's degree from North-Eastern Hill University, Shillong, India. He has done his M.Phil. from the Department of Statistics, North-Eastern Hill University, Shillong in the area of multi-state models in event history analysis.