Introduction to Statistical Methods
SOC-GA 2332

Lecture 11: Longitudinal Data Analysis

Siwei Cheng

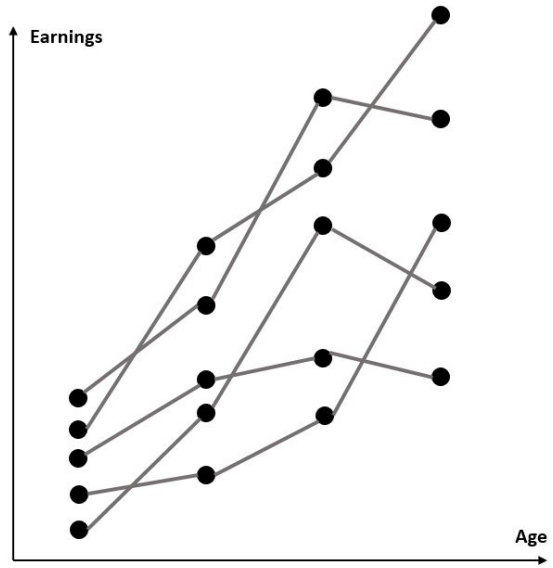
NEW YORK UNIVERSITY

Introduction to Longitudinal Data

- Repeated cross-sectional data provide long-term data, gives the same survey to different samples over time.
- Longitudinal data (panel data) track the same sample at different points in time. The sample can consist of individuals, households, establishments...
- See illustration below:

(a) Macrolevel Analysis

Data required: synthetic cohort from repeated cross-sectional data (e.g. CPS)



(b) Microlevel Analysis

Data required: longitudinal data (e.g. PSID)

Examples of Longitudinal Data

- Examples of longitudinal data: National Longitudinal Survey of Youth (NLSY), Panel Study of Income Dynamics (PSID), Wisconsin Longitudinal Study (WLS), National Longitudinal Study of Adolescent to Adult Health (Add Health)...
- Some advantages of longitudinal data: Longitudinal data allow for the measurement of within-sample change over time, enable the measurement of the duration of events, and record the timing of various events.

Introduction to Longitudinal Data

- Example research questions that can be addressed with longitudinal data (try to think about why they should be better studied with longitudinal data):
  1. Is longer duration of unemployment associated with lower likelihood of re-employment?
  2. What factors determine transitions into and out of marriage?
  3. Does having a daughter rather than a son promote gender egalitarian attitudes among fathers?
  4. How does family income affect children's cognitive and non-cognitive development?
  5. Do black and white workers of the same educational attainment experience the same rate of wage growth over life?
  6. Do people tend to befriend their friends' friend?

Displaying Longitudinal Data

- What does a longitudinal dataset look like?
- There are two ways to display longitudinal data:

i		..... X_{ij}		
id	sex	inc80	inc81	inc82
1	0	5000	5500	6000
2	1	2000	2200	3300
3	0	3000	2000	1000

i	j		X_{ij}
id	year	sex	inc
1	80	0	5000
1	81	0	5500
1	82	0	6000
2	80	1	2000
2	81	1	2200
2	82	1	3300
3	80	0	3000
3	81	0	2000
3	82	0	1000

- The "reshape" command in STATA will help you transform one data form to another. In R, you can do this with the "melt" and "gather" commands.

Longitudinal Data Analysis Methods

In this lecture, we will give a (very brief) introduction to the following models for longitudinal data:

- First difference model
- Fixed effect model
- Random effect model
- Models for Event Occurrence

First Difference Model with Two-period Panel Data

- Let's begin with the simplest kind of longitudinal data: data with **two** time periods.
- Let Y be county crime rate, and X be county unemployment rate.
- Suppose we only have cross-sectional data, we can run a regression like this:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

- But since we have two-period data for Time 1 and Time 2, we actually have two equations:

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + \epsilon_{i1} \tag{2}$$
$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \epsilon_{i2} \tag{3}$$

First Difference Model with Two-period Panel Data

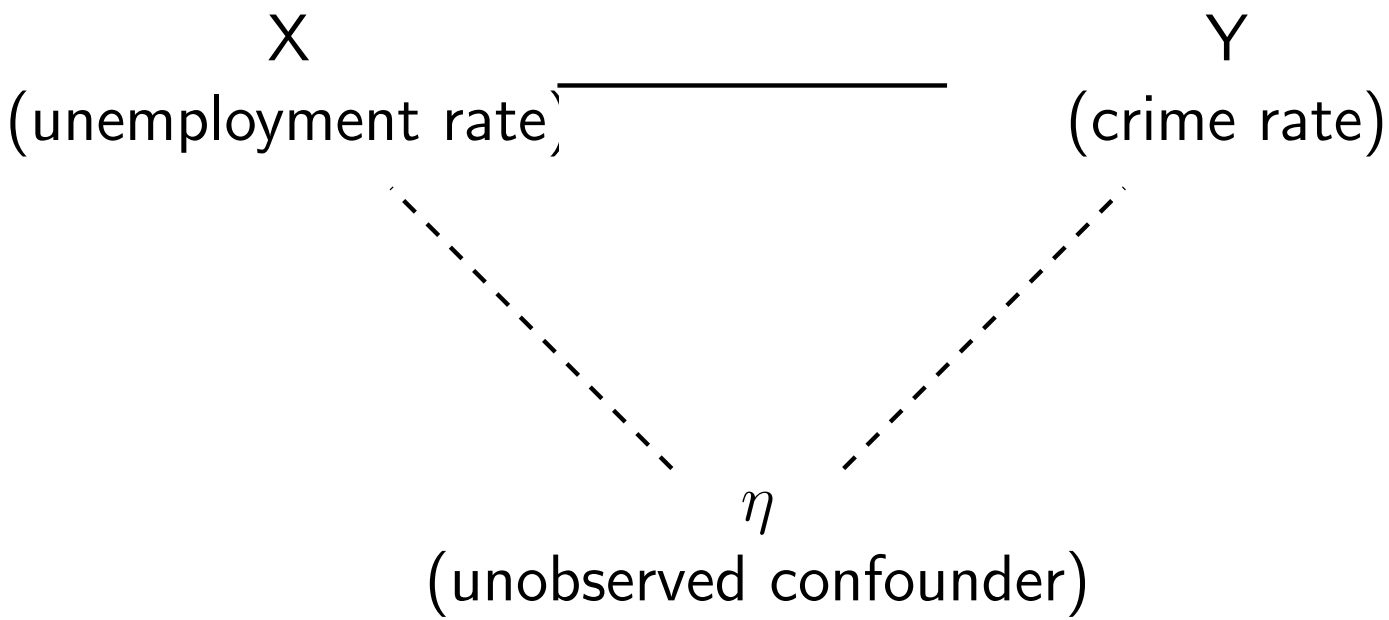
- Now we can take the difference between the above two equations to generate the first-difference estimator:

$$\Delta y_i = \beta_1 \Delta x_i + \Delta \epsilon_i \tag{4}$$

- Why do we want to do this first difference model?
- The advantage of the first-difference model is thatthe effect of county-level unobserved confounders can be canceled out by taking the difference.

First Difference Model with Two-period Panel Data

- To see this, let's assume that there exists an unobserved confounder η_i such that the relationship among the variables look like this:



First Difference Model with Two-period Panel Data

- To express this in an equation:

$$y_i = \beta_0 + \beta_1 x_i + \eta_i + \epsilon_i \tag{5}$$

- This is undesirable, because without observing η_i, our new residual term, $\eta_i + \epsilon_i$, will be **correlated** with the explanatory variable X.
- This violates the regression assumption, and also undermines our ability to draw causal conclusions from the analysis.
- The first difference model offers an approach to addressing this when we have longitudinal, as apposed to cross-sectional, data.

First Difference Model with Two-period Panel Data

- Now, we re-write the two-period model:

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + \eta_i + \epsilon_{i1} \tag{6}$$
$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \eta_i + \epsilon_{i2} \tag{7}$$

- So, what happens if we take their differences?

$$\Delta y_i = \beta_1 \Delta x_i + \Delta \epsilon_i \tag{8}$$

- As you will see, the effect of the confounder η disappeared.
- Hence, having two periods of observations will allow us to net out the impact of person-fixed unobserved confounders.

The Fixed Effect Model

The Fixed Effect Model

- Now, let's consider a more general type of model for analyzing longitudinal data: the **fixed-effect** model.
- First, let us draw a distinction between **time-varying** and **time-invariant** variables. Time-varying variables are denoted by terms such as Y_{it} or X_{it}, and time-invariant variables are denoted by terms such as Y_i or X_i.
- Examples of time-varying variables include age (for a person), annual birth rate (for a country), number of employees (for firms that change size); Examples of time-invariant variables include gender and genetic variables (for a person) and geographic location (for a country).

The Fixed Effect Model

- We do not need the subscript t in the time-invariant variables because they will be the same (i.e. constant over time) for unit i.
- When we consider a longitudinal dataset, we always think about what variables are time-varying and what variables are time-invariant. Longitudinal models deal with predicting or explaining variations in time-varying variables.
- Consider a simple case of the fixed-effect model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \eta_i + \epsilon_{it} \tag{9}$$

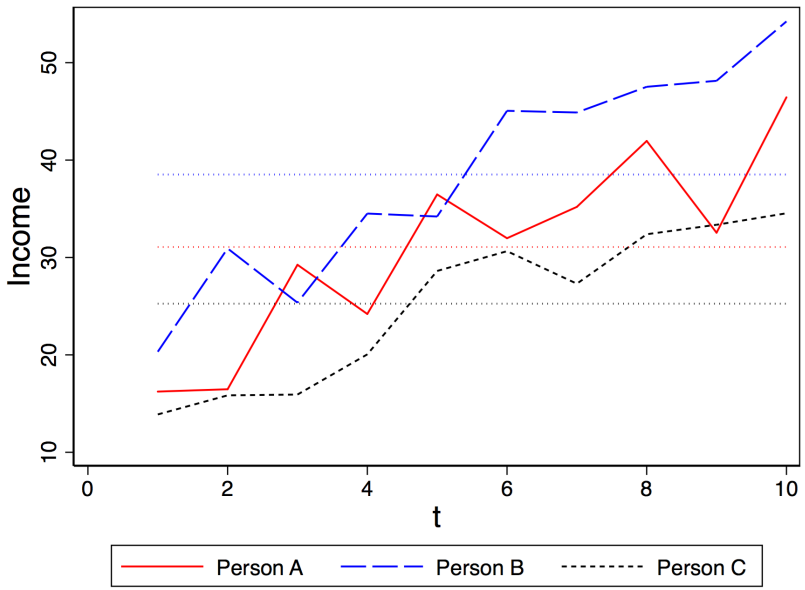
- Note that Y_{it}, X_{1it}, and ϵ_{it} are time-varying variables, and η_i is time-invariant variable.

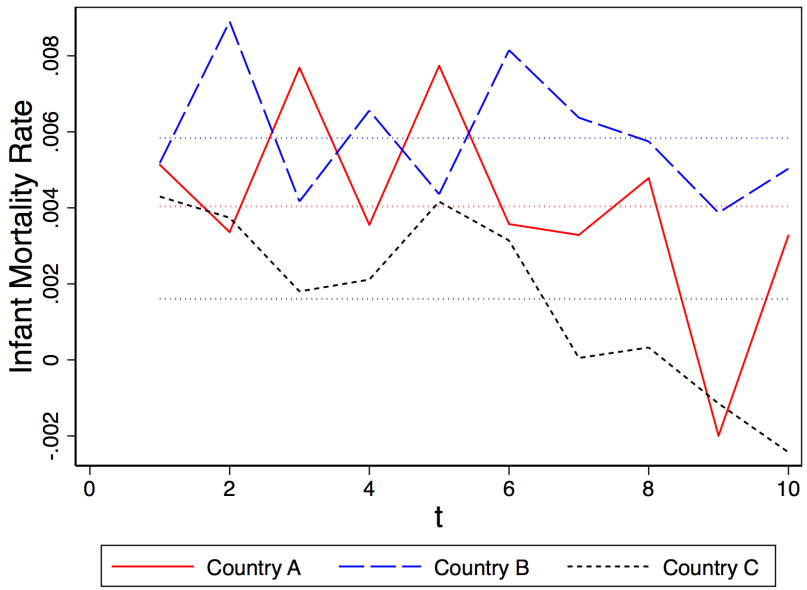
The Fixed Effect Model

- Examples of a fixed effect model:

$$\text{Hours of Housework}_{it} = \beta_0 + \beta_1 \text{Marital Status}_{it} + \beta_2 \text{Age}_{it} + \eta_i + \epsilon_{it} \tag{10}$$

- Some graphic illustrations:





The Fixed Effect Model

- When estimating the fixed-effect model, we assign a "fixed effect" for each person to capture the effect of all time-invariant variables within each person. So, if you have 300 observations in your sample, then you will have 300 "fixed effects."
- The fixed-effect model will tell us the effect of time-varying IVs on the time-varying DV. For example, effect of marital status on hours of housework, effect of GDP on infant mortality rate.
- For example, in predicting income, the fixed-effect model assigns a "person fixed-effect" to absorb income differences due to the *between-person* differences in income, and then estimates the effect of within-person *changes* in time varying covariates on time-varying income.

The Fixed Effect Model

- The fixed-effect estimator is sometimes also called the "within estimator." To see this, recall the equation:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \eta_i + \epsilon_{it} \tag{11}$$

- Take the average on both sides:

$$\bar{Y}_{it} = \beta_0 + \beta_1 \bar{X}_{it} + \bar{\eta}_i + \bar{\epsilon}_{it} \tag{12}$$

- Subtract Equation (12) from (11):

$$(Y_{it} - \bar{Y}_{it}) = \beta_1 (X_{it} - \bar{X}_{it}) + (\epsilon_{it} - \bar{\epsilon}_{it}) \tag{13}$$

- Equation (13) describes the within-estimator version of the FE model. (Equation (12) is also called the between-estimator.)

The Fixed Effect Model

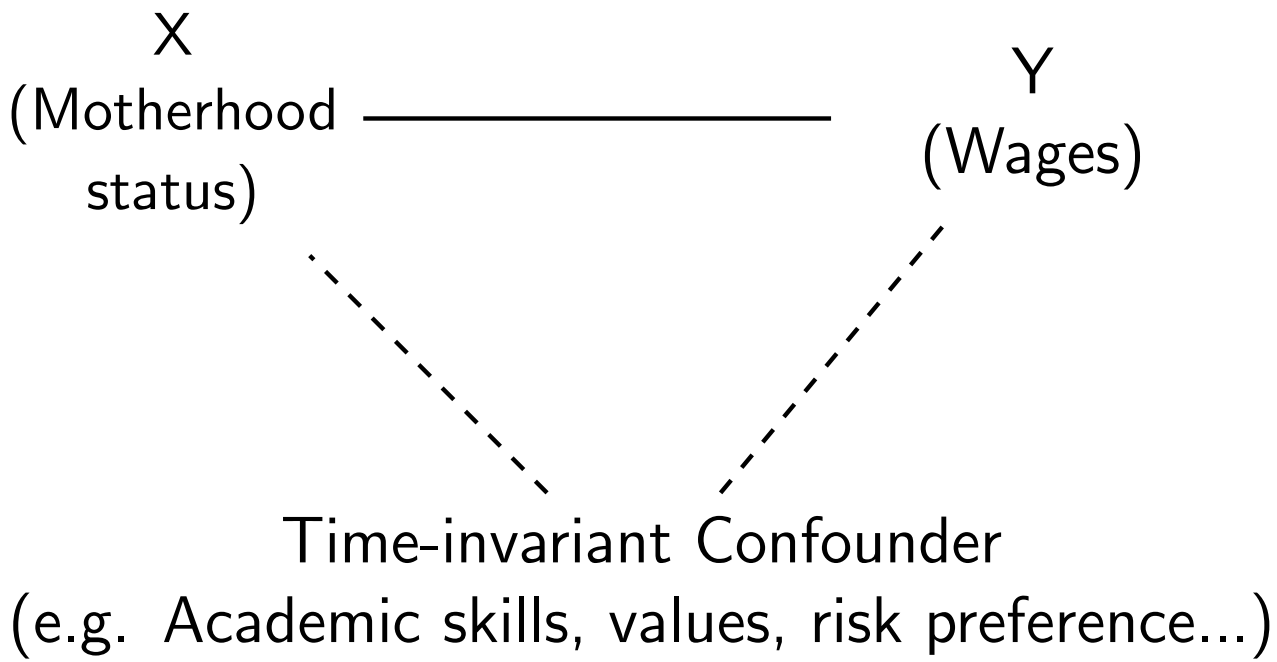
- Another rather obvious way to estimate the model with fixed effects is to include a dummy variable for every observation unit.
- In what types of scenarios are the FE model most useful? FE model is helpful if you are interested in within-unit variations and you also worry about the potential "confounding" effect of time-invariant unobserved variables.
- For example, the effect of marriage on earnings can be estimated by comparing the earnings of those who are married to those who are single (cross-sectional analysis), or by comparing the earnings of the same person before and after marriage (FE model).

Application of the Fixed Effect Model

- Application of fixed-effect models: Example from Budig and England (2001), which examines the effect of motherhood on wages.
- Normally, wages are lower among mothers than non-mothers in the cross-sectional data. However, The authors noted that:
- *"It is possible that there is no causal effect of motherhood on wages, but rather that some of the same individual characteristics that cause lower earnings for mothers also lead to childbearing at higher rates."*

Application of the Fixed Effect Model

- Person fixed-effect model helps the authors control for unchanging characteristics of the person that are unmeasured but have additive effects on earnings.
- For example:



Application of the Fixed Effect Model

- An important question to think about: what kind of confounders cannot be controlled away by the inclusion of person fixed effects?
- For example, what if women decide to become pregnant when they see a period of low wages coming (e.g. recession)?

Application of the Fixed Effect Model

- See their results:

Table 2. Unstandardized Coefficients for the Effect of Total Number of Children (Continuous Variable) on Women's Hourly Wage (ln), from Fixed-Effects Models and OLS Models: NLSY, 1982 to 1993

Control Variables in Model	Fixed-Effects Model	OLS Model
Gross (no controls)	–.068** (.004)	–.081** (.002)
Marital status	–.073** (.004)	–.081** (.002)
Marital status and human capital variables [a]	–.047** (.004)	–.018** (.002)
Marital status, human capital variables, and job characteristics [b]	–.037** (.004)	–.012** (.002)

Notes: OLS models include age and year, each in linear, squared, and cubed form. Numbers in parentheses are standard errors. Standard errors in OLS models were corrected using the Huber-White

The Random Effect Model

The Random Effect Model

- Consider the equation:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + ... + \beta_k X_{kit} + \gamma_i + \epsilon_{it} \quad (14)$$

- The random effect model assumes that the person-fixed unobserved variable γ_i is **uncorrelated** with the explanatory variables at any time point:

$$Corr(X_{jit}, \gamma_i) = 0, \forall t = 1, 2, ..., T; j = 1, 2, ..., k. \quad (15)$$

The Random Effect Model

- To put this assumption differently, in a treatment effect context, we assume that there is **no selection on the unit-specific effect**. For example, if something like motivation is both unmeasured and fixed over time, then the random effect model assumes that the selection into treatment cannot be related to motivation.

- There are different ways to estimate the random effect model. For example, we can use the Maximum Likelihood Estimator, which assumes that the random effect γ_i and error term ϵ_{it} follow independent normal distributions. Under this normality assumption, we will be able to estimate the distribution of the random effects as well as generate predicted values for γ_i for individuals in the data.

The Random Effect Model

- Fixed-effect of random-effect?
- As you probably noticed, the "unbiasedness" of the coefficients in the random effect model relies on the assumption as specified in Equation (15), while the fixed effect model allows arbitrary correlation between γ_i and the explanatory variables. In this sense, FE model can be a more convincing tool.
- However, if your key explanatory variable is constant over time, we cannot use FE to estimate its effect on DV... (Why?) In this case, we still prefer the RE model.
- The Hausman test (available in both R and STATA) can be used to test whether the key RE assumption in Equation (15) is rejected.

Multilevel Analysis and Growth Curve Models

Multilevel Analysis and Growth Curve Models

- Multilevel analysis deals with the scenarios in which observations are "nested" within clusters. For example, individuals nested within schools, children nested within families (sibling model), neighborhoods nested within states, and person-year observations nested within people.
- Multilevel analysis allows the regression coefficients to vary by these "clusters."

Multilevel Analysis and Growth Curve Models

- For example, suppose we are interested in the test scores (Y_{ij}) for student i in school j, instead of estimating:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij} \tag{16}$$

- We can estimate a cluster-specific intercept:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \epsilon_{ij} \tag{17}$$

- We can further allow the slope (i.e. effect of X) to be cluster-specific:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \tag{18}$$

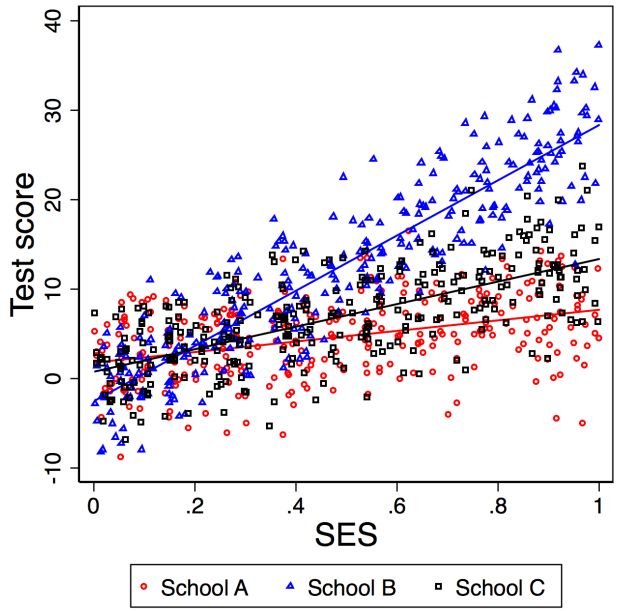
- In other words, we allow the intercept and effects of variable X to differ by j.

Multilevel Analysis and Growth Curve Models

- For example, if you are interested in the following equation predicting the test score for student i in school j:

$$SCORE_{ij} = \beta_{0j} + \beta_{1j} SES_{ij} + \epsilon_{ij} \tag{19}$$

- This equation says that the school may affect the overall level of test score (random intercept), as well as the effect of family SES on scores (random slope).



Multilevel Analysis and Growth Curve Models

- Now, let's go back to the context of **longitudinal models**.
- Longitudinal data can be considered as person-year observations nested within each person.
- For example, if we are interested in the growth trajectory of income over time, we can write Level 1 of the multilevel model as:

$$Income_{it} = \beta_{0i} + \beta_{1i} \cdot Time + \epsilon_{it} \tag{20}$$

- Level 1 above is around the person-year observations.
- We then model Level 2 around the person-specific intercepts and slopes:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} \cdot X_i + u_{0i} \tag{21}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} \cdot X_i + u_{1i} \tag{22}$$

Multilevel Analysis and Growth Curve Models

- See an example in Cheng (2014):

TABLE 3
ESTIMATED COEFFICIENTS FROM MULTILEVEL GROWTH CURVE MODELS PREDICTING LOG HOURLY WAGE

	Model 1	Model 2
Coefficients predicting baseline wage (β_0):		
Constant intercept (γ_{00})	1.952***	1.926***
	(.011)	(.013)
Gender (γ_{01}) (reference: male):		
Female	−.252***	−.205***
	(.011)	(.015)
Race (γ_{02}) (reference: white):		
Hispanic	−.011	.007
	(.013)	(.017)
Black	−.118***	−.083***
	(.013)	(.019)
Educational attainment (γ_{03}) (reference: high school or less):		
Some college but less than four years	.331***	.302***
	(.012)	(.016)
At least four years of college	.734***	.723***
	(.017)	(.018)
Coefficients predicting wage growth rate β_1:		
Constant slope (γ_{10})	.052***	.058***
	(.002)	(.002)
Gender (γ_{11}) (reference: male):		
Female		−.010***
		(.002)
Race (γ_{12}) (reference: white):		
Hispanic		−.004
		(.002)
Black		−.008**
		(.002)
Educational attainment (γ_{13}) (reference: high school or less):		
Some college but less than four years		.006*
		(.003)
At least four years of college		.002
		(.004)
Other coefficient:		
Squared experience (β_2)	−.002***	−.002***
	(.000)	(.000)
Variance components:		
var(u_0)	.637***	.636***
	(.082)	(.081)

Multilevel Analysis and Growth Curve Models

- In this paper, a general framework is derived to examine the implications of heterogeneous intercept and slope for the trend in wage inequality over the life course.
- Consider the following model for predicting log wages:

$$lnY_{it} = \lambda_i + \theta_i t + e_{it}. \tag{23}$$

- By taking variances of both sides of Equation (23), we can decompose the total variance of hourly wage at time t, Var_t, into four separable and distinct components:

$$Var_t = Var(lnY_t) = \underbrace{Var(\lambda)}_{V1} + \underbrace{t^2 \cdot Var(\theta)}_{V2} + \underbrace{2t \cdot Cov(\lambda, \theta)}_{V3} + \underbrace{Var(e_{it})}_{V4}. \tag{24}$$

A Very Brief Introduction to Models for Event Occurrence

A Very Brief Introduction to Models for Event Occurrence

- Another type of statistical models for longitudinal data considers *changes in status*, *transitions*, or the *occurrence* of events.
- For example, social demographers are interested in the determinants for the transition into marriage, divorce, and death. Policy scholars are interested in the determinants for the transition into and out of unemployment, job changes, prisons, or mobility between neighborhoods.

A Very Brief Introduction to Models for Event Occurrence

- There are generally two approaches for the analysis of transition data.
- **Discrete-time** approach deals with events that occur at fixed time points rather than at any possible time along a continuum. For example, election may take place every 2 or 4 years, graduation generally takes place at the end of school terms...But there are also cases where we choose to treat events as happening discrete times. e.g. marriage, job displacement...
- **Continuous-time** approach deals with events that can happen in a continuous time (e.g. mortality, or job changes in very detailed weekly data)...

A Very Brief Introduction to Models for Event Occurrence

- Today, let's briefly introduce the **discrete-time event history** mode.
- For simplicity, we consider **absorbing states** - states that, once entered, cannot be left. e.g. death, ever-married.
- We define a variable D_{ij} to indicate the **event occurrence** for person i in time j. If the event has occurred, then we denote $D_{ij} = 1$, and for all the time points where the event had not yet occurred as $D_{ij} = 0$.
- The discrete-time event history model predicts the **discrete-time hazard rate**, which can be expressed as the probability that the event occurs conditional on not having occurred previously:

$$p_{ij} = Pr(D_{ij} = 1 | D_{i(j-1)} = 0, D_{i(j-2)} = 0 ... D_{i1} = 0) \qquad (25)$$

A Very Brief Introduction to Models for Event Occurrence

- We model the probability p_{ij} just as what we did for a logit model:

$$p_{ij} = \frac{exp(\alpha_j + \beta X_{ij})}{1 + exp(\alpha_j + \beta X_{ij})} \tag{26}$$

- Note that this is equivalent to writing:

$$logit(p_{ij}) = \alpha_j + \beta X_{ij} \tag{27}$$

- Models like this are often used to study how X affects the likelihood of making a transition.

Example: Transition into Marriage

- See an example in Sweeney (2002), where the author models the determinants for the transition into marriage among women of different cohorts.

Table 2. Coefficients for Women from the Logistic Regression of Transition to First Marriage on Selected Independent Variables: National Longitudinal Surveys of Youth (NLSY) and Young Women (NLSW)

	White Women				Black Women			
	Early Cohort		Late Cohort		Early Cohort		Late Cohort	
Independent Variable	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Earnings (log)	.07	(.04)	.19*†	(.04)	.01	(.06)	.22*†	(.05)
No earnings	−.25	(.13)	−.38*	(.13)	−.19	(.17)	−.59*	(.15)
Educational Attainment								
Less than 12 years	.24*	(.12)	.36*	(.11)	−.01	(.15)	−.04‡	(.15)
13 to 15 years	.41*	(.11)	.23*	(.10)	−.27‡	(.23)	.03	(.13)
16 or more years	.42*	(.16)	.41*	(.11)	.85*	(.26)	.37*	(.19)
Currently enrolled in school	−.87*	(.10)	−.71*	(.09)	−.25‡	(.16)	−.09‡	(.14)
Currently employed	.03	(.09)	−.01	(.09)	.14	(.16)	.08	(.13)
Family Background Variables								
Two-parent family at age 14	−.16	(.11)	.02	(.08)	.22‡	(.14)	.14	(.10)
Family head's job is professional/managerial	−.11	(.09)	−.06	(.08)	−.09	(.36)	.07	(.21)
Mother's education:								
12 years	−.06	(.09)	.00	(.08)	−.12	(.19)	−.06	(.11)
13 or more years	−.21	(.12)	−.21*	(.10)	−.25	(.31)	−.31	(.18)
Residence