Modeling Uncertainty in Event Occurrence for Discrete-Time Survival

Katherine E. Masyn, Ph.D.
University of California at Davis
kmasyn@ucdavis.edu

CILVR Inaugural Conference
University of Maryland at College Park
May 18, 2006

Overview

- Introduction to discrete-time survival analysis in a latent variable framework
- Missingness for single indicators of event occurrence
- Measurement error for single and multiple indicators of event occurrence

Discrete time-to-event data

- Event history: A record of ***if*** **and *when*** an event occurred for each individual in a sample during a finite observation period.
- Discrete time:

  1) The timing of an event is continuous but is only recorded for an *interval* of time, e.g., age (in years) of first alcohol use.

  2) The timing of an event is itself discrete, e.g., grade retention.

Survival probability

Let *T* be the time interval of the event where $T \in \{1,2,\ldots,J\}$

$P_S(j)$, called the ***survival probability***, is defined as the probability of “surviving” *beyond* time interval *j*, i.e., the probability that the event occurs after interval *j*:

$$P_S(j) = P(T > j).$$

Hazard probability

$P_h(j)$, called the ***hazard probability***, is defined as the probability of the event occurring in the time interval *j*, provided it has not occurred prior to *j*:

$$P_h(j) = P(T = j \mid T \geq j).$$

Essentially, *$P_h(j)$* is the probability of the event occurring in time interval *j* among those at-risk in *j*.

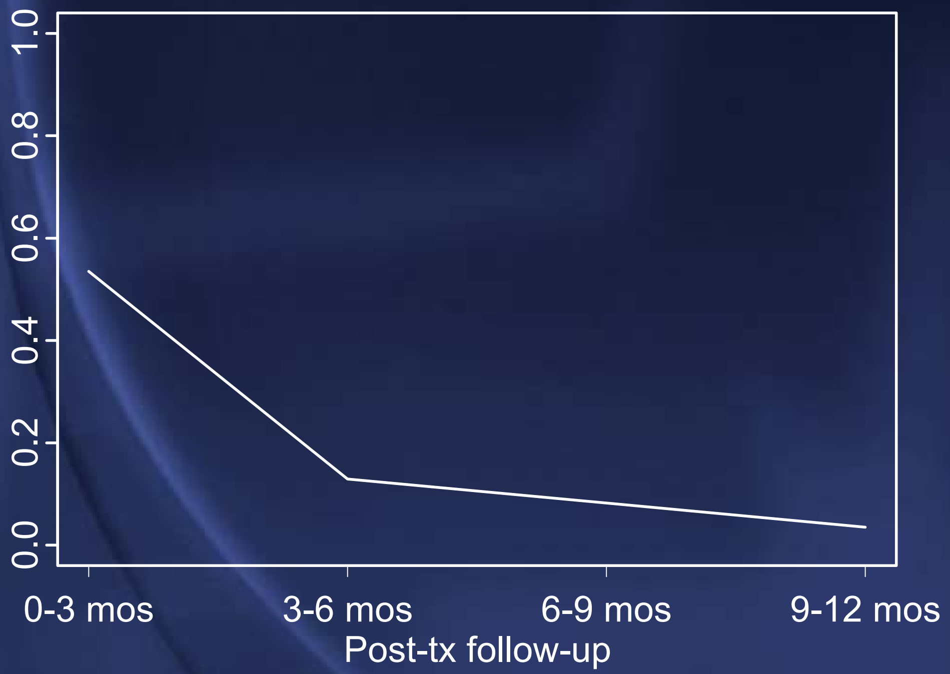
The relationship between *$P_S(j)$* and *$P_h(j)$* is given by

$$P_S(j) = P(T > j) =$$

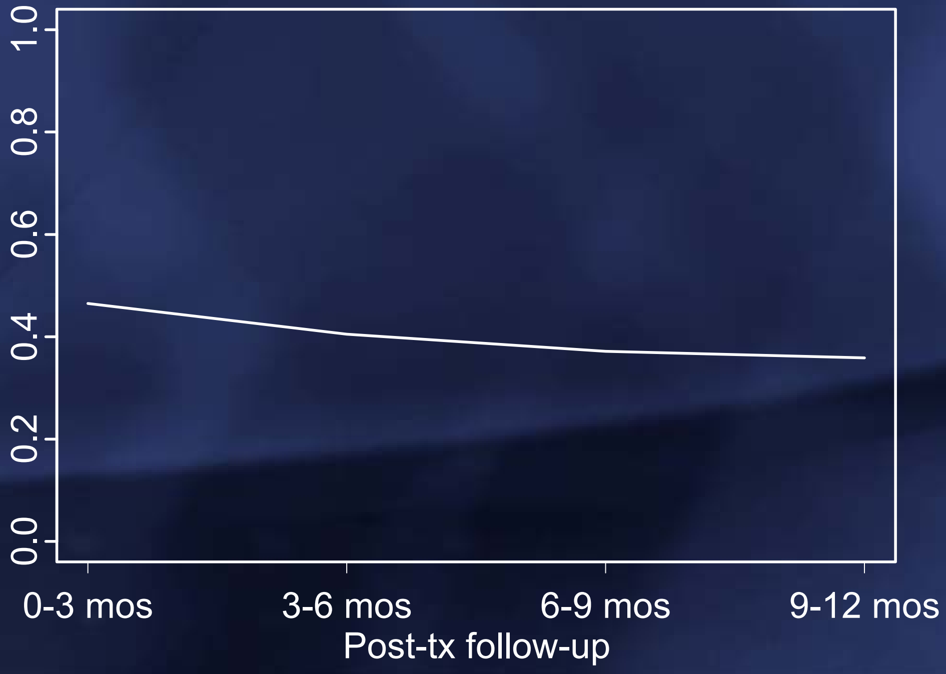
$$P(T > j \mid T \geq j) \times$$
$$P(T > j - 1 \mid T \geq j - 1) \times \ldots$$
$$P(T > 1 \mid T \geq 1) =$$
$$\prod_{a=1}^{j} (1 - P_h(a))$$

Most survival models are specified in terms of the hazard probabilities.

Drinking hazard probability



Sobriety survival probability



Likelihood for complete data

$$L_i = (P_h(T_i)) \prod_{a=1}^{T_i - 1} (1 - P_h(a))$$

$$= \prod_{a=1}^{T_i} (\Pr(E_a = e_{ia}))$$

$$\text{where } e_{ij} = \begin{cases} 1 \text{ if } T_i = j \\ 0 \text{ if } T_i > j \end{cases}$$

$$\widehat{P}_h(j) = \widehat{\Pr}(E_j = 1)$$

Censoring

- Missing data is endemic in longitudinal studies; survival studies are no exception.

- Various mechanisms for missing data in the survival context are referred to under the unifying term, *censoring*, indicating that the **event times** for some subjects are unknown to the researcher.

- ***Right censoring***: Occurs when a subject in the sample has *not* experienced the event of interest at the end of the observation period. It is assumed that the event eventually occurs sometime after the end of the study.
- ***Left censoring***: Occurs when a subject in the sample has experienced the event of interest *prior* to the onset of observation.
- ***Interval censoring***: Occurs when a subject is only known to have experienced the event of interest within a given time interval but the exact time is unknown.

- The most typical kind of missing data are right-censored and this type of missingness is the easiest to deal with in the data analysis.
- Censoring can be either *noninformative* or *informative* (analogous to *ignorable* and *nonignorable* in missing data terms). In conventional survival analysis, censoring is assumed to be *noninformative* which means that the distribution of censoring times is independent of event times, conditional on the set of observed covariates.

Observed data likelihood

$$L_i = \begin{cases} (P_h(T_i)) \prod_{a=1}^{T_i - 1} (1 - P_h(a)) \text{ if } T_i \le C_i \\ \prod_{a=1}^{C_i - 1} (1 - P_h(a)) \text{ if } T_i > C_i \end{cases}$$

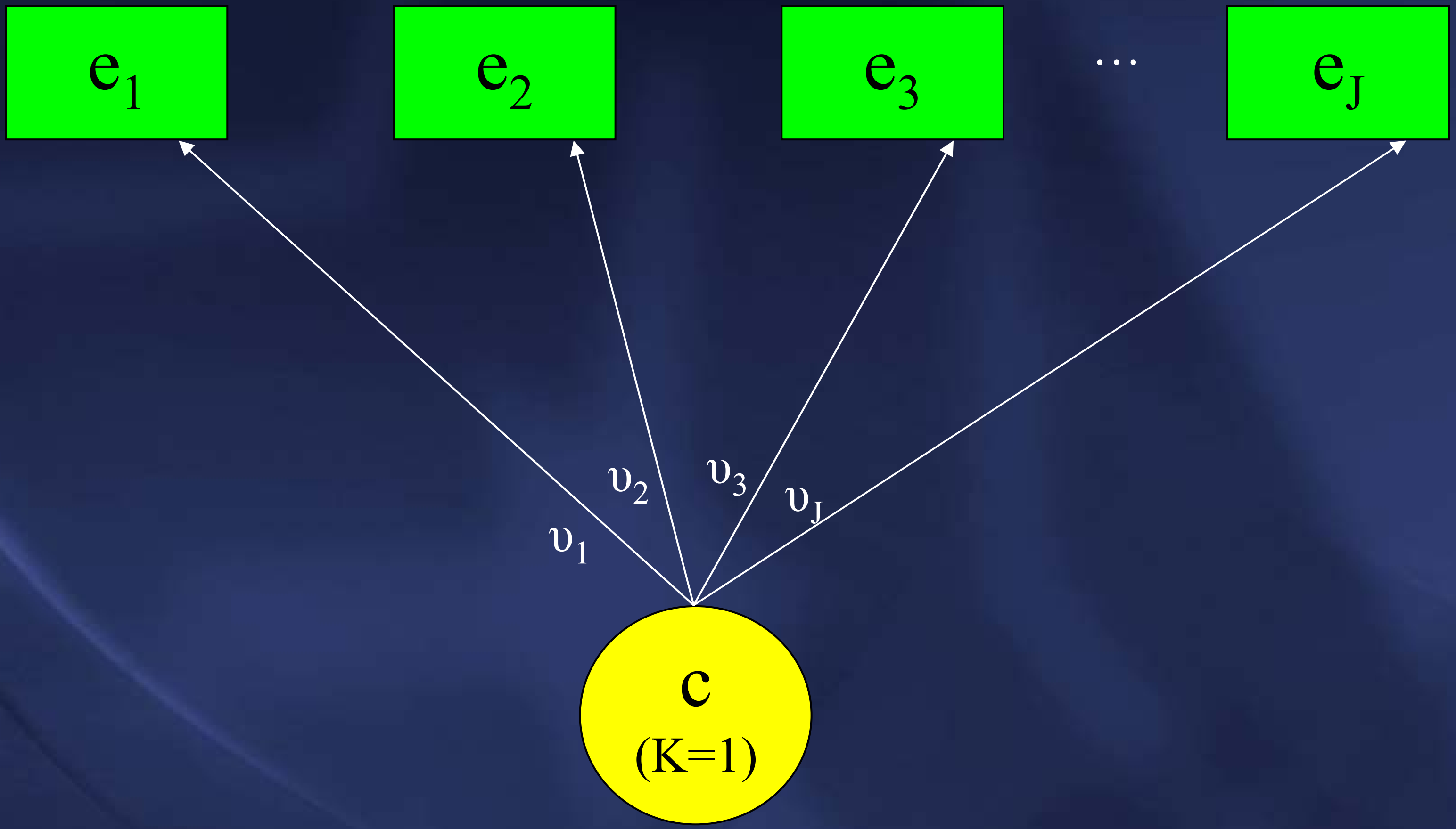
where T_i is the event time and C_i is the right-censoring time. T_i is only observed if $T_i \le C_i$.

$$L_i = \prod_{a \in \{1,..,J : e_{ia} \neq .\}} \Pr(E_a = e_{ia})$$

where $J = \max(\min(T_i, C_i), \forall i)$

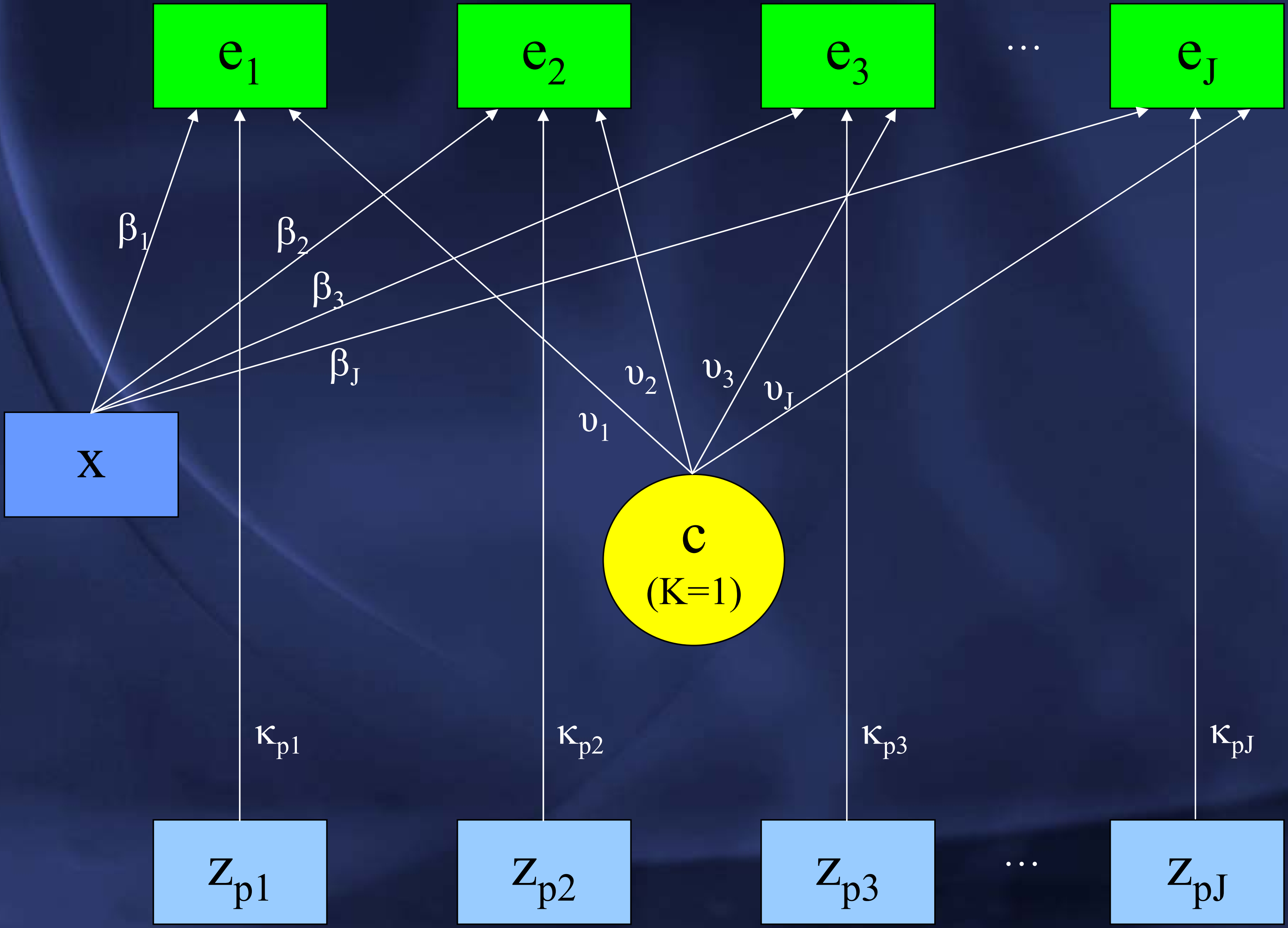
$$\text{and } e_{ij} = \begin{cases} 1 \text{ if } T_i = j \\ 0 \text{ if } T_i, C_i > j \\ . \text{ if } T_i < j \text{ or } C_i \leq j \end{cases}$$

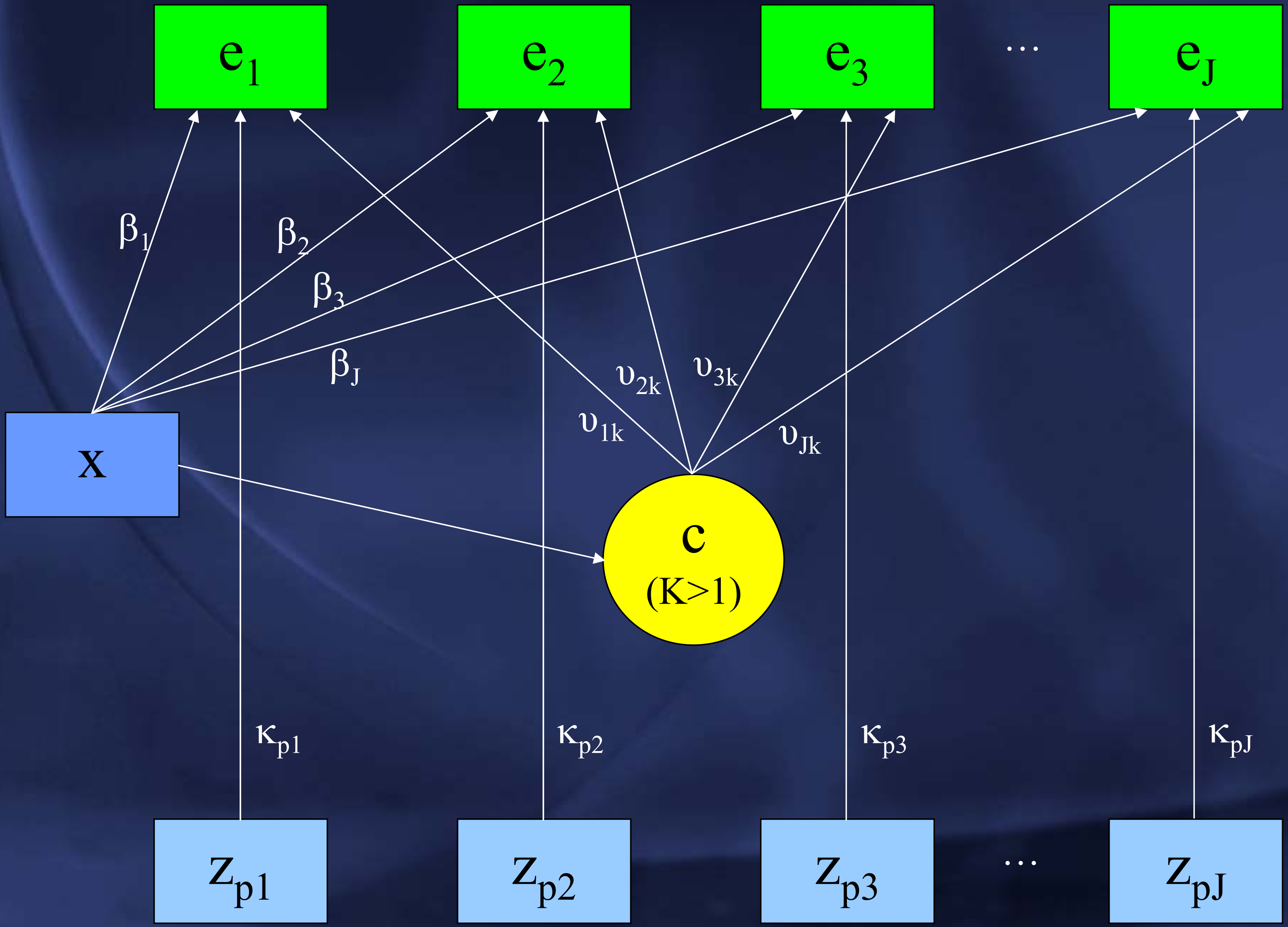
i	e_1	e_2	e_3	e_4	e_5
1	0	0	1	.	.
2	0	0	.	.	.
3	0	0	0	0	0



$$\widehat{P}_h(j) = \widehat{\Pr}(E_j = 1)$$

Maximum likelihood estimation under MAR


e_1
e_2
e_3
…
e_J
β_1
β_2
β_3
β_J
υ_1
υ_2
υ_3
υ_J
x
c
(K=1)
κ_{p1}
κ_{p2}
κ_{p3}
κ_{pJ}
Z_{p1}
Z_{p2}
Z_{p3}
…
Z_{pJ}


e_1
e_2
e_3
…
e_J
β_1
β_2
β_3
β_J
υ_{2k}
υ_{3k}
υ_{1k}
υ_{Jk}
x
c
(K>1)
κ_{p1}
κ_{p2}
κ_{p3}
κ_{pJ}
Z_{p1}
Z_{p2}
Z_{p3}
…
Z_{pJ}

Select references

- Masyn, K.E. (2003). *Discrete-time survival mixture analysis for single and recurrent events using latent variables.* Unpublished doctoral dissertation, University of California, Los Angeles.
- Muthén, B. & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, 30(1), 27-58.
- Muthén, L., & Muthén, B. (2006). Mplus User's Guide (Version 4.0). Los Angeles: Muthén & Muthén. (http://www.statmodel.com)
- Singer, J.D. & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.
- Vermunt, J.K. (2002). A general latent class approach to unobserved heterogeneity in the analysis of event history data. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 383-407). Cambridge: Cambridge University Press.

Missingness for single indicators of event occurrence

Risk status and event occurrence

- Estimating the hazard probability for a given time period depends on knowing who is *at-risk* at the beginning of the time period and who experiences the event during the time period.
- For (non-recurring) single or competing risks, anyone who has not experienced an event is considered *at-risk*.

- In classical survival analysis, it is assumed that event occurrence (the “if”) and, therefore, risk status, can be accurately determined even if the exact *timing* of the event (the “when”) is unknown or cannot be accurately and/or precisely assessed.
- If the occurrence of the event precludes further observation of the individual then if an individual misses a follow-up but returns at a later observation occasion, it could be inferred that the event had *not* occurred and that the individual is still at-risk.

- If event occurrence does not preclude further observation then if an individual misses a follow-up but returns at a later observation occasion, it could be ascertained where the event occurred during the absence from the study. If so, the individual’s event time would be interval censored. If not, the individual would still be at-risk.

- Sometimes, event occurrence does not bar subsequent observations but information about event occurrence during the period of absence from the study is missing. In this case, the risk status of the individual when he/she returns to the study is unknown.

Example

- Event: Age of first alcohol use
- Intake question: Have you ever used?
  - Only individuals answering “no” are at-risk for first use. The rest are left-censored.
- Yearly follow-up: How much alcohol have you consumed in the past year?
  - The first year an at-risk individual has an answer other than “none” marks the age of the event and termination of risk.

- If an individual, still at-risk, misses one or more years of follow-up prior to termination of risk, there is no way of knowing (from the given data) whether the event occurred during the one or more years that they were not questioned about yearly use and, thus, his/her risk status upon study reentry is unknown.

Example

- Event: Grade of first school removal (suspension or expulsion)
- Observation begins at T=0, when all subjects are first at-risk for the event.
- Yearly follow-up: Examination of school records in participating school districts for occurrence of school removal.

- If a child, still at-risk, moves out of the study area for one or more grades and then returns to the study area, it is unknown whether the child experienced a school removal during his/her time outside the study area and, thus, his/her risk status upon study reentry is unknown.

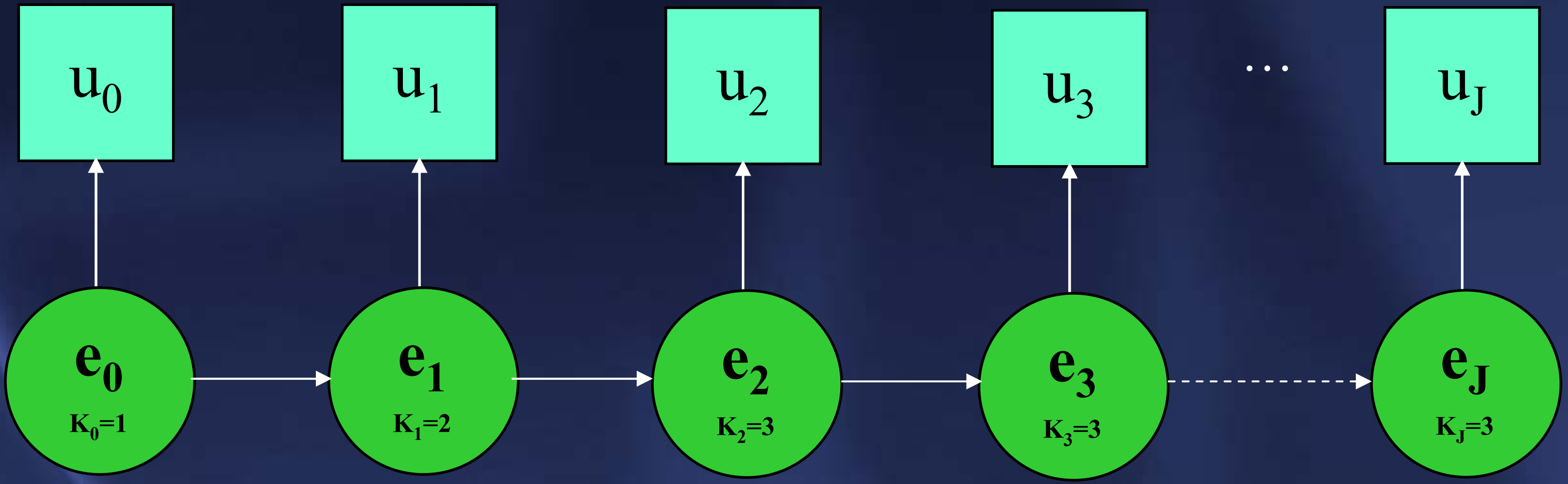
i	e_1	e_2	e_3	e_4	e_5
1	0	0	1	.	0
2	0	0	.	0	1
3	0	.	.	1	0

Solutions

- Right-censor at-risk individuals at the end of the last period of observation *before* study absence.
  - Will not bias results as long as the (unobserved) event time is independent of the timing of the study absence.
  - Doesn't use any observation following study reentry that could assist in the determination of whether the event occurred during the absence.
  - Prevents inclusion of those individuals in any simultaneous modeling of other outcomes conditional on event occurrence, e.g., use following onset, time between school removals, etc.

i	e_1	e_2	e_3	e_4	e_5
1	0	0	1	.	.
2	0	0	.	.	.
3	0	.	.	.	.

- Treat event occurrence as a transition between partially latent risk states
  - At T=0, all subjects are known to be in an *at-risk* state, e.g., never used alcohol, never been to school, etc.
  - All subjects have a non-zero probability of experiencing the event in the first time period and transitioning out of the *at-risk* state and into a non-risk state, i.e., a state represented by having experienced the event.
  - Subjects may return to a state that *looks* the same at the initial *at-risk* state in terms of observable behavior but is different because they are, by definition, no longer *at-risk*, e.g., a "never-user" initiates use and then returns to a state of non-use.



$e_j = 0$: Never-used

$e_j = 1$: Onset/recurrence

$e_j = 2$: Non-use

Measurement model:

$\Pr(u_0 = 0 \mid e_0 = 0) = 1$

$\Pr(u_1 = 0 \mid e_1 = 0) = 1$

$\Pr(u_1 = 1 \mid e_1 = 1) = 1$

$\Pr(u_j = 1 \mid e_j = 1) = 1$

$\Pr(u_j = 0 \mid e_j = 0 \text{ or } 2) = 1 \quad \text{for } j \geq 2$

Transition model:

Restrictions for T=0 to T=1:

τ_{01}	$e_1 = 0$	$e_1 = 1$	$e_1 = 2$
$e_0 = 0$	$1 – P_h(1)$	$P_h(1)$	.
$e_0 = 1$	.	.	.
$e_0 = 2$	.	.	.

Transition model:

Restrictions for T=1 to T=2:

τ_{12}	$e_2 = 0$	$e_2 = 1$	$e_2 = 2$
$e_1 = 0$	$1–P_h(2)$	$P_h(2)$	0
$e_1 = 1$	0	*	*
$e_1 = 2$	.	.	.

Transition model:

Restrictions for T=j to T=j+1 (j≥2):

$\tau_{j,j+1}$	$e_{j+1} = 0$	$e_{j+1} = 1$	$e_{j+1} = 2$
$e_j = 0$	$1–P_h(j+1)$	$P_h(j+1)$	0
$e_j = 1$	0	*	*
$e_j = 2$	0	*	*

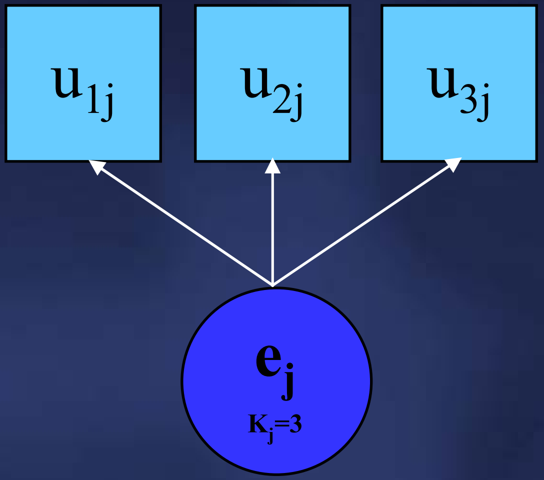
i	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$
1	$u_1=0$	$u_2=0$	$u_3=1$	$u_4=.$	$u_5=0$
	$e_1=0$	$e_2=0$	$e_3=1$	$e_4=1/2$	$e_5=2$
2	$u_1=0$	$u_2=0$	$u_3=.$	$u_4=0$	$u_5=1$
	$e_1=0$	$e_2=0$	$e_3=0/1$	$e_4=0/2$	$e_5=1$
3	$u_1=0$	$u_2=.$	$u_3=.$	$u_4=1$	$u_5=0$
	$e_1=0$	$e_2=0/1$	$e_3=0/1/2$	$e_4=1$	$e_5=2$

Measurement error for single and multiple indicators of event occurrence

Multiple indicators of event occurrence

- In some applications, event occurrence is inferred through indirect observation of the presence/absence of one or more symptoms that are used collectively (e.g., behavior checklist) to arrive at a "definitive" clinical diagnosis.
  - Time from first alcohol use to alcohol use disorder (AUD)
  - Time from treatment to AUD relapse
  - Onset age of first depressive episode
  - Duration of depressive episode

Quantifying error



- Symptom sensitivity: $P(u_{pj} = 1 \mid e_j = 1)$
- Symptom specificity: $P(u_{pj} = 0 \mid e_j \neq 1)$

The trouble with error

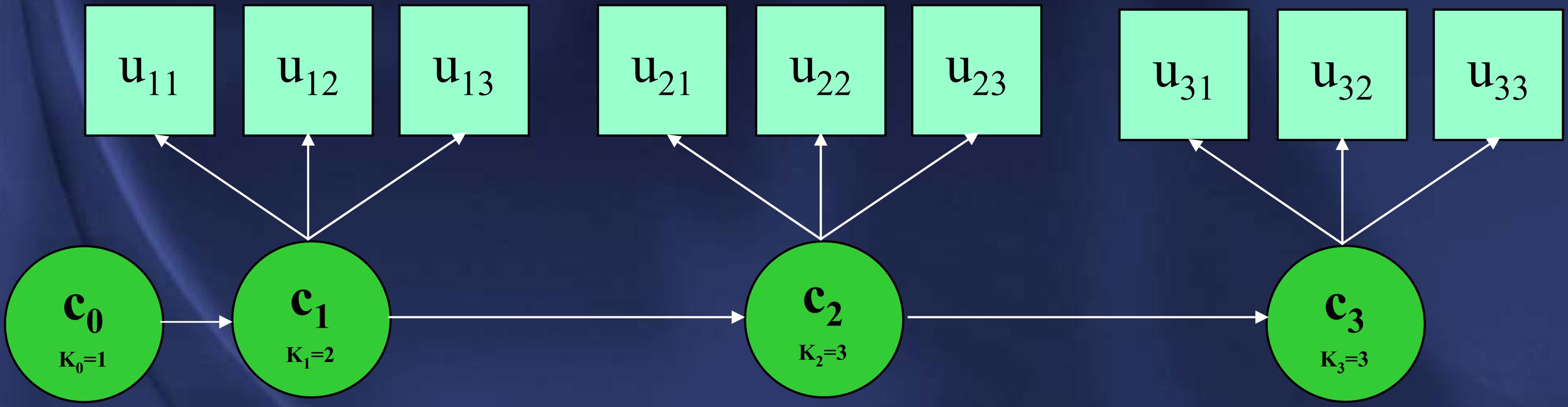
- Ignoring measurement error on event occurrence (and, thus, risk status) can results in *either* upward- *or* downward-biased hazard probability estimates.
- The impact of measurement error on the baseline hazard estimates depends on
  - Number of symptoms
  - Sensitivity and specificity of each symptom
  - “True” baseline rate

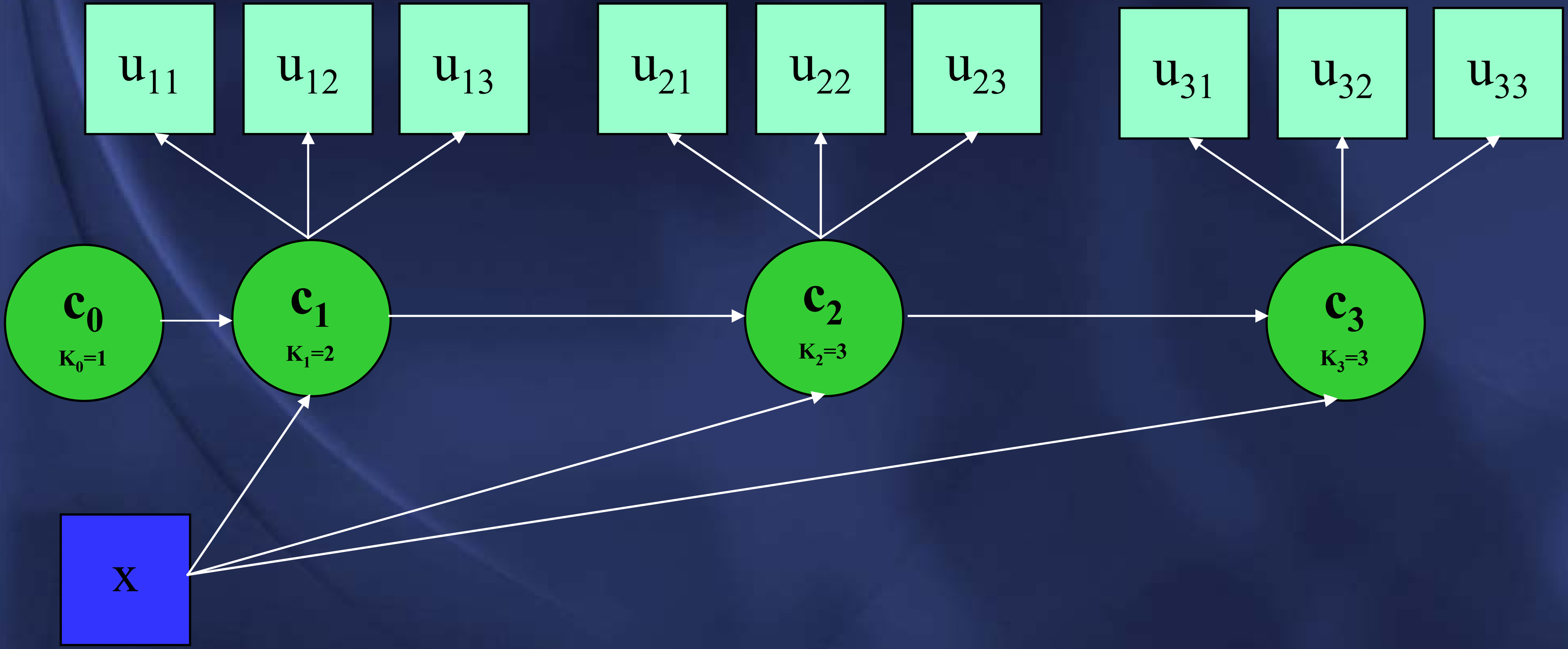
The trouble with error for the baseline hazard probability estimates:

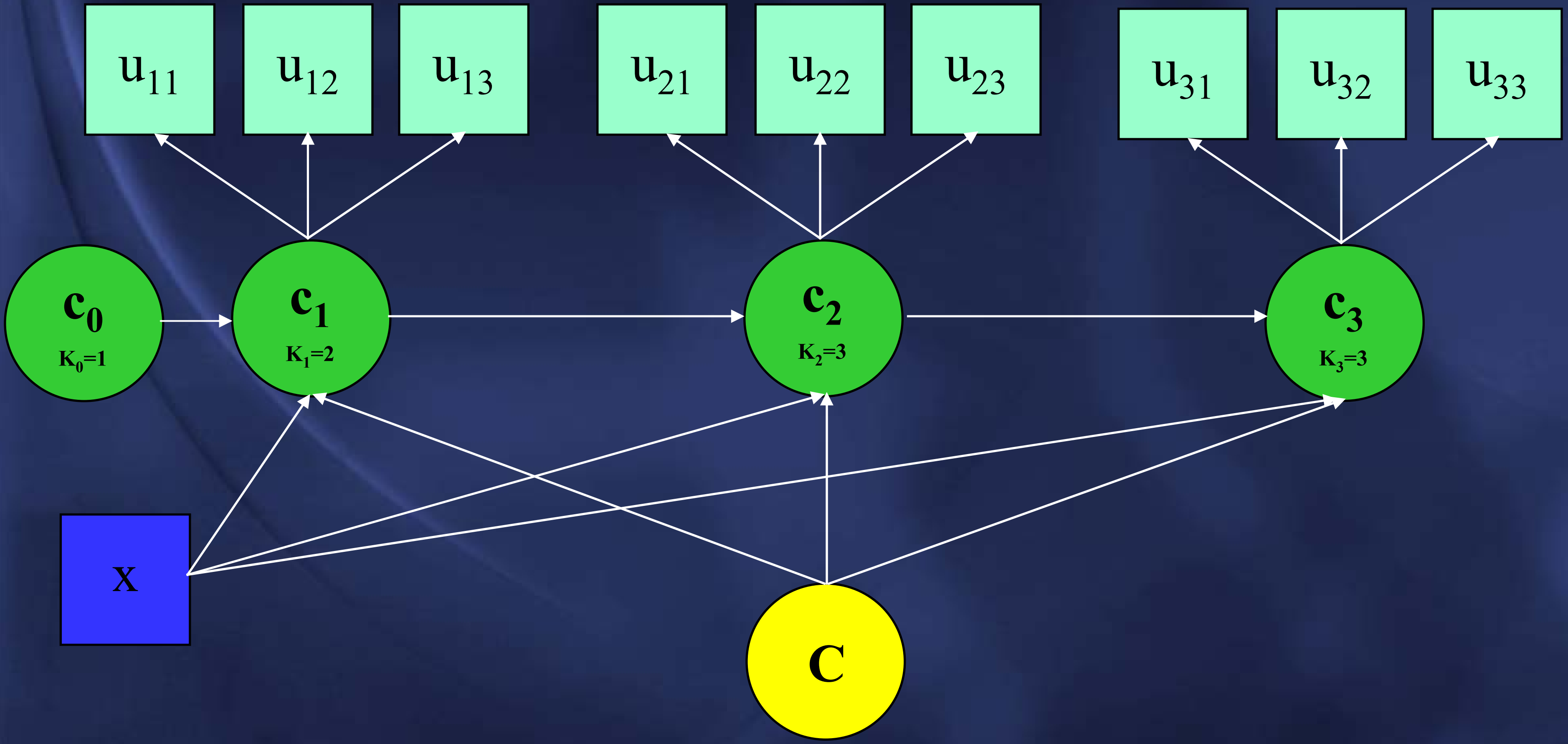
Symptom error		Baseline hazard probability		
Sensitivity	Specificity	**0.05**	**0.10**	**0.20**
1.0	**1.0**	**0.05**	**0.10**	**0.20**
0.9	**1.0**	**0.05**	**0.09**	**0.18**
1.0	**0.9**	**0.15**	**0.19**	**0.28**

The trouble with error for hazard odds ratio estimates:

Symptom error		Hazard odds ratio	
Sensitivity	Specificity	hOR=1.5	hOR=2.0
1.0	1.0	1.50	2.00
0.9	1.0	1.35	2.00
1.0	0.9	1.13	1.34


u_{11}
u_{12}
u_{13}
u_{21}
u_{22}
u_{23}
u_{31}
u_{32}
u_{33}
c_0
$K_0=1$
c_1
$K_1=2$
c_2
$K_2=3$
c_3
$K_3=3$


u_{11}
u_{12}
u_{13}
u_{21}
u_{22}
u_{23}
u_{31}
u_{32}
u_{33}
c_0
$K_0=1$
c_1
$K_1=2$
c_2
$K_2=3$
c_3
$K_3=3$
x


u_{11}
u_{12}
u_{13}
u_{21}
u_{22}
u_{23}
u_{31}
u_{32}
u_{33}
c_0 $K_0=1$
c_1 $K_1=2$
c_2 $K_2=3$
c_3 $K_3=3$
x
C