Multilevel Event History Analysis: Manifest and Latent Variable Modeling Approaches

Tenko Raykov
Michigan State University

Acknowledgments:
Prof. Dr. H.-P. Blossfeld,
European University Institute

***Citation of this booklet*: Raykov, 2014, Multilevel event history analysis: Manifest and latent variable approaches (short course). Florence, Italy: European University Institute.**

Plan:

1. **Resources and what this short course is about.**

2. **A brief review of needed fundamental concepts and relationships from event history analysis (EHA) and multilevel modeling (MLM).**

3. **In the beginning: What to take care of before one gets started with multilevel EHA (MEHA).**

4. **A manifest variable modeling approach to multilevel EHA (Day 1).**

5. **A latent variable modeling approach to multilevel EHA (Day 2).**

6. **Analysis of time-to-event (TTE) data from nationally representative samples (Day 2).**

7. **Conclusion.**

1. Resources and what this short course is about

MEHA is a relatively recent area of EHA and MLM. Unlike conventional/traditional EHA and MLM, there is substantially less literature (and published research) available currently on MEHA.

<u>Literature of relevance to workshop</u>:

Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis* (Second Edition). New York: Springer.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent linear and mixed models*. Boca Raton, FL: Chapman & Hall.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications.* New York: Taylor & Francis.

Muthén, L. K., & Muthén, B. (2012). *Mplus user's guide.* Los Angeles, CA: Muthén & Muthén.

<u>Software used</u>:

- **Stata,**
- **M*plus*.**

Further excellent software and (some) literature on MEHA are available as well.

<u>Aims of this workshop</u>:

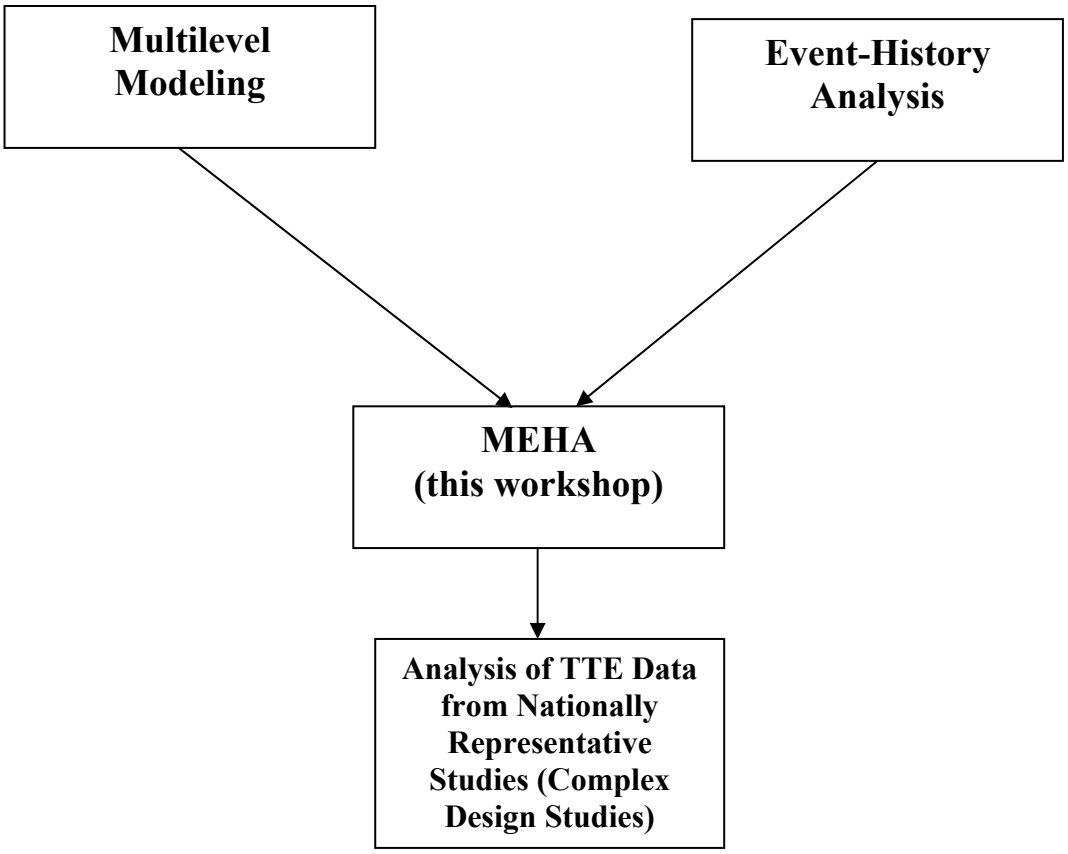
It is application-oriented but with a coherent discussion of theoretical issues involved, at a relatively non-technical level, and with some advanced features.

***Disclaimers*:**

1) **This is effectively a 'second' (rather than 'first', or introductory) workshop in EHA and/or multilevel modeling. Thus it assumes sufficient familiarity with both EHA and MLM, at the level of an introductory course in EHA and a short course in MLM. (A brief review, to brush up your memory of important facts for the rest, is provided in the next section of the workshop.)**

2) **Data sets are used in the workshop only for method illustration purposes, with no substantive conclusions aimed at (other than direct interpretation of obtained analytic results).**

3) **Pragmatically, this workshop 'marries' EHA and MLM, and a possible 'offspring' is the possibility to analyze time-to-event (TTE) data from nationally representative samples (last part of Day 2).**



What this workshop is about

The short course is concerned with

- **Event-history analysis (sociology and political science),**
- **Survival analysis (medicine),**
- **Duration analysis (economics),**
- **Failure analysis ('hard' sciences),**
- **Reliability analysis (engineering),**
- **Insurance claim (loans, defaults, payouts) analysis (business),**
- **or, more generally, Time-to-event analysis,**

when observations are nested/clustered/coming from hierarchies of units.

Statistically, the common underlying theme is how to analyze/model data from *positive* (non-negative) random variables stemming from observations that are *clustered* in some way.

Why can't we do it with what we already know, viz. with conventional or standard (traditional/classical) statistical methods?

Standard/traditional/conventional/classical applied statistical modeling does *not* handle the above setting because:

- **Regression analysis (standard) assumes independence,**
- **Multilevel modeling (standard) does not deal with data, which are not 'complete', like censored data (and that are *not* missing data),**
- **EHA (standard) does not deal with clustered data.**

Why interested in clustered data?

These data sets are richer in terms of information they contain. Also, oftentimes they 'naturally' arise in empirical research.

2. A brief review of needed fundamental concepts and relationships from event-history analysis and multilevel modeling

2.1. When and why should we consider EHA?

EHA = a set of methods for analyzing TTE data.

***Main question*: Does an *event* (marriage, child birth, unemployment) occur, and if so when? How is the time *T* that elapses until event occurs affected by/related to other personal variables? That is, *what are the characteristics that 'make' some people experience the event later/sooner than others*?**

***Question* 1: When should we consider EHA?**

***Answer*: If the research question contains 'whether' (an event occurs) and 'when', we're likely to need EHA/TTE analysis.**

***Question* 2: Why is EHA different from standard/classical/ traditional/conventional statistical methods?**

***Answer*: Because of censoring (partially missing information).**

Specifically, some persons experience the event during the study, while *others* (*i*) *don't experience it by the end of the study,* (*ii*) *are lost to follow up, or* (*iii*) *withdraw from the study*. They are called *censored* (observations).

For the rest of this short course, given a set of covariates used in a considered model or modeling effort, we assume non-informative right-censoring (in discrete-time TTE analysis, abbreviated 'DT-EHA', possibly with interval censoring).

2.2. Main concepts in EHA

0. *Time to event*: This is a random variable, T, defined at the individual observation level, such that $T > 0$ ($T \geq 0$). Its cumulative distribution function (cdf) is denoted $F(t)$ and its probability density function (pdf), assumed existing, is denoted $f(t)$ ($t \geq 0$ in the rest of this workshop).

1. *Survival function* (SF, denoted $S(t)$):

(2.1) $$S(t) = \Pr(T > t) = 1 - F(t).$$

2. *Hazard function* (HF, denoted $h(t)$; $h(t) \geq 0$ for all t):

(2.2) $$h(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

for the case of continuous time EHA (CT-EHA; see below for the discrete time case).

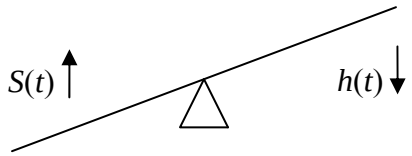
3. Either of the above three - SF, HF and F - *uniquely* characterizes the distribution of T. Formulas lead from one to the other (in the continuous time case as follows):

(2.3) $$h(t) = f(t)/S(t) = -S'(t)/S(t) \text{ , and}$$

(2.4) $$S(t) = \exp(-\int_0^t h(u)du) = \exp(-A(t)) \text{ ,}$$

where $A(t) = \int_0^t h(u)du$ is called *cumulative hazard*.

4. The hazard function is conceptually focused on failing, while the survival function is so on surviving, and they play a 'see-saw' game:



5. In discrete time EHA (DT-EHA), the counterpart of the hazard function is defined as follows:

(2.5) $$HF(j) = h_j = Pr(T = j \mid T > j - 1),$$

where *Pr* denotes probability and *T* is the time to event that is measured discretely, with *P* 'observational periods', $j = 1, \ldots, P$; note this *conditional* probability!

That is, $HF(j) = h_j$ = (conditional) probability to experience the event in time period *j* (i.e., observation interval *j*) *given* that it was not experienced before ($j = 1, \ldots, P$).

<u>Note</u>: $0 \leq h_j \leq 1$ in DT-EHA, since it is a probability. This is a main difference to continuous-time EHA (abbr. CT-EHA), where the hazard (HF(*t*)) can be any non-negative number (at a given point in time, *t*).

6. In DT-EHA, the SF is defined as follows ($j = 1, \ldots, P$):

(2.6) $$S_j = \Pr(T > j).$$

That is, SF = probability of surviving through period *j*, i.e., not experiencing the event up until period *j* and during that period (see next note).

Note: While the HF is a conditional probability, SF is an *un*conditional probability (in DT-EHA). Keep also in mind that in CT-EHA the hazard function is a conditional density, i.e., a conditional pdf.

7. **In DT-EHA, the relationship between HF and SF (for a given time period, say *j*th) is:**

(2.7) $S_j = (1-h_1).(1-h_2). \ldots . (1-h_j)$ **(*j* = 1, …, *P*).**

2.4. Fundamental notions of regression analysis and multilevel analysis

2.4.1. Regression analysis (general linear model)

In conventional regression analysis (RA), most popular is the ordinary least squares (OLS) setup (with *k* 'predictors', independent variables, explanatory variables, or regressors, generally referred to as 'covariates'; *k* > 0):

(2.8) $\underline{Y} = X \underline{\beta} + \underline{e}$,

where $\underline{e} \sim \aleph(0, \sigma^2 I_n)$ **(ℵ = normal distribution, *n* = sample size), and observations are assumed *in*dependent of each other. All these are called *OLS assumptions*. (Underlining is used to signal vector throughout this booklet.)**

Then the OLS parameter estimator, denoted $\hat{\underline{\beta}}$, and associated variance matrix (squared standard errors for the model parameters are along its main diagonal) are:

(2.9) $\hat{\underline{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\underline{\mathbf{Y}}$, **and**

(2.10) $\mathrm{Var}(\hat{\underline{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}\ (=: \mathbf{H}^{-1})$,

where

(2.11) s^2 **= sum of squared residuals/(*n*-*k*-1).**

When the OLS assumptions do <u>not</u> hold, (i) the *OLS estimator* (2.9) above is *<u>still</u> consistent* (also with non-normality); but (ii) the *standard errors* (SEs) for its elements, in Equation (2.10), are *no more accurate*.

Hence, the OLS SEs can<u>not</u> be trusted then for carrying out significance tests and obtaining confidence intervals for parameters of interest (e.g., with hierarchical data).

In those cases, one can use the "sandwich estimator" for the SEs (Huber/White estimator), yielding robust SEs:

(2.12) $\mathbf{V}_r(\hat{\underline{\beta}}) = \mathbf{H}^{-1}\,\mathbf{C}\,\mathbf{H}^{-1}$.

In Equation (2.12) that presents the robust SEs for each element of the parameter vector $\underline{\beta}$ in the original model (2.8) (viz., their squares lie along the main diagonal of this matrix $\mathbf{V}_r$), C is a correction matrix based on the observed raw residuals that is worked out by the used software (e.g., Stata).

If the OLS assumptions hold, the right-hand sides of Equations (2.10) and (2.12) produce each a consistent estimator of the covariance matrix of the model parameters in $\underline{\beta}$, but the OLS-based covariance matrix in Equation (2.10) is more efficient (i.e., yields smaller SEs).

If the OLS assumptions do *not* hold, only Equation (2.12) – i.e., only that sandwich formula – yields a consistent estimator of the covariance matrix of the parameters (and their SEs), i.e., *only* the sandwich estimator is consistent.

As usual in applications of statistics (regardless of methodological framework), there is the following important trade-off to keep in mind when dealing with assumption violations:

***Trade-off*: Robustness vs. Efficiency.**

2.4.2. Multilevel modeling

A multilevel model extends the OLS setup to that of clustered data (i = level-1 units, j = level-2 units, etc.).

How? – Start with the conventional RA model (note 'double-indexing' next; $i = 1, \ldots, n_j$; $j = 1, \ldots, J$):

(2.13) $$Y_{ij} = b_0 + b_1 X_{1,ij} + \ldots + b_p X_{p,ij} + e_{ij}$$

In MLM, a 'counterpart' of this model is

(2.14) $$Y_{ij} = b_{0j} + b_1 X_{1,ij} + \ldots + b_p X_{p,ij} + e_{ij} ,$$

which is referred to as *random-intercept model* (RIM), or

(2.15) $$Y_{ij} = b_{0j} + b_{1,j} X_{1,ij} + \ldots + b_{p,j} X_{p,ij} + e_{ij} ,$$

called the *random regression model* (RRM).

What makes the differences to standard RA, is the way the error term in the last 2 equations are treated in MLM (see earlier w/shop, 2013).

When the response variable(s) Y is highly discrete (binary), move to its mean that is then probability (binary Y) and postulate the following model:

(2.16) $$\text{logit}(p_{ij}) = b_{0j} + b_1 X_{1,ij} + \ldots + b_p X_{p,ij} ,$$

(RIM with discrete outcome, RIMDO)

or

(2.17) $$\text{logit}(p_{ij}) = b_{0j} + b_{1j} X_{1,ij} + \ldots + b_{pj} X_{p,ij}$$

(RRM with discrete outcome, RRMDO).

In an empirical study, bear in mind that one may need to restrict the number of random slopes to the largest number sustainable by the given data set, in order to avoid numerical difficulties (for a given data set).

The preceding discussion in this section provides a basis and a reference point to keep in mind while moving on with the following matters discussed in this course.

With this brief review, we are now ready to embark on studying multilevel EHA, after paying next the due attention however to "first things first".

3. In the beginning: What to take care of before one gets started with MEHA

This workshop assumes that the following very important matters have been already resolved appropriately, based on the research question(s) and available substantive knowledge in a given subject-matter domain of EHA application.

I am mentioning them explicitly next, since not resolving properly even one of these, is a prescription for incorrect (in addition to frustrating) EHA analyses, whether in single-level or multi-level contexts.

In particular, this workshop will NOT be concerned with - and thus will NOT aim at giving answers to - any of the following BIG NINE issues of EHA:

1. **Deciding who to study (and obtaining a representative sample from the target population).**
2. **Defining possible states and the target event (as well as possible nesting units, i.e., level-2 units, if present).**
3. **Identifying the beginning of time (origin).**
4. **Selecting the length of the data collection process (overall study period).**
5. **Choosing intervals for data collection (in DT-EHA settings in particular).**
6. **Reconstructing event histories in retrospective studies (instead, one is well advised to conduct prospective studies whenever possible).**
7. **Minimizing attrition.**
8. **Determining how many people to study.**
9. **Which covariates to handle as time-invariant and/or which covariates as time-varying.**

It is of paramount importance that each and every of these issues be resolved properly *before* one starts with EHA (MEHA). This resolution should be accomplished using all available substantive/ theoretical knowledge in the subject-matter domain of relevance, as well as in accordance with considered research question(s) and the specific study design/characteristics.

In particular, *if at a subsequent point in time/analysis or modeling stage one is having difficulties using the method(s) that follow in this workshop, then chances are that some of the BIG NINE (or closely related issues specific to the study) have not been dealt with properly/adequately*.

The criteria for having properly handled each and every of the BIG NINE are the following: (a) the research question(s), (b) all available knowledge in the subject-matter domain of concern, and (c) the particular study design characteristics and related features.

Pertinent aspects of how each of the Big Nine is resolved in EHA (single-level), in the context of a research question(s) of relevance, is typically discussed in an introductory course to EHA (which this one is not meant to be).

4. A manifest variable modeling approach

In this main workshop section, we discuss initially the setting of discrete-time EHA (DT-EHA), which is arguably the more often applicable framework in the social and behavioral sciences. Then we will move on to the continuous time framework (CT-EHA, in Day 2).

Throughout this section, we assume that *explanatory variables (covariates) are measured without error*. This is in fact (still) a routinely made assumption. We will relax it in the next main Section 5 (Day 2).

4.1. An important question at the 'border' of single- and two-level EHA

Suppose we have data either in the continuous or discrete time setting, and we decide to impose on the time elapsed since start of the study (small) successive interval windows until 'covering':

(a) the event for each person observed, or
(b) his/her last available observation, or
(c) the end point of the study.

In this way, for each subject, we create a number of consecutive records, viz. as many as the number of such successive observation windows needed, to 'cover' the event for him/her, his/her last observation available, or the end of the overall study period.

The question then is, *how should we analyze the resulting 'vertical' data set* (univariate, stacked, strung-out data set)?

The answer is in the following discussion. This set up, we will see soon, is actually one where MEHA is of direct relevance (incl. in particular the case of time-varying covariates; see below in this booklet).

4.1.1. Data used

To get us started, I should like to commence with an empirical example, using data contained in the file 'prom1.dta'.

This is a data set resulting from a study of how long it took *n* = 301 assistant professors at research universities in the US to get promotion to associate professor (usually with tenure, but the latter will not be of relevance for us in the remainder of this section).

We read in first the data set and check to see its variables, as well as the actual data of the first 10 persons say, as follows (Stata commands are given in this w/shop handout in red while output produced is provided following them in black and slightly smaller size, both in font Courier New; see below for variable notation and names):

```
. use "C:\T E A C H\Workshops\EUI\EHA\prom1.dta", clear

. d

Contains data from C:\T E A C H\Workshops\Italy\SA\prom1.dta
  obs:           301
 vars:             3                          30 Sep 2013 11:48
 size:         3,612
-------------------------------------------------------------------------------
storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g
dur             float  %9.0g
event           float  %9.0g
-------------------------------------------------------------------------------
Sorted by:

. list in 1/10
```

	id	dur	event
1.	1	10	0
2.	2	4	1
3.	3	4	0
4.	4	10	0
5.	5	7	1
6.	6	6	1
7.	7	10	0
8.	8	6	1
9.	9	4	0
10.	10	5	1

***Note*: The variable notation used is as follows: id = identifier for subject (row), 'dur' = number of years it took him/her to get promotion or become censored (i.e., # years he/she was observed for/in the study); event = promotion (value = 1) or (right-)censored (value = 0); note that 'event' is the 'status' variable and 'dur' the time to event here, in conventional EHA/survival/duration analysis nomenclature.**

I should like to stress that the last presented data format is what may be considered the 'usual format' in which we get EHA-related data in empirical research. (It typically takes considerably less time to produce/enter a data set in this format than in any other format, e.g., the following one.)

This format is often referred to as 'flat', 'horizontal', or 'multivariate' format, of relevance to some software (but Stata, for MEHA).

One of the first questions we ask when dealing with EHA, is what the *hazard* for the event of interest is. This is because as we know well by now, hazard is a central concept of relevance in EHA (e.g., Blossfeld, Golsch, & Rohwer, 2007, *EHA w/ Stata*).

For our empirical example under consideration, we can get the hazard for experiencing the event (promotion) as follows.

```
. ltable dur event, hazard noadjust
```

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf.	Int.]
1	2	301	0.0033	0.0033	0.0033	0.0033	0.0001	0.0123
2	3	299	0.0067	0.0047	0.0033	0.0033	0.0001	0.0123
3	4	292	0.0645	0.0143	0.0582	0.0141	0.0339	0.0890
4	5	263	0.2139	0.0243	0.1597	0.0246	0.1151	0.2115
5	6	211	0.4113	0.0297	0.2512	0.0345	0.1882	0.3232
6	7	149	0.5931	0.0303	0.3087	0.0455	0.2260	0.4041
7	8	96	0.7245	0.0282	0.3229	0.0580	0.2194	0.4461
8	9	59	0.7945	0.0262	0.2542	0.0656	0.1423	0.3981
9	10	42	0.8288	0.0248	0.1667	0.0630	0.0670	0.3109
10	11	29	0.8524	0.0241	0.1379	0.0690	0.0376	0.3023

From these analytic results, we see that the estimated hazard reaches its maximum in year 7 and then declines. (This finding can be explained by the fact that in the US, the assistant-to-associate professor promotion usually occurs in the 7th year after commencing work at one's current university, but for now this is a tangential finding.)

***Note*. If interested in the survival (survivor) function, use the same command, dropping 'hazard' as a subcommand. (We will not pursue this function here, though, as it less interesting than the hazard function in this example.)**

To respond to the starting question we posed above (see italicized red question on p. 16), we need to conduct a particular data restructuring that facilitates obtaining an answer to that question, which restructuring is described next.

4.1.2. The needed data reformatting

This data management activity is achieved by 'expanding' the data set, which re-expresses each person's data by as many rows as the number of years he/she was observed (we have data on for him/her), i.e., for each year he/she was at risk for the event in question (promotion).

The resulting data format is often referred to as 'univariate', 'vertical', 'stacked', 'long' format, and is typically used in multilevel modeling analyses (e.g., with Stata or SAS).

We accomplish this re-formatting as follows.

```
. expand dur
(1440 observations created)
```

Note that no output is produced by this command, since the only activity was this 'expansion' of the original data set (with no information lost or added to it).

To interpret meaningfully the result of this activity, we need to create a new variable for the year (the observational window of our original data collection procedure), which is achieved this way:

```
. by id, sort: gen yr = _n
```

We can now see the results of our above activities conducted thus far on the original data set by listing this data re-expression for the first 3 persons say:

```
. list id dur yr event if id<4
```

	id	dur	yr	event
1.	1	10	1	0
2.	1	10	2	0
3.	1	10	3	0
4.	1	10	4	0
5.	1	10	5	0
6.	1	10	6	0
7.	1	10	7	0
8.	1	10	8	0
9.	1	10	9	0
10.	1	10	10	0
11.	2	4	1	1
12.	2	4	2	1
13.	2	4	3	1
14.	2	4	4	1
15.	3	4	1	0
16.	3	4	2	0
17.	3	4	3	0
18.	3	4	4	0

As can be seen, the variable ‘event’ was treated in this re-formatting as a time-invariant measure, but no information of relevance to us has been lost thereby (or added, for that matter) in the resulting expanded data set. (Hence, in this long-formatted file, ‘event’ does not really have the meaning of ‘event’ anymore that it had before the expansion we just carried out - so don’t pay attention to variable ‘event’ here.)

For our purposes below, in order to be in a position to respond to the critical question of interest (p. 16, bottom), we must have an explicit variable saying if promotion occurred or not in any given year (of observation for a particular person). Hence, let’s generate it – and then see the result of this activity:

```
. gen y=0

. replace y=event if yr==dur
(217 real changes made)

. list id dur yr event y if id<4
```

	id	dur	yr	event	y
1.	1	10	1	0	0
2.	1	10	2	0	0
3.	1	10	3	0	0
4.	1	10	4	0	0
5.	1	10	5	0	0
6.	1	10	6	0	0
7.	1	10	7	0	0
8.	1	10	8	0	0
9.	1	10	9	0	0
10.	1	10	10	0	0
11.	2	4	1	1	0
12.	2	4	2	1	0
13.	2	4	3	1	0
14.	2	4	4	1	1
15.	3	4	1	0	0
16.	3	4	2	0	0
17.	3	4	3	0	0
18.	3	4	4	0	0

On this data set, we can also estimate the hazards for each of the 10 years in question (viz. for the maximal amount of time that each person could have been observed for; compare with the 'Hazard' column of the life-table we got earlier on the original data set, which hazard estimates were presented on p. 19):

```
. tab yr y, row
```

Key
frequency
row percentage

yr	y 0	y 1	Total
1	300	1	301
	99.67	0.33	100.00
2	298	1	299
	99.67	0.33	100.00
3	275	17	292
	94.18	5.82	100.00
4	221	42	263
	84.03	15.97	100.00
5	158	53	211
	74.88	25.12	100.00
6	103	46	149
	69.13	30.87	100.00
7	65	31	96
	67.71	32.29	100.00
8	44	15	59
	74.58	25.42	100.00
9	35	7	42
	83.33	16.67	100.00
10	25	4	29
	86.21	13.79	100.00
Total	1,524	217	1,741
	87.54	12.46	100.00

(I should like to note in passing here that this hazard estimation will become much more 'exciting' when we add covariates in the following model, as done later in this section. Also, since we have the same results in the 2nd last column here as in the 'Hazard' column on p. 19, this data expansion was properly carried out and we trust it next.)

Now we're ready to begin responding specifically to the analysis question asked at the beginning of this section 4.1. (I should underscore that the data set of concern is now in the form that's referred to in that question on p. 16.)

The direct answer to that question is the following:

Use logistic regression with a dependent variable being 'y' and independent variables being the dummies for the event occurring at year *t* (*t* = 1, …, 10 here; *i* = 1, …, 301 = sample size), dropping the first of these indicators to avoid multicollinearity:

$$\textbf{(4.1)} \qquad \text{logit}[P(y_{ti} = 1|d_{ti})] = \beta_0 + \beta_2\, d_{2,ti} + \ldots + \beta_{10}\, d_{10,ti}\,.$$

What's behind this analysis?

We are actually carrying out with it a form of two-level EHA modeling, by using single-level modeling of a binary response (denoted above 'y'), while accounting for the fact that the repeated observations for successive years (across the newly created rows per subject) are in fact nested within person.

(Stata internally creates these dummies and automatically omits the first of them in the analyses carried out below; see next subsection.)

With respect to this modeling, we need to consider 2 cases next.

4.2. The case of no covariates

To conduct the two-level modeling discussed, with no covariates, all we need to do is use conventional single-level modeling, i.e., logistic regression with the Stata 'logit' command (see further below for the case with covariates, covered in the next subsection 4.3):

```
. logit y i.yr

Iteration 0:   log likelihood = -654.73963
Iteration 1:   log likelihood = -561.40534
Iteration 2:   log likelihood =  -532.9051
Iteration 3:   log likelihood = -529.43483
Iteration 4:   log likelihood = -529.15799
Iteration 5:   log likelihood = -529.15641
Iteration 6:   log likelihood = -529.15641

Logistic regression                                     Number of obs   =       1741
                                                        LR chi2(9)      =     251.17
                                                        Prob > chi2     =     0.0000
Log likelihood = -529.15641                             Pseudo R2       =     0.1918

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          yr |
           2 |    .006689   1.416576     0.00   0.996     -2.76975    2.783128
           3 |   2.920224   1.032372     2.83   0.005     .8968116    4.943637
           4 |   4.043289   1.015710     3.98   0.000     2.052533    6.034045
           5 |   4.611479   1.014165     4.55   0.000     2.623753    6.599205
           6 |   4.897695   1.017242     4.81   0.000     2.903937    6.891452
           7 |   4.963382   1.025171     4.84   0.000     2.954084     6.97268
           8 |   4.627643   1.045336     4.43   0.000     2.578822    6.676463
           9 |   4.094344   1.083864     3.78   0.000     1.970009    6.218679
          10 |   3.871201   1.137248     3.40   0.001     1.642236    6.100166
             |
       _cons |  -5.703782   1.001665    -5.69   0.000     -7.66701   -3.740555
------------------------------------------------------------------------------
```

We see here a fairly strong contribution of year 7 to the probability of event, as we could expect (given that the cycle of promotion in the US is 7 years, as mentioned earlier).

But, while looking at these results, we may also wish to ask the following query that almost immediately pops up. We used here the conventional (rather than formally a multilevel) modeling approach, specifically logistic regression (and the Stata command 'logit' rather than 'melogit' that is its multilevel counterpart).

Here's that question:

***Query*: By doing this single-level modeling/analysis, didn't we forget about - if not ignored completely - the fact that the successive rows/observations per subject were in fact nested within him/her?**

Answer - No, we didn't!

How come?

This is because we actually took care of this nesting, by using the above dummies d_2, …, d_{10}, which are interrelated among themselves (by their construction as such; see next).

We can take a quick look at these dummies, to see why:

```
. list i.yr in 1/20
```

	1b. yr	2. yr	3. yr	4. yr	5. yr	6. yr	7. yr	8. yr	9. yr	10. yr
1.	0	0	0	0	0	0	0	0	0	0
2.	0	1	0	0	0	0	0	0	0	0
3.	0	0	1	0	0	0	0	0	0	0
4.	0	0	0	1	0	0	0	0	0	0
5.	0	0	0	0	1	0	0	0	0	0
6.	0	0	0	0	0	1	0	0	0	0
7.	0	0	0	0	0	0	1	0	0	0
8.	0	0	0	0	0	0	0	1	0	0
9.	0	0	0	0	0	0	0	0	1	0
10.	0	0	0	0	0	0	0	0	0	1
11.	0	0	0	0	0	0	0	0	0	0
12.	0	1	0	0	0	0	0	0	0	0
13.	0	0	1	0	0	0	0	0	0	0
14.	0	0	0	1	0	0	0	0	0	0
15.	0	0	0	0	0	0	0	0	0	0
16.	0	1	0	0	0	0	0	0	0	0
17.	0	0	1	0	0	0	0	0	0	0
18.	0	0	0	1	0	0	0	0	0	0
19.	0	0	0	0	0	0	0	0	0	0
20.	0	1	0	0	0	0	0	0	0	0

Hence *we have actually conducted here multilevel discrete-time EHA*, specifically *by employing a 'conventional' logistic regression approach* (and thus the single-level logistic regression command 'logit'), while accounting for the within-person dependencies of the rows of the expanded data set through the interrelationships between these dummies, d_2, …, d_{10}.

Next, if we want to get the estimated hazards, they're nothing but predicted probabilities within this model (the last one fitted; they're listed next for the first 3 say subjects):

```
. predict est_haz, pr

. list id yr est_haz y event if id<4
```

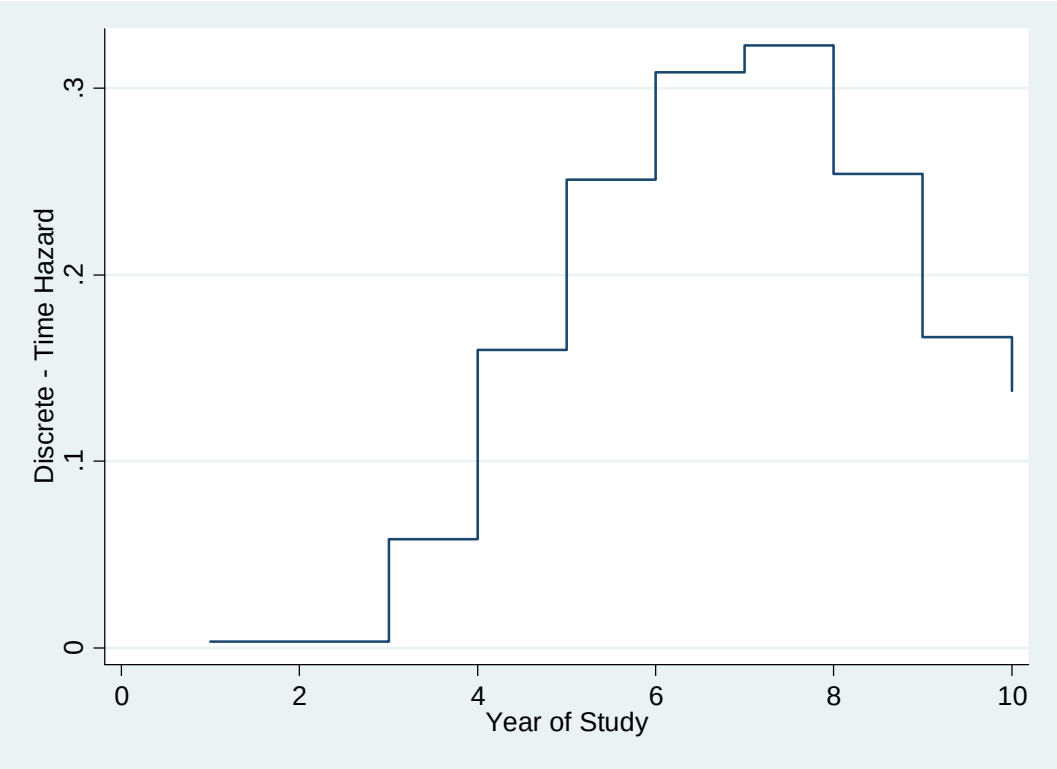
	id	yr	est_haz	y	event
1.	1	1	.0033223	0	0
2.	1	2	.0033445	0	0
3.	1	3	.0582192	0	0
4.	1	4	.1596958	0	0
5.	1	5	.2511848	0	0
6.	1	6	.3087248	0	0
7.	1	7	.3229167	0	0
8.	1	8	.2542373	0	0
9.	1	9	.1666667	0	0
10.	1	10	.1379310	0	0
11.	2	1	.0033223	0	1
12.	2	2	.0033445	0	1
13.	2	3	.0582192	0	1
14.	2	4	.1596958	1	1
15.	3	1	.0033223	0	0
16.	3	2	.0033445	0	0
17.	3	3	.0582192	0	0
18.	3	4	.1596958	0	0

Notice that these are the same hazards as estimated earlier from the original data set and presented on p. 19 and then again on p. 23.

Hence, we have not added or lost any information from our sample with the analysis just conducted (consisting in fitting model (4.1)).

Let's see how this estimated hazard looks like graphically, for someone with observations across all 10 years involved in this study (like the first subject; the resulting hazard plot is in the form of a stairway/step-function):

```
. twoway (line est_haz yr if id==1, connect(stairstep)),
> legend(off) xtitle(Year of Study) ytitle (Discrete-Time Hazard)
```



In conclusion of this subsection, note that by fitting the above model (4.1), we have in fact conducted essentially a *no-assumption, two-level DT-EHA*.

This is because we did not impose any functional form upon the relationship between the hazard, on the one hand, and time (year), on the other hand.

The reason is that we used all information in the dummies that represent uniquely time (year).

4.3. Covariates included in the model (time-invariant covariates)

If we want to study the time-to-event in the context of several other measures, i.e., controlling for individual differences on the latter, we add them in the above logistic regression model in Equation (4.1), leading to the extended model

$$\text{(4.2)} \qquad \text{logit}(P(y_{ti} = 1|d_{ti}, \underline{x}) = \beta_0 + \beta_2\, d_{2,ti} + \ldots + \beta_{10}\, d_{10,ti} + \gamma_1\, x_{1,i} + \ldots + \gamma_p\, x_{p,i} \qquad (p > 0).$$

The first part/line on the right-hand side of Equation (4.2) represents the *baseline hazard* (BH), i.e., the hazard if all covariates were 0.

In the 2nd part/line of Equation (4.2), we assume that the effects of the covariates are linear and additive, as well as parameterized in the corresponding partial regression coefficients $\gamma_1, \ldots, \gamma_p$.

As earlier in this section, we can fit this model using formally a single-level logistic regression analysis approach, and the pertinent command 'logit' in Stata.

To exemplify, for the empirical study under consideration, suppose we want to include as covariates (i.e., control for individual differences on) the following available measures in its data set:

- **selectivity of the university, called 'undgrad' in the data file 'prom2.dta';**

- **whether the professor has a Ph.D. degree from a medical university (college), called 'phdmed' there; and**

- **a measure of the prestige/reputation of the PhD awarding institution, called 'phdprest' in the data file 'prom2.dta'.**

We accomplish this aim as follows.

```
. logit y i.yr undgrad phdmed phdprest

Iteration 0:   log likelihood = -654.73963
Iteration 1:   log likelihood = -557.63219
Iteration 2:   log likelihood = -528.19908
Iteration 3:   log likelihood = -524.62994
Iteration 4:   log likelihood = -524.34451
Iteration 5:   log likelihood = -524.34273
Iteration 6:   log likelihood = -524.34273

Logistic regression                                     Number of obs   =       1741
                                                        LR chi2(12)     =     260.79
                                                        Prob > chi2     =     0.0000
Log likelihood = -524.34273                             Pseudo R2       =     0.1992

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          yr |
           2  |   .0063043   1.416703     0.00   0.996    -2.770383    2.782991
           3  |   2.923036   1.032552     2.83   0.005      .899272    4.946801
           4  |   4.051509   1.015925     3.99   0.000     2.060334    6.042685
           5  |   4.619307   1.014410     4.55   0.000     2.631099    6.607514
           6  |   4.924625   1.017614     4.84   0.000     2.930139    6.919111
           7  |   4.992068   1.025703     4.87   0.000     2.981728    7.002409
           8  |    4.69053   1.046322     4.48   0.000     2.639776    6.741284
           9  |   4.167769   1.085192     3.84   0.000     2.040833    6.294706
          10  |   3.964635   1.139095     3.48   0.001      1.73205    6.197219
              |
     undgrad  |   .1576609     .06007     2.62   0.009      .039926    .2753959
      phdmed  |  -.0950034   .1665181    -0.57   0.568    -.4213728    .2313661
    phdprest  |   .0650372   .0854488     0.76   0.447    -.1024394    .2325138
        _cons |   -6.67189   1.081551    -6.17   0.000    -8.791692   -4.552088
------------------------------------------------------------------------------
```

The last fitted model actually assumes that the difference in log odds for the event in question, for any two given individuals, is the same regardless of time, i.e., is constant (over time; see Equation (4.2)).

We can see this easily by plotting also the predicted log odds of say person 1 and person 4 (both having all 10 observations in the expanded data file), and inspect them visually.

This we achieve as follows. We begin by obtaining the predicted log odds (and visually examining them – see last column of the following listing of say the first 20 observations):

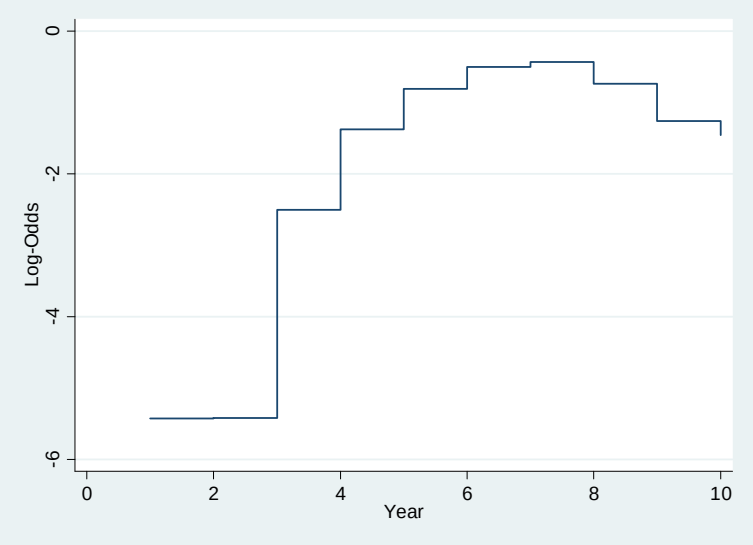
```
. predict lo, xb

. list in 1/20
```

	dur	event	undgrad	phdmed	phdprest	id	yr	y	lo
1.	10	0	7	0	2.21	1	1	0	-5.424531
2.	10	0	7	0	2.21	1	2	0	-5.418227
3.	10	0	7	0	2.21	1	3	0	-2.501495
4.	10	0	7	0	2.21	1	4	0	-1.373022
5.	10	0	7	0	2.21	1	5	0	-.8052251
6.	10	0	7	0	2.21	1	6	0	-.4999065
7.	10	0	7	0	2.21	1	7	0	-.4324633
8.	10	0	7	0	2.21	1	8	0	-.7340013
9.	10	0	7	0	2.21	1	9	0	-1.256762
10.	10	0	7	0	2.21	1	10	0	-1.459897
11.	4	1	6	0	2.21	2	1	0	-5.582192
12.	4	1	6	0	2.21	2	2	0	-5.575888
13.	4	1	6	0	2.21	2	3	0	-2.659156
14.	4	1	6	0	2.21	2	4	1	-1.530683
15.	4	0	4.95	0	2.21	3	1	0	-5.747736
16.	4	0	4.95	0	2.21	3	2	0	-5.741432
17.	4	0	4.95	0	2.21	3	3	0	-2.8247
18.	4	0	4.95	0	2.21	3	4	0	-1.696227
19.	10	0	2	1	4.54	4	1	0	-6.156303
20.	10	0	2	1	4.54	4	2	0	-6.149999

Now that we have these predicted log-odds, let's plot them correspondingly:

```
. twoway (line lo yr if id==1, connect(stairstep) lpatt(solid)),
xtitle(Year) ytitle(Log-Odds)
```

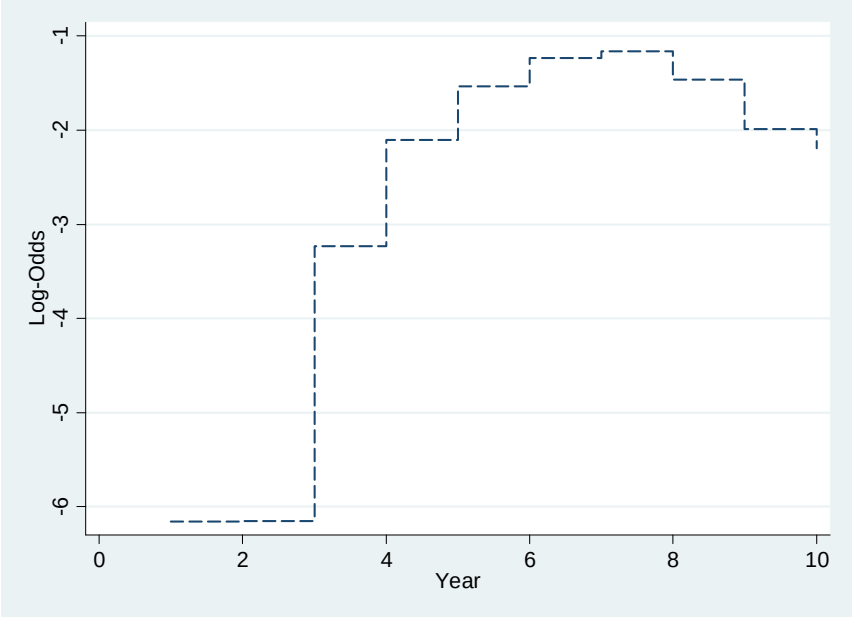
This leads to the following plot for person 1.



For subject 4, we proceed in complete analogy:

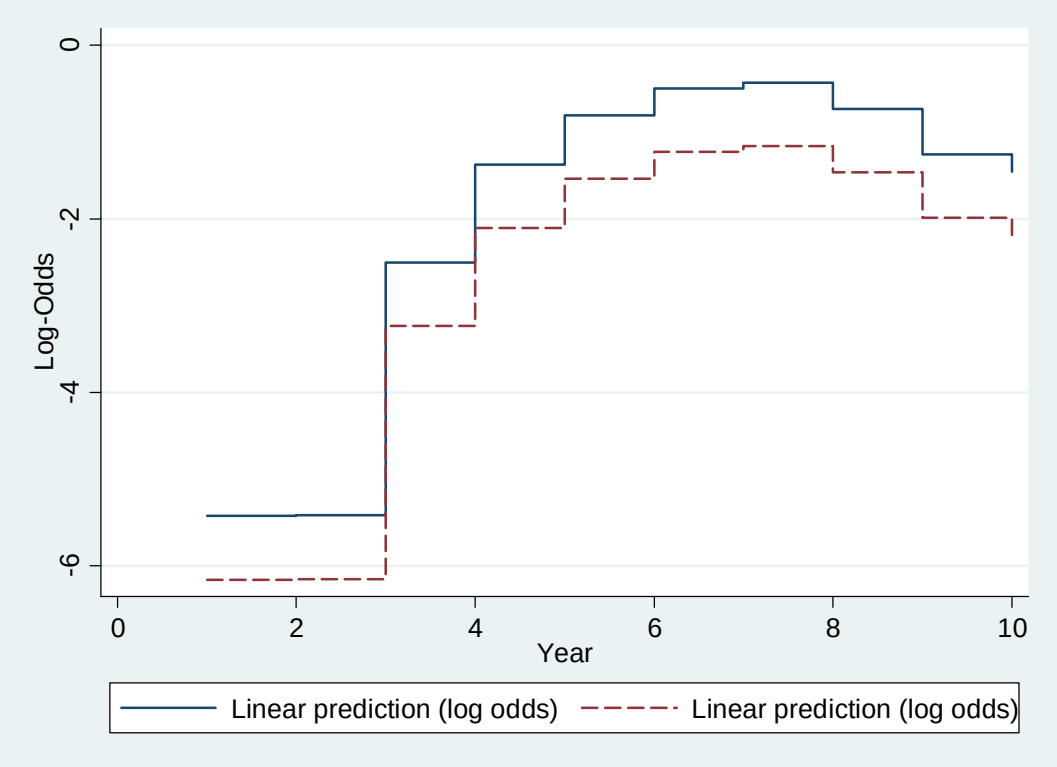
```
. twoway (line lo yr if id==4, connect(stairstep) lpatt(dash)),
xtitle(Year) ytitle(Log-Odds)
```

His/her resulting plot is as follows.



And, of course, we can overlay the two plots, to see their parallelism even more easily (see next page for resulting plot):

```
. twoway (line lo yr if id==1, connect(stairstep) lpatt(solid))
(line lo yr if id==4, connect(stairstep) lpatt(dash)), xtitle(Year)
ytitle(Log-Odds)
```



Due to this parallelism, the fitted model (4.2) is called a *proportional odds model* (in the context of EHA).

4.4. Examining odds ratios for event

Since we were able, as discussed earlier, to 'reduce' the initial problem/question about EHA (see p. 16) for the setting under consideration to a logistic regression model fitting, we can also examine the odds ratios associated with individual predictors, as follows.

```
. logit y i.yr undgrad phdmed phdprest, or
```

```
Iteration 0:   log likelihood = -654.73963
Iteration 1:   log likelihood = -557.63219
Iteration 2:   log likelihood = -528.19908
Iteration 3:   log likelihood = -524.62994
Iteration 4:   log likelihood = -524.34451
Iteration 5:   log likelihood = -524.34273
Iteration 6:   log likelihood = -524.34273

Logistic regression                                     Number of obs   =       1741
                                                        LR chi2(12)     =     260.79
                                                        Prob > chi2     =     0.0000
Log likelihood = -524.34273                             Pseudo R2       =     0.1992

------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          yr |
           2 |   1.006324   1.425663     0.00   0.996      .062638    16.16731
           3 |   18.59767   19.20306     2.83   0.005     2.457813    140.7241
           4 |   57.48415   58.39956     3.99   0.000     7.848588    421.0219
           5 |   101.4237   102.8852     4.55   0.000     13.88902    740.6396
           6 |   137.6377    140.062     4.84   0.000     18.73024    1011.421
           7 |   147.2406   151.0251     4.87   0.000     19.72187    1099.278
           8 |   108.9109   113.9559     4.48   0.000     14.01007    846.6475
           9 |   64.57125   70.07219     3.84   0.000     7.697014    541.6966
          10 |     52.701   60.03145     3.48   0.001     5.652226    491.3809
             |
     undgrad |   1.170769    .070328     2.62   0.009     1.040734    1.317052
      phdmed |   .9093699   .1514265    -0.57   0.568     .6561454    1.260321
    phdprest |   1.067199   .0911909     0.76   0.447     .9026328    1.261768
       _cons |    .001266   .0013692    -6.17   0.000      .000152    .0105452
------------------------------------------------------------------------------
```

As we can see from the last analytic result section, given the remaining 2 covariates, a unit increase in the prestige of the Ph.D. awarding institution is associated with a 7% (rounded off) increase in the odds of being promoted *in any given year* - assuming this has not already occurred. However, this increase is not significant.

Similarly we can interpret the odds associated with the other covariates. In particular, for given other 2 covariates, these odds of being promoted in any given year increase by 17% (and significantly) for every unit increase in selectivity of the undergraduate institution where the professor is employed.

While the log-odds curves above demonstrate the proportional odds feature of the model quite well, they're difficult to interpret as overall curves per se.

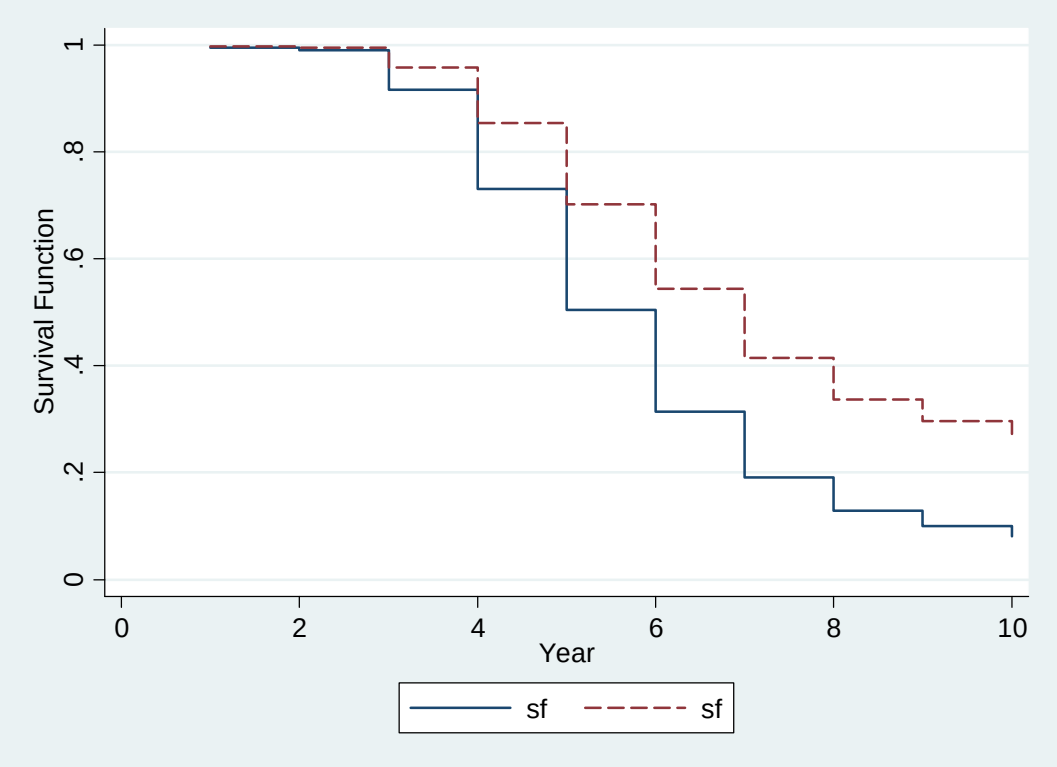
Instead, we may wish to see how the 'survival' of the professors develops over time and compares across them.

We can obtain these survival curves using their formal definition in the current DT-EHA context, employing the command 'invlogit', and plot them easily as follows (see Section 2, Equation (2.7) for *S*(*t*)):

```
. gen ln_1_m_haz = ln(1-invlogit(lo))

. by id (yr), sort: gen ln_sf = sum(ln_1_m_haz)

. gen sf = exp(ln_sf)
```

```
. twoway (line sf yr if id==1, connect(stairstep) lpatt(solid))
(line sf yr if id==4, connect(stairstep) lpatt(dash)), xtitle(Year)
ytitle(Survival Function)
```

This set of commands produces the following joint plot of the survival functions for person 1 and person 4 (having each all 10 consecutive observations possible in this study).



From the last plot, we can see for instance two interesting findings:

1. **Assistant professors who have the covariate values of professor 1 (viz. undgrad = 7, phdmed = 0, and phdprest = 2.21), have a 50% chance of being promoted by year 5.**

2. **Assistant professors with covariate values of professor 4 (lower 'undgrad' and 'phdprest' values) have a 50% chance of being promoted by year 7.**

4.5. Including time-varying covariates

Thus far in section 4, we have only considered time-invariant covariates. In particular, in our empirical example, the working university prestige, that of degree awarding institution, and the type of degree, were all constants (over time).

Quite often, however, research questions in EHA involve covariates that do not remain constant across the course of a study.

For instance, still in the same example, it would be of interest to ask the following questions: what's the effect of publications during the years up to promotion application, and what's the effect of the citations of the candidate?

The preceding discussion in the section allows also to include this type of covariates, once we have generated the expanded data set.

For our empirical example, we proceed as follows (see data in file 'prom3.dta').

```
. use "C:\T E A C H\Workshops\EUI\EHA\prom3.dta", clear

. d

Contains data from C:\T E A C H\Workshops\Italy\SA\prom3.dta
  obs:           301
 vars:            26                          1 Oct 2013 11:56
 size:        31,304
-------------------------------------------------------------------------------
storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g
dur             float  %9.0g
event           float  %9.0g
```

```
undgrad         float  %9.0g
phdmed          float  %9.0g
phdprest        float  %9.0g
art1            float  %9.0g
art2            float  %9.0g
art3            float  %9.0g
art4            float  %9.0g
art5            float  %9.0g
art6            float  %9.0g
art7            float  %9.0g
art8            float  %9.0g
art9            float  %9.0g
art10           float  %9.0g
cit1            float  %9.0g
cit2            float  %9.0g
cit3            float  %9.0g
cit4            float  %9.0g
cit5            float  %9.0g
cit6            float  %9.0g
cit7            float  %9.0g
cit8            float  %9.0g
cit9            float  %9.0g
cit10           float  %9.0g
------------------------------------------------------------------------------
Sorted by:
```

Let's first take a look at the original data (above data set), say for the first 10 persons (including for instance only the yearly number of articles variables):

```
. list id-art10 in 1/10
```

	id	dur	event	undgrad	phdmed	phdprest	art1	art2	art3	art4	art5	art6	art7	art8	art9	art10
1.	1	10	0	7	0	2.21	0	0	2	2	2	2	2	2	2	2
2.	2	4	1	6	0	2.21	8	10	14	18	.	.	.	.	.	.
3.	3	4	0	4.95	0	2.21	0	0	0	2	.	.	.	.	.	.
4.	4	10	0	2	1	4.54	2	3	3	3	3	4	6	6	6	6
5.	5	7	1	5	1	2.15	1	1	1	2	2	3	5	.	.	.
6.	6	6	1	4.95	0	4.54	0	2	3	5	5	6	.	.	.	.
7.	7	10	0	6	1	4.54	5	5	5	5	5	5	5	6	6	6
8.	8	6	1	4	1	2.96	3	4	7	8	8	9	.	.	.	.
9.	9	4	0	4	1	1.63	8	8	10	11	.	.	.	.	.	.
10.	10	5	1	5	1	2.96	0	1	1	1	2	.	.	.	.	.

Well, this is not the format we need the data in, in order to include time-varying covariates like number of published articles per year.

As we mentioned earlier, what we need now are the data in a different format, the 'long' format, where each year has its own raw of data.

We re-format our data as follows, so that each year until promotion application is presented in a single row then:

```
. reshape long art cit, i(id) j(year)
(note: j = 1 2 3 4 5 6 7 8 9 10)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                      301   ->    3010
Number of variables                  26   ->       9
j variable (10 values)                    ->   year
xij variables:
                    art1 art2 ... art10   ->   art
                    cit1 cit2 ... cit10   ->   cit
-----------------------------------------------------------------------------
```

This is indeed the required format, as we check its first say 20 rows:

```
. list in 1/20

     +-------------------------------------------------------------------+
     | id   year   dur   event   undgrad   phdmed   phdprest   art   cit |
     |-------------------------------------------------------------------|
  1. |  1      1    10       0         7        0       2.21     0     0 |
  2. |  1      2    10       0         7        0       2.21     0     0 |
  3. |  1      3    10       0         7        0       2.21     2     1 |
  4. |  1      4    10       0         7        0       2.21     2     1 |
  5. |  1      5    10       0         7        0       2.21     2     1 |
     |-------------------------------------------------------------------|
```

```
     |-------------------------------------------------------------------|
  6. |  1      6    10       0         7        0       2.21     2     1 |
  7. |  1      7    10       0         7        0       2.21     2     1 |
  8. |  1      8    10       0         7        0       2.21     2     1 |
  9. |  1      9    10       0         7        0       2.21     2     1 |
 10. |  1     10    10       0         7        0       2.21     2     1 |
     |-------------------------------------------------------------------|
 11. |  2      1     4       1         6        0       2.21     8    27 |
 12. |  2      2     4       1         6        0       2.21    10    44 |
 13. |  2      3     4       1         6        0       2.21    14    57 |
 14. |  2      4     4       1         6        0       2.21    18    63 |
 15. |  2      5     4       1         6        0       2.21     .     . |
     |-------------------------------------------------------------------|
 16. |  2      6     4       1         6        0       2.21     .     . |
 17. |  2      7     4       1         6        0       2.21     .     . |
 18. |  2      8     4       1         6        0       2.21     .     . |
 19. |  2      9     4       1         6        0       2.21     .     . |
 20. |  2     10     4       1         6        0       2.21     .     . |
     +-------------------------------------------------------------------+
```

This reshaping we did has actually created, as we see, a few unnecessary rows, viz. for each person whose original 'duration' value is less than 10.

We can readily dispense with these excess rows (viz. those with 'missings' in the last 2 columns of the just expanded data set, i.e., with 'missings' on the articles' and citations' variables, 'art' and 'cit' respectively):

```
. drop if year>dur
(1269 observations deleted)

. list in 1/20
```

```
     +-------------------------------------------------------------------+
     | id   year   dur   event   undgrad   phdmed   phdprest   art   cit |
     |-------------------------------------------------------------------|
  1. |  1      1    10       0         7        0       2.21     0     0 |
  2. |  1      2    10       0         7        0       2.21     0     0 |
  3. |  1      3    10       0         7        0       2.21     2     1 |
  4. |  1      4    10       0         7        0       2.21     2     1 |
  5. |  1      5    10       0         7        0       2.21     2     1 |
```

```
     |-------------------------------------------------------------------|
  6. |  1      6    10       0         7        0       2.21     2     1 |
  7. |  1      7    10       0         7        0       2.21     2     1 |
  8. |  1      8    10       0         7        0       2.21     2     1 |
  9. |  1      9    10       0         7        0       2.21     2     1 |
 10. |  1     10    10       0         7        0       2.21     2     1 |
     |-------------------------------------------------------------------|
 11. |  2      1     4       1         6        0       2.21     8    27 |
 12. |  2      2     4       1         6        0       2.21    10    44 |
 13. |  2      3     4       1         6        0       2.21    14    57 |
 14. |  2      4     4       1         6        0       2.21    18    63 |
 15. |  3      1     4       0      4.95        0       2.21     0     0 |
     |-------------------------------------------------------------------|
 16. |  3      2     4       0      4.95        0       2.21     0     0 |
 17. |  3      3     4       0      4.95        0       2.21     0     0 |
 18. |  3      4     4       0      4.95        0       2.21     2     2 |
 19. |  4      1    10       0         2        1       4.54     2    11 |
 20. |  4      2    10       0         2        1       4.54     3    17 |
     +-------------------------------------------------------------------+
```

In order to proceed with our planned analyses (consisting in fitting the model defined by Equation (4.3) below), we still need our earlier 'y' variable, however.

As you recall, that variable was very important as it was telling us the 'status' for each of the observations/rows in the last version/format of the original data file, 'prom3.dta' (and then list the first 20 say observations):

```
. gen y=0
. replace y=event if year==dur
. list in 1/20
```

```
     +-------------------------------------------------------------------+
     | id   year   dur   event   undgrad   phdmed   phdprest   art   cit   y |
     |-------------------------------------------------------------------|
  1. |  1      1    10       0         7        0       2.21     0     0   0 |
  2. |  1      2    10       0         7        0       2.21     0     0   0 |
  3. |  1      3    10       0         7        0       2.21     2     1   0 |
  4. |  1      4    10       0         7        0       2.21     2     1   0 |
  5. |  1      5    10       0         7        0       2.21     2     1   0 |
```

6.	1	6	10	0	7	0	2.21	2	1	0
7.	1	7	10	0	7	0	2.21	2	1	0
8.	1	8	10	0	7	0	2.21	2	1	0
9.	1	9	10	0	7	0	2.21	2	1	0
10.	1	10	10	0	7	0	2.21	2	1	0
11.	2	1	4	1	6	0	2.21	8	27	0
12.	2	2	4	1	6	0	2.21	10	44	0
13.	2	3	4	1	6	0	2.21	14	57	0
14.	2	4	4	1	6	0	2.21	18	63	1
15.	3	1	4	0	4.95	0	2.21	0	0	0
16.	3	2	4	0	4.95	0	2.21	0	0	0
17.	3	3	4	0	4.95	0	2.21	0	0	0
18.	3	4	4	0	4.95	0	2.21	2	2	0
19.	4	1	10	0	2	1	4.54	2	11	0
20.	4	2	10	0	2	1	4.54	3	17	0

With all this having been accomplished now, if we want to include in our model also the number of articles and citations for each year, we extend the last fitted model in Equation (4.2) as follows:

$$\text{(4.3)} \qquad \text{logit}(P(y_{ti} = 1|d_{ti}) = \beta_0 + \beta_2\, d_{2,ti} + \ldots + \beta_{10}\, d_{10,ti} + \gamma_1\, x_{1,i} + \ldots + \gamma_p\, x_{p,i} + \gamma_{p+1}\, x_{p+1,ti} + \ldots + \gamma_{p+q}\, x_{p+q,ti} \quad (p, q > 0).$$

For our empirical example and interest (viz. studying time to first professorial promotion in the US), $p = 3$ and $q = 2$. Therefore, we fit the model in Equation (4.3) as follows (requesting also the odds ratios):

```
. logit y i.year undgrad phdmed phdprest art cit, or
```

```
Iteration 0:   log likelihood = -654.73963
Iteration 1:   log likelihood =  -544.9642
Iteration 2:   log likelihood =  -510.2932
```

```
Iteration 3:   log likelihood =  -505.4005
Iteration 4:   log likelihood = -505.22193
Iteration 5:   log likelihood =  -505.2219

Logistic regression                                     Number of obs   =       1741
                                                        LR chi2(14)     =     299.04
                                                        Prob > chi2     =     0.0000
Log likelihood =  -505.2219                             Pseudo R2       =     0.2284

------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
          2  |   .9194398   1.302999    -0.06   0.953     .0571781    14.78486
          3  |   15.39302   15.91448     2.64   0.008     2.029021     116.778
          4  |   44.88887   45.67199     3.74   0.000     6.110646    329.7541
          5  |   74.30607   75.52769     4.24   0.000     10.13512    544.7784
          6  |   94.16784   96.12298     4.45   0.000     12.73586     696.269
          7  |   99.32770   102.2799     4.47   0.000     13.19986    747.4318
          8  |   69.66594   73.28581     4.03   0.000     8.863192    547.5841
          9  |   39.63958   43.44556     3.36   0.001     4.625938    339.6709
         10  |   34.72460   39.86698     3.09   0.002     3.659155     329.529
             |
     undgrad |    1.18396   .0730565     2.74   0.006     1.049092    1.336166
      phdmed |   .7999923   .1369606    -1.30   0.192       .57195    1.118957
    phdprest |   .9714253   .0876035    -0.32   0.748     .8140437    1.159234
         art |   1.075675   .0192744     4.07   0.000     1.038553    1.114123
         cit |   .9998082   .0012797    -0.15   0.881     .9973031     1.00232
       _cons |   .0012647   .0013733    -6.15   0.000     .0001505    .0106243
------------------------------------------------------------------------------
```

We see that once fixing the other four covariates, the number of articles variable is significant, but not that of number of citations.

We see also that the odds for promotion in year 6 are over 102 times those in year 2 say.

4.5. Multilevel EHA via robust modeling accounting for clustering effect

When the observations in a DT-EHA context are nested or clustered themselves within higher-order, level-2 units, one can account for it by conducting robust modeling (see main section 2 of this workshop, specifically subsection 2.4.1).

This is achieved by invoking robust standard errors using the subcommand `vce(cluster <name of unit-2 identifier>)` (see below).

By doing this, one requests the sandwich estimator of the parameter covariance matrix we indicated earlier in this workshop (see subsection 2.4.1).

In this way, one relies on the consistency of the associated parameter estimator, as well as resulting corrected standard errors. (They obviously affect statistical tests as well as confidence intervals.)

We illustrate by returning to the preceding example context and using the pertinent (larger) data set 'prom3c.dta' containing as last variable the state within which the current university of the studied professor is located.

(Note that upon reading in this data file, we need to do the same data management activities as in the last subsection, 4.4, which I am skipping next.)

This two-level DT-EHA analysis via robust modeling (upon appropriate expansion and management of the data set used), is achieved in the following way.

```
. logit y i.year undgrad phdmed phdprest art cit, or vce(cluster
state)

Iteration 0:   log pseudolikelihood = -654.73963
Iteration 1:   log pseudolikelihood =  -544.9642
Iteration 2:   log pseudolikelihood =  -510.2932
Iteration 3:   log pseudolikelihood =  -505.4005
Iteration 4:   log pseudolikelihood = -505.22193
Iteration 5:   log pseudolikelihood =  -505.2219

Logistic regression                               Number of obs   =       1741
                                                  Wald chi2(14)   =   23749.93
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood =  -505.2219                 Pseudo R2       =     0.2284

                                  (Std. Err. adjusted for 16 clusters in state)
------------------------------------------------------------------------------
             |               Robust
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |
          2  |   .9194398   1.347998    -0.06   0.954     .0519481    16.27335
          3  |   15.39302   15.30892     2.75   0.006     2.191657    108.1122
          4  |   44.88887   47.34988     3.61   0.000     5.678979    354.8191
          5  |   74.30607   74.00434     4.33   0.000     10.55065    523.3225
          6  |   94.16784   91.19875     4.69   0.000      14.1104    628.4431
          7  |    99.3277    106.112     4.30   0.000     12.23854    806.1415
          8  |   69.66594   62.93302     4.70   0.000     11.85998    409.2201
          9  |   39.63958   31.88358     4.57   0.000     8.193629    191.7705
         10  |    34.7246   35.06921     3.51   0.000     4.797204    251.3543
             |
     undgrad |    1.18396   .0540938     3.70   0.000     1.082546    1.294874
      phdmed |   .7999923   .1782725    -1.00   0.317     .5168943     1.23814
    phdprest |   .9714253   .0875869    -0.32   0.748     .8140709    1.159195
         art |   1.075675   .0211751     3.71   0.000     1.034963    1.117988
         cit |   .9998082   .0013047    -0.15   0.883     .9972542    1.002369
       _cons |   .0012647   .0013601    -6.20   0.000     .0001537    .0104087
------------------------------------------------------------------------------
```

Note that the parameter estimates are the same (as they should be) as in the last fitted model (subsection 4.4). However, their standard errors are slightly larger (as they also should be). This is due to the clustering effect of professors within states (which we didn't account for in our preceding analyses).

(The extent to which these robust SEs are larger than the SEs obtained with that modeling approach discussed last in subsection 4.4, depends on the strength of the clustering effect.)

4.6. Summary

In this section 4. of the workshop, we have conducted first multilevel DT-EHAs responding to the 'border' question of how to analyze a given data set after its expansion described earlier.

We have carried out thereby (i) no-covariate modeling, then (ii) analysis with time-invariant covariates, followed by (iii) analysis with time-varying covariates, as well as with both types of covariates (for details, see subsections 4.1 through 4.4).

Subsequently, we have conducted (iv) two-level DT-EHA including (a) time-invariant as well as (b) time-varying covariates *and* (c) accounting for the clustering effects of studied persons within higher-order units (for details, see subsection 4.5).

We move next to multilevel EHA with latent variables, using a highly popular EHA model, the Cox proportional hazards model.

5. A latent variable modeling approach to MEHA

The discussion thus far in the workshop was concerned with observed (manifest) variables involved in an EHA. Thereby, the assumption of *error-free covariates* (placed in the vector $\underline{x}$ say) has been made throughout, as routinely in conventional (widely used, traditional) EHA - whether in single-level or multilevel modeling.

This assumption is particularly important for the popular Cox PH model (Cox regression). However, it's equally relevant to any discrete-time EHA as covered in the workshop or used in empirical social and behavioral research (unless the following type of developments are pursued – currently rather rarely, if even worth mentioning, in terms of frequency in EHA applications).

This assumption of error-free covariates, is rarely satisfied in typical empirical settings in the social and behavioral sciences.

In the present Section 5 of the workshop, we relax this requirement of error-free covariates, for the purpose of carrying out multilevel EHA. We begin our discussion with the CT-EHA setup, where the Cox PH is one of the most frequently used - if not most celebrated - of modeling approaches.

5.1. Cox regression extension with fallible covariates

Suppose that in a continuous time EHA setting we have a set of available error-free covariates, collected in the vector $\underline{x}$, in addition to measures of duration as well as status of studied subjects (sample from a population under consideration). We assume also that we have access to another set of measures administered to the same persons, with each measure having been evaluated with error.

As discussed in detail in the measurement related literature (e.g., Raykov & Marcoulides, 2011, *Introduction to Psychometric Theory*), a useful modeling framework for handling error-prone (i.e., fallible) measures is provided by the comprehensive latent variable modeling (LVM) methodology.

This methodology encompasses models developed in terms of latent variables, each with multiple indicators. But what are latent variables and indicators? We begin this discussion with a definition and examples. (See, e.g., Raykov & Marcoulides, 2006, *A first course in structural equation modeling*, for more details.)

***Latent variable* (construct, factor, trait; LV) = An indirectly observable, 'hidden' random variable that *captures the commonality across similar behavioral manifestations*, and which has individual realizations in each subject of the sample (or population for that matter) that are however not observed.**

How to identify a LV?

<u>Simple rule</u>: Theoretical concept ~ LV.

<u>Examples</u>: Attitude, conservatism, liberalism, alienation, social cohesiveness, motivation, (specific) ability, intelligence, anxiety, depression, motivation.

<u>Note</u>: Most theories in the social sciences are developed in terms of LVs, as one can consider the theoretical concepts such variables (LVs - like we just pointed out).

The *indicators* of a LV are observed variables that can be used as proxies for the LV, which are typically however error-prone, i.e., contain error (pure random measurement error, with possibly added error due to invalidity of measurement).

A simple example is provided by the questions in a given part of a survey evaluating a particular trait (e.g., conservatism/liberalism) or construct.

Alternatively, one can obtain more than one indicator by randomly splitting a homogeneous/unidimensional scale (the set of items in the scale) evaluating the LV in question, such as say voting sophistication or attitude toward a particular political issue.

Like in the experimental sciences where one collects multiple replications of a given experiment, we want multiple indicators of each LV (at least 2 indicators, and preferably more; they are usually obtained as mentioned above).

Latent variable modeling

The methodology using latent variable models, is referred to as *latent variable modeling* (LVM). As an example, a widely used such model (within LMV), is that of factor analysis:

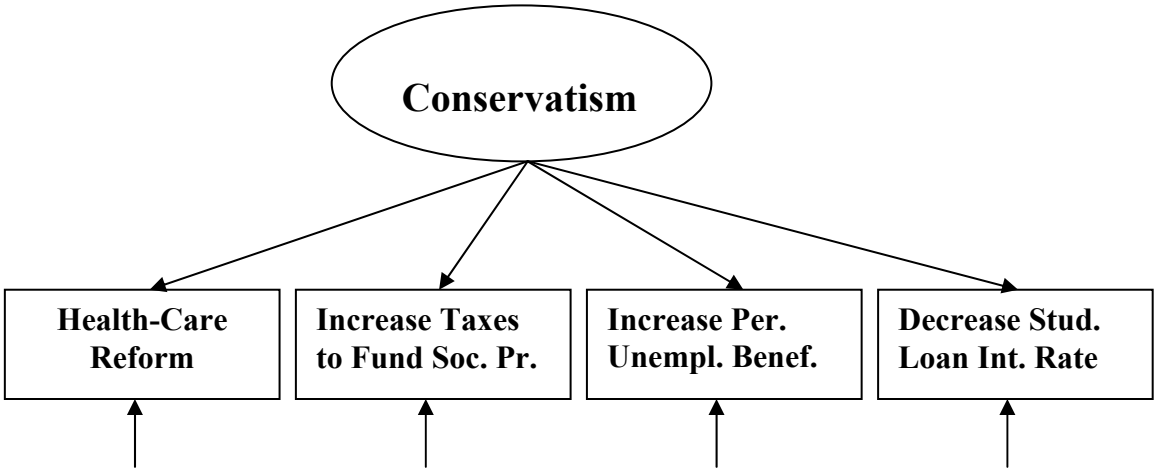
(5.1) $$y = \alpha + \Lambda\eta + \varepsilon ,$$

where

y = set/vector of observed variables (e.g., alienation items/subscale scores, and those of a social support inventory; α = associated intercepts),
η = set of common factors/latent variables (e.g., alienation and social support),
Λ = factor loading matrix,
ε = set of unique factors (residuals).

For example, suppose one were interested in evaluating political conservatism. This main theoretical concept in political science, can be considered a latent variable, which manifests itself for instance in subjects' responses to certain (suitably chosen) questions in a survey/questionnaire.

With this in mind, the following would be a factor analysis model worth considering then, if using the below questions in a subscale from a larger questionnaire say (yes/no questions: do you agree with …; or Likert-type questions: strongly agree through strongly disagree with…).



In this widely used path-diagram notation in applications of LVM (e.g., Bollen, 1989, *Structural equations with latent variables*), a circle/ellipse denotes a latent variable, viz. the common source of variance for the observed/manifest variables that are in the squares/rectangles.

Similarly, short 1-way arrows denote error terms, and 1-way arrows symbolize the assumption that the variable at its beginning plays explanatory role for the variable at its end.

After this introductory example of a factor analysis model, in the remainder of this section the following general model will be instrumental as part of an even more comprehensive model of CT-EHA and an extension of the Cox's PH model:

(5.2) $$\underline{y} = \underline{\alpha} + \Lambda\, \underline{\eta} + \Gamma\, \underline{x} + \underline{\varepsilon}\,, \text{ and}$$
$$\underline{\eta} = \underline{\nu} + B\, \underline{\eta} + \Delta\, \underline{x} + \underline{\zeta}\,,$$

where in addition to the notation utilized in Equation (5.1),

B = matrix of linear (regression-like) coefficients for the relationships between the latent variables,
$\underline{\zeta}$ = set/vector of residual terms (latent disturbances) in these equations,
$\underline{x}$ = (vector/set of) covariates measured without error (e.g., some demographics)
$\underline{\alpha}$, $\underline{\nu}$ = pertinent intercepts (vector),
Γ, Δ = regression-like coefficient matrices for $\underline{x}$.

In general, there is *no* need to make the normality assumption for the observed variables (conditional upon covariates), or for the latent variables (incl. error terms). Specifically, a *robust maximum likelihood* method of model fitting and testing, as well as parameter estimation, can be used (if no piling of cases at end of scale for observed variable or highly discrete items; alternative maximum likelihood possible then).

A survival analysis modeling framework for time to event with latent explanatory variables

This model is based on the assumption that time to event is related (also) to latent explanatory variables, each measured with a set of indicators.

***General model*:**

(5.3) $g(T) = f(x, y, \eta)$

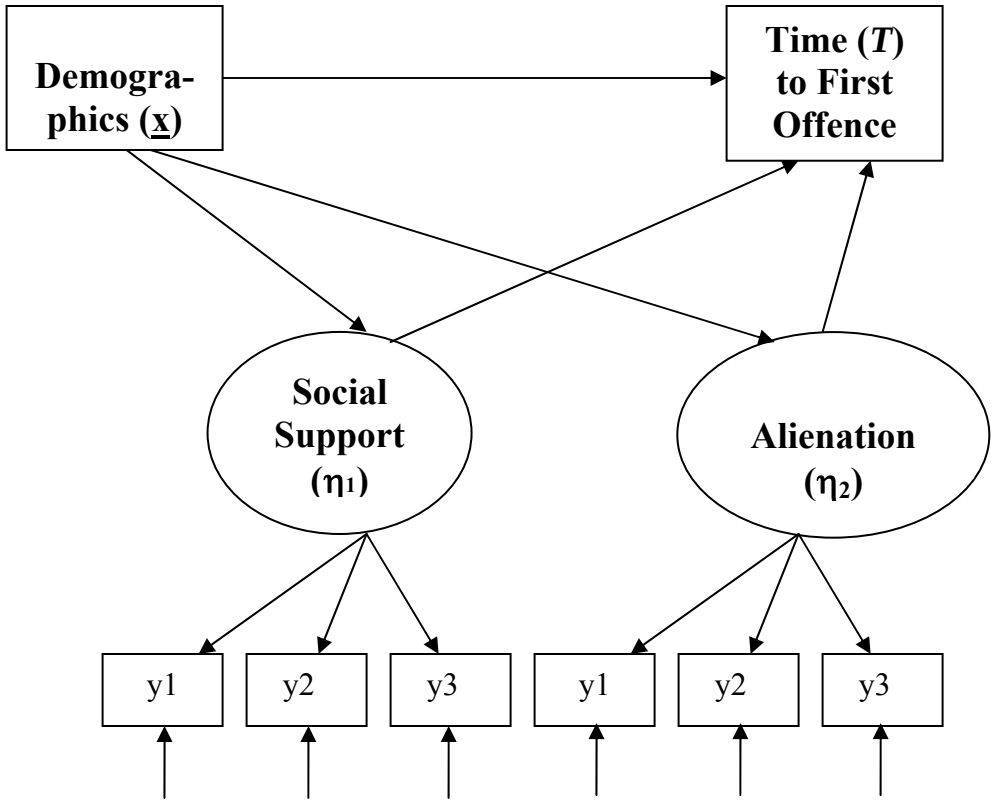
$$y = \alpha + \Lambda \eta + \Gamma x + \varepsilon ,$$

$$\eta = \nu + B \eta + \Delta x + \zeta ,$$

where in addition to the notation (and assumptions) in Equations (5.2),

T **= time to event,**
***g*(.)** **= function to be specified (usually can take *g*(*t*) = *h*(*t*), the hazard function, as used next).**

As an example, consider a conceptual model of the relationship between time to offence and other variables (cf. Larson, 2005, *Biometrics*; in the figure next, latent disturbance covariance are not explicitly presented, to avoid graphical clutter).



In the following subsections, we will deal first with a model closely related to the last presented one, which will be applied later on empirical data.

How to fit this type of models to data? The *main idea* (e.g., Larson, 2005) is as follows:

(i) consider η as a set of variables, on which all subjects have *missing values*, and then
(ii) use the *Expectation-Maximization* (EM) *algorithm* for estimating the unknown parameters.

This is achieved within the framework of the following Cox regression extension.

The extended multilevel Cox regression model for fallible covariates with multiple indicators

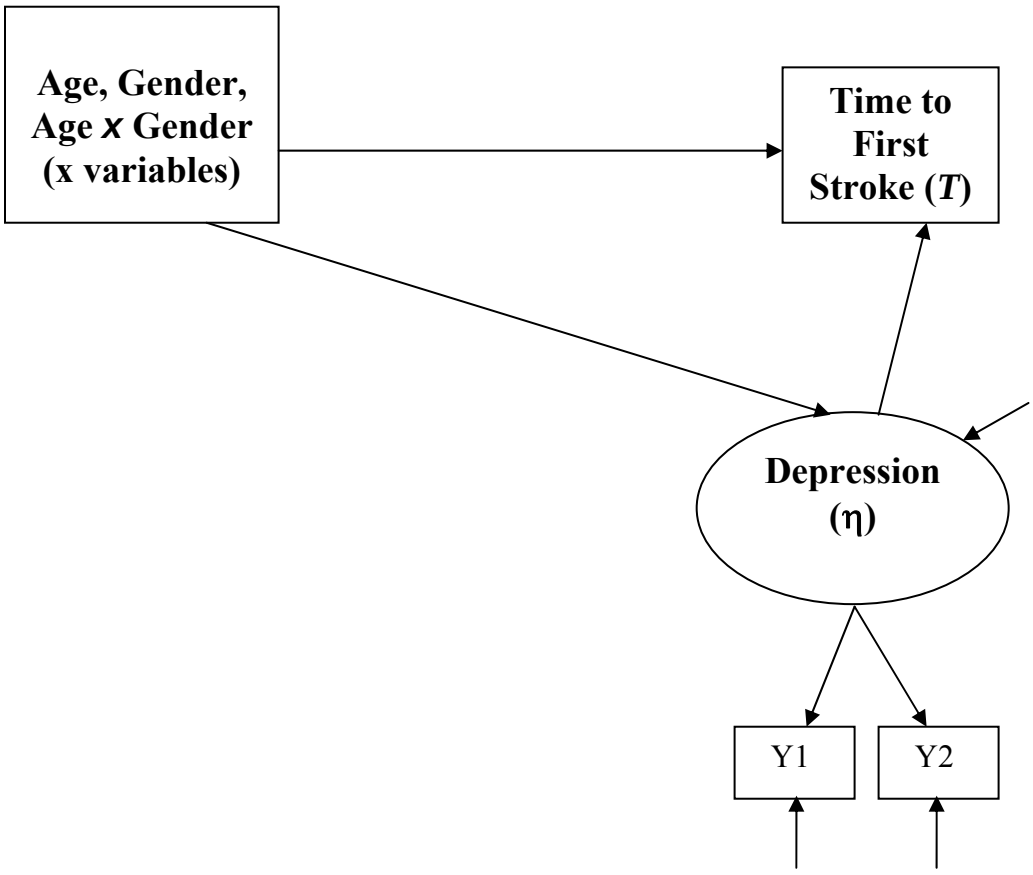
To achieve the above aim, we extend the popular Cox PH model in the following way. To the error-free covariates, $\underline{x}$, we add now the latent ones, $\underline{\eta}$, whose fallible measures/manifestations are $\underline{y}$, as discussed in the preceding section (see Equations (5.2) and (5.3)).

We assume that the latent covariates' measures, $\underline{y}$, are not affected by the error-free covariates, the $\underline{x}$. (In ability testing, this is referred to as *no differential item functioning*; e.g., Larsen, 2005.)

This is a plausible assumption, typically used in applications, and testable utilizing a latent variable model (e.g., Raykov, Lee, & Marcoulides, 2013).

For instance, consider the following model that addresses an increasingly popular hypothesis in gerontology and geriatrics, the vascular depression hypothesis. Accordingly, depression plays a warning role for an impending stroke in adults and elderly persons.

In this model, notice that there are no paths connecting any of the *x* variables with any of the *y* variables; similarly there are no paths connecting *Y* variables with time to event, *T*:



Now, the *extended Cox PH model* is (individual subscript suppressed):

(5.4) $$h(t, \underline{x}, \underline{y}) = h_0(t) \exp(\sum_{l=1}^{r} \beta_l x_l + \sum_{j=1}^{q} \gamma_j \eta_j) \ .$$

To fit this model to data and estimate its parameters, the above main idea is used. This procedure is implemented in the popular LVM software Mplus (Muthén & Muthén, 2012).

To illustrate, consider the last model graphically presented above. Its Mplus input file is as follows. (I will discuss in detail its individual commands in the workshop; a comment is indicated by an exclamation mark within a row of this command file.)

```
TITLE:      EXTENDED COX PROPORTIONAL HAZARDS MODEL
            (INCL. LATENT PREDICTORS WITH MULTIPLE INDICATORS).

DATA:       FILE = <name of raw data file>;

VARIABLE:   NAMES = GENDER T_TO_STR AGE CENSOR DEP1 DEP2;
            SURVIVAL = T_TO_STR(ALL);
            TIMECENSORED = CENSOR(1=NOT 0=RIGHT);
            ! CLUSTER = <CLUSTER VARIABLE NAME>; -cluster effect
            ! MISSING = ALL(-999);
            ! WEIGHT = WEIGHT_VARIABLE;
            ! CLASSES = C(#);

ANALYSIS:   BASEHAZARD = OFF;
            ALGORITHM = INTEGRATION;
            !TYPE = COMPLEX; MIXTURE; ! accounts for clustr eff.

MODEL:      DEPRESSN BY DEP1-DEP2;
            DEPRESSN ON AGE GENDER; !AGEXGEN;
            T_TO_STR ON DEPRESSN AGE GENDER; !AGEXGEN;
```

At this point it is worth mentioning that the last command file also indicates further possible extensions of the method, specifically to settings with:

- **missing data (missing at random),**
- **clustered data (see next subsection),**

- **interactions of variables (observed measures),**
- **hierarchical EHA data (see clustering variable and complex analysis requested, which invokes the robust estimation procedure mentioned earlier, and next subsection), and**
- **finite mixtures.**

5.2. Multilevel continuous-time EHA

The last example showed how the Cox PH model could be fit to single-level event-history data, using M*plus*.

When the survival data are of hierarchical nature—e.g., respondents coming from different cities (and hence being *nested* within them)—one can use an approach that is 'borrowed' from complex survey sampling, the *pseudo-maximum likelihood* (PML) estimator with *robust* estimates of parameter standard errors.

This method is implemented in M*plus* and provides parameter estimates and in particular standard errors that account for the lack of independence among the persons (subjects) within higher-order units (e.g., cities, interviewers, families).

The standard errors are obtained as the square-rooted main diagonal elements of the following matrix (cf. Section 2 of this workshop):

(5.5)
$$\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]^{-1}\left\{\sum_{i=1}^{n}\frac{\partial \log L_i}{\partial\theta}\left(\frac{\partial \log L_i}{\partial\theta}\right)'\right\}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]^{-1},$$

where *L* = data likelihood and (simplifying some notation):

(5.6)
$$L_i = h_i^{1-\delta_i} S_i \ ,$$

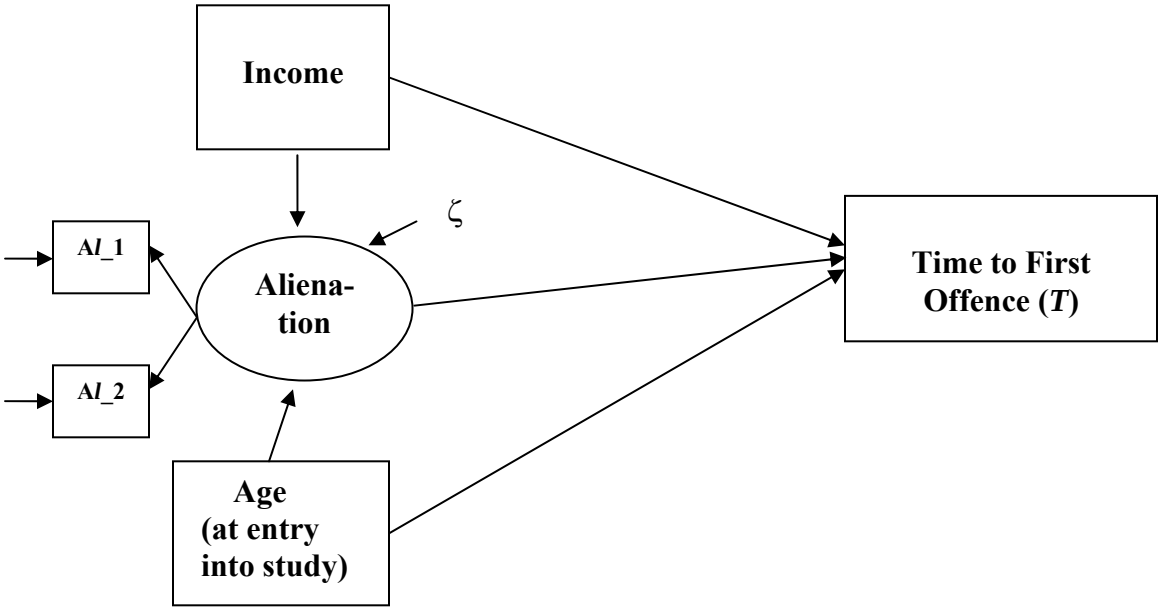
with δ_i being the subject censoring indicator ($i = 1, \ldots, n$).

We demonstrate this approach next.

Hierarchical EHA analysis with a latent predictor(s), accounting for clustering effect

To exemplify this approach, for the sake of demonstration in this subsection we use two examples.

***Example* 1. Suppose we were interested in studying time to offence committed by adults (nested) in neighborhoods in a large urban area, and have data on their income and age at entry into the study as well as two questions related to alienation. A model of interest would be as follows.**



We wish to pay special attention to the fact that the adults studied are *nested* within neighborhoods, and want to account for this effect.

I should also like to note the presence of the latent predictor in our model, Alienation. This is not a regular (and in fact still very rare) feature of EHA models in the social and behavioral sciences.

In actual fact, this example embodies an extension of conventional EHA in what are effectively up to six directions:

- **(a) from single-level to two-level data;**
- **(b) using a latent construct as an explanatory variable that has multiple indicators (at least two);**
- **(c) accounting (as opposed to ignoring) measurement error in predictor variables (and thus avoiding inconsistent parameter estimates resulting otherwise);**
- **(d) using the disturbance term associated with the latent predictor, Alienation, to represent individual differences in event propensity;**
- **(e) applicability to settings with missing data (MAR) on the latent explanatory variables' indicators (also w/ auxiliary variables);**
- **(f) including possibly further levels of data hierarchy—e.g., as in three-level or higher-level data settings (incl. stratification, sampling weights, etc.).**

I should further like to point out that we may wish to use the observed alienation subscale scores A*l*_1 and A*l*_2 directly as explanatory variables themselves, rather than as indicators of the Alienation construct in the last model considered.

In that case, we would end up with *inconsistent* parameter estimates (in addition to possible multicollinearity and related numerical issues).

To fit this model, we use the following approach with M*plus*.

```
TITLE:      MULTILEVEL EHA WITH A LATENT COVARIATE.

DATA:       FILE = ALIENATN.DAT;

VARIABLE:   NAMES = ID INCOME TT1STOFF AGE CSTATUS AL_DEP1 AL_2 N_HOOD;
            USEV = INCOME TT1O AGE CSTATUS AL_DEP1 AL_2 N_HOOD;
            SURVIVAL = TT1STOFF(ALL);
            TIMECENSORED = CSTATUS(1=NOT 0=RIGHT);
            CLUSTER = N_HOOD;

ANALYSIS:   BASEHAZARD = OFF; ! REQUESTS FRAMEWORK OF COX REGRESSION.
            TYPE = COMPLEX; !   ACCOUNTS FOR NESTING WITHIN N'HOODS.

MODEL:      ALIEN BY AL_1 AL_2; ! factor model, latent predictor.

            ALIEN ON AGE INCOME; ! latent predictor explained itself
                                 ! in terms of age and INCOME.

            TT1STOFF ON ALIEN AGE INCOME; ! main rel'ship of interest.
```

This yields the following results.

```
INPUT READING TERMINATED NORMALLY

SUMMARY OF DATA AND ANALYSIS

Number of clusters                                       27

Number of groups                                          1
Number of observations                                  504

Number of dependent variables                             3
Number of independent variables                           2
Number of continuous latent variables                     1


Observed dependent variables

  Continuous
   AL_1     AL_2

  Time-to-event (survival)
   TT1STOFF
```

Observed independent variables
INCOME AGE

Continuous latent variables
ALIEN

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Loglikelihood

H0 Value	-2637.836
H0 Scaling Correction Factor for MLR	0.922

Information Criteria

Number of Free Parameters	11
Akaike (AIC)	5297.672
Bayesian (BIC)	5344.120
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	5309.205

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ALIEN	BY				
AL_1		1.000	0.000	999.000	999.000
AL_2		1.542	0.761	2.025	0.043
ALIEN	ON				
AGE		-0.001	0.016	-0.063	0.950
INCOME		0.087	0.067	1.288	0.198
TT1STOFF	ON				
ALIEN		0.091	0.019	4.789	0.002
TT1STOFF	ON				
AGE		-0.029	0.010	-3.036	0.002
INCOME		-1.803	0.301	-5.988	0.000
Intercepts					
AL_1		11.684	0.359	32.581	0.000
AL_2		16.311	0.567	28.778	0.000
Residual Variances					
AL_1		0.600	0.790	0.759	0.448
AL_2		0.230	1.916	0.120	0.905
ALIEN		1.608	0.781	2.059	0.040

These results suggest the following interpretation (under the model):

(a) age decreases the hazard for time to offence, controlling for income and alienation;
(b) income decreases the hazard for time to offence, controlling for age and alienation; and
(c) controlling for age and income, alienation enhances the hazard of offence.

Further interpretations of the above results will be offered during the pertinent workshop session.

For the next example, data are provided in the file 'INMATES.DAT' - note, this is NOT a Stata file, and keep in mind its different extension.

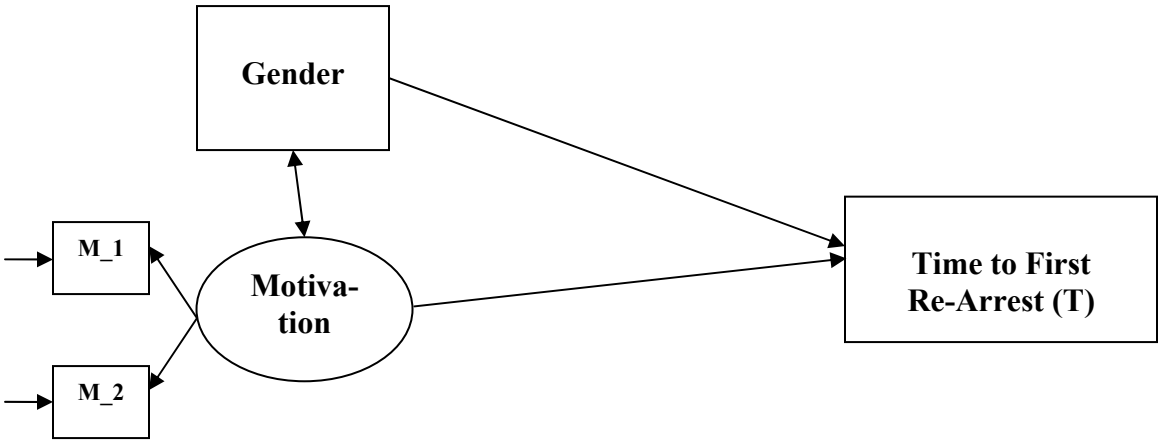
***Example* 2. A study of the time it takes from prison release to first re-arrest, for $n = 3120$ inmates released within a given month from $J = 851$ prisons.**

With respect to data on this example study, we have access in addition to data on their income and gender as well as motivation to integrate back into civil life.

A model of interest would be the one displayed in the next figure (next page).

Like in the last model (see last figure with the alienation construct), we need to account for the clustering effect of inmates in correctional facilities. Hence, we are dealing here with a two-level (multilevel) CT-EHA accounting for nesting.

We fit this CT-MEHA model as follows with Mplus (see command file presented after figure next).




```
TITLE:      MULTILEVEL EHA WITH A LATENT COVARIATE.
            EXTENDED COX PH MODEL.
            DATA ON RELEASED INMATE RE-ARREST.

DATA:       FILE = INMATES.DAT;

VARIABLE:   NAMES = FACILID INMID TIME EVENT GENDER
             MOTIVN1 MOTIVN2;
            USEV = TIME EVENT GENDER MOTIVN1 MOTIVN2;
            SURVIVAL = TIME(ALL);
            TIMECENSORED = EVENT(1=NOT 0=RIGHT);
            CLUSTER = FACILID;

ANALYSIS:   BASEHAZARD = OFF; ! REQUESTS FRAMEWORK OF COX REGRESSION.
            TYPE = COMPLEX; !   ACCOUNTS FOR NESTING WITHIN FACILITIES.

MODEL:      MOT BY MOTIVN1 MOTIVN2; ! factor model, latent predictor.
            TIME ON MOT GENDER;
            MOT WITH GENDER;
```

The following results are then obtained.

```
Mplus VERSION 7
MUTHEN & MUTHEN
4/1/2014   9:20 PM


INPUT INSTRUCTIONS

  TITLE:      MULTILEVEL EHA WITH A LATENT COVARIATE.
              EXTENDED COX PH MODEL.
              DATA ON RELEASED INMATE RE-ARREST.

  DATA:       FILE = INMATES.DAT;

  VARIABLE:   NAMES = FACILID INMID TIME EVENT GENDER
               MOTIVN1 MOTIVN2;
              USEV = TIME EVENT GENDER MOTIVN1 MOTIVN2;
              SURVIVAL = TIME(ALL);
              TIMECENSORED = EVENT(1=NOT 0=RIGHT);
              CLUSTER = FACILID;

  ANALYSIS:   BASEHAZARD = OFF; ! REQUESTS FRAMEWORK OF COX REGRESSION.
              TYPE = COMPLEX; !   ACCOUNTS FOR NESTING WITHIN FACILITIES.

  MODEL:      MOT BY MOTIVN1 MOTIVN2; ! factor model, latent predictor.
              TIME ON MOT GENDER;
              MOT WITH GENDER;




INPUT READING TERMINATED NORMALLY


MULTILEVEL EHA WITH A LATENT COVARIATE.
EXTENDED COX PH MODEL.
DATA ON RELEASED INMATE RE-ARREST.

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        3120

Number of dependent variables                                    3
Number of independent variables                                  1
Number of continuous latent variables                            1

Observed dependent variables

  Continuous
   MOTIVN1     MOTIVN2

  Time-to-event (survival)
   TIME

Observed independent variables
   GENDER
```

```
Continuous latent variables
   MOT


Variables with special functions

  Cluster variable      FACILID


  Time-censoring variables
   EVENT


Estimator                                                MLR
Information matrix                                  OBSERVED
Optimization Specifications for the Quasi-Newton Algorithm for
Continuous Outcomes
  Maximum number of iterations                           100
  Convergence criterion                            0.100D-05
Optimization Specifications for the EM Algorithm
  Maximum number of iterations                           500
  Convergence criteria
    Loglikelihood change                           0.100D-02
    Relative loglikelihood change                  0.100D-05
    Derivative                                     0.100D-02
Optimization Specifications for the M step of the EM Algorithm for
Categorical Latent variables


  Number of M step iterations                              1
  M step convergence criterion                     0.100D-02
  Basis for M step termination                     ITERATION
Optimization Specifications for the M step of the EM Algorithm for
Censored, Binary or Ordered Categorical (Ordinal), Unordered
Categorical (Nominal) and Count Outcomes
  Number of M step iterations                              1
  M step convergence criterion                     0.100D-02
  Basis for M step termination                     ITERATION
  Maximum value for logit thresholds                      15
  Minimum value for logit thresholds                     -15
  Minimum expected cell size for chi-square        0.100D-01
Optimization algorithm                                   EMA
Integration Specifications
  Type                                              STANDARD
  Number of integration points                           15
  Dimensions of numerical integration                     1
  Adaptive quadrature                                    ON
Base Hazard                                              OFF
Cholesky                                                 OFF

Input data file(s)
  INMATES.DAT

Input data format  FREE
```

SUMMARY OF DATA

Number of clusters 851

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters 11

Loglikelihood

H0 Value	-34289.580
H0 Scaling Correction Factor for MLR	18.3135

Information Criteria

Akaike (AIC)	68601.160
Bayesian (BIC)	68667.661
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	68632.710

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
MOT BY				
MOTIVN1	1.000	0.000	999.000	999.000
MOTIVN2	1.493	6.541	0.228	0.819
TIME ON				
MOT	0.139	0.855	0.163	0.871
TIME ON				
GENDER	-0.002	0.001	-3.232	0.001
MOT WITH				
GENDER	0.538	1.804	0.298	0.766
Means				
GENDER	94.528	4.668	20.248	0.000
Intercepts				
MOTIVN1	0.516	0.009	59.718	0.000
MOTIVN2	26.095	0.210	124.245	0.000
Variances				
GENDER	16043.476	5554.184	2.889	0.004
MOT	0.059	0.251	0.236	0.814
Residual Variances				
MOTIVN1	0.190	0.251	0.757	0.449
MOTIVN2	42.964	1.506	28.523	0.000

```
QUALITY OF NUMERICAL RESULTS
     Condition Number for the Information Matrix              0.380E-06
       (ratio of smallest to largest eigenvalue)
```

The last presented analysis result section suggests that within gender, motivation does not affect time to re-arrest. Further, given motivation, gender does affect this time, however. (Additional interpretation of command file and output will be provided during the pertinent workshop session.)

Hierarchical multilevel EHA modeling clustering effect: The Cox regression model with frailty

As indicated earlier in the w/shop, the introduction of the frailty random effect allows one to model the clustering effect in a two-level DT-EHA setting. In the preceding discussion in this section 5.2, we have been only accounting for this effect rather than modeling it.

Like in that earlier DT-EHA setting, it is also possible to specifically model the clustering effect using frailty in the multilevel CT-EHA context, by appropriately extending the Cox PH model.

We do it in this subsection. To this end, we introduce a random intercept in this model, leading to the following equation for the hazard function:

(5.7) $$h(t, \underline{x}_{ij}, \underline{y}_{ij}) = h_0(t)\exp(\beta_{0j} + \sum_{k=1}^{r}\beta_k x_{k,ij} + \sum_{s=1}^{q}\gamma_j \eta_{s,ij})$$

where $i = 1, \ldots, n_j$ = sample size within level2 unit, i.e., #level-1 units in the jth level-2 unit, and $j = 1, \ldots, J$ = number of level-2 units.

With simple algebra, from the right-hand side of (5.7) we obtain

(5.8) $$h(t, \underline{x}_{ij}, \underline{y}_{ij}) = \exp(\beta_{0j}) h_0(t) \exp(\sum_{k=1}^{r} \beta_k x_{k,ij} + \sum_{s=1}^{q} \gamma_j \eta_{s,ij}),$$

and hence

(5.9) $$h(t, \underline{x}_{ij}, \underline{y}_{ij} \mid \beta_{0j}) = \alpha h(t, \underline{x}_{ij}, \underline{y}_{ij}) \qquad (\alpha = \exp(\beta_{0j})).$$

Equation (5.9) can also be taken as a definition of frailty, viz. as a random effect that once fixed in fact multiplies the hazard function for the event (e.g., Kleinbaum & Klein, 2005).

This two-level Cox regression model (i.e., a Cox PH model with frailty) is fitted using the same general approach discussed earlier in this section 5.2, and is also implemented in Mplus.

I illustrate with the following empirical example of a two-level Cox PH model. In a study of $n = 977$ students, nested in $J = 108$ schools in a given state, the number of days they participated in soccer training was recorded, as well as the income of their family and a measure collected of the size of the school they were enrolled in.

We would like to see if there is a relationship between income and time being coached in soccer, in relation to the size of the school.

We use for this aim the two-level Cox PH model with frailty, thus (i) modeling the clustering effect by including a random intercept, and thereby (ii) relating it to the size of the school, i.e., a level-2 variable.

I should like to stress that with this approach, we are interested in a particular cross-level effect as well, viz. in the effect of a level-2 variable (school size) upon the level-1 outcome variable.

We use for this aim the following Mplus input file.

```
TITLE:      TWO-LEVEL COX PROPORTIONAL HAZARD MODEL
            WITH FRAILTY.

DATA:       FILE = 2L_COX_PH.DAT;

VARIABLE:   NAMES = TIME INCOME SIZE EVENT SCHOOLID;
            CLUSTER = SCHOOLID;
            WITHIN = INCOME;
            BETWEEN = SIZE;
            SURVIVAL = TIME(ALL);
            TIMECENSORED = EVENT(0 = NOT 1 = RIGHT);

ANALYSIS:   TYPE = TWOLEVEL;
            BASEHAZARD = OFF;

MODEL:      %WITHIN% ! LEVEL-1 MODEL IN MLM VERNACULAR
            TIME ON INCOME;

            %BETWEEN% ! LEVEL-2 MODEL IN MLM VERNACULAR
            TIME ON SIZE;
            TIME;
```

We obtain the following results.

```
INPUT READING TERMINATED NORMALLY


TWO-LEVEL COX PROPORTIONAL HAZARD MODEL
WITH FRAILTY.

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         977

Number of dependent variables                                    1
Number of independent variables                                  2
Number of continuous latent variables                            0

Observed dependent variables

  Time-to-event (survival)
   TIME
```

```
Observed independent variables
   INCOME      SIZE

Variables with special functions

  Cluster variable      SCHOOLID

  Time-censoring variables
   EVENT

  Within variables
   INCOME

  Between variables
   SIZE

Estimator                                                MLR
Information matrix                                  OBSERVED
Optimization Specifications for the Quasi-Newton Algorithm for
Continuous Outcomes
  Maximum number of iterations                           100
  Convergence criterion                            0.100D-05
Optimization Specifications for the EM Algorithm
  Maximum number of iterations                           500
  Convergence criteria
    Loglikelihood change                           0.100D-02
    Relative loglikelihood change                  0.100D-05
    Derivative                                     0.100D-02


Optimization Specifications for the M step of the EM Algorithm for
Categorical Latent variables
  Number of M step iterations                              1
  M step convergence criterion                     0.100D-02
  Basis for M step termination                     ITERATION
Optimization Specifications for the M step of the EM Algorithm for
Censored, Binary or Ordered Categorical (Ordinal), Unordered
Categorical (Nominal) and Count Outcomes
  Number of M step iterations                              1
  M step convergence criterion                     0.100D-02
  Basis for M step termination                     ITERATION
  Maximum value for logit thresholds                      15
  Minimum value for logit thresholds                     -15
  Minimum expected cell size for chi-square        0.100D-01
Optimization algorithm                                   EMA
Integration Specifications
  Type                                              STANDARD
  Number of integration points                           15
  Dimensions of numerical integration                     1
  Adaptive quadrature                                    ON
Base Hazard                                              OFF
Cholesky                                                  ON

Input data file(s)
  2L_COX_PH.DAT
Input data format  FREE
```

SUMMARY OF DATA

Number of clusters	108

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters	3

Loglikelihood

H0 Value	115.136
H0 Scaling Correction Factor for MLR	1.0403

Information Criteria

Akaike (AIC)	-224.273
Bayesian (BIC)	-209.619
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	-219.147

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Within Level				
TIME ON				
INCOME	0.426	0.043	9.826	0.000
Between Level				
TIME ON				
SIZE	0.215	0.079	2.722	0.006
Residual Variances				
TIME	0.488	0.109	4.495	0.000

QUALITY OF NUMERICAL RESULTS

Condition Number for the Information Matrix (ratio of smallest to largest eigenvalue)	0.304E+00

We see from this analysis result section that income enhances the risk of experiencing the event (leaving the soccer coaching program).

Similarly, in larger schools the hazard of leaving this program is higher.

These findings may be explained with the relatively high injury rate from soccer (in the US), as well as the availability of a multitude of competing sport coaching programs especially in larger schools.

Current limitations of used Cox regression extension

While the discussed multilevel CT-EHA approach is rather attractive, it is still almost in its 'infancy' (but that's a possible 'plus point' as well), and the following are its limitations at present:

1. **No direct test of the PH assumption. (Indirect test possible.)**
2. **No residuals available (yet).**
3. **No interactions of latent variables, or of latent and observed variables. ('Detour' possible.)**
4. **No omnibus test of (overall) model fit available (related to 2.)**

We are now ready to move on to an alternative approach to MEHA with discrete time.

5.3. Multilevel discrete-time EHA with latent variables

First we need to note that another approach to multilevel DT-EHA, which is an alternative to that discussed in Section 4 of this workshop, is available within the LVM modeling framework.

I should like to point out here that this alternative approach presents a number of important modeling opportunities. Perhaps the most noteworthy of them is the possibility of inclusion of fallible covariates measured by multiple (error-prone) indicators.

To describe this LVM approach to multilevel DT-EHA, let j_i be the last time period of data collection for the *i*th person in a sample of subjects from a studied population.

Thereby, if a study consists of P time periods, obviously $j_i < P$ or $j_i = P$ ($i = 1, \ldots, n$). That is, in that period j_i, the *i*th subject

- **(a) experienced the event, or**
- **(b) the study ended, or**
- **(c) he/she was lost to follow up (withdrew from study).**

In order to carry out multilevel DT-EHA via LVM, denote an individual's data for the *j*th time period as follows:

$u_j = 1$ if he/she experiences the event, or
$= 0$ if he/she doesn't.

Thus, every subject has a vector of j_i elements, which are 0 or 1 ($i = 1, \ldots, n$).

To produce equal length data records for each individual, which we need in order to be in a position to analyze them via LVM subsequently, we do the following 'data extension' (cf. Section 4).

If $j_i < P$, add $(P - j_i)$ symbols for missing data, such as say '-999', to complete the data record (for the *i*th subject). Then each subject will have a data vector $\underline{u}_i = (u_1, \ldots, u_P)$ full with 0's, 1's, and/or -999's, depending on his/her survival experience.

Following are a few data presentation examples for this type of data formatting.

<u>Examples</u>: A study consisting of 8 time periods, e.g., the successive semesters (half-years) in high school. Suppose we were interested in first drug use while in high school.

1. A subject who never used drugs will have this data record:

(0 0 0 0 0 0 0 0)

2. A person who used drug in 1st semester of junior year:

(0 0 0 0 1 -999 -999 -999)

3. A person who drops out of the study after ending sophomore year:

(0 0 0 0 -999 -999 -999 -999) .

Main assumption of DT-EHA (cf. Section 4): The pattern of these missing data (censoring) is ignorable. That is, the reason for an individual being censored is unrelated to his/her event status following it (dropping out of the study; this is akin to the missing at random, MAR, assumption).

As mentioned earlier in the workshop, this ignorability assumption is often referred to as non-informative censoring (NIC; Section 4).

It is sufficient that NIC is ensured *conditional* on covariates included in a model under consideration.

With this data arrangement, we can apply the same way of reasoning re. DT-EHA as in Section 4, but using Mplus that allows any of the predictors to be fallible, i.e., to include in the model latent variables with multiple indicators.

In this LVM approach to DT-EHA, each of the above event indicators in the observed vector u of event indicators is considered a coarse (binary) measurement of an underlying propensity to experience the event at that time interval (observation window).

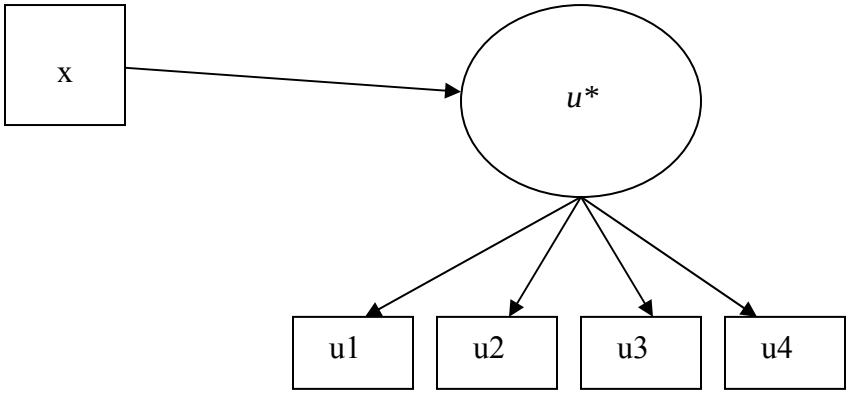
The latent variable that is their common factor, denoted *u in the figure below, represents subject susceptibility to experience the event during the study period.**

(This is the same variable denoted θ in item-response theory/models; e.g., Raykov & Marcoulides, 2011.)

***Example*: Consider the setting of DT-EHA, with <u>x</u> encompassing a set of observed variables and/or latent variables with multiple indicators each (with subjects nested in level-2 units).**

To exemplify, suppose we were interested in studying time to first use of cannabis in high-school, with 'x' below being family income, for students nested in several neighborhoods in a large urban area.

The model of interest would then be as follows.



To fit this model, accounting for students' nesting within neighborhoods, i.e., to carry out two-level DT-EHA with LVM (accounting for the clustering effect), we need the following M*plus* input file:

```
TITLE:      TWO-LEVEL DT-EHA USING LVM - ACCOUNTING
            FOR SUBJECT NESTING.

DATA:       FILE = CANABIS.DAT;
```

```
VARIABLE:   NAMES = U1-U4 X N_HOOD;
            CATEGORICAL = U1-U4;
            MISSING = ALL(999);
            CLUSTER = N_HOOD;

ANALYSIS:   TYPE = COMPLEX;
            ESTIMATOR = MLR; ! 'COUNTERPART' OF VCE(ROBUST).

MODEL:      F BY U1-U4;
            F ON X;
            F@0;
```

The following results are then obtained.

```
INPUT READING TERMINATED NORMALLY

TWO-LEVEL DT-EHA USING LVM - ACCOUNTING
FOR SUBJECT NESTING.

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                         348

Number of dependent variables                                    4
Number of independent variables                                  1
Number of continuous latent variables                            1

Observed dependent variables

  Binary and ordered categorical (ordinal)
   U1          U2          U3          U4

Observed independent variables
   X

Continuous latent variables
   F

Variables with special functions

  Cluster variable      N_HOOD

Estimator                                                      MLR
Information matrix                                        OBSERVED
Optimization Specifications for the Quasi-Newton Algorithm for
Continuous Outcomes
  Maximum number of iterations                                 100
  Convergence criterion                                 0.100D-05
Optimization Specifications for the EM Algorithm
  Maximum number of iterations                                 500
```

```
  Convergence criteria
    Loglikelihood change                                    0.100D-02
    Relative loglikelihood change                           0.100D-05
    Derivative                                              0.100D-02
Optimization Specifications for the M step of the EM Algorithm for
Categorical Latent variables
  Number of M step iterations                                       1
  M step convergence criterion                              0.100D-02
  Basis for M step termination                              ITERATION
Optimization Specifications for the M step of the EM Algorithm for
Censored, Binary or Ordered Categorical (Ordinal), Unordered
Categorical (Nominal) and Count Outcomes
  Number of M step iterations                                       1
  M step convergence criterion                              0.100D-02
  Basis for M step termination                              ITERATION
  Maximum value for logit thresholds                               15
  Minimum value for logit thresholds                              -15
  Minimum expected cell size for chi-square                 0.100D-01
Maximum number of iterations for H1                              2000
Convergence criterion for H1                                0.100D-03
Optimization algorithm                                           EMA
Integration Specifications
  Type                                                     STANDARD
  Number of integration points                                     15
  Dimensions of numerical integration                               0
  Adaptive quadrature                                              ON
Link                                                            LOGIT
Cholesky                                                         ON

Input data file(s)
  CANABIS.DAT
Input data format  FREE


SUMMARY OF DATA

     Number of missing data patterns             7

     Number of clusters          9


COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100


     PROPORTION OF DATA PRESENT FOR U


           Covariance Coverage
              U1            U2            U3            U4
              ________      ________      ________      ________
 U1             0.514
 U2             0.448         0.934
 U3             0.339         0.747         0.747
 U4             0.230         0.537         0.537         0.537
```

UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES

U1		
Category 1	0.872	156.000
Category 2	0.128	23.000
U2		
Category 1	0.800	260.000
Category 2	0.200	65.000
U3		
Category 1	0.719	187.000
Category 2	0.281	73.000
U4		
Category 1	0.754	141.000
Category 2	0.246	46.000

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters	8

Loglikelihood

H0 Value	-468.315
H0 Scaling Correction Factor for MLR	0.7051

Information Criteria

Akaike (AIC)	952.630
Bayesian (BIC)	983.448
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	958.069

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
F	BY				
U1		1.000	0.000	999.000	999.000
U2		1.157	0.617	1.876	0.061
U3		0.857	0.475	1.802	0.072
U4		1.867	1.122	1.663	0.096
F	ON				
X		0.445	0.220	2.022	0.043
Thresholds					
U1$1		2.024	0.158	12.823	0.000
U2$1		1.448	0.077	18.694	0.000

```
    U3$1               0.919      0.113      8.157      0.000
    U4$1               1.058      0.151      7.002      0.000

 Residual Variances
    F                  0.000      0.000    999.000    999.000


QUALITY OF NUMERICAL RESULTS

     Condition Number for the Information Matrix 0.259E-02
       (ratio of smallest to largest eigenvalue)
```

As we can see from the last analysis result section, income increases significantly students' susceptibility to using cannabis (as far as the hazard for first use is concerned).

I will discuss the remaining parts of the output in the pertinent workshop session.

6. Analysis of time-to-event data from nationally representative samples

Social and political scientists are becoming increasingly more involved in *nationally representative studies* (NRSs; e.g., HRS, NHANES, PIRLS, PISA, TIMMS, to name only a few).

A main reason for this interest is the growing realization that studied human populations in these and cognate disciplines are becoming increasingly more diverse and heterogeneous.

NRSs are examples of *complex design studies* (CDSs; Heeringa, West, & Berglund, 2010, *Applied survey data analysis*). They are also called at times *complex sample studies*. CDSs usually take a considerable period of time to conceptualize, plan, and conduct - mostly done by professional *sampling* statisticians - before the data collection process can begin.

The unique feature of a CDS is the use of *primary sampling units* (PSUs; not infrequently counties), within *strata* that which represent relatively homogeneous clusters where further sampling is conducted (or neighborhoods/households/respondents) and subsequent data collection is proceeded with.

Another major feature of a CDS is that it contains *subject specific weights* – reflecting the number of subjects in the population of interest that each individual can be thought of representing in the available sample.

These weights are also called *probability weights* or *sampling weights*, and are worked out by sampling statisticians before the data analysis. They are typically supplied/come with the study data set in question (e.g., upon downloading it from pertinent website).

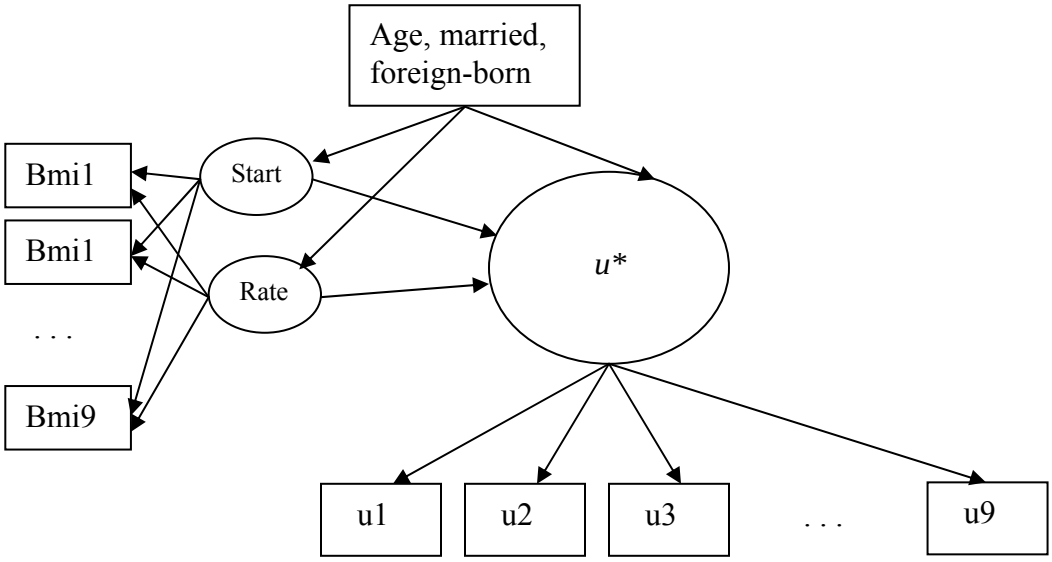
Analysis of CDS data represents also an application of multilevel modeling (due to subjects being in general nested within PSUs). When the main response variable is time to an event, we are dealing with multilevel EHA of CDS data, the subject of this Section 6.

This analysis is facilitated by recent advances in statistical modeling, incl. in particular the software Mplus and Stata, thus allowing straight-forward survey data analysis.

I illustrate this discussion on multilevel DT-EHA with survey data using two empirical examples using data from the increasingly popular Health and Retirement Study (HRS) in the US (data are obtainable from this website: http://hrsonline.umich.edu; *n* = 5,961 retired women involved in the first, and *n* = 9479 retirees in the second example).

***Example* 1: How are the developmental trajectories of BMI related to the susceptibility to death in retired women in the USA, controlling for age, education, marital status, an local vs. foreign born?**

The data are contained in file 'HRS-BMI-DTH.DAT' and the diagram of the model of interest is as follows.



To fit this model, we make use of the following Mplus command file (I will discuss in detail its input commands in the pertinent workshop session; in the last figure, latent disturbance covariances are omitted, to avoid graphical clutter).

```
TITLE:      MULTILEVEL DT-EHA ON HRS DATA (BMI + COVARIATES → DTH).

DATA:       FILE = HRS_BMI_DTH.dat;

VARIABLE:   NAMES = hrs female yob cyob race white black
                hisp other nonmar
                foreign eduyear ceduyear educ educ3 h1-h9
                dead1-dead9
                died1-died9 bmi1-bmi9
                weight92 strata seunit;
            USEV =  bmi1-bmi9 weight92 strata seunit
                    died2-died9 nonmar foreign educ
                    age;
            MISSING = ALL(-999);
            WEIGHT = WEIGHT92;
              STRA = STRATA;
            CLUSTER = SEUNIT;
            CATEGORICAL = DIED2-DIED9;

DEFINE:     SEUNIT = 1000*STRATA + SEUNIT;
            AGE = 1992-YOB;

ANALYSIS:   TYPE = COMPLEX;
            ESTIMATOR = MLR;

MODEL:      I S | BMI1@-4 BMI2@-3 BMI3@-2 BMI4@-1
            BMI5@0 BMI6@1 BMI7@2 BMI8@3 BMI9@4;
            BMI1-BMI9(1);

            I S ON AGE NONMAR FOREIGN EDUC;

            SD BY DIED2-DIED9;
            SD@0;

            SD ON I S AGE NONMAR FOREIGN EDUC;
```

The results obtained then are as follows.

```
INPUT READING TERMINATED NORMALLY

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        5169

Number of dependent variables                                   17
Number of independent variables                                  4
Number of continuous latent variables                            3

Observed dependent variables

  Continuous
   BMI1        BMI2        BMI3        BMI4        BMI5        BMI6
   BMI7        BMI8        BMI9

  Binary and ordered categorical (ordinal)
   DIED2       DIED3       DIED4       DIED5       DIED6       DIED7
   DIED8       DIED9

Observed independent variables
   NONMAR      FOREIGN     EDUC        AGE

Continuous latent variables
   SD          I           S

Variables with special functions

  Stratification        STRATA
  Cluster variable      SEUNIT
  Weight variable       WEIGHT92

Estimator                                                      MLR
Information matrix                                        OBSERVED
Optimization Specifications for the Quasi-Newton Algorithm for
Continuous Outcomes
  Maximum number of iterations                                 100
  Convergence criterion                                0.100D-05
Optimization Specifications for the EM Algorithm
  Maximum number of iterations                                 500
  Convergence criteria
    Loglikelihood change                               0.100D-02
    Relative loglikelihood change                      0.100D-05
    Derivative                                         0.100D-02
Optimization Specifications for the M step of the EM Algorithm for
Categorical Latent variables
  Number of M step iterations                                    1
  M step convergence criterion                         0.100D-02
  Basis for M step termination                         ITERATION
Optimization Specifications for the M step of the EM Algorithm for
Censored, Binary or Ordered Categorical (Ordinal), Unordered
Categorical (Nominal) and Count Outcomes
  Number of M step iterations                                    1
  M step convergence criterion                         0.100D-02
  Basis for M step termination                         ITERATION
```

```
  Maximum value for logit thresholds                          15
  Minimum value for logit thresholds                         -15
  Minimum expected cell size for chi-square            0.100D-01
Maximum number of iterations for H1                         2000
Convergence criterion for H1                         0.100D-03
Optimization algorithm                                       EMA
Integration Specifications
  Type                                                  STANDARD
  Number of integration points                                15
  Dimensions of numerical integration                          2
  Adaptive quadrature                                         ON
Link                                                       LOGIT
Cholesky                                                     OFF

Input data file(s)
  HRS_BMI_DTH.dat
Input data format  FREE


SUMMARY OF DATA

     Number of missing data patterns           237
     Number of strata                   52
     Number of clusters                104


COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100


     PROPORTION OF DATA PRESENT


           Covariance Coverage
              DIED2         DIED3         DIED4         DIED5         DIED6
              ________      ________      ________      ________
________
 DIED2          0.998
 DIED3          0.984         0.984
 DIED4          0.952         0.952         0.952
 DIED5          0.922         0.922         0.922         0.922
 DIED6          0.887         0.887         0.887         0.887         0.887
 DIED7          0.846         0.846         0.846         0.846         0.846
 DIED8          0.797         0.797         0.797         0.797         0.797
 DIED9          0.765         0.765         0.765         0.765         0.765
 BMI1           0.998         0.984         0.952         0.922         0.887
 BMI2           0.899         0.899         0.877         0.852         0.824
 BMI3           0.855         0.855         0.855         0.832         0.805
 BMI4           0.815         0.815         0.815         0.813         0.789
 BMI5           0.766         0.766         0.766         0.766         0.765
 BMI6           0.734         0.734         0.734         0.734         0.734
 BMI7           0.713         0.713         0.713         0.713         0.713
 BMI8           0.677         0.677         0.677         0.677         0.677
 BMI9           0.653         0.653         0.653         0.653         0.653
 NONMAR         0.998         0.984         0.952         0.922         0.887
 FOREIGN        0.998         0.984         0.952         0.922         0.887
```

EDUC	0.998	0.984	0.952	0.922	0.887
AGE	0.998	0.984	0.952	0.922	0.887

Covariance Coverage

	DIED7	DIED8	DIED9	BMI1	BMI2
DIED7	0.846				
DIED8	0.797	0.797			
DIED9	0.765	0.765	0.765		
BMI1	0.846	0.797	0.765	1.000	
BMI2	0.789	0.749	0.720	0.899	0.899
BMI3	0.771	0.737	0.710	0.855	0.821
BMI4	0.760	0.735	0.708	0.815	0.779
BMI5	0.737	0.720	0.694	0.766	0.733
BMI6	0.734	0.718	0.692	0.734	0.702
BMI7	0.713	0.713	0.689	0.713	0.681
BMI8	0.677	0.677	0.677	0.677	0.648
BMI9	0.653	0.653	0.653	0.653	0.622
NONMAR	0.846	0.797	0.765	1.000	0.899
FOREIGN	0.846	0.797	0.765	1.000	0.899
EDUC	0.846	0.797	0.765	1.000	0.899
AGE	0.846	0.797	0.765	1.000	0.899

Covariance Coverage

	BMI3	BMI4	BMI5	BMI6	BMI7
BMI3	0.855				
BMI4	0.781	0.815			
BMI5	0.732	0.739	0.766		
BMI6	0.696	0.699	0.695	0.734	
BMI7	0.674	0.675	0.668	0.683	0.713
BMI8	0.643	0.642	0.637	0.644	0.653
BMI9	0.618	0.621	0.616	0.620	0.623
NONMAR	0.855	0.815	0.766	0.734	0.713
FOREIGN	0.855	0.815	0.766	0.734	0.713
EDUC	0.855	0.815	0.766	0.734	0.713
AGE	0.855	0.815	0.766	0.734	0.713

Covariance Coverage

	BMI8	BMI9	NONMAR	FOREIGN	EDUC
BMI8	0.677				
BMI9	0.624	0.653			
NONMAR	0.677	0.653	1.000		
FOREIGN	0.677	0.653	1.000	1.000	
EDUC	0.677	0.653	1.000	1.000	1.000
AGE	0.677	0.653	1.000	1.000	1.000

PROPORTION OF DATA PRESENT FOR U

Covariance Coverage

	DIED2	DIED3	DIED4	DIED5	DIED6
DIED2	0.998				
DIED3	0.984	0.984			
DIED4	0.952	0.952	0.952		
DIED5	0.922	0.922	0.922	0.922	
DIED6	0.887	0.887	0.887	0.887	0.887
DIED7	0.846	0.846	0.846	0.846	0.846
DIED8	0.797	0.797	0.797	0.797	0.797
DIED9	0.765	0.765	0.765	0.765	0.765

Covariance Coverage

	DIED7	DIED8	DIED9
DIED7	0.846		
DIED8	0.797	0.797	
DIED9	0.765	0.765	0.765

PROPORTION OF DATA PRESENT FOR Y

Covariance Coverage

	BMI1	BMI2	BMI3	BMI4	BMI5
BMI1	1.000				
BMI2	0.899	0.899			
BMI3	0.855	0.821	0.855		
BMI4	0.815	0.779	0.781	0.815	
BMI5	0.766	0.733	0.732	0.739	0.766
BMI6	0.734	0.702	0.696	0.699	0.695
BMI7	0.713	0.681	0.674	0.675	0.668
BMI8	0.677	0.648	0.643	0.642	0.637
BMI9	0.653	0.622	0.618	0.621	0.616
NONMAR	1.000	0.899	0.855	0.815	0.766
FOREIGN	1.000	0.899	0.855	0.815	0.766
EDUC	1.000	0.899	0.855	0.815	0.766
AGE	1.000	0.899	0.855	0.815	0.766

Covariance Coverage

	BMI6	BMI7	BMI8	BMI9	NONMAR
BMI6	0.734				
BMI7	0.683	0.713			
BMI8	0.644	0.653	0.677		
BMI9	0.620	0.623	0.624	0.653	
NONMAR	0.734	0.713	0.677	0.653	1.000

FOREIGN	0.734	0.713	0.677	0.653	1.000
EDUC	0.734	0.713	0.677	0.653	1.000
AGE	0.734	0.713	0.677	0.653	1.000

Covariance Coverage

	FOREIGN	EDUC	AGE
FOREIGN	1.000		
EDUC	1.000	1.000	
AGE	1.000	1.000	1.000

UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES

DIED2		
Category 1	0.988	5109.082
Category 2	0.012	59.926
DIED3		
Category 1	0.987	5040.855
Category 2	0.013	65.929
DIED4		
Category 1	0.983	4856.425
Category 2	0.017	86.494
DIED5		
Category 1	0.977	4691.556
Category 2	0.023	108.150
DIED6		
Category 1	0.970	4473.005
Category 2	0.030	140.195
DIED7		
Category 1	0.981	4322.044
Category 2	0.019	82.686
DIED8		
Category 1	0.967	4027.461
Category 2	0.033	135.818
DIED9		
Category 1	0.966	3870.980
Category 2	0.034	137.356

THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters 34

Loglikelihood

H0 Value	-92054.420
H0 Scaling Correction Factor for MLR	2.2519

Information Criteria

Akaike (AIC)	184176.840
Bayesian (BIC)	184399.554
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	184291.514

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
I |				
BMI1	1.000	0.000	999.000	999.000
BMI2	1.000	0.000	999.000	999.000
BMI3	1.000	0.000	999.000	999.000
BMI4	1.000	0.000	999.000	999.000
BMI5	1.000	0.000	999.000	999.000
BMI6	1.000	0.000	999.000	999.000
BMI7	1.000	0.000	999.000	999.000
BMI8	1.000	0.000	999.000	999.000
BMI9	1.000	0.000	999.000	999.000
S |				
BMI1	-4.000	0.000	999.000	999.000
BMI2	-3.000	0.000	999.000	999.000
BMI3	-2.000	0.000	999.000	999.000
BMI4	-1.000	0.000	999.000	999.000
BMI5	0.000	0.000	999.000	999.000
BMI6	1.000	0.000	999.000	999.000
BMI7	2.000	0.000	999.000	999.000
BMI8	3.000	0.000	999.000	999.000
BMI9	4.000	0.000	999.000	999.000
SD BY				
DIED2	1.000	0.000	999.000	999.000
DIED3	0.630	0.218	2.884	0.004
DIED4	0.381	0.211	1.800	0.072
DIED5	0.798	0.200	4.000	0.000
DIED6	0.931	0.200	4.656	0.000
DIED7	0.805	0.192	4.197	0.000
DIED8	0.676	0.148	4.577	0.000
DIED9	0.680	0.165	4.131	0.000
SD ON				
I	-0.012	0.013	-0.931	0.352
S	-1.370	0.215	-6.364	0.000
I ON				
AGE	-0.088	0.028	-3.126	0.002
NONMAR	0.782	0.155	5.042	0.000
FOREIGN	-0.581	0.271	-2.142	0.032
EDUC	-0.563	0.062	-9.052	0.000

S ON				
AGE	**-0.017**	**0.002**	**-7.027**	**0.000**
NONMAR	**-0.005**	**0.021**	**-0.224**	**0.822**
FOREIGN	**-0.019**	**0.032**	**-0.611**	**0.541**
EDUC	**0.012**	**0.006**	**2.138**	**0.032**
SD ON				
AGE	**0.084**	**0.021**	**4.083**	**0.000**
NONMAR	**0.596**	**0.140**	**4.242**	**0.000**
FOREIGN	**-0.730**	**0.210**	**-3.468**	**0.001**
EDUC	**-0.313**	**0.057**	**-5.491**	**0.000**
Intercepts				
BMI1	**0.000**	**0.000**	**999.000**	**999.000**
BMI2	**0.000**	**0.000**	**999.000**	**999.000**
BMI3	**0.000**	**0.000**	**999.000**	**999.000**
BMI4	**0.000**	**0.000**	**999.000**	**999.000**
BMI5	**0.000**	**0.000**	**999.000**	**999.000**
BMI6	**0.000**	**0.000**	**999.000**	**999.000**
BMI7	**0.000**	**0.000**	**999.000**	**999.000**
BMI8	**0.000**	**0.000**	**999.000**	**999.000**
BMI9	**0.000**	**0.000**	**999.000**	**999.000**
I	**33.723**	**1.613**	**20.906**	**0.000**
S	**1.059**	**0.131**	**8.053**	**0.000**
Thresholds				
DIED2$1	**8.268**	**1.306**	**6.330**	**0.000**
DIED3$1	**6.634**	**1.199**	**5.535**	**0.000**
DIED4$1	**5.371**	**0.932**	**5.760**	**0.000**
DIED5$1	**6.719**	**1.084**	**6.200**	**0.000**
DIED6$1	**6.922**	**1.112**	**6.226**	**0.000**
DIED7$1	**6.889**	**1.124**	**6.130**	**0.000**
DIED8$1	**5.798**	**0.811**	**7.150**	**0.000**
DIED9$1	**5.748**	**0.943**	**6.095**	**0.000**
Residual Variances				
BMI1	**3.294**	**0.154**	**21.435**	**0.000**
BMI2	**3.294**	**0.154**	**21.435**	**0.000**
BMI3	**3.294**	**0.154**	**21.435**	**0.000**
BMI4	**3.294**	**0.154**	**21.435**	**0.000**
BMI5	**3.294**	**0.154**	**21.435**	**0.000**
BMI6	**3.294**	**0.154**	**21.435**	**0.000**
BMI7	**3.294**	**0.154**	**21.435**	**0.000**
BMI8	**3.294**	**0.154**	**21.435**	**0.000**
BMI9	**3.294**	**0.154**	**21.435**	**0.000**
SD	**0.000**	**0.000**	**999.000**	**999.000**
I	**29.023**	**1.120**	**25.921**	**0.000**
S	**0.204**	**0.014**	**14.895**	**0.000**

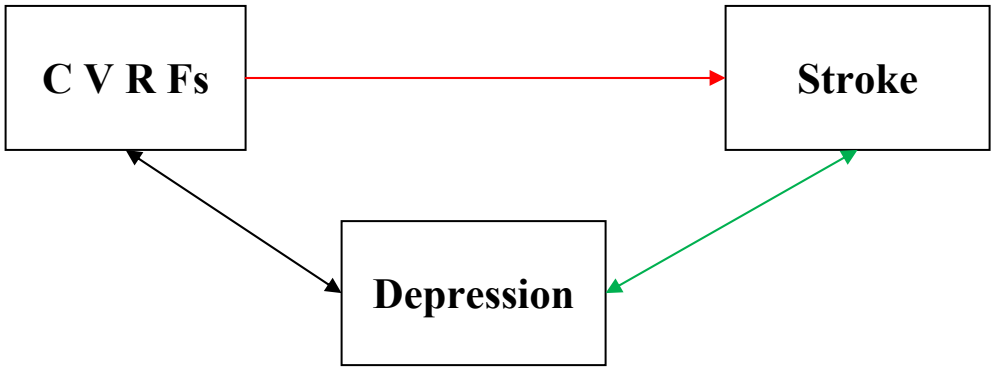
According to the presented modeling results, there's no effect of starting position (women's BMI at start of the study, 1992) but a negative effect of the rate of increase in it.

We will discuss in more detail this finding during the workshop, due to an important empirical 'law' related to subjects' weight, which is involved here (found in demographics a few years ago).

***Example* 2: Here we wish to address the vascular depression hypothesis (VDH), stipulating unique predictive power of depression with respect to subject susceptibility to experience a (first) stroke, over and above cardio-vascular and cardio-cerebral risk factors (CCRFs).**

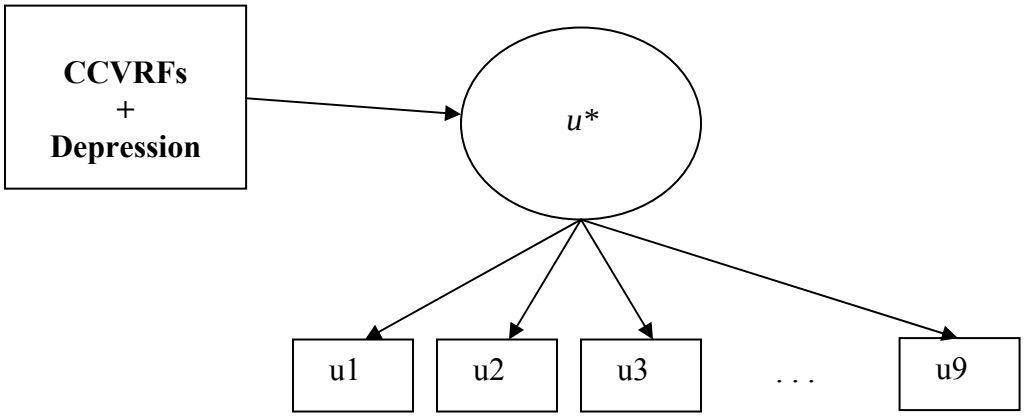
The conceptual model of the VDH is as follows.

Conceptual model of the VDH
(*not* a causal model; *no* path assumed causal):



The model of interest postulates susceptibility to stroke (*u in section 5.4) as the dependent variable, which is latent and with indicators being the event indicator at each of 9 waves of the HRS.**

In addition, this variable is regressed upon a number of CCRV and depression. A simple diagram of this model can be readily obtained as follows, noticing that 'x' there is actually a set of CCRFs and depression.



We fit this multilevel DT-EHA model with the following command file that is only a minor modification of the last used Mplus input file.

```
TITLE:      MULTILEVEL DT-EHA ON HRS DATA (BMI + COVARIATES → DTH).

DATA:       FILE = HRS_BMI_DTH.dat;

VARIABLE:   NAMES = hrs female yob cyob race white black
                hisp other married
                foreign edyrs ceduyear educ educ3 h1-h9
                dead1-dead9
                died1-died9 bmi1-bmi9
                cesd hi_bp diabetes cancer lung_dis heart_dis mental_hlth
                arthritis smoke age_by_g edyrs
                weight92 strata seunit;
            USEV =  bmi1 weight92 strata seunit female
                    died2-died9 married edyrs
                    age cesd hi_bp diabetes cancer lung_dis heart_dis
                    mental_hlth arthritis smoke age_by_g;
            MISSING = ALL(-999);
```

```
            WEIGHT = WEIGHT92;
              STRA = STRATA;
            CLUSTER = SEUNIT;
            CATEGORICAL = DIED2-DIED9;

DEFINE:     SEUNIT = 1000*STRATA + SEUNIT;
            AGE = 1992-YOB;
            AGE_BY_G = AGE*FEMALE;

ANALYSIS:   TYPE = COMPLEX;
            ESTIMATOR = MLR;

MODEL:      SD BY DIED2-DIED9;
            SD@0;

            SD ON AGE MARRIED FEMALE EDYRS BLACK HISP OTHER
             cesd hi_bp diabetes cancer lung_dis heart_dis mental_hlth
             arthritis smoke age_by_g bmi1;
```

Results obtained thereby are as follows (see above diagram, for model, appropriately extended here):

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
STROKE BY				
STROKE1	1.000	0.000	999.000	999.000
STROKE2	1.103	0.244	4.525	0.000
STROKE3	0.901	0.076	11.859	0.000
STROKE4	0.917	0.110	8.351	0.000
STROKE5	0.692	0.115	6.002	0.000
STROKE6	0.724	0.130	5.550	0.000
STROKE7	0.564	0.107	5.279	0.000
STROKE8	0.575	0.099	5.842	0.000
STROKE9	0.458	0.076	6.042	0.000
STROKE ON				
AGE	0.224	0.055	4.102	0.000
CESD	0.434	0.104	4.172	0.000
HI_BP	2.444	0.493	4.960	0.000
DIABETES	2.212	0.608	3.636	0.000
CANCER	0.565	0.809	0.699	0.485
LUNG_DIS	1.325	0.591	2.241	0.025
HEART_DIS	3.373	0.717	4.704	0.000
MENTAL_HLTH	0.613	0.554	1.106	0.269
ARTHRITIS	0.826	0.373	2.210	0.027
SMOKE	0.979	0.235	4.174	0.000
FEMALE	-1.115	0.449	-2.481	0.013
AGE_BY_G	0.060	0.334	0.179	0.858

BLACK	1.042	0.434	2.402	0.016
HISP	0.502	0.698	0.720	0.472
OTHER	0.500	1.169	0.427	0.669
MARRIED	-1.246	0.738	-1.688	0.091
EDYRS	-0.131	0.057	-2.312	0.021
BMI1	0.009	0.030	0.298	0.766

In this multiple testing setting, to work out which null hypotheses from the last analysis result section is to be rejected, we use the increasingly popular Benjamini-Hochberg (1995) FDR-procedure (see, e.g., Raykov, Lichtenberg, & Paulson, 2012, *SEM*, for an R-function accomplishing it).

This R-function yields here

$r = 6$, $T = 0$ (*p*-value threshold; software reported).

That is, to be rejected are the 6 hypotheses with smallest *p*-values, i.e., $p < .001$ here (as reported by software; these are for the covariates highlighted in red on preceding page).

Hence, according to these findings, depression provides *unique* predictive power with regard to propensity of getting a first stroke, even after controlling for all above covariates.

I should like to stress that this unique predictive power of depression holds with respect in particular to the 7 main medical conditions, as well as race, gender, age, age by gender interaction, marital status, education, BMI. That is, this predictive power of depression is *over and above* that of these 7 main medical conditions, 6 demographics, and BMI.
We will discuss further the above output in the pertinent workshop session.

7. Conclusion

Multilevel EHA is a relatively 'emerging' field in the social and political sciences, and is currently associated with substantial computational demands and time when fitting complicated models. The next several years will likely see a great increase in the interest in MEHA and associated improvement of the numerical/ computational situation.

We currently have 2 main avenues to follow in order to conduct MEHA (whether with continuous or discrete time). One is via accounting for the nesting effects (of any order), using robust/sandwich estimator of standard errors. This is a fairly straightforwardly utilized approach to modeling.

The second is via explicitly modeling of the clustering effect, and we used discussed such models, incl. multilevel Cox PH with frailty, as well as multilevel DT-EHA via the CLL model with frailty.

Limitations of MEHA lie primarily in the requirement for large samples, and methods for sample size determination are currently in their 'infancy'.

Similarly, I should like to emphasize that more complex MEHA and related models are readily developed and fitted along the lines discussed in this short course (incl. in particular those covered in Day 2). To this end, use the same methodological principles utilized in our preceding discussion; for the more complex models, employ as building blocks the models we focused on there.

Thereby, it should be kept in mind that due to the heavy computational burden, at present much more complex models than those covered in this course are more likely to be associated with numerical issues of various sorts.

This limitation is likely to be dealt with in the following years.

The next few years will see major progress in MEHA.

With this and all preceding in mind, I am hopeful that this short course will prove useful for you until then, and possibly (well) beyond.

Happy MEHA-ing!