Historical demography: From a crossroad to an interdisciplinary exercise

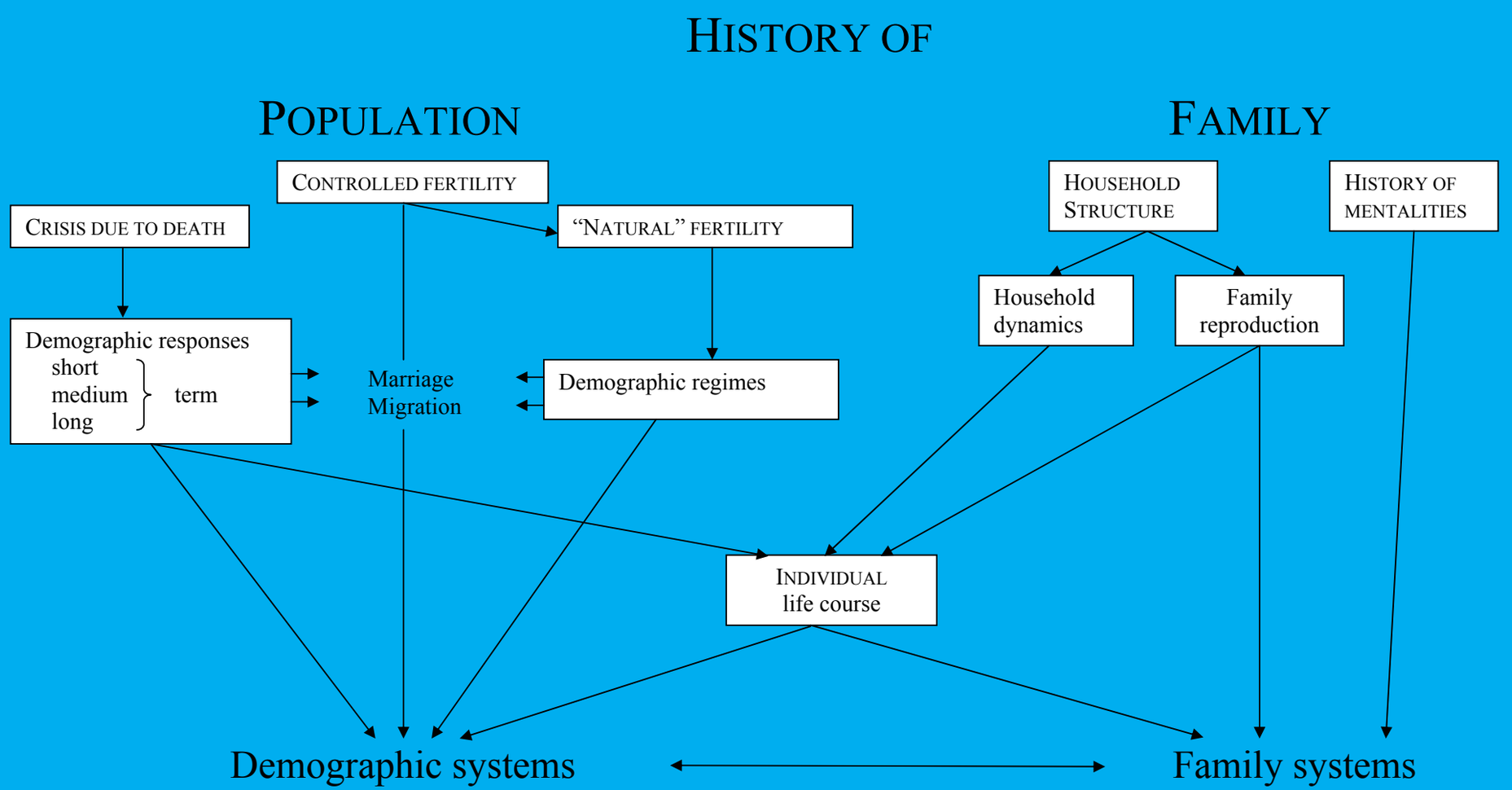
Some methodological issues

Michel Oris and Gilbert Ritschard, LaboDemo, University of Geneva
October 2003

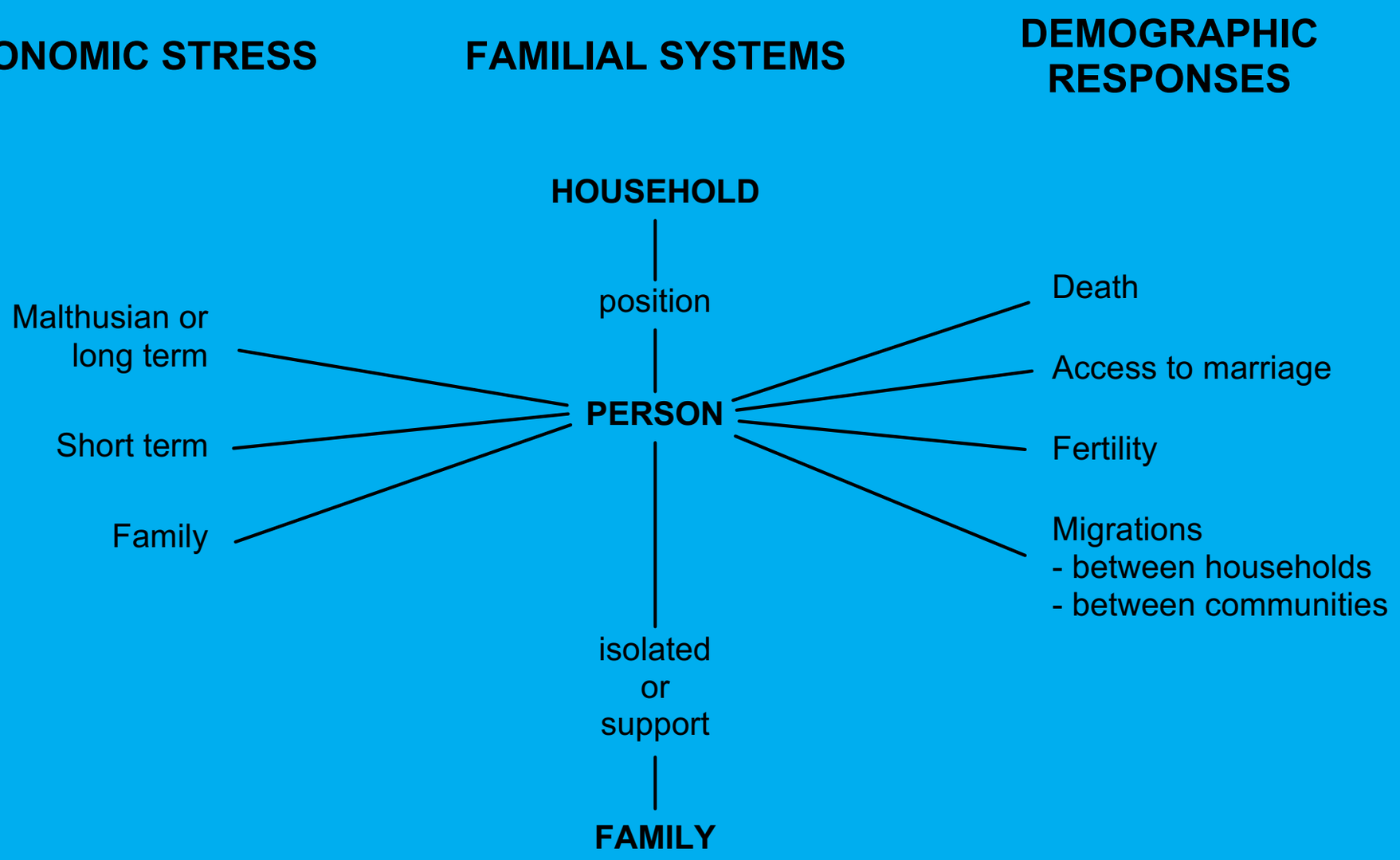
Table of Content

http://mephisto.unige.ch

1 Population and Family History: How they are Intertwined



EurAsia Project for a Comparative Population and Family Study



2 Individual and contextual effects

Formal setting of the interdisciplinary approach

i individual, subject, case
g group, context, family, ...
T time until event (death, recovering from stress, ...)

Event history approach: how can we explain differences in the hazard rate function $h(t) = p(T = t|T \geq t)$?

- predictable heterogeneity with covariates
  - individual covariates x_{ig} (education level, age when exposed to stress, sex, ...)
  - shared contextual covariates x_g (size of family, place of residence, household income, ...)
- Shared unpredictable heterogeneity (random effect resulting from unobserved covariates)

Issues

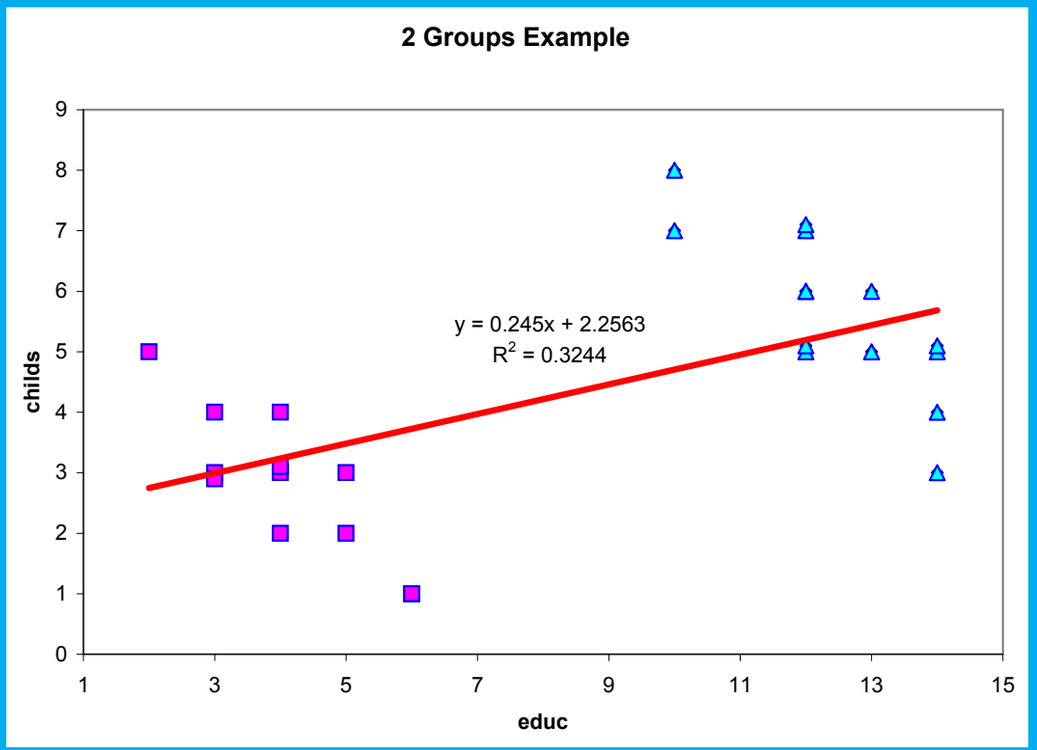
- Correlation between individuals sharing a context, an environment
  - What are we estimating? Need for robust standard errors of estimates
- Controlling the shared heterogeneity
  - Multilevel models

Other issues: multiple events

- competitive events (death, recovering from stress)
- simultaneous study of correlated events (death of mother, death of child)
  $\Rightarrow$ multivariate models

3 Models with clustered data

In the presence of clusters, the estimates $\hat{\beta}$ concern the relationships at the cluster level, rather than at the individual level



Ignoring the clusters, the observed discrepancy is at individual level

Robust standard errors

Lin and Wei (1989), Prentice et al. (1981), Therneau and Grambsch (2000)

Grouped Jackknife: variance of the $\hat{\beta}_{(g)}$ obtained by leaving out one group (cluster) $g = 1, \ldots, G$ at a time.

Can be expressed as a "sandwich" estimate of the variance

$$V_{sdw} = V\, B\, V$$

with V estimate of the variance matrix of $\hat{\beta}$ assuming independent observations and B correction factor

(Available with Stata, SAS, S-Plus)

Robust standard error estimates,

- do not affect the coefficient estimates
- affect their significance

Ignoring the dependence within cluster $\Rightarrow$ under- or overestimates of the standard errors (Kish and Frankel, 1974)

Cox model, return after emigration, Sart 1812-1846 (1143 obs, 5 years)

covariate	hazard ratio	sig. (indep)	sig. (cluster)
CHILD_FHD	.43	.027	.035
DEST_ARD	6.52	.000	.000
DEST_RUI	3.36	.002	.004
SINGLE	3.37	.013	.013

CHILD_FHD child of family head, DEST_ARD to Ardennes, DEST_RUI to rural or industrial regions

For a full scale example, see for instance Beeking et al. (2002) .

4 Modeling group heterogeneity

Robust standard errors useful if we are interested in the population average model (essentially the between cluster effects).

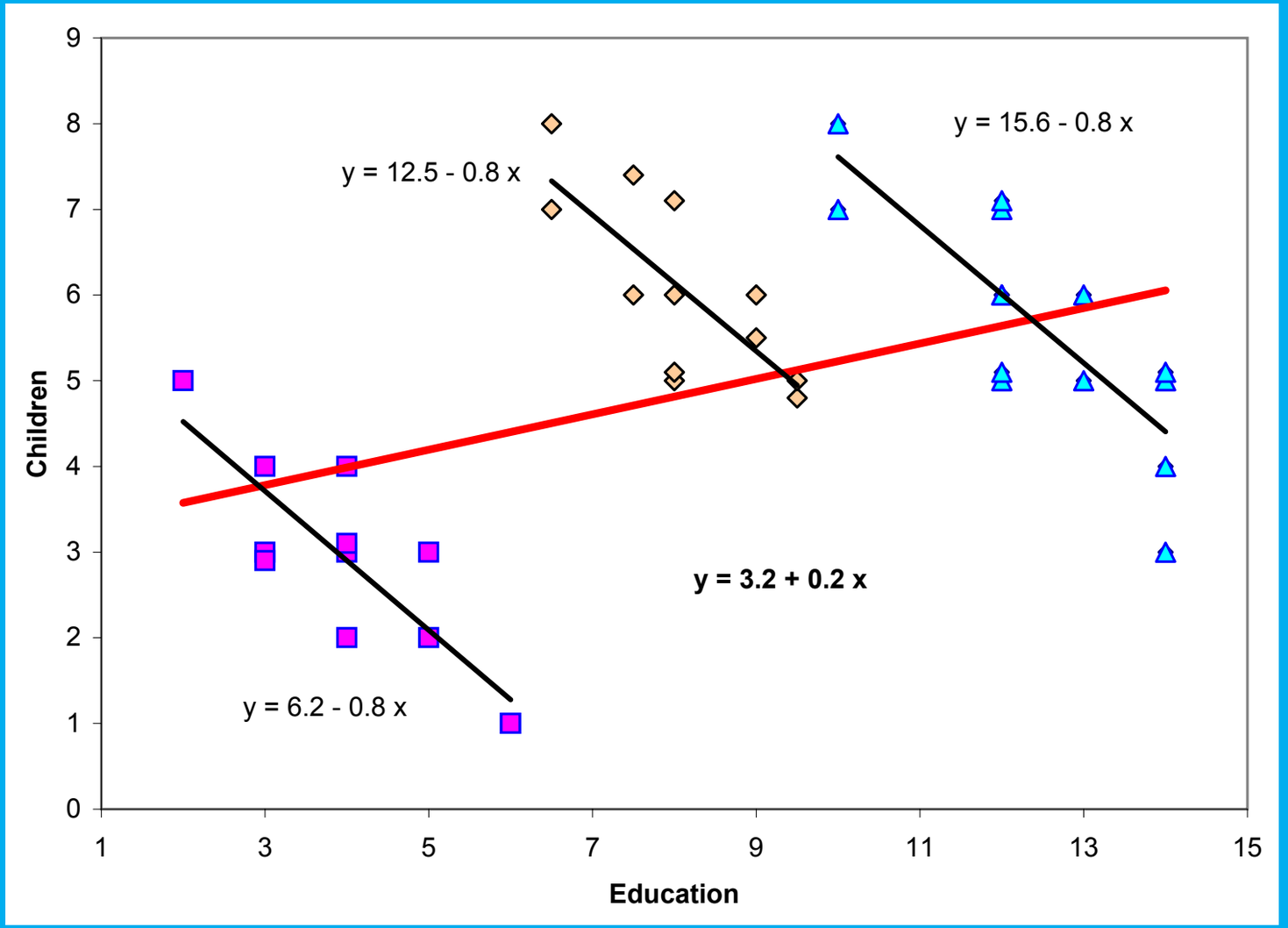
Do not help to describe the within cluster effects.

How can we control for the between cluster effects?

$\Rightarrow$ multilevel models

4.1 Principle of multilevel modeling

Goldstein (1995), Courgeau and Baccaïni (1998)



Some alternative linear models for G groups $g = 1, \ldots, G$

	model	err. stdev	# coef.	
m1	$y_{ig} = a + bx_{ig} + u_{ig}$	σ	$2+1$	average model
m2	$y_{ig} = a_g + b_g\, x_{ig} + u_{ig}$	$\sigma_1 \ldots \sigma_G$	$G(2+1)$	independent
m3	$y_{ig} = a_g + b_g\, x_{ig} + u_{ig}$	σ	$2G+1$	seemingly indep.
m4	$y_{ig} = a_g + bx_{ig} + u_{ig}$	σ	$G+2$	dummies
m5	$y_{ig} = (a + u_{ag}) + (b + u_{bg})x_{ig} + u_{ig}$	$\sigma_a, \sigma_b, \sigma$	$2+3$	random effects
m6	$y_{ig} = (a + u_{ag}) + bx_{ig} + u_{ig}$	σ_a, σ	$2+2$	shared frailty

4.2 Event history analysis: hazard rates

T time until an event (death) occurs.

The distribution of T is equivalently characterized by

- density: $f(t)$ (If T is discrete, $f(t) = p(T = t)$)
- survival function: $S(t) = p(T \geq t)$
- hazard: $h(t) = f(t|T \geq t)$

$$h(t) = \frac{f(t)}{S(t)}$$

Covariates and predictable heterogeneity

Continuous proportional hazard model (parametric: exponential, Weibull, Gompertz; semi-parametric Cox model):

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}'\beta)$$

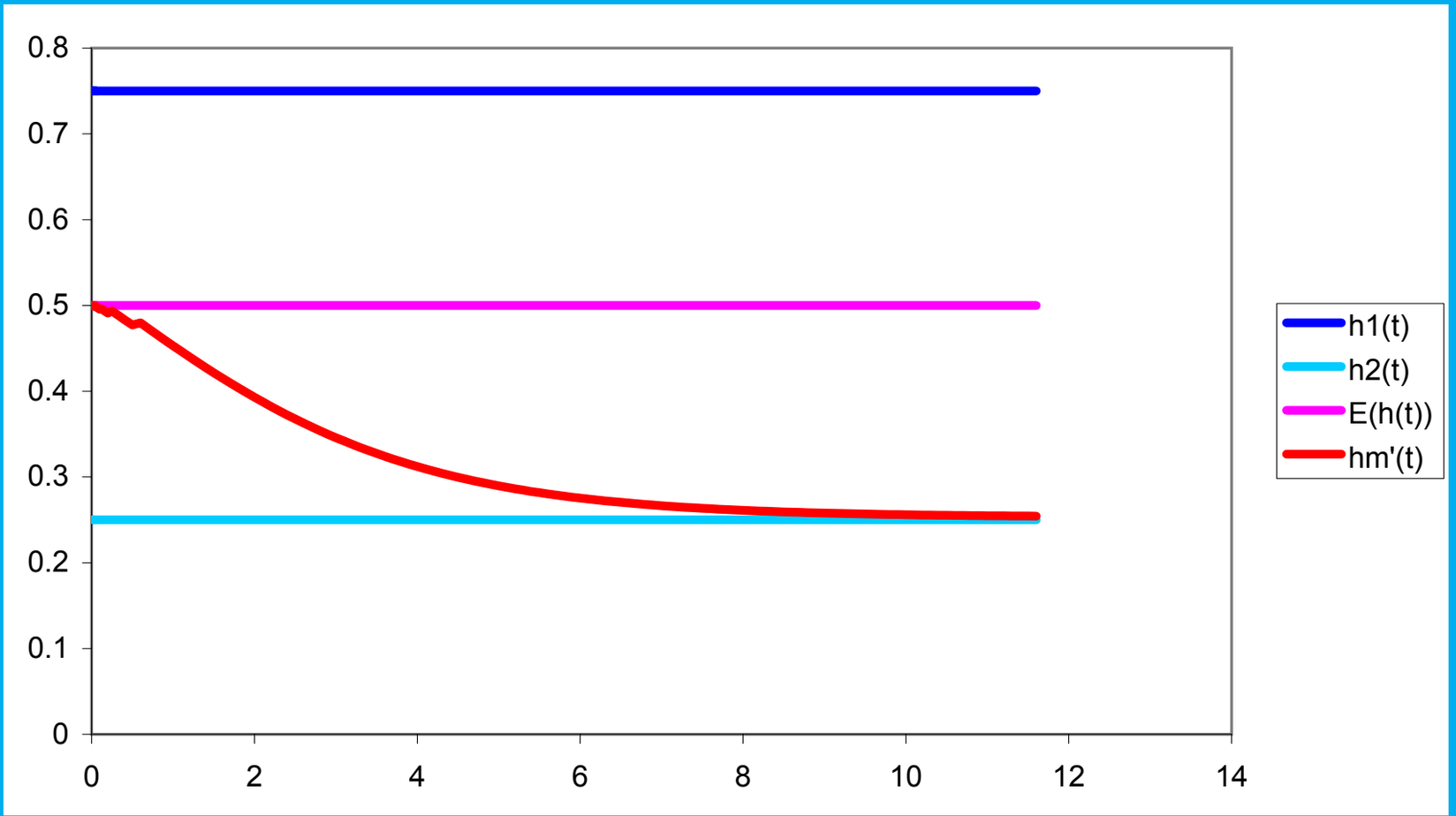
Accelerated failure model (duration model: Weibull, log-logistic, log-normal, ...):

$$T(\mathbf{x}) = T_0\exp(\mathbf{x}'\beta)$$

Discrete hazard model: proportional hazard odd ratios:

$$\frac{h(t|\mathbf{x})}{1-h(t|\mathbf{x})} = \frac{h_0(t)}{1-h_0(t)}\exp(\mathbf{x}'\beta)$$

4.3 Constant hazard rates by groups and population average



$h_g(t) = \nu_g h(t)$ with $E(\nu_g) = 1 \quad \Rightarrow \quad E(h_g(t)|t) = h(t)$

4.4 Shared Frailty (Continuous time)

$$\boxed{h(t|\mathbf{x}_{ig}) = \nu_g \, h_0(t) \exp(\mathbf{x}'_{ig}\beta)}$$

with ν_g random variable with $\mathrm{E}(\nu_g) = 1$ and $\mathrm{Var}(\nu_g) = \theta$

Mainly for technical reasons, Gamma or log-normal distributions are usually assumed

Gamma density $\gamma(r, \lambda)$:

$$f(\nu) = \frac{\lambda}{\Gamma(r)}(\lambda\nu)^{r-1}e^{\lambda\nu}$$

$\mathrm{E}(\nu) = r/\lambda$ and $\mathrm{Var}(\nu) = r/\lambda^2$.
Hence, we have $\mathrm{E}(\nu) = 1$ for $r = \lambda$, and then $\mathrm{Var}(\nu) = \theta = 1/\lambda$.

Available in Stata (Cox since ver 8), S-Plus, (SAS ?)

Example

Cox model, return after emigration, Sart 1812-1846 (1143 obs, 5 years)

	without frailty		with frailty	
covariate	hazard ratio	sig. (indep)	hazard ratio	sig.
CHILD_FHD	.43	.027	.38	.047
DEST_ARD	6.52	.000	9.78	.000
DEST_RUI	3.36	.002	5.76	.001
SINGLE	3.37	.013	4.54	.010
θ			6.42	.000

2ΔLogLik $= 12.95$ (chi-2(1)).

CHILD_FHD child of family head, DEST_ARD to Ardennes, DEST_RUI to rural or industrial regions

See Alter et al. (2001) for a full scale example.

Cox model for Return within 5 years after emigration, Sart 1812-1900, $n = 5351$

	coefficient		hazard ratio		p-value in %		
	basic	frailty	basic	frailty	basic	robust	frailty
Economic ratio	1.02	0.30	2.76	1.35	0.2	3.8	45.0
Man	-0.28	-0.18	0.76	0.83	0.1	0.2	5.6
Single	0.40	0.52	1.49	1.68	1.2	1.2	0.3
Born in Ardennes	0.25	0.17	1.29	1.18	4.1	15.0	28.0
Age when Leaving	0.01	0.00	1.01	1.00	12.0	17.0	62.0
To Ardennes	destination reference category						
To rural	-0.32	-0.60	0.73	0.55	5.7	14.0	0.2
To urban/indust.	-0.07	-0.23	0.93	0.79	50.0	68.0	6.8
To other	-1.25	-1.25	0.29	0.29	0.0	0.0	0.0
Head or spouse of	parenthood reference category						
Child of head	0.02	-0.25	1.02	0.78	89.0	90.0	19.0
Other parenthood	0.12	-0.27	1.13	0.76	54.0	56.0	26.0
No parenthood	-0.50	-0.54	0.61	0.58	6.7	7.3	9.0
Standard deviation $\sqrt{\theta}$ of family effect				1.75			0.0

4.5 Discrete time: logistic regression

Alternatively, we can use discrete time approach and use logistic regression.

Logistic models are special cases of Generalized Linear Models (GLM). Hence, multilevel logistic regression is available whenever multilevel GLM is implemented.

Barber et al. (2000), show how to estimate a model with several random effects with the HLM (Bryk et al., 1996) and MLN (Goldstein et al., 1998) softwares.

5 Conclusion: What about data mining?

- Interdisciplinary research: mix focuses, mix aggregation levels

  $\Rightarrow$ multi-level modeling

  Cox model: Available tools for the shared frailty

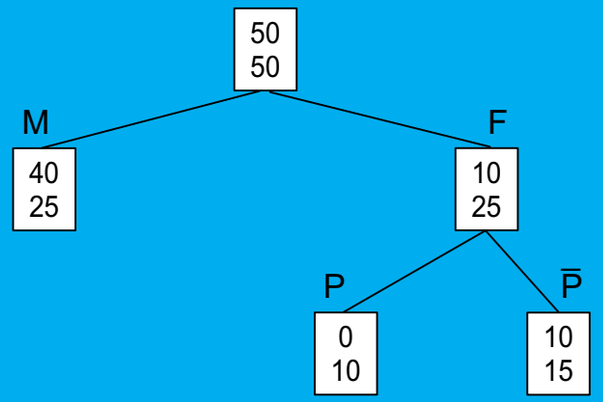
  Lack of operational tools for full random effect hazard models

  $\Rightarrow$ discrete time model (logistic)

- Other important issues
  - Competitive events
  - Multivariate models for analysing how different events are intertwined

What about data mining tools?

- Mining frequent sequences, association between subsequences
  What kind of sequence of demographic events is paired with a given sequence of individual development stages?
- Induction trees:



  - The response could be cluster of hazard functions, or durations
  - Using the family or family type as a predictor, trees can show how the effect of other factors may be conditioned by the family
    See the related work by Breiman (2001) (survival trees)

References

Alter, G., M. Oris, and G. Broström (2001). The family and mortality: A case study from rural Belgium. *Annales de la démographie historique 1*, 11–31.

Barber, J. S., S. A. Murphy, W. G. Axinn, and J. Maples (2000). Discrete-time multilevel hazard analysis. In M. E. Sobel and M. P. Becker (Eds.), *Sociological Methodology*, Volume 30, pp. 201–235. New York: The American Sociological Association.

Beeking, E., F. van Poppel, and A. C. Liefbroer (2002). Parental death and death of the child. See Derosas and Oris (2002), pp. 231–260.

Breiman, L. (2001). Satistical modeling: The two cultures (with discussion). *Statistical Science 16*(3), 199–231.

Bryk, A., S. W. Raudenbush, and R. Congdon (1996). *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2l and HLM/3l Programs*. Chicago: Scientific Software International.

Courgeau, D. and B. Baccaïni (1998). Multilevel analysis in the social sciences. *Population: An English Selection 10*(1), 39–71.

Derosas, R. and M. Oris (Eds.) (2002). *When Dad Died*. Bern: Peter Lang.

Goldstein, H., J. Rasbash, I. Plewis, D. Draper, W. Browne, M. Yang, G. Woodhouse, and M. Haely (1998). A user guide to MLwiN. Technical report, Multilevels Models Project, London.

Goldstein, H. (1995). *Multilevel Statistical Models*. New York: Halsted Press.

Kish, L. and M. R. Frankel (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Societey, Series B 36*, 1–37.

Lin, D. Y. and L. J. Wei (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association 84*, 1074–1078.

Prentice, R. L., B. J. Williams, and A. V. Peterson (1981). On the regression analysis of multivariate failure time data. *Biometrika 68*, 373–379.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data*. New York: Springer.