Event history analysis

An introduction

Alessandro Di Nallo

Università Bocconi

Université de Lausanne, 2 November 2022

PART III

Models with unobserved heterogeneity (or ‘frailty’)

Unobserved heterogeneity ('frailty')

- Frailty is a term introduced by biostatisticians
- Underlying concept: Each individual has intrinsically different risk of 'failure' (transition from a state to another)
- Standard ('no frailty') models: $\theta(t, X)$
- Frailty models: $\theta_\nu(t, X) \equiv \theta(t, X|\nu) = \theta(t, X)\nu$
- In the formulas, the frailty is expressed by a multiplicative scaling factor, ν
- ν is a random variable taking on positive values, with the mean normalised to one (for identification reasons) and finite variance σ^2.
- By assumption, ν is distributed independently of X and t

Unobserved heterogeneity (frailty)

For proportional hazard models, the frailty model can be re-written as

$$\theta_\nu(t,x) \equiv \theta(t,X|\beta,\nu) = \theta_0(t)exp(\beta X)\nu = \theta_0(t)exp(\beta X + u)$$

with $\nu = exp(u)$

or

$$\ln \theta_\nu(t,X|\nu) = \ln \theta_0(t) + \beta X + u$$

where $\theta_0(t)$ is the baseline hazard function and the 'error' term $u \equiv ln(\nu)$, which is random with a mean of zero

Unobserved heterogeneity (frailty)

For proportional hazard models, the frailty model can be re-written as

$\theta_\nu(t, x) \equiv \theta(t, X|\beta, \nu) = \theta_0(t)exp(\beta X)\nu = \theta_0(t)exp(\beta X + u)$

with $\nu = exp(u)$

or

$\ln \theta_\nu(t, X|\nu) = \ln \theta_0(t) + \beta X + u$

where $\theta_0(t)$ is the baseline hazard function and the 'error' term $u \equiv ln(\nu)$, which is random with a mean of zero

For tractability reasons, the choice of distribution of ν is typically limited to those that provide a closed form expression for the frailty survivor function

- For discrete time PH models (e.g., logistic), the most popular specifications are Gamma or Normal

Interpretation of frailty

The random variable ν, or equivalently u can be interpreted in different ways:

- The factor summarising the impact of '**omitted variables**' on the hazard rate – whether the missing regressors are intrinsically unobservable or simply unobserved in the data set to hand.
- The **individual-specific transition risk**. Within the distribution of ν, each individual (or firm, etc...) takes a distinct position.
- The **errors of measurement** in recorded regressors or recorded survival times

Implications of ignoring/including ν in survival models

The omission of ν (frailty or unobserved heterogeneity) has *bias* as main implication:

1. The non-frailty models over-estimate the degree of negative duration dependence in the (true) baseline hazard, and under-estimate the degree of positive duration dependence.

Implications of ignoring/including ν in survival models

The omission of ν (frailty or unobserved heterogeneity) has *bias* as main implication:

1. The non-frailty models over-estimate the degree of negative duration dependence in the (true) baseline hazard, and under-estimate the degree of positive duration dependence.
    - *Selection effect*. In the negative duration dependence case, observations with high ν values fail faster, other things equal, so the survivors at any given survival time are increasingly composed of observations with relatively low ν values and thence lower hazard rates.

Implications of ignoring/including ν in survival models

The omission of ν (frailty or unobserved heterogeneity) has *bias* as main implication:

1. The non-frailty models over-estimate the degree of negative duration dependence in the (true) baseline hazard, and under-estimate the degree of positive duration dependence.
   - *Selection effect*. In the negative duration dependence case, observations with high ν values fail faster, other things equal, so the survivors at any given survival time are increasingly composed of observations with relatively low ν values and thence lower hazard rates.
2. The proportionate effect of a given regressor on the hazard rate is no longer constant and independent of survival time
   - In the non-frailty PH model, the proportionate effect for regressor X_k is the fixed amount β_k

Implications of ignoring/including ν in survival models

The omission of ν (frailty or unobserved heterogeneity) has *bias* as main implication:

3. The presence of unobserved heterogeneity attenuates the proportionate response of the hazard to variation in each regressor at any survival time.
   - In short, the estimate of a positive (negative) β_k derived from the (wrong) no-frailty model will underestimate (overestimate) the 'true' estimate.

Frailty models in STATA

Overview of frailty models in Stata

- Continuous-time models: `streg [depvar] [indepvar], frailty(.)`
- Discrete-time models:
  - Logistic model with Normal ν: use `xtlogit`
  - Cloglog model with Gamma ν:
    1. `ssc install pgmhaz8`
    2. use `xtcloglog`

Continuous-time (Weibull) frailty models in Stata

```
use bc.dta

stset t, f(dead)
```

```
Survival-time data settings

         Failure event: dead!=0 & dead<.
Observed time interval: (0, t]
     Exit on or before: failure

--------------------------------------------------------------------------
         80  total observations
          0  exclusions
--------------------------------------------------------------------------
         80  observations remaining, representing
         58  failures in single-record/single-failure data
   1,257.07  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =        35
```

Continuous-time (Weibull) frailty models in Stata

Model 1: `streg age smoking dietfat, d(weib) nohr nolog`

_t	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.559197	.0563239	9.93	0.000	.4488042	.6695899
smoking	1.649311	.3276501	5.03	0.000	1.007128	2.291493
dietfat	2.222411	.2404553	9.24	0.000	1.751128	2.693695
_cons	-45.97988	4.634153	-9.92	0.000	-55.06265	-36.89711
/ln_p	1.431728	.0978872	14.63	0.000	1.239872	1.623583
p	4.185925	.4097485			3.455172	5.071228
1/p	.2388958	.0233848			.1971909	.2894212

The risk of death is positively associated with :

- patient's age ($\beta^{m1}_{age} > 0$)
- smoking ($\beta^{m1}_{smoking} > 0$)
- average weekly calorific intake ($\beta^{m1}_{dietfat} > 0$)

Continuous-time (Weibull) frailty models in Stata

Model 2: `streg age smoking, d(weib) nohr nolog`

_t	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.1644213	.0149837	10.97	0.000	.1350538	.1937888
smoking	.9056537	.3061656	2.96	0.003	.3055801	1.505727
_cons	-11.20242	.9989083	-11.21	0.000	-13.16024	-9.244594
/ln_p	.3633523	.0955797	3.80	0.000	.1760195	.5506852
p	1.438142	.1374573			1.192461	1.734441
1/p	.6953414	.0664605			.5765546	.8386016

`dietfat` is **omitted** → unobserved heterogeneity (omitted variable **bias**)

However, unobserved heterogeneity is not 'modelled' → no explicit treatment of individuals' frailty

- corr(`age, dietfat`) $> 0 \rightarrow \beta^{m2}_{age} > \beta^{m1}_{age}$
- corr(`smoking, dietfat`) $> 0 \rightarrow \beta^{m2}_{smoking} > \beta^{m1}_{smoking}$

The βs of this model are over-estimated, hence biased.

Continuous-time (Weibull) frailty models in Stata

Model 3: `streg age smoking, d(weib) nohr nolog frailty(gamma)`

_t	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.3893002	.0934984	4.16	0.000	.2060467	.5725537
smoking	1.025521	.5225054	1.96	0.050	.0014291	2.049613
_cons	-23.8082	5.204923	-4.57	0.000	-34.00966	-13.60674
/ln_p	1.087761	.222261	4.89	0.000	.6521376	1.523385
/lntheta	.3307466	.5250758	0.63	0.529	-.698383	1.359876
p	2.967622	.6595867			1.91964	4.587727
1/p	.3369701	.0748953			.2179729	.520931
theta	1.392007	.7309092			.4973889	3.895711

```
LR test of theta=0: chibar2(01) = 22.57                Prob >= chibar2 = 0.000
```

`theta` (θ) is the component of frailty (aka ν) represented by a Gamma

θ is statistically significant ($p<0.05$)

In M1, persons' differences in 'calories intake' was captured by `dietfat`

In M3, part of that 'effect' is represented by $\theta \rightarrow$ unobserved heterogeneity is 'modelled'

Continuous-time (Weibull) frailty models in Stata

Model 3: `streg age smoking , d(weib) nohr nolog frailty(gamma)`

_t	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.3893002	.0934984	4.16	0.000	.2060467	.5725537
smoking	1.025521	.5225054	1.96	0.050	.0014291	2.049613
_cons	-23.8082	5.204923	-4.57	0.000	-34.00966	-13.60674
/ln_p	1.087761	.222261	4.89	0.000	.6521376	1.523385
/lntheta	.3307466	.5250758	0.63	0.529	-.698383	1.359876
p	2.967622	.6595867			1.91964	4.587727
1/p	.3369701	.0748953			.2179729	.520931
theta	1.392007	.7309092			.4973889	3.895711

LR test of theta=0: chibar2(01) = 22.57   Prob >= chibar2 = 0.000

The inclusion of 'frailty' reduces the bias of the βs of the prior model:

- $\beta^{m2}_{age} > \beta^{m3}_{age}$
- $\beta^{m2}_{smoking} > \beta^{m3}_{smoking}$

Continuous-time (Weibull) frailty models in Stata

Model 3: `streg age smoking, d(weib) nohr nolog frailty(gamma)`

_t	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.3893002	.0934984	4.16	0.000	.2060467	.5725537
smoking	1.025521	.5225054	1.96	0.050	.0014291	2.049613
_cons	-23.8082	5.204923	-4.57	0.000	-34.00966	-13.60674
/ln_p	1.087761	.222261	4.89	0.000	.6521376	1.523385
/lntheta	.3307466	.5250758	0.63	0.529	-.698383	1.359876
p	2.967622	.6595867			1.91964	4.587727
1/p	.3369701	.0748953			.2179729	.520931
theta	1.392007	.7309092			.4973889	3.895711

LR test of theta=0: chibar2(01) = 22.57 Prob >= chibar2 = 0.000

The inclusion of 'frailty' reduces the bias of the βs of the prior model:

- $\beta^{m2}_{age} > \beta^{m3}_{age}$
- $\beta^{m2}_{smoking} > \beta^{m3}_{smoking}$

Continuous-time (Weibull) frailty models in Stata

Model 4: `streg age smoking dietfat, d(weib) nolog nohr frailty(gamma)`

_t	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	.5592066	.05632	9.93	0.000	.4488214	.6695918
smoking	1.649354	.327641	5.03	0.000	1.007189	2.291518
dietfat	2.222451	.2404402	9.24	0.000	1.751197	2.693705
_cons	-45.98067	4.633826	-9.92	0.000	-55.0628	-36.89853
/ln_p	1.431747	.0978781	14.63	0.000	1.239909	1.623584
/lntheta	-15.92571	6058.646	-0.00	0.998	-11890.65	11858.8
p	4.186005	.4097183			3.4553	5.071235
1/p	.2388912	.0233822			.1971906	.2894105
theta	1.21e-07	.0007344			0	.

LR test of theta=0: chibar2(01) = 0.00 Prob >= chibar2 = 1.000

'Frailty' (θ) and `dietfat` are redundant.

People's heterogeneity is accounted for by two 'overlapping' components

The LR test of θ reveals that 'frailty' is not statistically significant.

Continuous-time (Weibull) frailty models in Stata

Bottomline

- ‘Frailty’ captures unobserved individuals’ characteristics that are not accounted for by any other covariate.
- If you suspect that your model is affected by omitted-variable bias, it is advisable to model ‘frailty’.
- Conversely, if your model is correctly specified (no OVB), you can run a model without specifying unobserved heterogeneity.

Discrete-time frailty models in Stata

Model with no 'frailty'

```
logit died logt age smoking dietfat, nolog
```

Model with 'frailty'

```
xtlogit died logt age smoking, nolog i(id)
```

The command `xtlogit` in combination with `i(.)` (the option that prompts STATA to compute the individual-specific frailty) models unobserved heterogeneity.

Discrete time models with time-varying covariates

Discrete time models with time-varying covariates

- So far, we have addressed models with **time-invariant** variables
- The advantage of discrete-time survival models is to include **time-varying** covariates.
  - A model studying the risk of death for cancer patients with 'age at the first observation' and 'survival time' implicitly considers age as a time-varying covariate: fixed *age* + time-varying *time*
  - What if characteristics other than (say) age change over time?
  - The nature of the transition process motivates the choice of time-varying covariates.
- Examples of transition processes are associated to *time-varying* covariates, such as educational attainment and income
  - Home-leaving
  - Parenthood
  - Separation

An example. Union dissolution

coupleid	pidp_m	pidp_f	wave	sep	union_dur~on	union_dur~ln	union	hiqual_f	age_xt_f	income	income_miss	age_homo	background
1	10023364	819478885	14	0	0	.	cohabitation	A level…	34	7.96421	0	Woman older	Intact
2	1003007	1003691	1	0	35	3.555348	marriage	99	64.91666	6.872128	0	Woman older	Intact
2	1003007	1003691	2	0	36	3.583519	marriage	99	66	6.917376	0	Woman older	Intact
3	10030124	206613245	14	Separation/Divorce	1	0	cohabitation	GCSE etc	24.75	7.420545	0	Man older	Non intact
4	1005047	1005731	2	0	38	3.637586	marriage	Degree	60.5	5.556828	0	Man older	Intact
4	1005047	1005731	3	0	39	3.663562	marriage	Degree	61.58333	5.934436	0	Man older	Intact
4	1005047	1005731	5	0	41	3.713572	marriage	Degree	63.58333	6.35792	0	Man older	Intact
4	1005047	1005731	6	0	42	3.73767	marriage	Degree	64.5	6.744059	0	Man older	Intact
4	1005047	1005731	7	0	43	3.7612	marriage	Degree	65.5	6.588748	0	Man older	Intact
4	1005047	1005731	10	0	46	3.828641	marriage	Degree	68.58334	7.007321	0	Man older	Intact
5	10057242	10064082	7	0	0	.	marriage	No qual	61	7.429323	0	Same age	Intact
5	10057242	10064082	8	0	1	0	marriage	No qual	62	6.027475	0	Same age	Intact
5	10057242	10064082	9	0	2	.6931472	marriage	No qual	63	5.554251	0	Same age	Intact
5	10057242	10064082	10	0	3	1.098612	marriage	No qual	64	6.914268	0	Same age	Intact
5	10057242	10064082	11	0	4	1.386294	marriage	No qual	65	6.192921	0	Same age	Intact
6	1006407	1007091	3	0	27	3.295837	marriage	Other h…	50.16667	0	0	Man older	Intact
6	1006407	1007091	4	0	27	3.295837	marriage	Other h…	50.75	6.987798	0	Man older	Intact
6	1006407	1007091	5	0	28	3.332205	marriage	Other h…	51.75	6.913508	0	Man older	Intact
6	1006407	1007091	7	0	30	3.401197	marriage	Other h…	53.66667	7.090847	0	Man older	Intact
6	1006407	1007091	9	0	33	3.496508	marriage	Other h…	55.83333	7.003835	0	Man older	Intact
6	1006407	1007091	11	0	34	3.526361	marriage	Other h…	57.75	7.550379	0	Man older	Intact
7	10070842	10077682	7	0	0	.	marriage	Other h…	77.58334	6.837644	0	Same age	Intact
7	10070842	10077682	8	0	1	0	marriage	Other h…	78.58334	6.8226	0	Same age	Intact
7	10070842	10077682	9	0	2	.6931472	marriage	Other h…	79.5	6.918468	0	Same age	Intact
8	6354611	1007775	6	0	1	0	cohabitation	Degree	29.33333	8.218207	0	Same age	Intact
8	6354611	1007775	7	0	1	0	marriage	Degree	30.08333	9.372092	0	Same age	Intact
8	6354611	1007775	8	0	3	1.098612	marriage	Degree	31.25	9.182489	0	Same age	Intact
8	6354611	1007775	9	Separation/Divorce	4	1.386294	marriage	Degree	32.16667	9.179215	0	Same age	Intact
9	10084442	10091282	7	0	0	.	marriage	GCSE etc	33.66667	7.265275	0	Same age	Intact
9	10084442	10091282	8	0	1	0	marriage	GCSE etc	34.66667	7.36941	0	Same age	Intact
9	10084442	10091282	9	0	2	.6931472	marriage	GCSE etc	35.58333	7.256486	0	Same age	Intact
9	10084442	10091282	10	0	3	1.098612	marriage	GCSE etc	36.58333	7.382798	0	Same age	Intact
9	10084442	10091282	11	0	4	1.386294	marriage	GCSE etc	37.58333	7.368526	0	Same age	Intact
10	1008447	1009131	1	[illegible]	0	[illegible]	marriage	Other q…	33.5	5.663863	0	Same age	Intact

identifier: `coupleid`

interval: `wave`

dependent variable: `sep`

time: `union duration`

time-varying covariates: `union, hiqual, age, income ...`

time-invariant covariataes: `age homo, background`

An example. Union dissolution

Use the command logit: `logit [depvar] [indepvar]`

sep	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
union_duration_ln	-.2081217	.0203799	-10.21	0.000	-.2480656	-.1681778
age_xt_f	-.0438752	.0070841	-6.19	0.000	-.0577597	-.0299907
c.age_xt_f#c.age_xt_f	.0001858	.0000725	2.56	0.010	.0000437	.000328
income	.1034592	.0464292	2.23	0.026	.0124597	.1944587
c.income#c.income	-.0106885	.0047767	-2.24	0.025	-.0200506	-.0013265
1.income_miss	-.0277949	.1400929	-0.20	0.843	-.302372	.2467821
hiqual_f						
Other higher	.1662028	.0737885	2.25	0.024	.02158	.3108256
A level etc	.2626498	.0633518	4.15	0.000	.1384826	.386817
GCSE etc	.3368424	.0622636	5.41	0.000	.214808	.4588768
Other qual	.4320462	.0814471	5.30	0.000	.2724129	.5916795
No qual	.2813151	.0836438	3.36	0.001	.1173762	.445254
99	.0771324	.0832087	0.93	0.354	-.0859537	.2402184
education_homo						
Man higher education	.027696	.0580988	0.48	0.634	-.0861755	.1415675
Woman higher education	.2191767	.0556055	3.94	0.000	.1101919	.3281615
Missing	.6644837	.0610671	10.88	0.000	.5447943	.784173
union						
cohabitation	1.072358	.0392017	27.35	0.000	.9955241	1.149192
_cons	-3.047207	.2065594	-14.75	0.000	-3.452056	-2.642359

An example. Union dissolution

Use the command logit: `logit [depvar] [indepvar]`

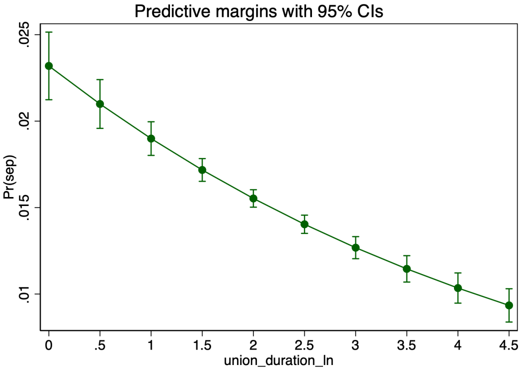
sep	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
union_duration_ln	-.2081217	.0203799	-10.21	0.000	-.2480656	-.1681778
age_xt_f	-.0438752	.0070841	-6.19	0.000	-.0577597	-.0299907
c.age_xt_f#c.age_xt_f	.0001858	.0000725	2.56	0.010	.0000437	.000328
income	.1034592	.0464292	2.23	0.026	.0124597	.1944587
c.income#c.income	-.0106885	.0047767	-2.24	0.025	-.0200506	-.0013265
1.income_miss	-.0277949	.1400929	-0.20	0.843	-.302372	.2467821
hiqual_f						
Other higher	.1662028	.0737885	2.25	0.024	.02158	.3108256
A level etc	.2626498	.0633518	4.15	0.000	.1384826	.386817
GCSE etc	.3368424	.0622636	5.41	0.000	.214808	.4588768
Other qual	.4320462	.0814471	5.30	0.000	.2724129	.5916795
No qual	.2813151	.0836438	3.36	0.001	.1173762	.445254
99	.0771324	.0832087	0.93	0.354	-.0859537	.2402184
education_homo						
Man higher education	.027696	.0580988	0.48	0.634	-.0861755	.1415675
Woman higher education	.2191767	.0556055	3.94	0.000	.1101919	.3281615
Missing	.6644837	.0610671	10.88	0.000	.5447943	.784173
union						
cohabitation	1.072358	.0392017	27.35	0.000	.9955241	1.149192
_cons	-3.047207	.2065594	-14.75	0.000	-3.452056	-2.642359

$\ln(t)$: the risk of union dissolution decreases ($\beta < 0$) as time passes

An example. Union dissolution

Graph of $h_0(t)$ as a function of $\ln(t)$

```
margins, at(union_duration_ln=(0(0.5)4.5))
marginsplot
```



The time scale (x-axis) is not straightforward (expressed in log).

You can choose other functions (for example the linear and quadratic terms) to have an easier interpretation of time.

An example. Union dissolution

Use the command logit: `logit [depvar] [indepvar]`

sep	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
union_duration_ln	-.2081217	.0203799	-10.21	0.000	-.2480656	-.1681778
age_xt_f	-.0438752	.0070841	-6.19	0.000	-.0577597	-.0299907
c.age_xt_f#c.age_xt_f	.0001858	.0000725	2.56	0.010	.0000437	.000328
income	.1034592	.0464292	2.23	0.026	.0124597	.1944587
c.income#c.income	-.0106885	.0047767	-2.24	0.025	-.0200506	-.0013265
1.income_miss	-.0277949	.1400929	-0.20	0.843	-.302372	.2467821
hiqual_f						
Other higher	.1662028	.0737885	2.25	0.024	.02158	.3108256
A level etc	.2626498	.0633518	4.15	0.000	.1384826	.386817
GCSE etc	.3368424	.0622636	5.41	0.000	.214808	.4588768
Other qual	.4320462	.0814471	5.30	0.000	.2724129	.5916795
No qual	.2813151	.0836438	3.36	0.001	.1173762	.445254
99	.0771324	.0832087	0.93	0.354	-.0859537	.2402184
education_homo						
Man higher education	.027696	.0580988	0.48	0.634	-.0861755	.1415675
Woman higher education	.2191767	.0556055	3.94	0.000	.1101919	.3281615
Missing	.6644837	.0610671	10.88	0.000	.5447943	.784173
union						
cohabitation	1.072358	.0392017	27.35	0.000	.9955241	1.149192
_cons	-3.047207	.2065594	-14.75	0.000	-3.452056	-2.642359

age_xt_f : the risk of union dissolution decreases as women's age (linear term's $\beta < 0$) increases

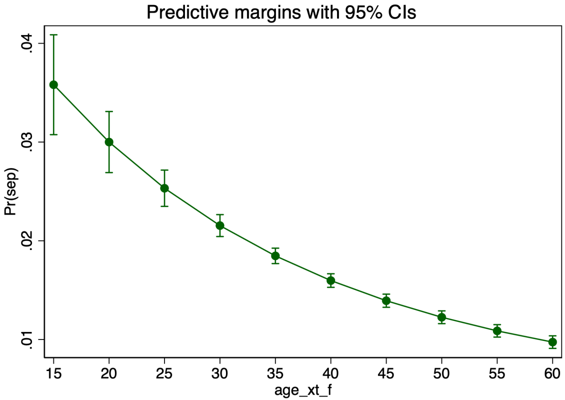
An example. Union dissolution

sep	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
union_duration_ln	-.2081217	.0203799	-10.21	0.000	-.2480656	-.1681778
age_xt_f	-.0438752	.0070841	-6.19	0.000	-.0577597	-.0299907
c.age_xt_f#c.age_xt_f	.0001858	.0000725	2.56	0.010	.0000437	.000328
income	.1034592	.0464292	2.23	0.026	.0124597	.1944587
c.income#c.income	-.0106885	.0047767	-2.24	0.025	-.0200506	-.0013265
1.income_miss	-.0277949	.1400929	-0.20	0.843	-.302372	.2467821
hiqual_f						
Other higher	.1662028	.0737885	2.25	0.024	.02158	.3108256
A level etc	.2626498	.0633518	4.15	0.000	.1384826	.386817
GCSE etc	.3368424	.0622636	5.41	0.000	.214808	.4588768
Other qual	.4320462	.0814471	5.30	0.000	.2724129	.5916795
No qual	.2813151	.0836438	3.36	0.001	.1173762	.445254
99	.0771324	.0832087	0.93	0.354	-.0859537	.2402184
education_homo						
Man higher education	.027696	.0580988	0.48	0.634	-.0861755	.1415675
Woman higher education	.2191767	.0556055	3.94	0.000	.1101919	.3281615
Missing	.6644837	.0610671	10.88	0.000	.5447943	.784173
union						
cohabitation	1.072358	.0392017	27.35	0.000	.9955241	1.149192
_cons	-3.047207	.2065594	-14.75	0.000	-3.452056	-2.642359

c.age_xt_f#c.age_xt_f : the effect of age (quadratic term's $\beta > 0$) is upwardly concave (parabola 'looks' up)

An example. Union dissolution

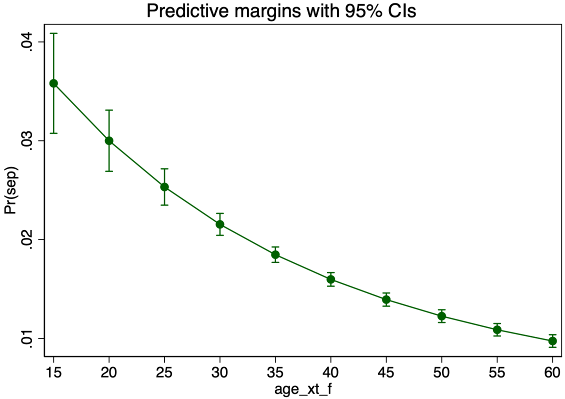
Predicted probability of union dissolution by woman's age



c.age_xt_f#c.age_xt_f : the effect of age (quadratic term's $\beta > 0$) is upwardly concave (parabola 'looks' up)

An example. Union dissolution

Predicted probability of union dissolution by woman's age



Please remember that all the predicted probabilities can be multiplied by 100 to be expressed in %.

For instance, 0.02 means **2**% on an **annual** basis, because the dataset is organized in year-long episodes (= each line corresponds to one year).

An example. Union dissolution

Use the command logit: `logit [depvar] [indepvar], or` to display the **odds ratios**

hiqual_f						
Other higher	1.180813	.0871304	2.25	0.024	1.021815	1.364551
A level etc	1.300371	.0823808	4.15	0.000	1.14853	1.472287
GCSE etc	1.400518	.0872013	5.41	0.000	1.239624	1.582296
Other qual	1.540406	.1254616	5.30	0.000	1.313129	1.807021
No qual	1.324871	.1108173	3.36	0.001	1.124542	1.560887
99	1.080185	.0898808	0.93	0.354	.9176367	1.271527
education_homo						
Man higher education	1.028083	.0597304	0.48	0.634	.9174332	1.152078
Woman higher education	1.245051	.0692317	3.94	0.000	1.116492	1.388413
Missing	1.943487	.1186831	10.88	0.000	1.724254	2.190595
union						
cohabitation	2.922262	.1145576	27.35	0.000	2.706142	3.155642

It is very useful especially when you deal with categorical variables.

The OR associated with a category shows you how much more ($OR > 1$) or less ($OR < 1$) likely the transition is for that category.

An example. Union dissolution

How to interpret the OR of a categorical variable.
The case of woman's education

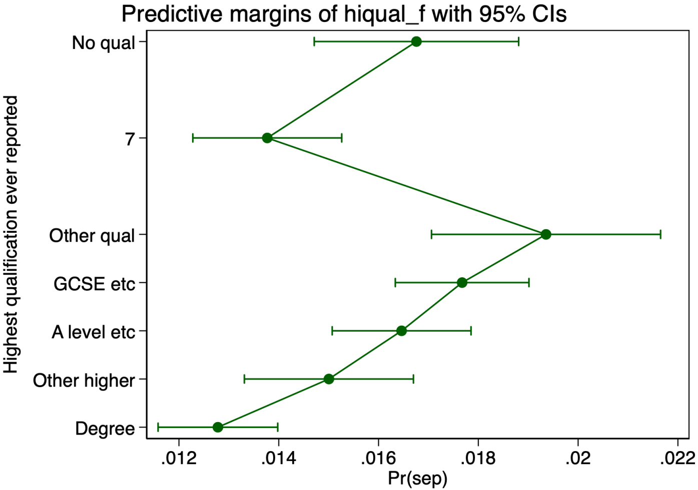
hiqual_f						
Other higher	1.180813	.0871304	2.25	0.024	1.021815	1.364551
A level etc	1.300371	.0823808	4.15	0.000	1.14853	1.472287
GCSE etc	1.400518	.0872013	5.41	0.000	1.239624	1.582296
Other qual	1.540406	.1254616	5.30	0.000	1.313129	1.807021
No qual	1.324871	.1108173	3.36	0.001	1.124542	1.560887
99	1.080185	.0898808	0.93	0.354	.9176367	1.271527

Examples:

- **Baseline**: 'Degree' (implicit odds ratio: 1)
- 'Other higher' = 1.18 → The odds of separating/divorcing for women holding another higher degree are 18% *higher* than women holding a degree
- 'No qual' = 1.32 → The odds of separating/divorcing for women holding no qualification are 32% *higher* than women holding a degree

An example. Union dissolution

How to interpret the OR of a categorical variable.
The case of woman's education



An example. Union dissolution

How to model **multiple** transitions in or out of unions?

Recurrent events

Use the command `xtlogit` jointly with the option `re` :

```
xtlogit [depvar] [indepvars], re
```

STATA processes more than one episode of (say) separation by attributing to each individual a *underlying* propensity to separate.

Does it remind you of ‘frailty’ ?

Indeed, a person that breaks up unions *more frequently* than others might have ‘unobserved characteristics’ (propensity to break up) that can be modelled with an individual-specific distribution (random effects `re`)

Competing risk models

Competing risk models

- So far, we have modelled time-to-event data and only *a single type of event* (hence, **binary**): 'failure' (e.g., time to: death, separation, birth, unemployment, promotion)
- Models in which there are different types of events – **multiple destinations** – are also of interest.
- For example, in a model of unemployment duration, you may wish to know about (a) *the time to exit from unemployment to a job*, and compare this with (b) *the time to exit from unemployment to economic inactivity*.
- We only consider the simplest case – the *independent* competing risk model.

Competing risk models

Multinomial discrete time event history model. An example about the job market.

To estimate the chances of :

- (a) a transition to a job
- (b) a transition out-of-labour market ('other')
- (c) staying in unemployment

... contextually

Competing risk models

Multinomial discrete time event history model. An example about the job market.

newid	t	status2	depvar	leftui	job	other	age	famresp	groupreg	status
12	15	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	16	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	17	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	18	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	19	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	20	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	21	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	22	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	23	Exit-other	unemployed	0	0	0	29	No responsibilities	Centre	ExitOth
12	24	Exit-other	Exit-other	1	0	1	29	No responsibilities	Centre	ExitOth
13	1	Exit-job	unemployed	0	0	0	25	Has responsibilities	Centre	ExitJob
13	2	Exit-job	unemployed	0	0	0	25	Has responsibilities	Centre	ExitJob
13	3	Exit-job	Exit-job	1	1	0	25	Has responsibilities	Centre	ExitJob
14	1	Exit-job	unemployed	0	0	0	52	Has responsibilities	North	ExitJob
14	2	Exit-job	unemployed	0	0	0	52	Has responsibilities	North	ExitJob
14	3	Exit-job	unemployed	0	0	0	52	Has responsibilities	North	ExitJob
14	4	Exit-job	Exit-job	1	1	0	52	Has responsibilities	North	ExitJob
15	1	censored	unemployed	0	0	0	31	No responsibilities	Centre	Exh-NoA
15	2	censored	unemployed	0	0	0	31	No responsibilities	Centre	Exh-NoA
15	3	censored	unemployed	0	0	0	31	No responsibilities	Centre	Exh-NoA

identifier: `newid`

unemployment duration: `t`

status2: `status at the end of period`

depvar of mlogit: `depvar`

depvar of logit: `leftui (job & other), job, other`

covariataes: `age, famresp, groupreg`

Competing risk models

Flashback. A discrete time EH model with a *binary* variable

```
Logistic regression                                     Number of obs = 11,234
                                                        LR chi2(8)    =  64.33
                                                        Prob > chi2   = 0.0000
Log likelihood = -2495.4163                             Pseudo R2     = 0.0127
```

leftui	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	-.0034686	.0048337	-0.72	0.473	-.0129425	.0060052
famresp	.0681293	.0854925	0.80	0.426	-.0994328	.2356915
tyentry	.6964857	.0934628	7.45	0.000	.513302	.8796693
reg1	.1435767	.2052072	0.70	0.484	-.258622	.5457754
reg2	.0696173	.2055525	0.34	0.735	-.3332581	.4724928
reg3	-.0692704	.2204173	-0.31	0.753	-.5012804	.3627396
reg4	-.0150869	.2120189	-0.07	0.943	-.4306364	.4004625
logt	.1662141	.0452708	3.67	0.000	.077485	.2549431
_cons	-3.440587	.2751854	-12.50	0.000	-3.979941	-2.901234

Dependent variable of logit `(leftui)`: leaving unemployment $(= 0, 1)$
Covariates :

- `famresp`: “No family reponsibilities”, “Some family reponsibilities”
- `reg*`: regions of residence (dummies)
- `tyentry`: contract type of prior job (“permanent $= 0$” , “temporary $= 1$”)

Competing risk models

A **multinomial** discrete time EH model

```
mlogit depvar age famresp tyentry reg1-reg4 logt, base(0)
```

depvar	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
unemployed	(base outcome)					
Exit_job						
age	-.0024503	.0055886	-0.44	0.661	-.0134037	.0085031
famresp	.0514627	.1006845	0.51	0.609	-.1458754	.2488008
tyentry	1.092581	.1242046	8.80	0.000	.8491446	1.336018
reg1	.2291151	.2438589	0.94	0.347	-.2488395	.7070698
reg2	.1313109	.2452359	0.54	0.592	-.3493426	.6119644
reg3	-.1784552	.2655201	-0.67	0.502	-.6988649	.3419546
reg4	-.0178443	.2516282	-0.07	0.943	-.5110265	.475338
logt	-.1947983	.052208	-3.73	0.000	-.2971241	-.0924725
_cons	-3.579488	.3239595	-11.05	0.000	-4.214437	-2.94454
Exit_other						
age	-.0084695	.0095718	-0.88	0.376	-.0272299	.010291
famresp	.0951781	.1600171	0.59	0.552	-.2184496	.4088058
tyentry	.1424349	.1666566	0.85	0.393	-.184206	.4690757
reg1	-.1239563	.3741337	-0.33	0.740	-.8572449	.6093324
reg2	-.1084249	.3704304	-0.29	0.770	-.8344551	.6176052
reg3	.2285222	.3886337	0.59	0.557	-.5331859	.9902302
reg4	.0045579	.3873852	0.01	0.991	-.7547031	.7638189
logt	1.931943	.1618052	11.94	0.000	1.61481	2.249075
_cons	-8.246268	.6537316	-12.61	0.000	-9.527559	-6.964978

Dependent variable of mlogit `(depvar)`: leaving unemployment (0; baseline); entering a job (1); out of labour force (2; e.g., retirement)

Competing risk models

A **multinomial** discrete time EH model

```
mlogit depvar age famresp tyentry reg1-reg4 logt, base(0)
```

Exit_job						
age	-.0024503	.0055886	-0.44	0.661	-.0134037	.0085031
famresp	.0514627	.1006845	0.51	0.609	-.1458754	.2488008
tyentry	1.092581	.1242046	8.80	0.000	.8491446	1.336018
reg1	.2291151	.2438589	0.94	0.347	-.2488395	.7070698
reg2	.1313109	.2452359	0.54	0.592	-.3493426	.6119644
reg3	-.1784552	.2655201	-0.67	0.502	-.6988649	.3419546
reg4	-.0178443	.2516282	-0.07	0.943	-.5110265	.475338
logt	-.1947983	.052208	-3.73	0.000	-.2971241	-.0924725
_cons	-3.579488	.3239595	-11.05	0.000	-4.214437	-2.94454
Exit_other						
age	-.0084695	.0095718	-0.88	0.376	-.0272299	.010291
famresp	.0951781	.1600171	0.59	0.552	-.2184496	.4088058
tyentry	.1424349	.1666566	0.85	0.393	-.184206	.4690757
reg1	-.1239563	.3741337	-0.33	0.740	-.8572449	.6093324
reg2	-.1084249	.3704304	-0.29	0.770	-.8344551	.6176052
reg3	.2285222	.3886337	0.59	0.557	-.5331859	.9902302
reg4	.0045579	.3873852	0.01	0.991	-.7547031	.7638189
logt	1.931943	.1618052	11.94	0.000	1.61481	2.249075
_cons	-8.246268	.6537316	-12.61	0.000	-9.527559	-6.964978

`tyentry`: men who had a temporary employment contract in their last job before u/e `(tyentry=1)` are much *more likely* ($\beta > 0$) to exit to a job than are men who had a permanent employment contract in their last job `(tyentry=0)`.

Competing risk models

A **multinomial** discrete time EH model

```
mlogit depvar age famresp tyentry reg1-reg4 logt, base(0)
```

Exit_job						
age	-.0024503	.0055886	-0.44	0.661	-.0134037	.0085031
famresp	.0514627	.1006845	0.51	0.609	-.1458754	.2488008
tyentry	1.092581	.1242046	8.80	0.000	.8491446	1.336018
reg1	.2291151	.2438589	0.94	0.347	-.2488395	.7070698
reg2	.1313109	.2452359	0.54	0.592	-.3493426	.6119644
reg3	-.1784552	.2655201	-0.67	0.502	-.6988649	.3419546
reg4	-.0178443	.2516282	-0.07	0.943	-.5110265	.475338
logt	-.1947983	.052208	-3.73	0.000	-.2971241	-.0924725
_cons	-3.579488	.3239595	-11.05	0.000	-4.214437	-2.94454
Exit_other						
age	-.0084695	.0095718	-0.88	0.376	-.0272299	.010291
famresp	.0951781	.1600171	0.59	0.552	-.2184496	.4088058
tyentry	.1424349	.1666566	0.85	0.393	-.184206	.4690757
reg1	-.1239563	.3741337	-0.33	0.740	-.8572449	.6093324
reg2	-.1084249	.3704304	-0.29	0.770	-.8344551	.6176052
reg3	.2285222	.3886337	0.59	0.557	-.5331859	.9902302
reg4	.0045579	.3873852	0.01	0.991	-.7547031	.7638189
logt	1.931943	.1618052	11.94	0.000	1.61481	2.249075
_cons	-8.246268	.6537316	-12.61	0.000	-9.527559	-6.964978

`tyentry`: the type of employment contract has no significant association ($\beta \approx 0$) with the hazard of exit from u/e for other reasons (Exit other).

Competing risk models

A **multinomial** discrete time EH model

```
mlogit depvar age famresp tyentry reg1-reg4 logt, base(0)
```

Exit_job						
age	-.0024503	.0055886	-0.44	0.661	-.0134037	.0085031
famresp	.0514627	.1006845	0.51	0.609	-.1458754	.2488008
tyentry	1.092581	.1242046	8.80	0.000	.8491446	1.336018
reg1	.2291151	.2438589	0.94	0.347	-.2488395	.7070698
reg2	.1313109	.2452359	0.54	0.592	-.3493426	.6119644
reg3	-.1784552	.2655201	-0.67	0.502	-.6988649	.3419546
reg4	-.0178443	.2516282	-0.07	0.943	-.5110265	.475338
logt	-.1947983	.052208	-3.73	0.000	-.2971241	-.0924725
_cons	-3.579488	.3239595	-11.05	0.000	-4.214437	-2.94454
Exit_other						
age	-.0084695	.0095718	-0.88	0.376	-.0272299	.010291
famresp	.0951781	.1600171	0.59	0.552	-.2184496	.4088058
tyentry	.1424349	.1666566	0.85	0.393	-.184206	.4690757
reg1	-.1239563	.3741337	-0.33	0.740	-.8572449	.6093324
reg2	-.1084249	.3704304	-0.29	0.770	-.8344551	.6176052
reg3	.2285222	.3886337	0.59	0.557	-.5331859	.9902302
reg4	.0045579	.3873852	0.01	0.991	-.7547031	.7638189
logt	1.931943	.1618052	11.94	0.000	1.61481	2.249075
_cons	-8.246268	.6537316	-12.61	0.000	-9.527559	-6.964978

`logt`: the hazard for exits to a job is declining with time on unemployment (`Exit job`), whereas the hazard for other types of exits is rising with time on unemployment (`Exit other`).

Competing risk models

Compare a logit (A.) and one outcome of the multinomial model (B.)

A. `logit job age famresp tyentry reg1-reg4 logt`

job	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	-.0023954	.0055883	-0.43	0.668	-.0133482	.0085575
famresp	.0497747	.1006708	0.49	0.621	-.1475364	.2470857
tyentry	1.093551	.1241995	8.80	0.000	.8501247	1.336978
reg1	.2316425	.2438092	0.95	0.342	-.2462147	.7094998
reg2	.1339491	.2451883	0.55	0.585	-.3466112	.6145093
reg3	-.1794071	.2654604	-0.68	0.499	-.6997001	.3408858
reg4	-.0160927	.2515802	-0.06	0.949	-.5091809	.4769955
logt	-.2101334	.05197	-4.04	0.000	-.3119927	-.1082741
_cons	-3.574223	.3239443	-11.03	0.000	-4.209142	-2.939304

B. `mlogit depvar age famresp tyentry reg1-reg4 logt, base(0)`

Exit_job						
age	-.0024503	.0055886	-0.44	0.661	-.0134037	.0085031
famresp	.0514627	.1006845	0.51	0.609	-.1458754	.2488008
tyentry	1.092581	.1242046	8.80	0.000	.8491446	1.336018
reg1	.2291151	.2438589	0.94	0.347	-.2488395	.7070698
reg2	.1313109	.2452359	0.54	0.592	-.3493426	.6119644
reg3	-.1784552	.2655201	-0.67	0.502	-.6988649	.3419546
reg4	-.0178443	.2516282	-0.07	0.943	-.5110265	.475338
logt	-.1947983	.052208	-3.73	0.000	-.2971241	-.0924725
_cons	-3.579488	.3239595	-11.05	0.000	-4.214437	-2.94454

Competing risk models

Bottomline

- A **logit** model is a 'safe' option for your survival analysis.
  - For instance, a transition from 'unemployment' to 'exit from unemployment'
- A **multinomial** model is 'conceptually' justified when the transition can be specified and you have proper data. For instance, a transition from 'unemployment'
  - to 'a job'
  - to 'retirement'
  - to 'maternal leave'
- A **logit** model for all the above transitions *separately* is also ok estimate-wise, although conceptually not accurate:
  - A logit model with a dependent variable such as 'transition to a job' ($= 1$) vs 'unemployment' ($= 0$) does not capture *other transitions* that may occur (e.g., to 'retirement' or 'maternal leave', etc.) and end up in the absorbing state 'unemployment'.

alessandro.dinallo@unibocconi.it