UNIVERSITY OF EDINBURGH
Business School

What seems right for all may not be right for any of us! Survival modelling with unobserved heterogeneity.

Credit Scoring & Credit Control XII Conference August 2011

Jonathan Crook, Mindy Leow, Tony Bellotti
Credit Research Centre, University of Edinburgh

The Problem

Simplistically:

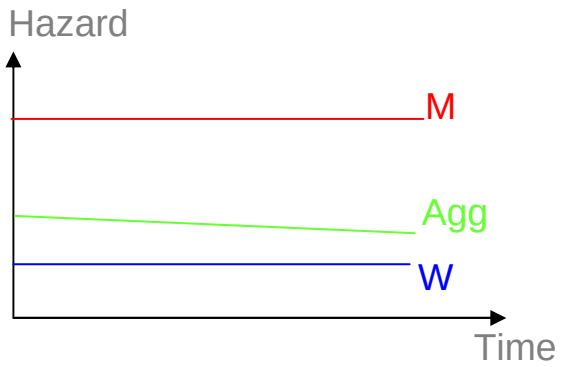
A hazard rate is the instantaneous probability of an event occurring in the next instant of time, given it has not happened before.

Consider a population of credit card holders from 2 regions each with a constant hazard rate of default:

Men: 0.4 (higher probability of defaulting next month),
Women 0.1 (lower probability of defaulting next month)

Population is 50% men, 50% women

Defaults in period:	period 1	period 2	period3	period 4
Take 100 men:	40	24	14.4	etc
Take 100 women:	10	9	8.1	etc
Aggregate	50	33	22.5	
No at risk	200	150	117	
As proportion of numbers at risk	0.25	0.22	0.192	



The simple aggregate hazard rate does not tell us about the individual hazard rate!

Theoretical considerations 1

Consider the hazard rate model: $\lambda(t_i \mid \mathbf{x}_i)$

Suppose the covariates do not perfectly explain the expected time of default, T

Common way to account for unobserved heterogeneity is by introducing a multiplicative unobserved random term:

$$\lambda(t_i \mid \mathbf{x}_i, \upsilon_i) = \lambda(t_i \mid \mathbf{x}_i)\upsilon_i \qquad \upsilon_i > 0 \qquad \forall \mathrm{i}$$

$$\mathrm{E}(\upsilon_i) = 1$$

Corresponding survival function:

$$S(t_i \mid \mathbf{x}_i, \upsilon_i) = [S(t_i, x_i)]$$

Implications

Failing to account for unobserved heterogeneity may result in

- Biased hazard rate with respect to t

as t increases hazard rate decreases by less when frailty omitted than when included.

Intuition:

Individuals with higher values of υ_I default before those with lower values of υ_I so average value of υ_I of those surviving must decrease – 'selection effect'.

Maths:

Aggregate hazard function is $$\lambda(t) = -\int \frac{\partial \ln S(t \mid \upsilon)}{\partial t} f(\upsilon) d\upsilon$$

Subs in Weibull survivor function: $$S(t \mid \upsilon) = \exp(-\gamma t^{\alpha})\upsilon$$

……we eventually ……get $$\lambda(t) = \alpha\gamma t^{\alpha-1} E(\upsilon \mid T \geq t)$$

+ve +ve - ve

It gets worse!!!

- Incorrect estimates of parameters (and so proportional impact)

If frailty is wrongly omitted then estimates of
* positive β_j will be lower than if frailty included
* negative β_j will be higher than if frailty included

In default model – this could alter the ranking of individuals!

Maths

Assume Weibull and write PH model as: $\ln \lambda(t \mid z) = \ln(zt^{\alpha-1}) + \ln \alpha$ where $z = \exp(\mathbf{x}^T \boldsymbol{\beta})$

Can derive:

$$\frac{\partial \ln(t \mid z)}{\partial x_j} = \beta_j$$

Now include frailty term

$$\lambda(t_i \mid \mathbf{x}_i, \upsilon_i) = \lambda_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \upsilon_i$$

$$\ln \lambda(t \mid z) = \ln(zt^{\alpha-1}) + \ln \alpha + \ln E(\upsilon \mid T \geq t)$$

Can derive

$$\frac{\partial \ln \lambda(t \mid z)}{\partial x_j} = \beta_j \left(1 - \frac{zt^{\alpha} \mathrm{Var}(\upsilon \mid T \geq t)}{E(\upsilon \mid T \geq t)}\right)$$

$$< \beta_j$$

Empirical Questions

How sensitive are the coefficient estimates to

- **the omission/inclusion of frailty?**
- **the distribution of the frailty term?**
- **the flexibility of the baseline hazard function?**

Discrete Time Survival Models

Let continuous time be divided into *A* intervals $[t_0,t_1),[t_1,t_2),.....,[t_{a-1},t_a),[t_a,\infty) \quad a=0,1,.....,A$

Defn of hazard function $\lambda^d(t_a,\mathbf{x}_i)=\Pr(t_{a-1}\le T<t_a \mid T\ge t_{a-1},x)$ Surv function $S^d(t_a,\mathbf{x}_i)=\Pr(T>t_{a-1}\mid \mathbf{x}_i)=\prod_{s=1}^{a-1}(1-\lambda^d(t_s\mid \mathbf{x}(t_s))$

Model 1 Discrete version of continuous PH Model: no frailty

From above: $\lambda^d(t_a\mid \mathbf{x}_i)=\dfrac{S(t_{a-1}\mid \mathbf{x}_i)-S(t_a\mid \mathbf{x}_i)}{S(t_{a-1}\mid \mathbf{x}_i)}$ If time were continuous: $S(t_a\mid \mathbf{x}_i)=\exp\left(-\int_0^t \lambda(s)ds\right)$

By substitution $\lambda^d(t_a\mid \mathbf{x}_i)=1-\exp\left(-\int_{t_{a-1}}^{t_a}\lambda(s)ds\right)$

Subs in the $\lambda(s)$ function of PH model $\lambda^d(t_a\mid \mathbf{x}_i)=1-\exp(-\exp(\ln\lambda_{0t_a}+\mathbf{x}(t_{a-1})^T\boldsymbol{\beta}))$

Notice $\lambda^d(t_a\mid \mathbf{x}_i)=\Pr(d_{t_a}=1\mid \mathbf{x}_i)$ where $\lambda_{0t_a}=\int_{t_{a-1}}^{t_a}\lambda_0(s)ds$ and $d_{t_a}=I((t_{a-1}\le T<t_a)\mid T\ge t_{a-1},\mathbf{x}_i)$

$$\Pr(d_{t_a}=1\mid \mathbf{x}_i)=F^E\left(g(t_a)+\mathbf{x}_i^T\beta\right) \qquad F^E(z)=1-\exp(-\exp(z))$$

Further Models

Model 2 Discrete version of continuous PH Model: with gamma frailty

$$\Pr(d_{t_a}=1\,|\,\mathbf{x}_i)=F^E\left(g(t_a)+\mathbf{x}_i^T\beta+\ln(\upsilon_i)\right) \qquad F^E(z)=1-\exp(-\exp(z)) \qquad \upsilon_i\sim\Gamma, E(\upsilon_i)=1, Var=\sigma_G^2$$

Model 3 Discrete version of continuous PH Model: with Gaussian frailty

$$\Pr(d_{t_a}=1\,|\,\mathbf{x}_i)=F^E\left(g(t_a)+\mathbf{x}_i^T\beta+\ln(\upsilon_i)\right) \qquad F^E(z)=1-\exp(-\exp(z)) \qquad \ln(\upsilon_i)\sim N(0,\sigma_N^2)$$

Model 4 Discrete Choice: no frailty

$$\Pr(d_{it_a}=1\,|\,\mathbf{x}_i)=F^L(g(t_a)+\mathbf{x}_i^T\boldsymbol{\beta}) \qquad F^L(z)=\exp(z)/(1+\exp(z))$$

Model 5 Discrete Choice: Gaussian frailty

$$\Pr(d_{it_a}=1\,|\,\mathbf{x}_i)=F^L(g(t_a)+\mathbf{x}_i^T\boldsymbol{\beta}+e_i) \qquad F^L(z)=\exp(z)/(1+\exp(z)) \qquad e_i\sim N(0,\sigma^2)$$

Modelling Choices

- Variables chosen in light of literature
- Observations in duration time 1& 3 deleted (cannot be 3 behind) – treated as left censored.
- Cases entering after duration time of 3 excluded (cautious)
- Two alternative baseline hazard functions:

Simple $g(t_a) = \alpha(\ln t)$

Flexible $g(t_a) = \alpha_1(\ln t) + \alpha_2(\ln t)^2 + \alpha_3 t + \alpha_4 t^2$

Data

- Random sample of credit cards issued late 1990s to mid 200s.
- Training sample: opened up to 31 December 2003: 35,000 accounts
- Test sample: opened after 31 December 2003: 8307 accounts
- Performance over time recorded
  time of default = first duration time observed to be 3 payments behind
- Macroeconomic variables from ONS: take differences to induce stationarity.

UNIVERSITY OF EDINBURGH
Business School

Comparison of models with different types of frailty: simple baseline

CREDIT RESEARCH CENTRE

	PH no	PH Normal	PH Gamma	Logit No	Logit Normal
ln durt	**-0.0205**	**0.5444****	**0.3083****	**-0.0204**	**0.6168****
ln income	**-0.0485****	**-0.0833****	**-0.0811****	**-0.0486****	**-0.0881****
age25_29	-0.1189	-0.2080*	-0.1998*	-0.1192	-0.2200*
age30_33	-0.0258	-0.0724	-0.0740	-0.0260	-0.0772
age34_37	0.0743	-0.0558	-0.0151	0.0743	0.0569
age38_41	-0.0355	-0.1414	-0.1663	-0.0357	-0.1516
age42_47	-0.0207	-0.0678	-0.1049	0.0206	-0.0744
age48_55	-0.2159**	-0.3807**	-0.3554**	-0.2163**	-0.4018**
age56_plus	-0.3731**	-0.6154**	-0.5742**	-0.3736**	-0.6479**
A1	**1.1276****	**1.6912****	**1.4496****	**1.1295****	**1.7796****
A2	**0.7466****	**1.1344****	**1.0069****	**0.7477****	**1.1936****
A3	**0.6917****	**1.0307****	**0.9139****	**0.6826****	**1.0833****
Employed	**0.2580****	**0.3999****	**0.3599****	**0.2584****	**0.4203****
Self employed	**0.4656****	**0.7311****	**0.6687****	**0.4665****	**0.7687****
Years at address	-0.0103**	-0.0158**	-0.0137**	-0.0103**	-0.0166**
Time with bank	-0.0026**	-0.0036**	-0.0031**	-0.0026**	-0.0037**
Number of cards	0.0061	-0.0118	-0.0197	0.0061	-0.0131
Δ base rate	-0.2667	-0.2883	-0.2509	-0.2672*	-0.2961
Δearnings	3.3321**	3.1818**	3.0616*	3.3373**	3.2241**
Δunemployment	0.0086*	0.0089**	0.0097**	0.0086*	0.0092**
Δhouse prices	-0.0000	-0.0000	-0.0000	0.0000	0.0000
Δtotal credit outsg	0.0000	-0.0000	-0.0000	0.0000	-0.0000
Δconsumer confidence	-0.0139	-0.0147	-0.0145	-0.0139	-0.0150
Constant	-5.5922**	-8.7299**	-5.8985**	-5.58975**	-9.1137**
Var (υi)		5.3981	5.4390		6.2581
Chibarsq(01)		46.73**	52.85**		50.37**
No obs (unbal N*T)	1,499,400	1,499,400	1,499,400	1,499,400	1,499,400
* = sig at 5%; ** at 1%					
	22a	20c	22b	20dd	20ee

UNIVERSITY OF EDINBURGH
Business School

Comparison of models with different types of frailty: flexible baseline

	PH no	PH Normal	PH Gamma	Logistic no	Logistic Normal
ln durt	**10.6313****	**11.7793****	**10.6153****	**10.6475****	**11.1165****
(ln durt)²	**-3.6309****	**-3.7946****	**-3.6151****	**-3.6363****	**-3.6825****
durt	**0.6475****	**0.6711****	**0.6532****	**0.6685****	**0.6599****
(durt)²	**-0.0029****	**-0.0030****	**-0.0029****	**-0.0029****	**-0.0030****
Ln income	-0.0472**	-0.0854**	-0.0644**	-0.0473**	-0.0746**
age25_29	-0.1107	-0.2159*	-0.1571	-0.1110	-0.1859
age30_33	-0.0172	-0.0784	-0.0446	-0.0174	-0.0613
age34_37	0.0845	0.0573	0.0525	0.0846	0.0602
age38_41	-0.0258	-0.1500	-0.0952	-0.0259	-0.1181
age42_47	0.0313	-0.0737	-0.0348	0.0312	-0.0487
age48_55	-0.2037*	-0.3955**	-0.2770**	-0.2041*	-0.3363**
age56_plus	-0.3624**	-0.6331**	-0.4638**	-0.3629**	-0.5485**
A1	**1.1260****	**1.7334****	**1.3076****	**1.1281****	**1.5490****
A2	**0.7461****	**1.1625****	**0.8889****	**0.7474****	**1.0402****
A3	**0.6951****	**1.0578****	**0.8070****	**0.6961****	**0.9472****
Employed	**0.2586****	**0.4121****	**0.3130****	**0.2590****	**0.3639****
Self employed	**0.4686****	**0.7532****	**0.5769****	**0.4695****	**0.6674****
Years at address	-0.0102**	-0.0162**	-0.0120**	-0.0102**	-0.0142**
Time with bank	-0.0026**	-0.0036**	-0.0029**	-0.0026**	-0.0033**
Number of cards	-0.0020	-0.0195	-0.0106	-0.0020	-0.0140
Δ base rate	-0.1856	-0.2380	-0.2187	-0.1861	-0.2320
Δearnings	2.8957*	2.8640*	2.8836*	2.9002*	2.9042*
Δunemployment	0.0111**	0.0111**	0.0112**	0.0111**	0.0113**
Δhouse prices	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
Δtotal credit outsg	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000
Δconsumer confidence	-0.0134	-0.0141	-0.0137	-0.0134	-0.0141
Constant	-16.7063**	-20.6517**	-16.5958**	-16.7206**	-19.0229**
Var (υ_i)		5.8758	2.7094		3.8532
Chibarsq(01)		24.10**	12.89**		18.61**
No obs	1499400	1499400	1499400	1499400	1499400
	21a	21c	21b	21dd	21ee

Predictive Performance

Training sample: accounts opened late 1990s to 31 December 2003

Test sample: accounts opened after 1 January 2004 (N=8307)

Frailty Models: predict survival probability when frailty term has mean value

For those that survived 12 months:	Survival probabilities 12 months
For those that defaulted or left the sample before 12 months:	Survival probabilities in most recent month

Cut-off such that
% predicted to default at or before 12 months = % observed to default at or before 12 months in training sample

Predict if good or bad at or before 12 months	compared with	observed if defaulted at or before 12 months

Percentage Correctly Classified

% of observed defaults in training sample = % of predicted defaults in training sample.

	No frailty	Gaussian	Gamma
	Flexible Baseline Hazard		
PH	91.67(0.40)	92.84(0.40)	92.74(0.40)
Logistic	92.75(0.04)	92.84(0.40)	
	Simple Baseline Hazard		
PH	93.20(0.80)	93.02(0.80)	92.68(0.80)
Logistic	93.19(0.40)	92.98(0.80)	

(% of bads)

Conclusions

Theoretically, omitting a frailty term when it should be included will make

- positive coefficients lower and negative coefficients higher.
- a declining (wrt time) hazard function flatter.

In credit card data

- frailty terms are statistically significant over a variety of models and frailty distributions
- in simple baseline hazard functions: the inclusion of a frailty term substantially alters the parameters of the baseline (alters the slope of the hazard function)
- in more flexible baseline hazard functions: the inclusion of a frailty term makes negligible difference to the parameters (of the baseline)
- in both simple and flexible baseline hazard functions the inclusion of frailty substantially alters the parameters of the time invariant covariates, but only negligibly alters the parameters of (external) time varying covariates.
- inclusion of frailty, regardless of distribution, does not improve overall predictive performance or proportion of observed bads that are correctly classified.