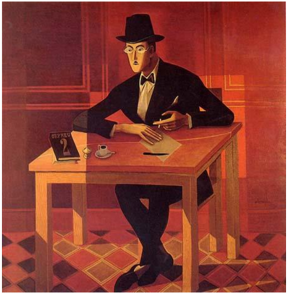
Microeconometrics
Duration Analysis

Pedro Portugal

Nova
School of Business and Economics

Spring 2025

Nem relógio parado, nem a falta
Da água em clepsidra, ou ampulheta cheia,
Tiram o tempo ao tempo



Poesia de Ricardo Reis

Goya *Saturn Devouring His Son* c. 1819–1823

Duration Analysis

- Econometric models of durations are models of the length of time spent in a given state before transition to another state
- A **state** is a classification of an individual entity at a point in time, **transition** is movement from one state to another, and a **spell** length or **duration** is the time spent in a given state
  - Example: regression analysis of the impact of higher unemployment benefit levels on the average length of an unemployment spell or the probability of transition out of unemployment
- **Duration analysis** or **transition analysis** is also called **survival analysis**, **failure time analysis**, **life time analysis** and **hazard analysis**

Duration Analysis

- Relationship with stochastic processes
  - Point Processes
  - Count Processes
  - Renewal Processes

- Connection with limited dependent variable models
  - Poisson Regression
  - Discrete Choice
  - Tobits

Duration Analysis: Applications

- Classical:
  - Survival time after heart transplant (Stanford)
  - Failure time of plane engines
  - Survival of mice after cancer-prone substance administration
  - Duration of political dictatorships
  - Strike duration

- Labor economics:
  - Unemployment duration
  - Duration of income-support programs
  - Job duration
  - Employment adjustment

Duration Analysis: Applications

- Industrial economics:
  - Firm closure (firm default)
  - Holding foreign firms
- Macroeconomics:
  - Business cycles
  - Investment regimes
- Finance:
  - Debt default

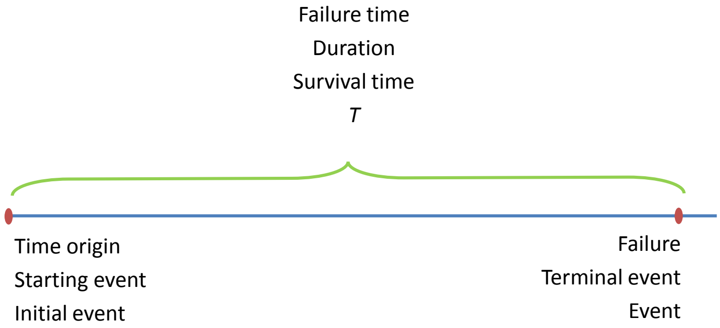
Duration Analysis

Duration in a state is a nonnegative random variable, denoted T, which in economic data is often a discrete random variable

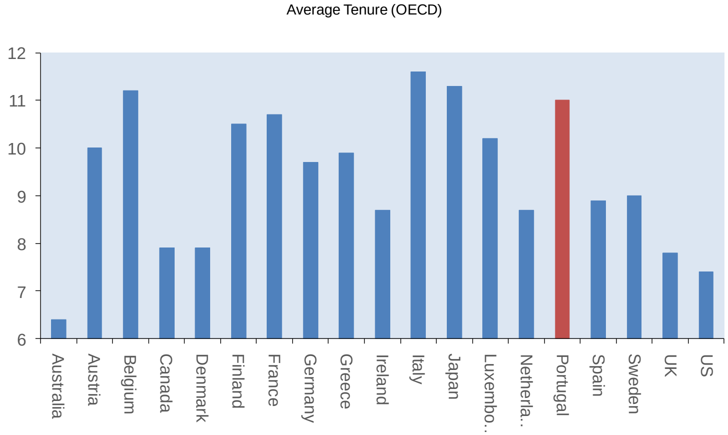
Extension of analysis to regression models:

- Censoring (completed spell is not completely observed)
- Inference conditional on time
- Models with time-varying regressors
- Discrete hazard models

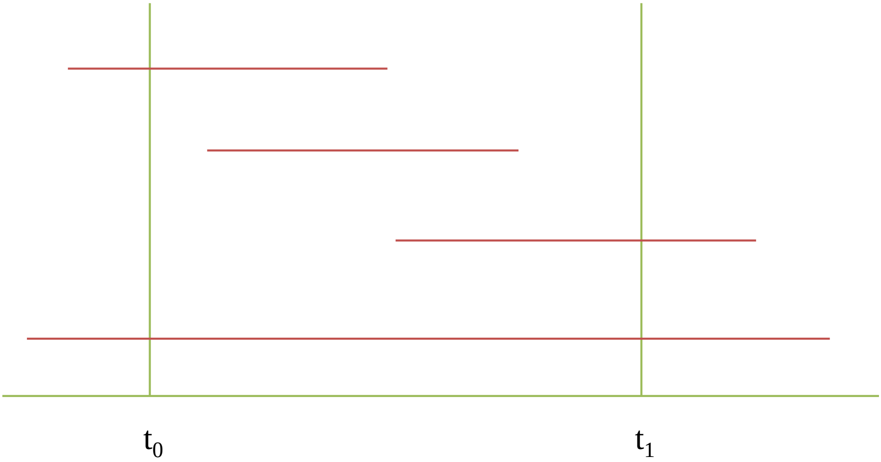
Duration


Failure time
Duration
Survival time
T
Time origin
Starting event
Initial event
Failure
Terminal event
Event

Duration Analysis: Single Risk Model - Job Duration



Duration Analysis: Censoring



Censoring

An individual's survival time in a state is censored if the date of transition into the state, or the date of transition out of the state, is not known exactly (only that before some date, or after some date)

- **Right censoring**: spell end date not observed (only know that total time in state $\geq$ time from start of spell to end of observation period)
- **Left censoring**: spell start date not observed

Truncation

Whereas censoring means that we don't know the exact length of a completed spell in total, truncation refers to whether or not we observe a spell or not in our data (sample selection on dependent variable):

- **Left truncation**: only those surviving a sufficient amount of time are included in the sample (e.g. stock sample with follow-up)

- **Right truncation**: only those with a transition by a particular time are included in the sample (e.g. sample from the outflow from a state)

Survivor, Hazard and Cumulative Hazard Function

The **cumulative distribution function** of T is denoted $F(t)$ and the **density function** is $f(t) = dF(t)/dt$

Then the probability that the duration or spell length is less than t is

$$F(t) = Pr[T \leq t] = \int_0^t f(s)ds$$

The probability that duration equals or exceeds t is called the **survivor function**, defined by

$$S(t) = Pr[T > t] = 1 - F(t)$$

Survivor, Hazard and Cumulative Hazard Function

The **hazard function** gives the instantaneous probability of leaving a state conditional on survival to time t, defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

We can verify that the hazard equals the change in log-survivor function,

$$\lambda(t) = -\frac{d \ln(S(t))}{dt}$$

The hazard $\lambda(t)$ specifies the distribution of T. In particular, integrating $\lambda(t)$ and using $S(0) = 1$ we can show that

$$S(t) = \exp\left(-\int_0^t \lambda(u) du \right)$$

Survivor, Hazard and Cumulative Hazard Function

The **cumulative hazard function** is defined by

$$\Lambda(t) = \int_0^t \lambda(s) ds = -\ln S(t)$$

or, equivalently,

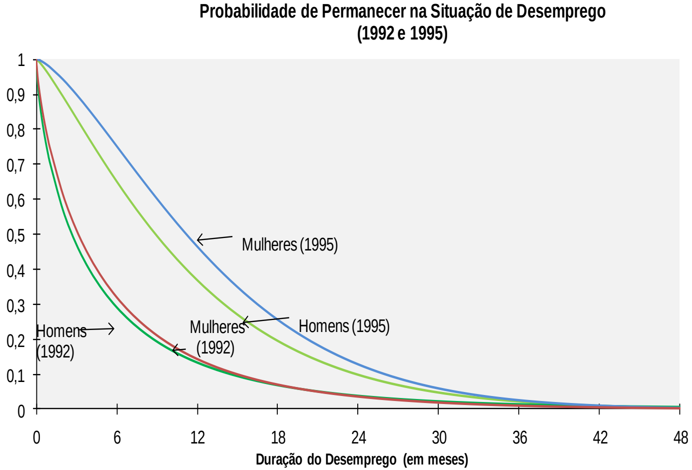
$$S(t) = \exp(-\Lambda(t))$$

Survival Analysis: Definitions of Key Concepts

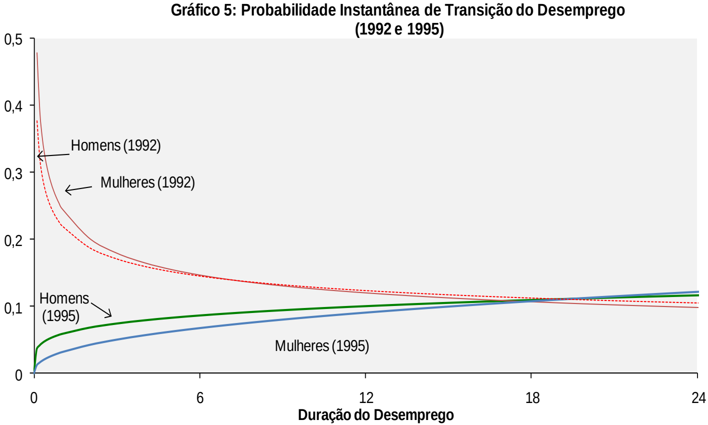
Function	Symbol	Definition	Relationships
Density	$f(t)$		$f(t) = dF(t)/dt$
Distribution	$F(t)$	$\Pr[T \leq t]$	$F(t) = \int_0^t f(s)ds$
Survivor	$S(t)$	$\Pr[T > t]$	$S(t) = 1 - F(t)$
Hazard	$\lambda(t)$	$\lim_{h\to 0} \frac{\Pr[t \leq T < t+h \mid T \geq t]}{h}$	$\lambda(t) = f(t)/S(t)$
Cumulative hazard	$\Lambda(t)$	$\int_0^t \lambda(s)ds$	$\Lambda(t) = -\ln S(t)$

Figure: Source: Cameron and Trivedi, 2005

Survival Function



Hazard Function



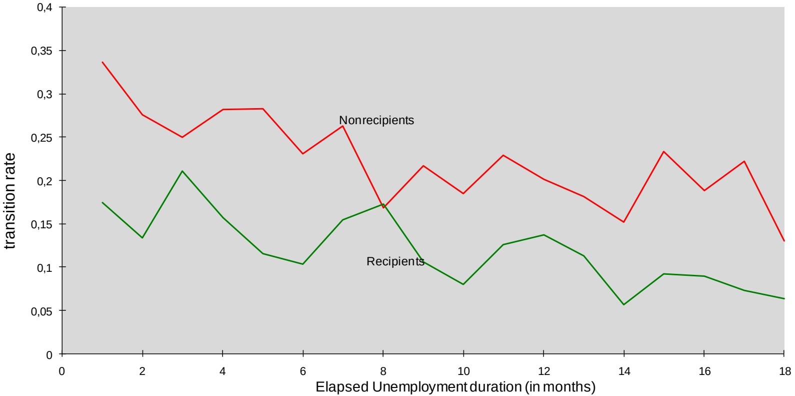
Hazard Function

"O conceito de taxa de quebras, sob o aspecto de força de mortalidade (Demografia) foi criado em 1757 por Soares de Barros e Vasconcelos, um estrangeirado...que o publicou no artigo "Loxodromia da Vida Humana", Mem. Real Academia de Sciencias de Lisboa, 1ª série, I, 1759. Soares de Barros e Vasconcelos usa $1/h(t)$ (chamada força da vida), que interpreta correctamente; note-se que $1/h(t)$ je um instrumento importante na Estatística dos Extremos. Só mais tarde Gompertz (1825) e Makeham (1860) redescobrem o conceito e o utilizam em Demografia e Actuariado."

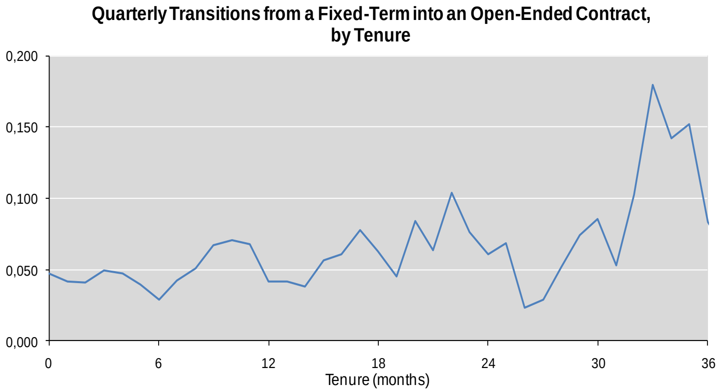
in Probabilidades e Estatística de Tiago de Oliveira

Empirical hazard function

Figure 1: Empirical Hazard Function by Unemployment Benefit Recipiency Status



Empirical hazard function



Residual Duration

$$\mu(t) = \int_t^{\infty} S(u)du$$

$$\mu_0 = \int_0^{\infty} S(t)dt$$

Exponential and Hazard Distributions: pdf, cdf, Survivor function, Hazard, Cumulative Hazard, Mean, and Variance

Function	Exponential	Weibull
$f(t)$	$\gamma \exp(-\gamma t)$	$\gamma\alpha t^{\alpha-1} \exp(-\gamma t^{\alpha})$
$F(t)$	$1 - \exp(-\gamma t)$	$1 - \exp(-\gamma t^{\alpha})$
$S(t)$	$\exp(-\gamma t)$	$\exp(-\gamma t^{\alpha})$
$\lambda(t)$	γ	$\gamma\alpha t^{\alpha-1}$
$\Lambda(t)$	γt	γt^{α}
$\mathrm{E}[T]$	γ^{-1}	$\gamma^{-1/\alpha}\Gamma(\alpha^{-1}+1)$
$\mathrm{V}[T]$	γ^{-2}	$\gamma^{-2/\alpha}[\Gamma(2\alpha^{-1}+1) - [\Gamma(\alpha^{-1}+1)]^2]$
γ, α	$\gamma > 0$	$\gamma > 0, \alpha > 0$

Figure: Source: Cameron and Trivedi, 2005

Exponential and Weibull Distributions

- The **exponential duration distribution** has a constant hazard rate γ, that does not vary with t (memoryless property of the exponential)
- It follows that

$$S(t) = \exp(-\int_0^t \gamma du) = \exp(-\gamma t)$$

- The density is

$$f(t) = -S'(t) = \gamma \exp(-\gamma t)$$

- and the cumulative hazard is

$$\Lambda(t) = -\ln S(t) = \gamma t$$

And,

$$\mu_0 = \frac{1}{\gamma}$$

Exponential and Weibull Distributions

- The **Weibull distribution** is less restrictive than the exponential, which is the special case $\alpha = 1$
- The Weibull has hazard $\lambda(t) = \gamma \alpha t^{\alpha - 1}$
  - which is monotonically increasing if $\alpha > 1$, and
  - monotonically decreasing if $\alpha < 1$
- Special case of the Proportional Hazards (PH) family, in which $\lambda(t)$ factors into a baseline component that depends only on t, $\lambda_0(t)$, and a second term (that can be parameterized as a function of covariates only)

Standard Parametric Models and Their Hazard and Survivor Functions[a][1]

Parametric Model	Hazard Function	Survivor Function	Type
Exponential	γ	$\exp(-\gamma t)$	PH, AFT
Weibull	$\gamma \alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$	PH, AFT
Generalized Weibull	$\gamma \alpha t^{\alpha-1} S(t)^{-\mu}$	$[1-\mu\gamma t^{\alpha}]^{1/\mu}$	PH
Gompertz	$\gamma \exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$	PH
Log-normal	$\frac{\exp\left(-(\ln t-\mu)^2/2\sigma^2\right)}{t\sigma\sqrt{2\pi}[1-\Phi((\ln t-\mu)/\sigma)]}$	$1-\Phi\left((\ln t-\mu)/\sigma\right)$	AFT
Log-logistic	$\alpha\gamma^{\alpha}t^{\alpha-1}/\left[(1+(\gamma t)^{\alpha})\right]$	$1/\left[1+(\gamma t)^{\alpha}\right]$	AFT
Gamma	$\frac{\gamma(\gamma t)^{\alpha-1}\exp[-(\gamma t)]}{\Gamma(\alpha)[1-I(\alpha,\gamma t)]}$	$1-I(\alpha,\gamma t)$	AFT

[a] All the parameters are restricted to be positive, except that $-\infty < \alpha < \infty$ for the Gompertz model.

Figure: Source: Cameron and Trivedi, 2005

[1] The function $\Gamma(.)$ is the gamma function

Maximum Likelihood Estimation

- The treatment is similar to that for the Tobit model
- For **uncensored observations** the contribution to the likelihood is $f(t|\mathbf{x},\theta)$
    - where θ is a $q \times 1$ parameter vector and $\mathbf{x}$ are regressors that can vary across subjects but do not vary over a spell for a given subject
- For **right-censoring observations** we know only that the duration exceeds t, so the contribution is

$$
\begin{aligned}
Pr[T > t] &= \int_t^\infty f(u|\mathbf{x},\theta)du \\
&= 1 - F(t|\mathbf{x},\theta) = S(t|\mathbf{x},\theta)
\end{aligned}
$$

where $S(.)$ is the survivor function.

Maximum Likelihood Estimation

- The density for the i^{th} observation can be written as

$$f(t_i|\mathbf{x}_i, \theta)^{\delta_i} S(t_i|\mathbf{x}_i, \theta)^{1-\delta_i}$$

where δ_i is a right-censoring indicator with $\delta_i = 1$ (no censoring) or $\delta_i = 0$ (right-censoring)

- The MLE $\widehat{\theta}$ maximizes the log-likelihood

$$\ln L(\theta) = \sum_{i=1}^{N} \left[\delta_i \ln f(t_i|\mathbf{x}_i, \theta) + (1 - \delta_i) \ln S(t_i|\mathbf{x}_i, \theta)\right]$$

Maximum Likelihood Estimation

- Since $\ln S(t) = -\Lambda(t)$ and $\ln f(t) = \ln(\lambda(t)S(t)) = \ln \lambda(t) + \ln S(t)$ we can alternatively write

$$\ln L(\theta) = \sum_{i=1}^{N} \left[\delta_i \ln \lambda(t_i|\mathbf{x}_i, \theta) + \Lambda(t_i|\mathbf{x}_i, \theta)\right]$$

- Given a mix of data, with durations that may be complete, truncated, or censored, each type of observation contributes a term to the likelihood function, and the full likelihood is formed by taking appropriate products of terms

Maximum Likelihood Estimation

complete durations:	$f(t)$
left-truncated at t_L $(t \geq t_L)$:	$f(t)/S(t_L)$
left-censored at t_{C_L} :	$1 - S(t_{C_L})$
right-censored at t_{C_R} :	$S(t_{C_R})$
right-truncated at t_{C_R} $(t \leq t_R)$:	$f(t_R)/[1 - S(t_R)]$
interval-censored at t_{C_L}, t_{C_R} :	$S(t_{C_L}) - S(t_{C_R})$

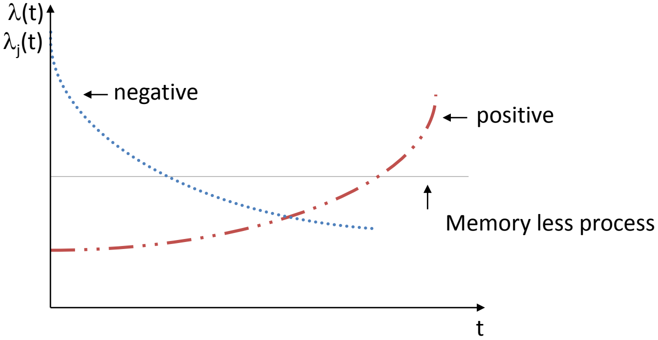
Maximum Likelihood Estimation

$$L = \prod_{i=1}^{N} \left[f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \right] \left[g(t_i)^{1-\delta_i} (1 - G(t_i))^{\delta_i} \right]$$

$$L = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

$$L = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(C_i)^{1-\delta_i}$$

Duration Dependence



Duration Dependence

Hazard function	Constant	Increasing	Decreasing	U Shaped	∩ Shaped
Exponential	●				
Weibull	●	●	●		
Log-normal					●
Log-logistic			●		●
Gompertz	●	●	●		
Pareto			●		
Gamma	●		●		●
Generalized gamma	●	●	●	●	●

Piecewise Constant Hazard Function

- Is an example of the proportional hazard model
  $\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\phi(\mathbf{x}, \beta)$

- Lets $\lambda_0(t, \alpha)$ be a step function with k segments so that

$$\lambda_0(t, \alpha) = e^{\alpha_j}, \quad c_{j-1} \leq t < c_j, \quad j = 1, ..., k$$

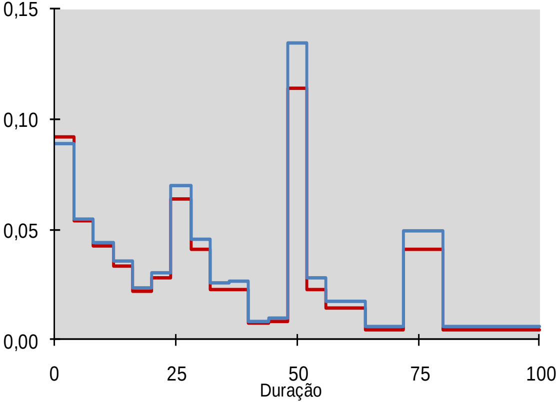
  where $c_0 = 0$, $c_k = \infty$, the other breakpoints $c_1, ..., c_{k-1}$ are specified, and the parameters $\alpha_1, ..., \alpha_k$ are to be estimated

- The parameters are exponentiated to ensure $\lambda_0(t, \alpha) > 0$

Piecewise Constant Hazard Function

$$\lambda_t = \begin{cases} \lambda_1 & 0 < t \leq c_1 \\ \lambda_2 & c_1 < t \leq c_2 \\ \lambda_3 & c_2 < t \leq c_3 \\ \lambda_4 & t > c_3 \end{cases}$$

Piecewise Constant Hazard Function



Proportional Hazards (PH) Model

The conditional hazard rate can be factored into separate functions of

$$\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\phi(\mathbf{x}, \beta)$$

where $\lambda_0(t, \alpha)$ is called the **baseline hazard** and is a function of t alone, and $\phi(\mathbf{x}, \beta)$ is a function of $\mathbf{x}$ alone. Usually $\phi(\mathbf{x}, \beta) = \exp(\mathbf{x}'\beta)$

All $\lambda(t|\mathbf{x})$ of this form are proportional to the baseline hazard, with scale factor $\phi(\mathbf{x}, \beta)$

The exponential, Weibull, and Gompertz regression models are PH models, since their hazards are $\exp(\mathbf{x}'\beta)$, $\exp(\mathbf{x}'\beta)\alpha t^{\alpha-1}$, and $\exp(\mathbf{x}'\beta)\exp(\alpha t)$, respectively

Accelerated Failure Time (AFT) Model

A regression model is specified for $\ln t = \mathbf{x}'\beta + u$ and different distributions for u lead to different AFT models

The term AFT arises because $t = \exp(\mathbf{x}'\beta)\nu$, where $\nu = e^u$, has hazard rate $\lambda(t|\mathbf{x}) = \lambda_0(\nu)\exp(\mathbf{x}'\beta)$, where $\lambda_0(\nu)$ does not depend on t

Substituting $\nu = t\exp(-\mathbf{x}'\beta)$ yields the hazard

$$\lambda(t|\mathbf{x}) = \lambda_0(t\exp(-\mathbf{x}'\beta))\exp(\mathbf{x}'\beta)$$

This is an acceleration of the baseline hazard $\lambda_0(t)$ if $\exp(-\mathbf{x}'\beta) > 1$ and a deceleration if $\exp(-\mathbf{x}'\beta) < 1$

Regression Analysis

- PH Model

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}'\beta)$$

$$S(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}'\beta)}$$

$$-\ln\Lambda(t|\mathbf{x}) = \mathbf{x}'\beta + \nu$$

- AFT Model

$$\ln(T|\mathbf{x}) = \mathbf{x}'\alpha + \sigma u$$

$$S(t|\mathbf{x}) = S_0(t\exp(-\mathbf{x}'\alpha))$$

$$\lambda(t|\mathbf{x}) = \lambda_0(t\exp(-\mathbf{x}'\alpha))\exp(\mathbf{x}'\alpha)$$

Two-Destination Piecewise-Constant Hazards Regression Models (n=9,451)

	Transition to:	
Variable	Employment	Inactivity
UB	-0,555	-0,511
	(0,067)	(0,165)
AGE GROUP		
25-29	-0,008	-0,434
	-(0,080)	(0,217)
30-34	-0,141	-0,820
	(0,022)	(0,306)
35-39	-0,238	0,550
	(0,119)	(0,351)
40-44	-0,098	0,034
	(0,118)	(0,311)
45-49	-0,246	-0,098
	(0,133)	(0,339)
50-54	-0,416	0,280
	(0,152)	(0,314)
55+	-0,972	0,244
	(0,159)	(0,310)
SCHOOLING	0,008	0,021
	0,008	(0,020)
TENURE	-0,023	0,008
	(0,005)	(0,009)
MARRIED	0,308	-0,194
	(0,078)	(0,214)
FIRSTJOB	-0,415	0,360
	(0,093)	(0,182)
LAYOFF	0,022	-0,642
	(0,090)	(0,241)
END FIXED	0,185	-0,124
	(0,062)	(0,172)
UNEMPLOYMENT RATE	-0,021	-0,137

	Quantile Regression				
	20th	50th	80th	Cox	AFT
Age	0.015	0.020	0.016	-0.013	0.021
(in years)	(0.006)	(0.004)	(0.003)	(0.002)	(0.003)
Gender	-0.004	0.111	0.082	-0.082	0.116
(male=1)	(0.165)	(0.116)	(0.085)	(0.057)	(0.087)
Race	-.290	-.345	-0.370	0.322	-0.481
White=1	(0.158)	(0.099)	(0.083)	(0.054)	(0.082)
Marital status	-0.323	-.214	-0.097	0.108	-0.189
(married=1)	(0.136)	(0.094)	(0.072)	(0.046)	(0.071)
Marital*Gender	0.495	0.625	0.451	-0.343	0.541
(married female=1)	(0.263)	(0.150)	(0.120)	(0.074)	(.112)
Schooling	-0.080	-0.016	-0.019	0.018	-0.031
(in years)	(0.024)	(0.016)	(0.013)	(0.008)	(0.012)
Tenure	-0.001	0.022	0.020	-0.009	0.014
(in years)	(0.013)	(0.008)	(0.005)	(0.003)	(0.005)
Unskilled	0.312	0.387	0.254	-0.200	0.330
(Unskilled=1)	(0.122)	(0.086)	(0.065)	(0.040)	(0.061)
Plant Closing	-0.668	-0.357	-0.164	0.179	-0.321
(Shutdown=1)	(0.123)	(0.072)	(0.057)	(0.034)	(0.053)
Informal Notice	-0.292	-0.081	-0.051	0.043	-0.082
(Notice=1)	(0.123)	(0.080)	(0.057)	(0.035)	(0.054)
Written Notice	-0.757	0.097	0.031	-0.014	-0.038
(Notice=1)	(0.394)	(0.196)	(0.111)	(0.078)	(.120)
Unemp. Rate	0.0931	0.122	0.123	-0.076	0.116
	(0.026)	(0.018)	(0.014)	(0.008)	(0.012)
Previous Wage	-0.261	0.032	0.009	0.014	-0.038
(in logs)	(0.119)	(0.077)	(0.069)	(0.037)	(0.057)
Constant	1.167	2.432	3.598		2.890.
	(0.160)	(0.122)	(0.098)		(0.106)
Scale Parameter					1.565
Shape Parameter					0.613

Specification Problems

Basic Model:

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}'\beta)$$

Time-varying regressors:

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}(t)'\beta)$$

Time-varying coefficients:

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}'\beta(t))$$

Specification Problems

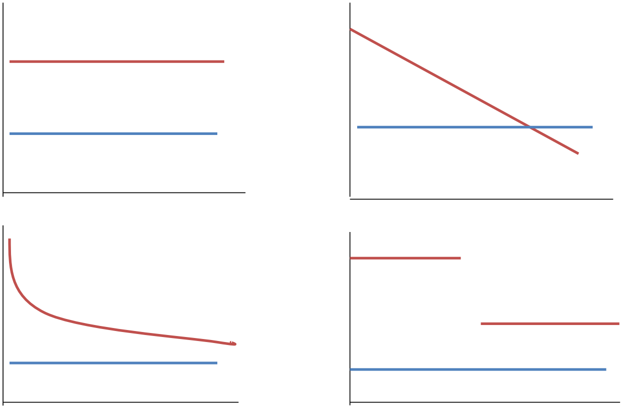
Time-varying regressors and coefficients:

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}(t)'\beta(t))$$

Unobserved Individual Heterogeneity:

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}'\beta)\nu$$

Time-Varying Coefficients



Time-Varying Coefficients and Regressors

Table 5. Regression results (N=118070)

Variable	Column 1		Column 2		Column 3		Column 4	
			Exponential decay		Linear decay		Exponential decay	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Size								
Initial	−0.336	0.005	−0.354	0.007	−0.353	0.007	−0.366	0.008
Decay			0.966	0.008	0.011	0.003	0.997	0.009
Change	−0.623	0.009	−0.616	0.009	−0.618	0.010	−0.655	0.009
College								
Initial	−0.472	0.054	−0.510	0.067	−0.507	0.064	−0.524	0.067
Decay			0.912	0.069	0.031	0.025	0.965	0.056
Change	−0.006	0.075	0.042	0.079	0.029	0.081	0.038	0.081
Entry rate								
Initial	0.643	0.085	0.693	0.095	0.684	0.097	0.764	0.099
Decay			1.028	0.048	0.003	0.037	1.063	0.040
Change	0.104	0.111	0.287	0.121	0.204	0.121	0.273	0.126
Concentration								
Initial	−0.415	0.071	−0.988	0.096	−0.649	0.094	−0.856	0.103
Decay			−0.007	0.164	0.183	0.031	−0.021	0.195
Change	−0.235	0.147	0.107	0.135	0.154	0.155	0.116	0.143
GDP growth								
Initial	−0.049	0.002	−0.050	0.002	−0.050	0.002	−0.052	0.002
Decay			0.979	0.026	0.002	0.001	0.992	0.025
Change	−0.031	0.002	−0.030	0.003	−0.028	0.003	−0.028	0.003
Exit rate	3.239	0.105	3.418	0.115	3.310	0.125	3.757	0.140
Gamma variance							0.283	0.056
LL	−158 292		−158 253		−158 263		−158 242	

Time-Varying Regressors

Continuous time, three intervals, complete duration

$$S(t_1|\mathbf{x_1})\frac{S(t_2|\mathbf{x_2})}{S(t_1|\mathbf{x_2})}\frac{f(t_3|\mathbf{x_3})}{S(t_2|\mathbf{x_3})}$$

Continuous time, three intervals, incomplete duration

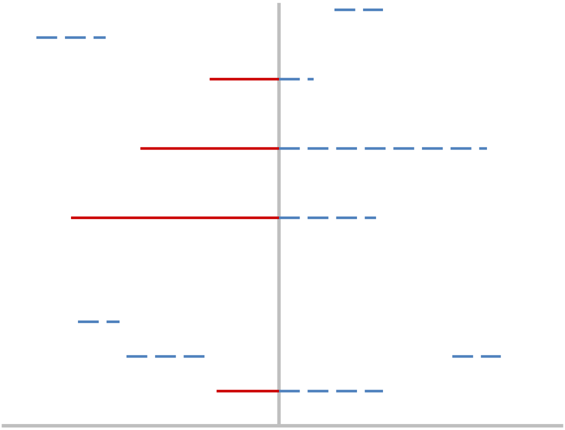
$$S(t_1|\mathbf{x_1})\frac{S(t_2|\mathbf{x_2})}{S(t_1|\mathbf{x_2})}\frac{S(t_3|\mathbf{x_3})}{S(t_2|\mathbf{x_3})}$$

Sampling Plan

Sampling plan is crucial to the construction of the likelihood function

- **Stock sampling**: sampling in the survey period from the stock of individuals who are in a given state
- **Flow sampling**: we sample those who enter the state during a particular interval
- **Observation over a fixed interval**
- **Destination**

Inquérito ao Emprego do INE (Stock Sampling)



Sampling Plan

- Flow Sampling

$$\prod_{i=1}^{N}[f_i(t)]^{\delta_i}[S_i(t)]^{1-\delta_i}$$

- Stock Sampling

$$\prod_{i=1}^{N}\frac{S_i(t)}{\mu_i}$$

- Observation over a fixed interval

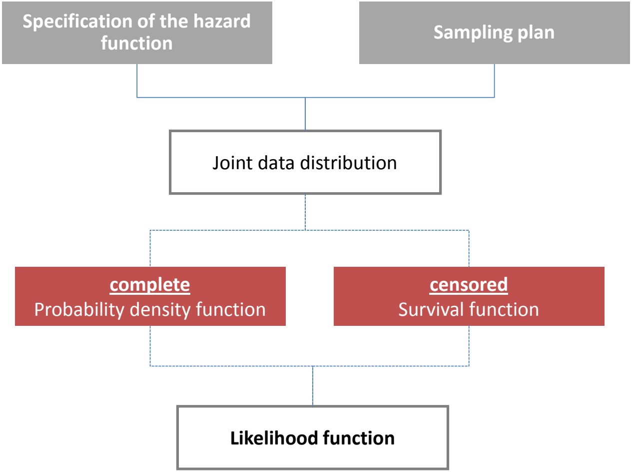
$$\prod_{i=1}^{N}\left[\frac{f_i(e+k)}{S_i(e)}\right]^{\delta_i}\left[\frac{S_i(e+l)}{S_i(e)}\right]^{1-\delta_i}$$

Sampling Plan

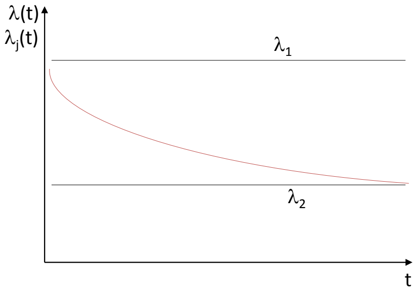
- Destination

$$\prod_{i=1}^{N} f_i(t)$$

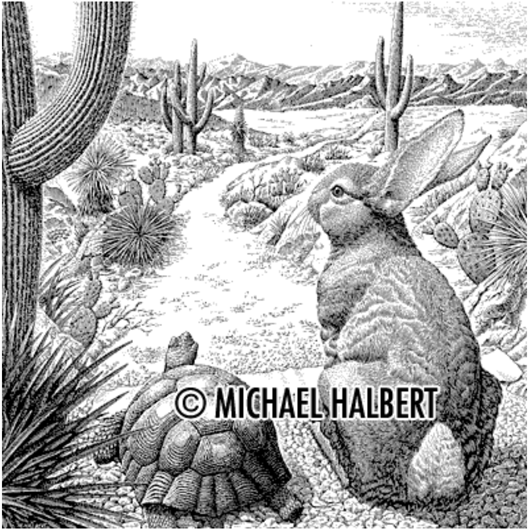
Summing up


Specification of the hazard function
Sampling plan
Joint data distribution
complete
Probability density function
censored
Survival function
Likelihood function

Unobserved Individual Heterogeneity



Unobserved Individual Heterogeneity



Unobserved Individual Heterogeneity

Mass Point Approach

$$L = \prod_{i=1}^{N} [pf_1(t_i)]^{\delta_i} [pS_1(t_i)]^{1-\delta_1} + [(1-p)f_2(t_i)]^{\delta_i} [(1-p)S_2(t_i)]^{1-\delta_i}, 0 \leq p \leq 1$$

$$L = \prod_{i=1}^{N} [p_1 f_1(t_i)]^{\delta_i} [p_1 S_1(t_i)]^{1-\delta_i} + [p_2 f_2(t_i)]^{\delta_i} [p_2 S_2(t_i)]^{1-\delta_i} + ... + \\ + [p_J f_J(t_i)]^{\delta_i} [p_J S_J(t_i)]^{1-\delta_i}$$

where $0 \leq p_j \leq 1, j = 1, ..., J$, and $p_1 + p_2 + ... + p_J = 1$

Unobserved Individual Heterogeneity

Parametric Approach

Consider the PH model with unobserved heterogeneity (extended to include a multiplicative term ν)

$$\lambda(t, X|\nu) = \lambda_0(t) \exp(\mathbf{x}'\beta)\nu, \nu > 0$$

The relationship between frailty and non-frailty survivor function is:

$$S(t, X|\nu) = [S(t, X)]^{\nu}$$

- Individuals with above-average ν exit relatively fast (higher hazard and shorter survival time), and vice-versa.

Work with a function $S_\nu(t, X) = S_\nu(t, X|\beta, \sigma^2)$. Given some probability density $g(\nu)$ for ν, then

$$S_\nu(t, X) = \int_0^\infty [S(t, X)]^\nu g(\nu) d\nu$$

Unobserved Individual Heterogeneity

Gamma Heterogeneity

When $g(\nu)$ is Gamma distribution, the frailty survivor function is:

$$S(t, X|\beta, \sigma^2) = [1 + \sigma^2 \Lambda(t, X)]^{-\frac{1}{\sigma^2}}$$

And the frailty hazard function is:

$$\lambda(t, X|\nu) = \lambda(t, X)[1 + \sigma^2 \Lambda(t|\mathbf{x})]^{-1}$$

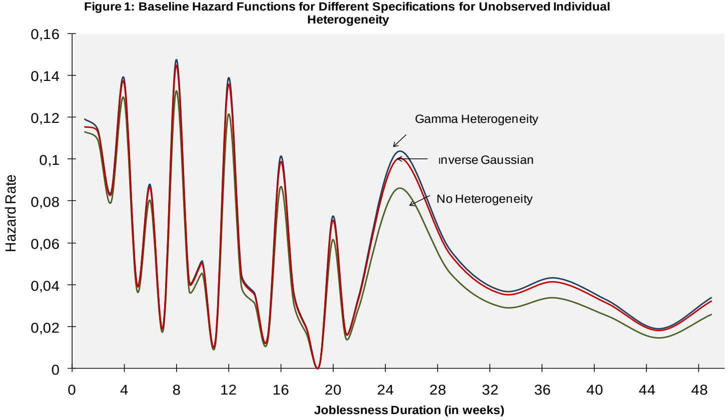
Gamma Heterogeneity: Example with the exponential distribution

$$S(t, X|\beta, \sigma^2) = [1 + \sigma^2 \gamma t \exp(\mathbf{x}'\beta)]^{-\frac{1}{\sigma^2}}$$

$$\lambda(t, X|\nu) = \gamma \exp(\mathbf{x}'\beta)[1 + \sigma^2 \gamma t \exp(\mathbf{x}'\beta)]^{-1}$$

$$\lambda(t) = \gamma[1 + \sigma^2 \gamma t]^{-1}$$

Unobserved Individual Heterogeneity

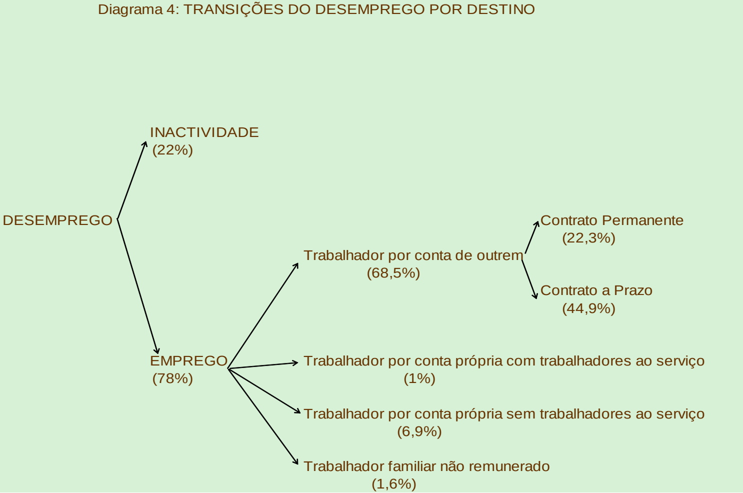


Competing Risk Models

- Three Risks
  - T_1 Full-time dependent worker
  - T_2 Part-time dependent worker
  - T_3 Self-employed
- Competing Risks

$$T = \min(T_1, ..., T_j), \qquad J = j$$

Transições no Mercado de Trabalho



Multiple Destinations

- Specific hazard functions for each destination

$$\lambda_j(t) = h_{0j}(t)\exp(\mathbf{x}'\beta_j)$$

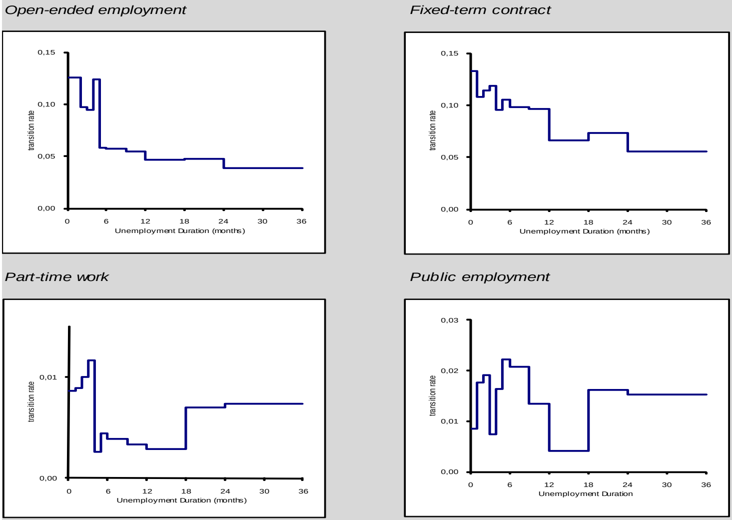
Then

$$\lambda(t) = \sum_{j=1}^{J} \lambda_j(t)$$

- Non-independent competing risks

$$\lambda_j(t) = \lambda_0(t)\exp(\mathbf{x}'\beta_j)\exp(\nu_j)$$

Baseline Hazard Functions by Destination State



Defective Risks

$$f^*(t) = \frac{f(t)}{1 - S(\infty)}$$

$$S^*(t) = \frac{S(t) - S(\infty)}{1 - S(\infty)}$$

$$\lambda^*(t) = \frac{f^*(t)}{S^*(t)} = \frac{f(t)}{S(t) - S(\infty)}$$

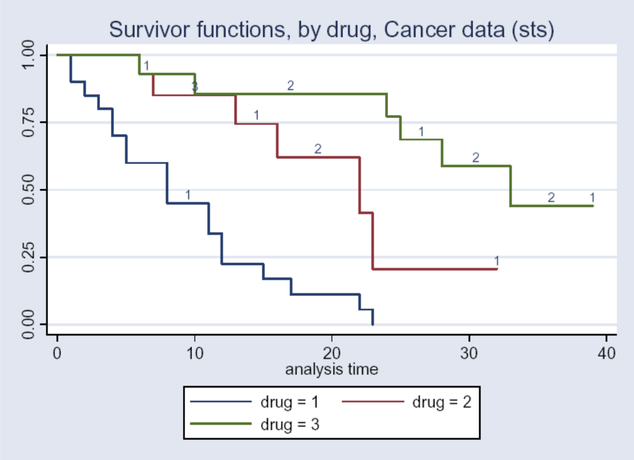
Defective Risks

$$f^*(t) = \frac{f(t)}{1-p}$$

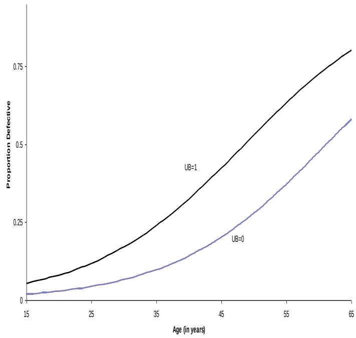
$$S^*(t) = \frac{S(t)-p}{1-p}$$

$$\lambda^*(t) = \frac{f(t)}{S(t)-p}$$

Degenerate Distributions

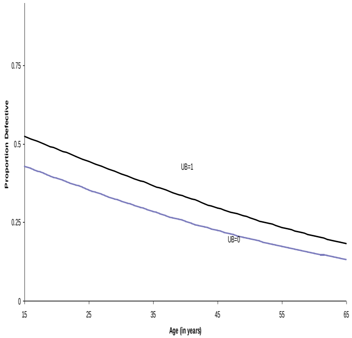


Defective Risks



Note: Vertical axis gives the proportion of U-E stayers (non-employable) for individuals w ith mean characteristics for the continous variables and reference category for the binary variables. Simulation values are obtained from the split-population equation estimates in Table 2, column 5.

Figure 10
Defective Risk into Employment by Age and UB Status



Note: Vertical axis gives the proportion of U-I stayers (not considering inactivity) for individua w ith mean characteristics for the continous variables and reference category for the binary variables.
Simulation values are obtained from the split-population equation estimates in Table 2, column

Figure 11
Defective Risk into Inactivity by Age and UB Status

Simulation

Simulations from the Split-Population Model

	Age=20 Years		Age=35 Years		Age=50 Years	
	UB=0	UB=1	UB=0	UB=1	UB=0	UB=1
Survival Rate after:						
3 months	0,629	0,792	0,672	0,841	0,790	0,920
12 months	0,257	0,466	0,321	0,574	0,474	0,723
36 months	0,050	0,140	0,092	0,250	0,218	0,435
Defective Risk:						
Employment	0,029	0,081	0,094	0,231	0,371	0,632
Inactivity	0,390	0,483	0,287	0,370	0,173	0,234
Median Duration: (in months)						
two destinations	5	11	7	16	11	28
until employment	7	14	7	21	24	na

Two-mass point approach

Transitions into employment

Variable	Specification				
	Complementary log–log (1)	Binomial-mixture model (2)	Split-population model (3)	Weibull-gamma (4)	Binomial-mixture model UB recipients (5)
UB	−0.694	−0.833	−0.833	−1.460	
	(0.171)	(0.190)	(0.190)	(0.309)	
YOUNG	−0.218	−0.277	−0.277	−0.219	
	(0.118)	(0.142)	(0.142)	(0.161)	
UB×YOUNG	0.531	0.625	0.625	0.597	0.477
	(0.230)	(0.255)	(0.255)	(0.307)	(0.240)
Log duration	−0.381	−0.254	−0.254		0.265
	(0.046)	(0.066)	(0.066)		(0.221)
Age					
29 or 30	−0.309	−0.501	−0.501	−0.159	−1.099
	(0.156)	(0.200)	(0.200)	(0.195)	(0.503)
34 or 35	−0.852	−1.143	−1.143	−0.552	−0.931
	(0.189)	(0.244)	(0.244)	(0.258)	(0.477)
39 or 40	−0.637	−0.995	−0.995	−0.142	−1.151
	(0.196)	(0.265)	(0.265)	(0.286)	(0.499)
44 or 45	−0.628	−0.941	−0.941	−0.198	−1.483
	(0.200)	(0.258)	(0.258)	(0.259)	(0.536)
49 or 50	−1.646	−2.064	−2.064	−1.119	−3.192
	(0.252)	(0.316)	(0.316)	(0.259)	(0.751)
54 or 55	−1.355	−1.708	−1.708	−0.802	−1.952
	(0.239)	(0.297)	(0.297)	(0.353)	(0.591)
Schooling	0.005	−0.009	−0.009	−0.011	−0.050
	(0.012)	(0.016)	(0.016)	(0.018)	(0.037)
Married	0.354	0.454	0.454	0.214	−0.129
	(0.133)	(0.161)	(0.161)	(0.190)	(0.364)
Handicapped	−0.502	−0.568	−0.568	−0.961	0.766
	(0.582)	(0.654)	(0.654)	(0.681)	(1.317)
Unemployment rate	−0.145	−0.175	−0.175	−0.055	−0.182
	(0.056)	(0.066)	(0.066)	(0.074)	(0.140)
Intercept	−3.069			−1.485	
	(0.137)			(0.255)	
Intercept for type 1 individuals		−13.879			−4.824
		(168.838)			(1.108)
Change in the intercept for type 2 individuals		11.541			2.342
		(168.834)			(1.108)
Intercept for susceptible individuals			−2.338		
			(0.240)		
Probability of being a type 1 individual		0.445			0.835
		(0.079)			(0.193)
Probability of being a type 2 individual		0.555			0.162
		(0.790)			(0.193)
Probability that a transition is never made			0.445		
			(0.079)		
Weibull parameter				0.764	
				(0.114)	
Gamma parameter				0.310	
				(0.152)	
Log-likelihood	−20063.4	−2061	−2061	−333.1	−578.8
n	35,368	35,368	35,368	2599	12,913

Semi-parametric (distribution-free) Estimation

- OLS
    - AFT – complete durations
    - Aggregate data
        - Survival Rates (Kaplan-Meier)
        - "Hazard Rates" (Life-Tables)
- Cox Model
    - Partial-Likelihood
    - Baseline Hazard
- Censored Quantile Regression
- Buckley-James

Kaplan-Meier Estimator

Define

- n_j to be the number of spells that end at time t_j
- r_j to equal the number of spells at risk at time t_{j-}, that is, just before time t_j

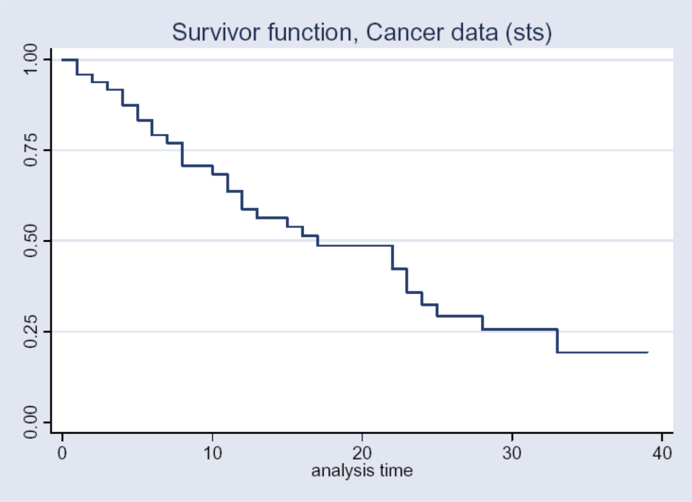
Since $\lambda_j = Pr[T = t_j | T \geq t_j]$, an obvious estimator of the hazard function is the number of spells ending at time t_j divided by the number at risk of failure at time t_{j-}

$$\widehat{\lambda}_j = \frac{n_j}{r_j}$$

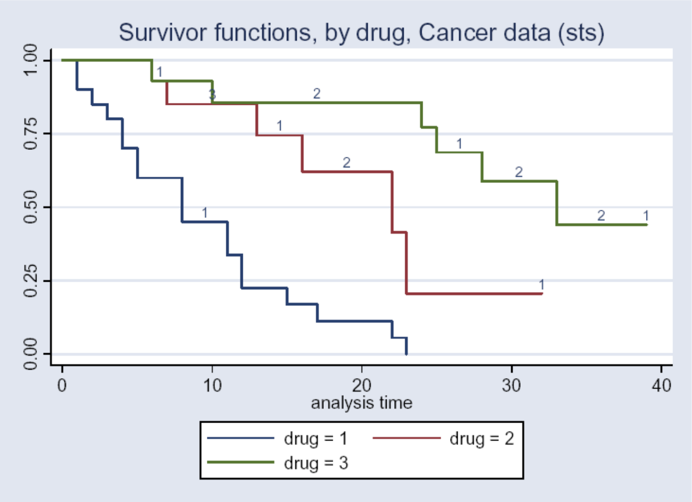
The Kaplan-Meier estimator of the survivor function is:

$$\widehat{S}(t) = \prod_{j|t_j \leq t} (1 - \widehat{\lambda}_j) = \prod_{j|t_j \leq t} \frac{r_j - n_j}{r_j}$$

Survivor function derived from the Cancer data using sts



Survivor function, stratified by drug (Cancer data), derived using sts



Cox Proportional Hazard Model

- Failure times are ordered and observations are categorized into those that are at risk at each failure time
- Define the risk set $R(t_j)$ to be the set of individuals who are at risk of failing just before the jth ordered failure
- Consider the probability of a particular at-risk spell ending at time t_j:

$$
\begin{aligned}
Pr[T_j = t_j | R(t_j)] &= \frac{Pr[T_j = t_j | T_j \geq t_j]}{\sum_{l \in R(t_j)} Pr[T_l = t_l | T_l \geq t_j]} \\
&= \frac{\lambda_j(t_j | \mathbf{x}_j, \beta)}{\sum_{l \in R(t_j)} \lambda_l(t_j | \mathbf{x}_l, \beta)} \\
&= \frac{\phi(\mathbf{x}_j, \beta)}{\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \beta)}
\end{aligned}
$$

Cox Proportional Hazard Model

To control for tied durations that are likely to occur when durations are grouped we need an adjustment. Breslow and Peto propose an approximation:

$$Pr[T_j = t_j | j \in R(t_j)] \simeq \frac{\prod_{m \in D(t_j)} \phi(\mathbf{x}_m, \beta)}{\left[\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \beta)\right]^{d_j}}$$

where $D(t_j)$ denotes the set of subjects that die at time t_j and d_j denotes the number that die at time t_j

Cox Proportional Hazard Model

Cox defined the **partial likelihood function** to be the joint product of $Pr[T_j = t_j | j \in R(t_j)]$ over the k ordered failure times. Then

$$L_p(\beta) = \prod_{j=1}^{k} \frac{\prod_{m \in D(t_j)} \phi(\mathbf{x}_m, \beta)}{\left[\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \beta)\right]^{d_j}}$$

Cox proposed estimation of β by minimizing the log partial likelihood function

$$\ln L_p = \sum_{j=1}^{k} \left[\sum_{m \in D(t_j)} \ln \phi(\mathbf{x}_m, \beta) - d_j \ln \left(\sum_{l \in R(t_j)} \phi(\mathbf{x}_l, \beta) \right) \right]$$

Cox Proportional Hazard Model

In the usual specification $\phi(\mathbf{x}, \beta) = \exp(\mathbf{x}'\beta)$. Then we can write:

$$L_p(\beta) = \prod_{j=1}^{k} \left[\frac{\exp(\mathbf{x}_j'\beta)}{\sum_{i \in R(t_j)} \exp(\mathbf{x}_i'\beta)} \right]$$

Consider a simple example with $t_1 < t_2 < t_3$. Then,

$$L_p(\beta) = \frac{h_0(t)\exp(\mathbf{x}_1'\beta)}{h_0(t)\exp(\mathbf{x}_1'\beta) + h_0(t)\exp(\mathbf{x}_2'\beta) + h_0(t)\exp(\mathbf{x}_3'\beta)} \frac{h_0(t)\exp(\mathbf{x}_2'\beta)}{h_0(t)\exp(\mathbf{x}_2'\beta) + h_0(t)\exp(\mathbf{x}_3'\beta)}$$

- Given the PH assumption the baseline hazard drops out and provides the basis for estimating β

Other Specification Problems

- Discrete time (grouped duration data)
- Multiple spells
- Multivariate Duration Models
- Specification Tests (Cox and Snell residuals)

Bibliography

Tony Lancaster, The Econometric Analysis of Transition Data, Cambridge University Press, 1990

John Kalbfleisch and Ross Prentice, The Statistical Analysis of Failure Time Data, Whiley, 1980

Terry Therneau and Patricia Grambsch, Modeling Survival Data, Springer, 2000

Philip Hougaard, Analysis of Multivariate Survival Data, Springer, 2000

Sem clepsidra ou relógio o tempo escorre
E nós com ele, nada o árbrito escravo
Pode contra o destino
Nem contra os deuses o desejo nosso
Hoje, quais servos com ausentes deuses
Na allheia casa, um dia sem juiz
Bebamos e comamos
Será amanhã o que aconteça

Tombai mancebos, o vinho em nobre taça
E o braço nu com que o entornais fique
No lembrando olhar
Uma estátua de homem apontando
Sim, heróis sê-lo-emos amanhã
Hoje adiemos . E na nossa taça
O roxo vinho transpareça
Depois - porque a noite nunca tarda

Poesia de Ricardo Reis