LABORATÓRIO NACIONAL
DE ENGENHARIA CIVIL

Phd Program in Transportation

Transport Demand Modeling

Carlos Roque
Postdoctoral Research Fellow at LNEC
(croque@lnec.pt)

Session 1

Hazard-Based Duration Models

Nonparametric, Semiparametric and Parametric models

INSTITUTO SUPERIOR TÉCNICO

FEUP

Outline of the Module on Hazard-Based Duration Models

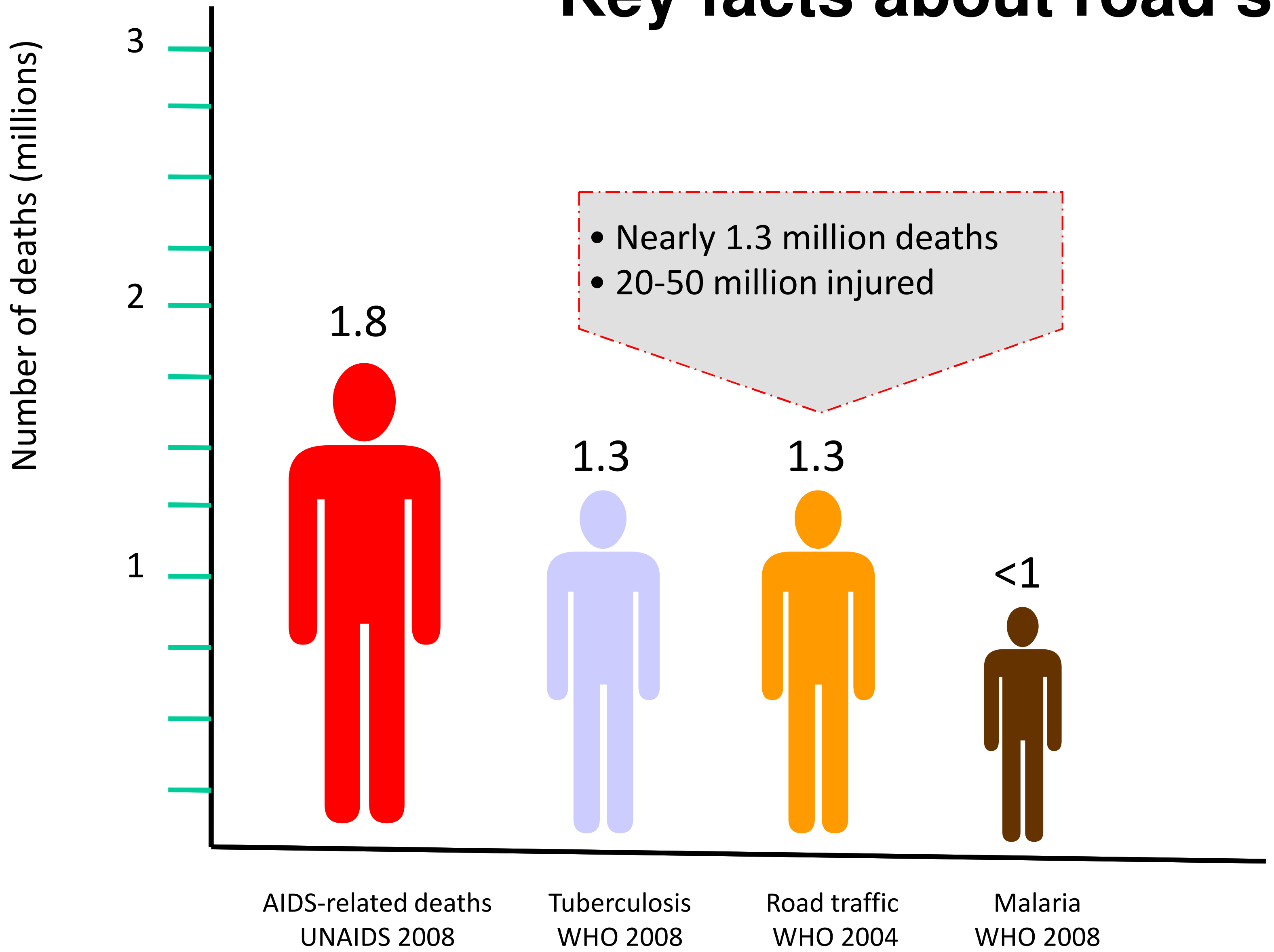
Our objectives for this session:

- ❑ Background information: statistical analysis of road safety data
- ❑ Characteristics of duration data
- ❑ Nonparametric models
- ❑ Semiparametric models
- ❑ Fully Parametric models
- ❑ Build your first Kaplan-Meier estimate (using R)
- ❑ Build your first Cox proportional-hazards model (using R)

How big is the road safety problem?

- Have you ever been injured in a crash?
- Have any of your family members or friends been injured or killed in a crash?
- Do you know someone who has been injured or killed in a crash?

Key facts about road safety

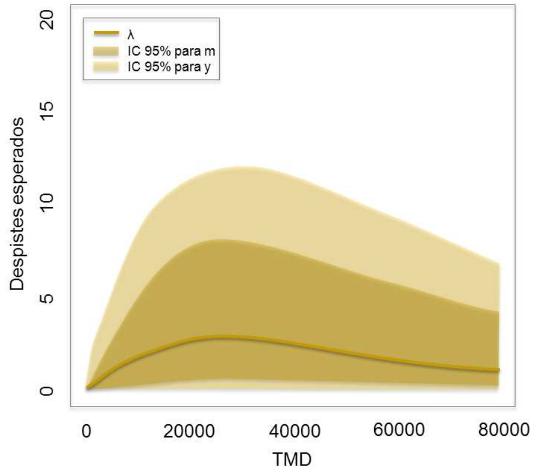


Source: http://www.who.int/roadsafety

Perspectives on road safety

Subjective safety	Normative safety	Objective safety

$$\lambda_i = \beta_0 TMD_i^{\beta_1} e^{\beta_2 \frac{TMD_i}{10000}} Extensão_i^{\beta_3}$$



Motorists' complaints
Experts' judgement

Standards compliance

Expected or actual crash frequency and severity

How safe is this road?

How do we evaluate alternatives?

Good news

- There is a lot of information on substantive safety

Sinalização Vertical

- Sinalização de Nós de Ligação
- Sinalização de Rotundas
- Sinalização de Cruzamentos e Entroncamentos
- Sinalização de Orientação - Sistema Informativo
- Instrução Técnica sobre a utilização da Sinalização de Mensagem Variável
- Sinalização Vertical - Características
- Princípios da Sinalização do Trânsito e Regimes de Circulação
- Sinalização Vertical - Critérios de Utilização
- Sinalização Vertical - Critérios de Colocação
- Destinos Principais e Pólos Não Classificados

Marcação Rodoviária

- Marcas Rodoviárias - Características Dimensionais, Critérios de Utilização e Colocação
- Marcas Rodoviárias - Dispositivos Retrorrefletores Complementares
- Sinalização de Proibição de Ultrapassagem

Projeto

- Norma de Traçado - Revisão
- Barreiras New-Jersey com Valeta Adjacente - Condições e Parâmetros de Segurança
- Dimensionamento de Rotundas - Documento síntese
- Auto-Estradas - Características Técnicas
- Medidas de Acalmia de Tráfego (Vol. 1) - Medidas Individuais Aplicadas-Atravessamentos de Localidades
- Medidas de Acalmia de Tráfego (Vol. 2) - Critérios para Definição dos Trechos de Intervenção
- Medidas de Acalmia de Tráfego (Vol. 3) - Tratamento das Zonas de Aproximação e Transição
- Medidas de Acalmia de Tráfego (Vol. 4) - Tratamento do Trecho Urbano em Atravessamentos de Localidade
- Medidas de Acalmia de Tráfego (Vol. 5) - Processo de Implementação e Monitorização das Intervenções

Pavimentação

- Construção e Reabilitação de Pavimentos - Agregados
- Construção e Reabilitação de Pavimentos - Indicadores de Estado de Conservação dos Pavimentos
- Construção e Reabilitação de Pavimentos - Reciclagem de Pavimentos
- Directivas para a concepção de pavimentos - Critérios de dimensionamento
- Construção e Reabilitação de Pavimentos - Ligantes Betuminosos

Manuais

- Recomendações para definição e sinalização de limites de velocidade máxima
- Área Adjacente à Faixa de Rodagem – Manual sobre Aspectos de Segurança
- Sistemas de Retenção Rodoviários - Manual de Aplicação
- Inspecções de Segurança Rodoviária - Manual de Aplicação

Guias de Procedimentos

Apresentação dos Projectos das Condições de Execução das Obras	Metodologia a seguir por Concessionárias em processos PCEO – Projectos das Condições de Execução das Obras, para intervenções com duração superior a 72 horas, previstos na Lei n.º 24/2007, de 18 de Julho e no seu D.R. n.º 12/2008, de 9 de Junho.
Auditorias de Segurança Rodoviária aos Projectos de Infra-estruturas Rodoviárias	Orientações técnicas sobre a realização de Auditorias de Segurança Rodoviária a projectos de infra-estruturas, definir o seu âmbito de aplicação e a forma como devem ser promovidas pelas entidades gestoras das vias.
Colocação de Sinalização Turístico-Cultural / Património em Auto-Estradas	Metodologia a seguir para instalação, na rede rodoviária nacional, de sinalização turístico-cultural.

Notas Técnicas

- Levantamento das Características dos Agregados produzidos em Portugal
- Ensaios Comparação Interlaboratorial Avaliação Sensibilidade à Água Misturas Betuminosas Compactadas
- InIR Apoia Instituto Sueco de Investigação das Estradas em Estudo sobre Acalmia de Tráfego
- Inquérito a utentes de estradas europeias - 2006
- Guia para as administrações rodoviárias intervenientes no processo de normalização
- Abordagem Integrada à Segurança de Túneis Rodoviários
- Integração dos Indicadores de Desempenho

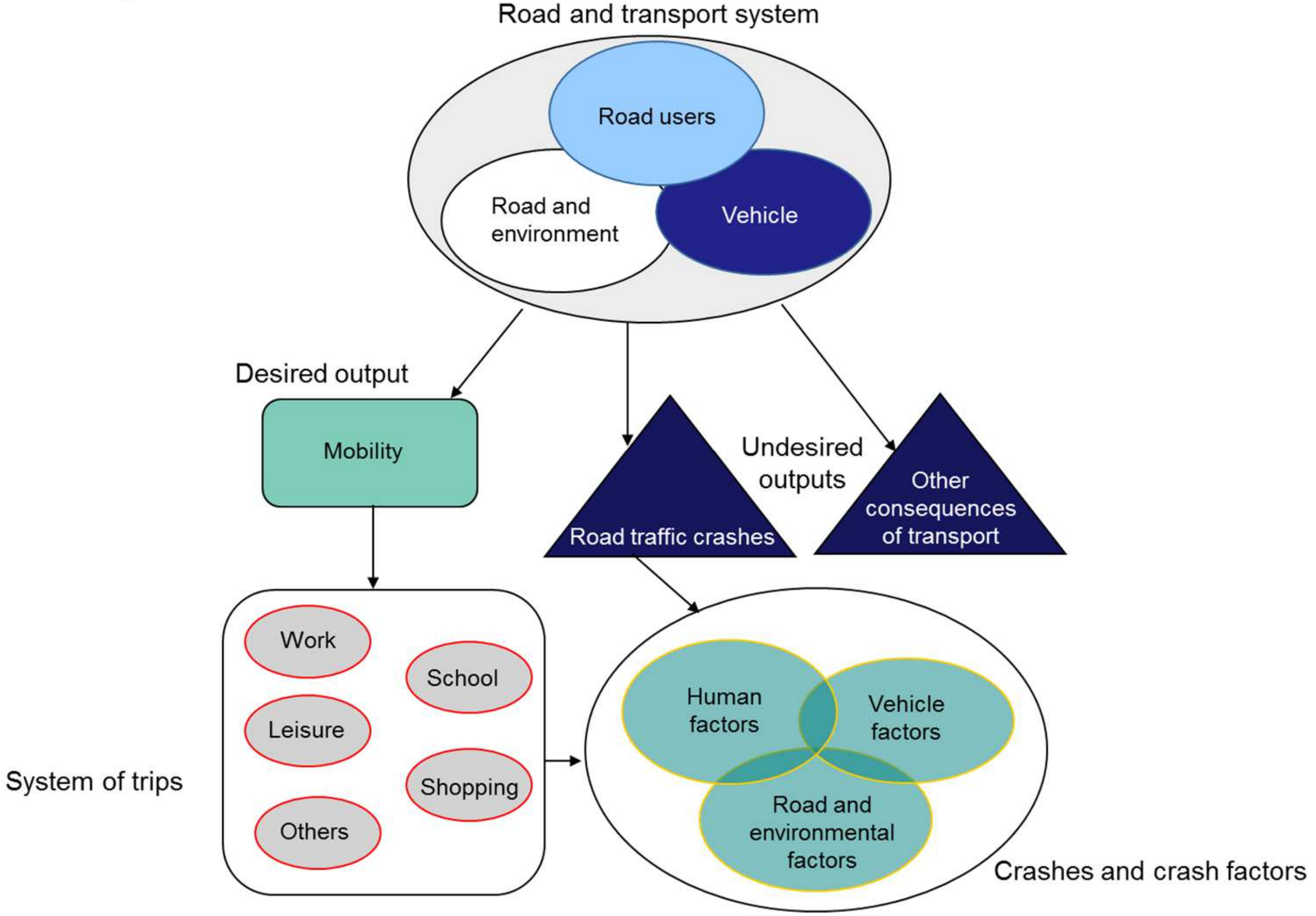
http://www.imt-ip.pt/sites/IMTT/Portugues/InfraestruturasRodoviarias/InovacaoNormalizacao/Paginas/DivulgacaoTecnica.aspx

Bad news

- Many study results are problematic
  - Poor study design & analysis
  - Highly variable results
  - Limited reproduction of results
  - Most sources are regarding normative safety

System approach

- Understand the system as a whole.
  - Understand interactions between different components.
  - Consider not only underlying factors, but also role of different agencies and actors in prevention efforts.



Major risk factors

- Factors influencing exposure to risk
  - economic factors
  - demographic factors
  - land-use planning practices
  - traffic mix
  - road function versus design and layout
- Risk factors influencing crash involvement
  - speed
  - alcohol or drugs
  - fatigue
  - gender
  - vehicle defects
  - age
  - vulnerable road users

Major risk factors

- Risk factors influencing crash severity
  - speed
  - seat-belts, child restraints
  - helmets
  - Non-crash protective roadside objects
  - insufficient vehicle crash protection
  - alcohol and other drugs
- Risk factors influencing post-crash outcome of injuries
  - delay in detecting crash
  - delay in transport to a health facility
  - fire resulting from collision
  - leakage of hazardous materials
  - alcohol and other drugs
  - rescue, extraction, evacuation
  - poor trauma care and rehabilitation

Measuring objective safety

- Why Analyze?
  - Identify crash-prone locations
  - Hoping that data analysis will suggest effective countermeasures
  - Evaluate the effectiveness of an implemented countermeasure
  - …

- Traditional Analysis Approaches:
  - Models of crash frequency over some specified time and space
  - Models of crash-injury severity (which is conditional the crash having occurred)
  - Some modeling approaches have combined the two (frequency and severity)

Crash Data Modeling

- Crash Frequency Models
  - Study crash frequency over some specified time and space
  - Various count-data and other methods have been used
  - Explanatory variables:
    - Traffic conditions
    - Roadway conditions
    - Weather conditions

- Crash Severity Models
  - Study injury severities of specific crashes
  - Various discrete-outcome and other methods have been used
  - Explanatory variables:
    - Traffic Conditions, Roadway conditions, Weather conditions
    - Specific crash data: Vehicle information, Occupant information, Crash specific characteristics

Data

- Traditional Crash Data
  - Available mostly from police and possibly other reports
  - Provide basic data on the characteristics of the crash
    - Road conditions
    - Estimates of injury severity
    - Occupant characteristics (age, gender)
    - Vehicle characteristics
    - Crash description, primary cause, etc.

- Emerging Data Sources
  - Data from driving simulators
  - Data from naturalistic driving
  - Data from automated vehicles
  - Data from other sources

Emerging data sources

- Naturalistic Driving Data
  - Extensively instrumented conventionally operated vehicles
- Simulator Data
  - Massive amounts of data collected from driving simulators
- Automated Vehicle Data
  - Including automated vehicle performance and response of drivers of conventional vehicles
- Others

Methodological Barriers

- Unobserved Heterogeneity
  - Many factors influencing the frequency and severity of crashes are simply not observed
- Endogeneity
  - Factors correlated with frequency and severity of crashes
- Temporal Correlation
  - Crashes in occurring near the same or similar times will share correlation due to unobserved factors associated with time (precise weather conditions, similar sun angle, etc.)
- Spatial Correlation
  - Crashes in close spatial proximity will share correlation due to unobserved factors associated with space (unobserved visual distractions, sight obstructions, etc.)
- Omitted Variables
  - Many crash frequency models use few explanatory variables (some only use traffic)

Duration Models

- In many instances, one encounters the need to study the elapsed time until the occurrence of an event or the duration of an event. Data such as these are referred to as duration data, and are encountered often in the field of transportation research.
  - Examples include the **time** until a vehicle accident occurs, the time between vehicle purchases, the time devoted to an activity (shopping, recreational, etc.), the time until the adoption of new transportation technologies, or the **distance** traveled until a vehicle stops.
- To study duration data, hazard-based models are applied to study the conditional probability of a time duration ending at some time *t*, given that the duration has continued until time *t*.
  - Hazard-based duration models can account for the possibility that the likelihood of a driver becoming involved in an accident may change over time.

- **Cumulative distribution function** *F(t)*:

$$F(t) = P(T < t)$$

  - where *P* denotes probability, *T* is a random time variable, and *t* is some specified time.

- The **density function** corresponding to this distribution function (the first derivative of the cumulative distribution with respect to time) is:

$$f(t) = \frac{dF(t)}{dt}$$

- And the **hazard function** is:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

  - where $h(t)$ is the conditional probability that an event will occur between time t and $t + dt$, given that the event has not occurred up to time t.

$$h(t) = \lim_{\delta \to 0} \frac{pr(t < T < t + \delta | T > t)}{\delta}$$

- The **cumulative hazard** H(t) is the integrated hazard function, and provides the cumulative rate at which events are ending up to or before time t.

$$H(t) = \int_0^t h(t)\, dt$$

Duration Models

- The **survivor function** (the probability that a duration is greater than or equal to some specified time *t*) is:

$$S(t) = P(T \geq t)$$

- If one of these functions is known any of the others are readily obtained.

$$S(t) = 1 - F(t) = 1 - \int_0^t f(t)dt = EXP\left[-H(t)\right]$$

$$f(t) = \frac{d}{dt}F(t) = h(t)EXP\left[-H(t)\right] = -\frac{d}{dt}S(t)$$

$$H(t) = \int_0^t h(t)\,dt = -LN\left[S(t)\right]$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = \frac{d}{dt}H(t)$$

Duration Models

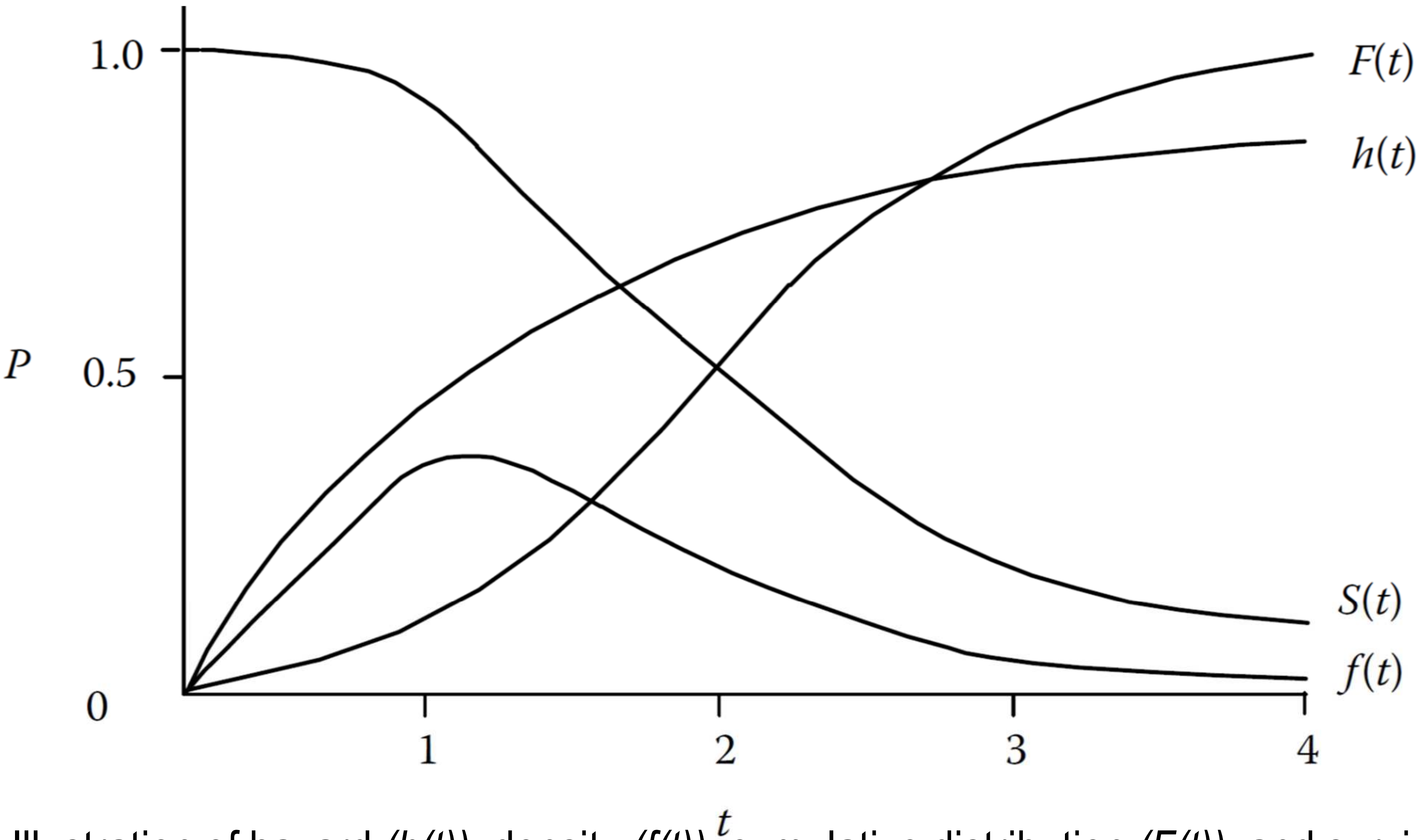


Illustration of hazard *(h(t))*, density *(f(t))*, cumulative distribution *(F(t))*, and survivor functions *(S(t))*.

Source: Washington *et al.* (2011)

Duration Models

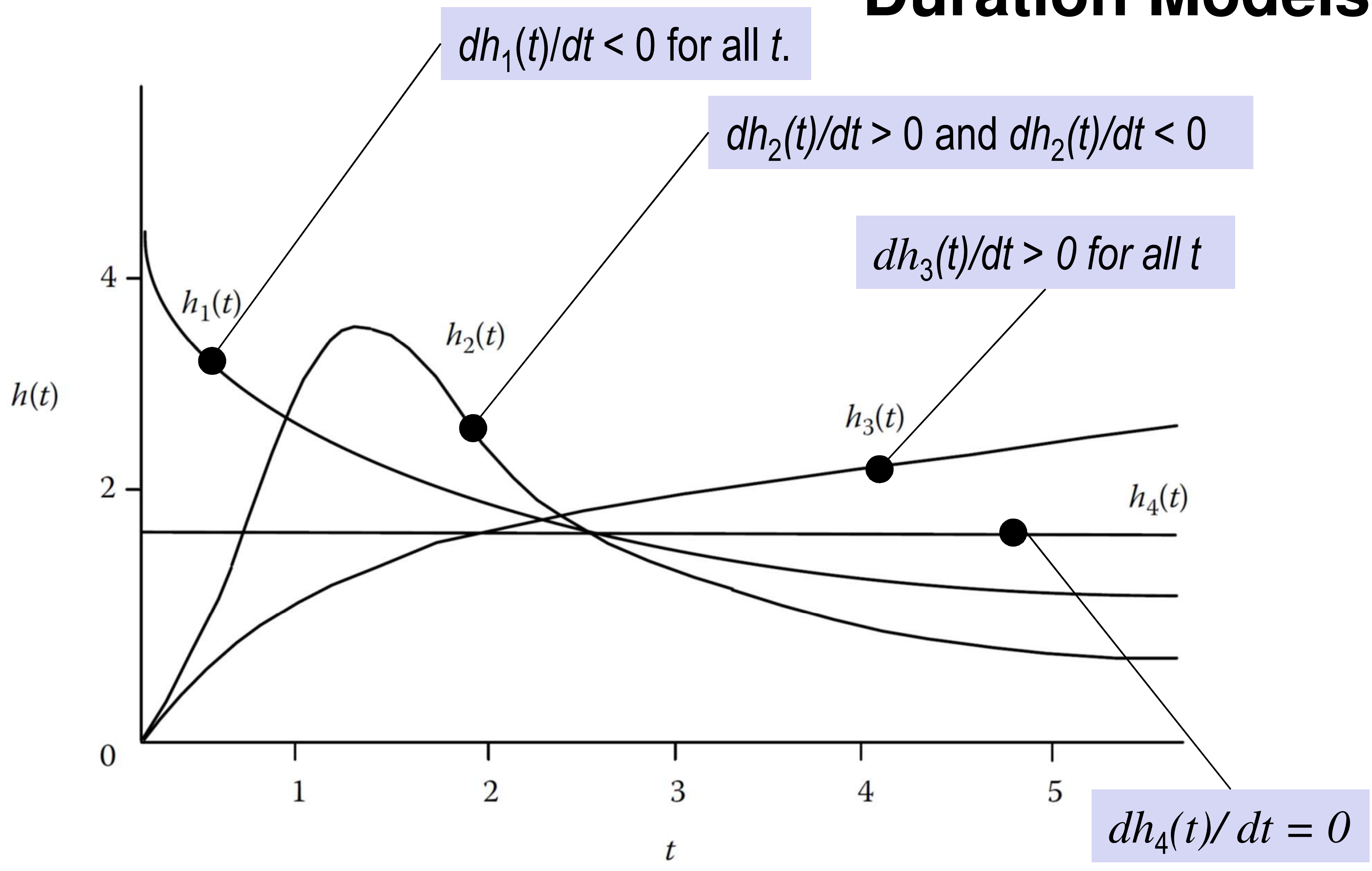


Illustration of four alternate hazard functions.

Source: Washington *et al.* (2011)

Duration Models

- In addition to duration dependence, hazard-based duration models account for the effect of covariates on probabilities.
  - Proportional hazards models
  - Accelerated lifetime models
- The **proportional-hazards approach** assumes that the covariates, which are factors that affect the probability that an event will occur, act multiplicatively on some underlying hazard function.

$$h(t\,|\,\mathbf{X}) = h_{o}(t)EXP(\boldsymbol{\beta}\mathbf{X})$$

  - where $h_o(t)$ denotes the underlying (or baseline) hazard function, **X** is the covariate vector and $\boldsymbol{\beta}$ is a vector of estimable parameters
  - $h(t)$ is separable into $h_o(t)$ and the effects of Xs
  - $h_o(t)$ depends on t but not on individual characteristics
  - Absolute differences in X → proportional differences in $h(t)$ ~scaling of $h_o(t)$

Duration Models

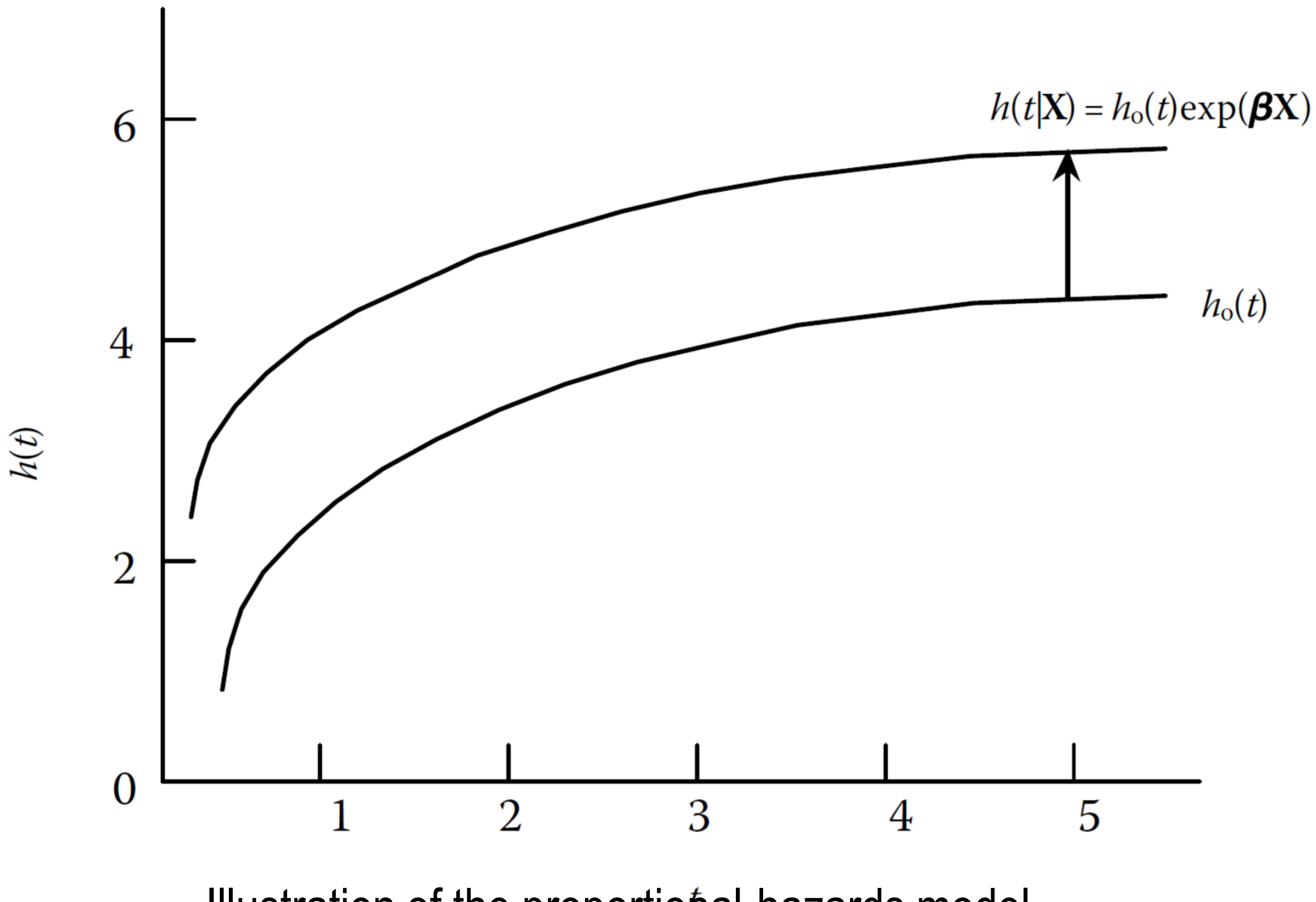


Illustration of the proportional-hazards model.

Source: Washington *et al.* (2011)

Duration Models

- The **accelerated lifetime method** assumes that the covariates rescale (accelerate) time directly in a baseline survivor function. This accelerated lifetime method again assumes covariates influence the process with the function *EXP*($\boldsymbol{\beta}$**X**). The **accelerated lifetime model** is written as

$$S(t|\mathbf{X}) = S_{o}[t\ EXP(\boldsymbol{\beta}\mathbf{X})]$$

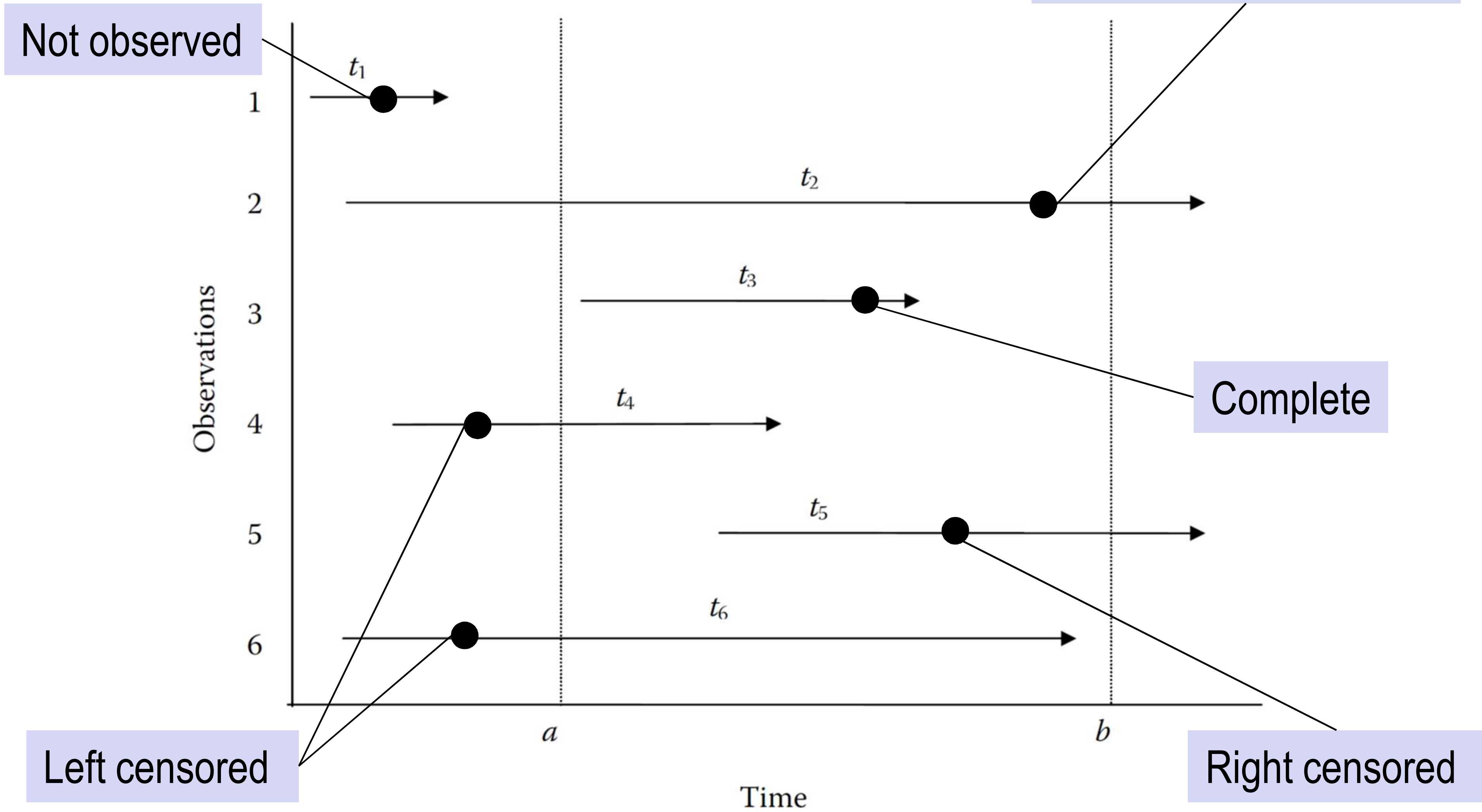
- which leads to the conditional hazard function

$$h(t|\mathbf{X}) = h_{o}[t\ EXP(\boldsymbol{\beta}\mathbf{X})]EXP(\boldsymbol{\beta}\mathbf{X})$$

  - where $h_o(t)$ denotes the underlying (or baseline) hazard function, t is the time, **X** is the covariate vector and $\boldsymbol{\beta}$ is a vector of estimable parameters

Characteristics of Duration Data

- Duration data are often left or right **censored**.



Source: Washington *et al.* (2011)

Characteristics of Duration Data

- Hazard-based models can readily account for right-censored data.
- Left-censored data creates a far more difficult problem because of the additional complexity added to the likelihood function.
- Another challenge may arise when a number of observations end their durations at the same time. This is referred to as the problem of tied data. Tied data can arise when data collection is not precise enough to identify exact duration-ending times. When duration exits are grouped at specific times, the likelihood function for proportional and accelerated lifetime models becomes increasingly complex.

INSTITUTO SUPERIOR TÉCNICO

FEUP

Nonparametric Models

- The **Kaplan-Meier method** (based on individual survival times) is the most widely applied nonparametric method in survival analysis
- The basic method for calculating survival probabilities using the Kaplan-Meier method begins by specifying the probability of surviving *r* years (without event *A* occurring) as the conditional probability of surviving *r* years given survival for *r*–1 years times the probability of surviving *r*–1 years (or months, days, minutes, etc.). In notation, the probability of surviving *k* or more years is given by

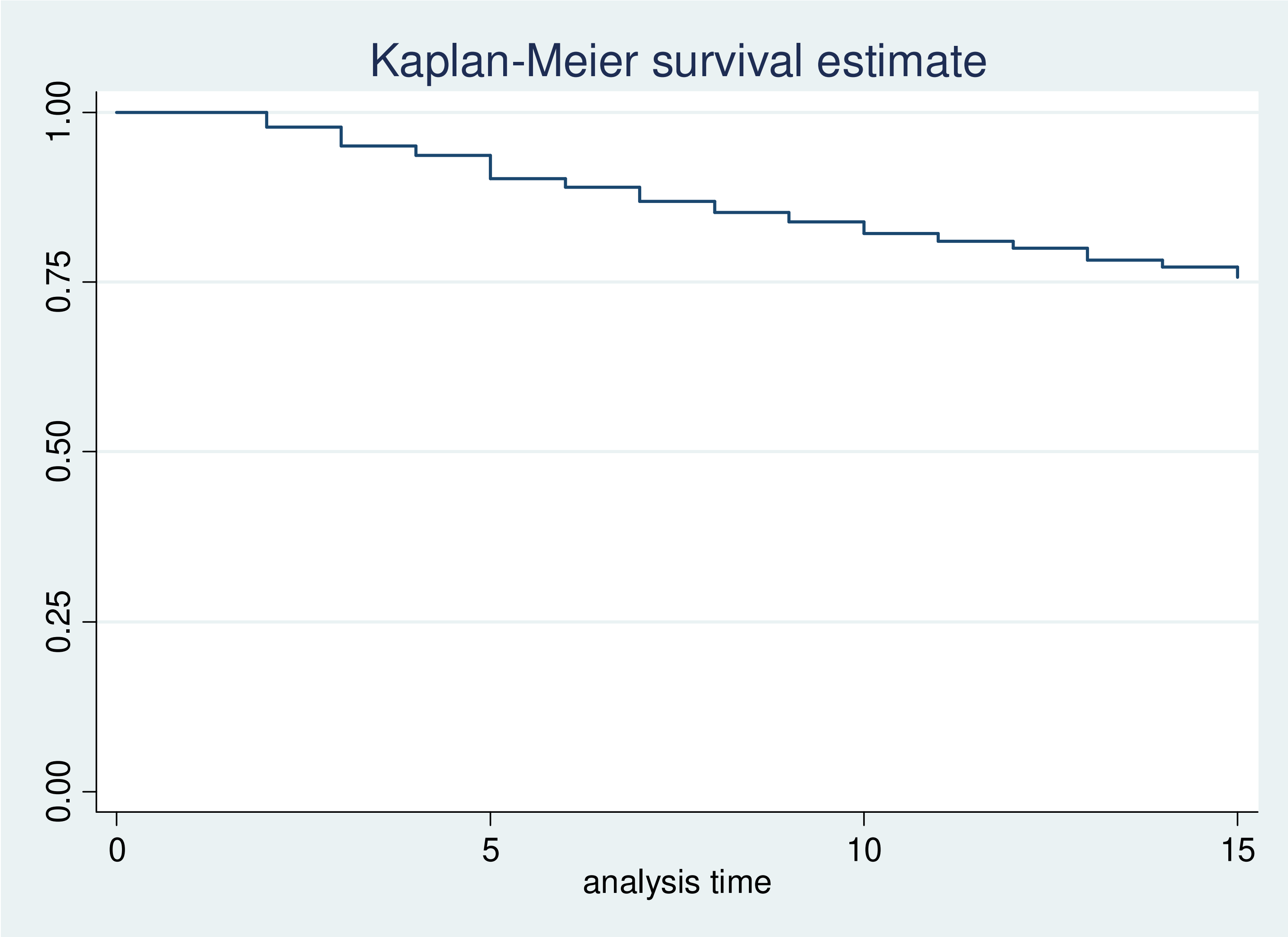
$$\hat{S}(k) = (p_k \mid p_{k-1}) \cdots (p_4 \mid p_2)(p_3 \mid p_2)(p_2 \mid p_1)(p_1)$$

  - where $(p_k \mid p_{k-1})$ is the proportion of observed subjects surviving to period *k*, given survival to period *k* – 1, and so on.

Nonparametric Models

- The Kaplan–Meier method provides useful estimates of survival probabilities and a graphical presentation of the survival distribution.
- It is the most widely applied nonparametric method in survival analysis.
- A few observations:
  - If the largest (survival) observation is right-censored, the Kaplan–Meier estimate is undefined beyond this observation.
  - If the largest observation is not right censored, then the Kaplan–Meier estimate at that time equals zero.
  - The median survival time cannot be estimated if more than 50% of the observations are censored and the largest observation is censored.
  - The Kaplan–Meier method assumes that censoring is independent of survival times. If this is false, the Kaplan–Meier method is inappropriate.

Nonparametric Models



Semiparametric Models

- Semiparametric models do not assume a distribution of duration times (like Weibull or exponential), although they do have a parametric assumption on the functional form of the covariates' influence on the hazard function (usually *EXP*(**β X**)).
- The **Cox proportional-hazards model** is semiparametric because *EXP*(**β X**) is used as the functional form of the covariate influence.
- Produces estimated hazard ratios (sometimes called rate ratios or risk ratios)
- Regression coefficients are on a log scale
  - Exponentiate to get hazard ratio

Semiparametric Models

- **Cox proportional-hazards model**

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \dots\dots + \beta_n x_{in})$$

- $h_i(t)$ is the hazard function for individual i
- $h_o(t)$ is the baseline hazard function and can take any form
- $X_{i1}, X_{i2} \dots X_{in}$ are the covariates
- $\beta_{i1}, \beta_{i2} \dots \beta_{in}$ are the regression coefficients estimated from the data
- PH assumption needed
- Estimate βs without estimating $h_0(t)$ → semi parametric model

Semiparametric Models

Cox proportional-hazards model

- If we divide both sides of the equation on the previous slide by $h_0(t)$ and take logarithms, we obtain:

$$\ln\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_n x_{in}$$

- We call $h_i(t)/h_0(t)$ the hazard ratio
- The coefficients β_{i1}, $\beta_{i2}\ldots\beta_{in}$ are estimated by Cox regression, and can be interpreted in a similar manner to that of multiple logistic regression
- $\exp(\beta_i)$ is the instantaneous relative risk of an event

Semiparametric Models

- This model is readily estimated using standard maximum likelihood methods.
  - If only one observation completes its duration at each time (no tied data), and no observations are censored, the partial log-likelihood is

$$LL = \sum_{i=1}^{I}\left[\boldsymbol{\beta}\mathbf{X}_i - \sum_{j \in R_i} EXP(\boldsymbol{\beta}\mathbf{X}_j)\right]$$

  - If no observations are censored and tied data are present with more than one observation exiting at time t_i, the partial log-likelihood is the sum of individual likelihoods of the n_i observations that exit at time t_i

$$LL = \sum_{i=1}^{I}\left[\boldsymbol{\beta}\sum_{j \in t_i}\mathbf{X}_j - ni\sum_{j \in R_i} EXP(\boldsymbol{\beta}\mathbf{X}_j)\right]$$

Semiparametric Models

- Cox regression assumptions
  - Assumption of proportional hazards
  - No censoring patterns
  - True starting time
  - Plus assumptions for all modelling
    - Sufficient sample size, proper model specification, independent observations, exogenous covariates, no high multicollinearity, random sampling, and so on.

Fully Parametric Models

- With fully parametric models, a variety of distributional alternatives for the hazard function have been used with regularity in the literature. These include gamma, exponential, Weibull, log-logistic, and Gompertz distributions, among others.
- The choice of any one of these alternatives is justified on theoretical grounds or statistical evaluation.
- The choice of a specific distribution has important implications relating not only to the shape of the underlying hazard, but also to the efficiency and potential biasedness of the estimated parameters.

Fully Parametric Models

Name	Hazard Function $h(t)$
Compound exponential	$h(t) = \frac{P}{t + (P/\lambda_0)}$
Exponential	$h(t) = \lambda$
Exponential with gamma heterogeneity	$h(t) = \frac{\lambda}{1 + \theta\lambda t}$
Gompertz	$h(t) = (P)EXP^{\lambda t}$
Gompertz–Makeham	$h(t) = \lambda_0 + \lambda_1 EXP^{\lambda_2 t}$
Log-logistic	$h(t) = \frac{(\lambda P)(\lambda t)^{P-1}}{1 + (\lambda t)^P}$
Weibull	$h(t) = (\lambda P)(\lambda t)^{P-1}$
Weibull with gamma heterogeneity	$h(t) = \frac{(\lambda P)(\lambda t)^{P-1}}{1 + \theta(\lambda t)^P}$

Some Commonly used Hazard Functions for Parametric Duration Models

Source: Washington *et al.* (2011)

Fully Parametric Models

❑ Exponential distribution

- With parameter $\lambda > 0$, the exponential density function is

$$f(t) = \lambda EXP(-\lambda t)$$

- with hazard,

$$h(t) = \lambda$$

- The equation above implies that this distribution's hazard is constant, (as illustrated by h_4(t)) .
- This means that the probability of a duration ending is independent of time and there is no duration dependence.

Fully Parametric Models

Weibull distribution

- With parameters λ > 0 and P > 0, the Weibull density function is

$$f(t) = \lambda P(\lambda t)^{P-1} EXP[-(\lambda t)^{P}]$$

- with hazard,

$$h(t) = (\lambda P)(\lambda t)^{P-1}$$

- As indicated in the equation above, if the Weibull parameter *P* is greater than one, the hazard is monotone increasing in duration (see $h_3(t)$);
- If *P* is less than one, it is monotone decreasing in duration (see $h_1(t)$)
- If *P* equals one, the hazard is constant in duration and reduces to the exponential distribution's hazard with $h(t) = \lambda$ (see $h_4(t)$).

Fully Parametric Models

Log-logistic distribution

- With parameters λ > 0 and P > 0, the log-logistic density function is

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

- with hazard,

$$h(t) = \frac{(\lambda P)(\lambda t)^{P-1}}{1+(\lambda t)^{P}}$$

- Equation above indicates that if $P < 1$, then the hazard is monotone decreasing in duration (see $h_1(t)$)
- If $P = 1$, then the hazard is monotone decreasing in duration from parameter λ;
- If $P > 1$, then the hazard increases in duration from zero to an inflection point, $t_i = (P - 1)^{1/P}/\lambda$, and decreases toward zero thereafter (see $h_2(t)$).

Example: Roadside safety analysis

- Roque, C. and Jalayer, M. 2018. Improving roadside design policies for safety enhancement using hazard-based duration modeling, Accident Analysis & Prevention, Volume 120, 2018, Pages 165-173.
  - The distance traveled by an errant vehicle in a ROR crash was modeled .
  - Two Cox mixed-effects regression models were developed.
  - Results confirmed the contribution of roadside obstacles to the distance travelled.
  - Results suggest that clear-zone distances proposed in guidelines should be evaluated.
  - This study can facilitate the appropriate planning and design of forgiving roadsides.

Example: Roadside safety analysis

Type of crash	Variable	Description	Mean (Std. Dev.)	Minimum	Maximum
Overturns	Distance traveled (ft)		60.008 (79.062)	0	999
	Roadway Variables				
	Speed limit (mph)	Speed limit at the location of the crash (mph)	52.598 (8.149)	20	70
Fixed-object crashes	Distance traveled (ft)		63.366 (98.021)	0	1421
	Roadway Variables				
	Speed limit (mph)	Speed limit at the location of the crash (mph)	53.631 (9.048)	20	70
	AADT (vpd)	Average Annual Daily Traffic	15333.880 (27383.540)	50	183000
	Shoulder Width (ft)	Paved shoulder width (Right)	6.496 (3.562)	0	22

Descriptive statistics of the continuous variables.

Source: Roque and Jalayer (2018)

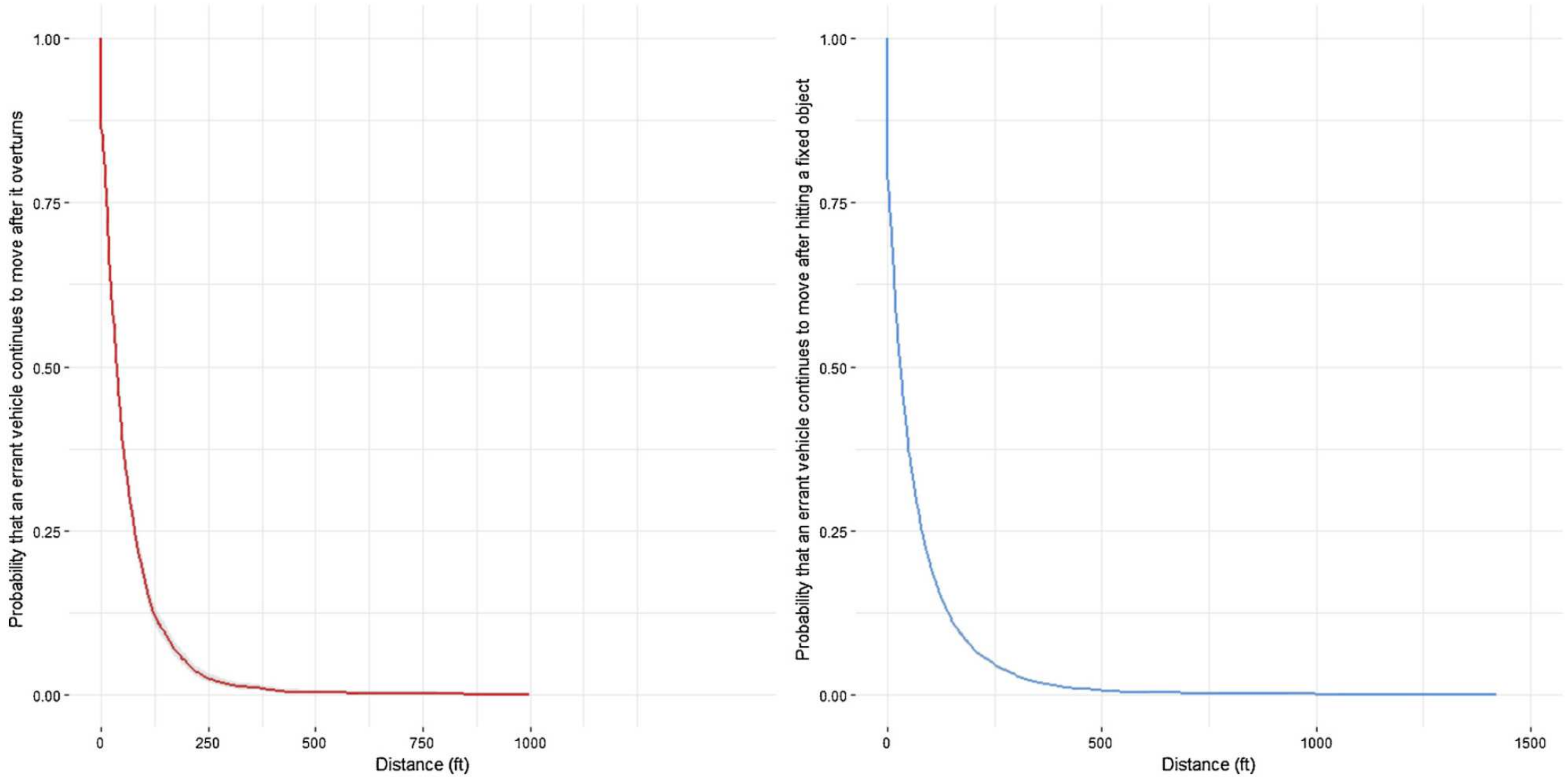
Example: Roadside safety analysis

Type of crash	Variable	Description	Percentage	Frequency
Overturns	**Seasonal Variables**			
	Clear weather	1 = if the crash occurred with clear weather conditions / 0 = otherwise	70.2% / 29.8%	1411 / 598
	Daylight	1 = if the crash occurred during daylight / 0 = otherwise	63.3% / 36.7%	1271 / 738
	Wet	1 = if the road surface was wet when the crash occurred / 0 = otherwise	13.5% / 86.5%	272 / 1737
	Vehicle Information			
	Airbag deploy	1 = if the vehicle's airbag was deployed when the crash occurred / 0 = otherwise	56.9% / 43.1%	1143 / 866
	Driver Characteristics			
	Normal condition	1 = if the physical condition of the driver when the crash occurred was apparently normal / 0 = otherwise	81.6% /18.4%	1640 / 369
	Driver PDO	1 = if no injury for the driver / 0 = otherwise	39.9% / 60.1%	801 / 1208
	Male	1 = if male driver / 0 = otherwise	73.4% / 26.6%	1474 / 535
Fixed-object crashes	**Seasonal Variables**			
	Clear weather	1 = if the crash occurred with clear weather conditions / 0 = otherwise	57.6% / 42.4%	10049 / 7408
	Roadway Variables			
	Rural	1 = if the crash occurred in a rural road / 0 = otherwise	90.3% / 9.7%	15763 / 1694
	Two-way	1 = if the crash occurred in a two-way, not divided road / 0 = otherwise	71.5% / 28.5%	12483 / 4974
	Crash Variables			
	Tree	1 = if first harmful event is collision with tree / 0 = otherwise	0.8% / 99.2%	148 / 17309
	Non-breakaway pole	1 = if first harmful event is collision with luminaire pole non-breakaway / 0 = otherwise	0.2% / 99.8%	31 / 17426
	Breakaway pole	1 = if first harmful event is collision with luminaire pole breakaway / 0 = otherwise	0.1% / 99.9%	10 / 17447
	Sign non-breakaway	1 = if first harmful event is collision with sign non-breakaway / 0 = otherwise	1.2% / 98.8%	212 / 17245
	Guardrail	1 = if first harmful event is collision with guardrail face on shoulder / 0 = otherwise	0.7% / 99.3%	122 / 17335
	Bridge rail	1 = if first harmful event is collision with bridge rail face / 0 = otherwise	0.3% / 99.7%	57 / 17400
	Curb/Median	1 = if first harmful event is collision with traffic island curb or median / 0 = otherwise	0.3% / 99.7%	55 /17402
	Ditch	1 = if first harmful event is collision with ditch / 0 = otherwise	1.0% / 99.0%	167 / 17290
	Vehicle Information			
	Front of the vehicle	1 = if the point of contact of the vehicle was its central front/ 0 = otherwise	10.3% / 89.7%	1801 / 15656
	Airbag deploy	1 = if the vehicle's airbag was deployed when the crash occurred / 0 = otherwise	65.6% / 34.4%	11456 / 6001
	Driver Characteristics			
	Normal condition	1 = if the physical condition of the driver when the crash occurred was apparently normal / 0 = otherwise	77.7% /22.3%	13563 / 3894
	Driver PDO	1 = if no injury for the driver / 0 = otherwise	64.1% / 35.9%	11188 / 6269
	Ejection	1 = if occupant not ejected in the crash / 0 = otherwise	97.6% / 2.4%	17041 / 416
	Male	1 = if male driver / 0 = if female driver	61.1% / 38.9%	10673 / 6784

Descriptive statistics of the categorical variables.

Source: Roque and Jalayer (2018)

Example: Roadside safety analysis



Kaplan–Meier estimate of the distance traveled for overturns and fixed-object crashes.

Source: Roque and Jalayer (2018)

Example: Roadside safety analysis

- In the Cox proportional-hazards model, the hazard ratio (HR) is a measure of the relative importance of the explanatory variables concerning hazard, while controlling for distance.
- The HR is often used to interpret results predicted by the Cox proportional-hazards model and can be obtained by the exponentiation of each regression coefficient.
- Specifically, the HR indicates the time rate of stopping at any distance during the study period, compared to that of the reference category.
  - If HR=1, then the explanatory variable in the model does not affect and does not change the baseline hazard, $h_0(\delta)$.
  - If HR < 1, then the time rate of stopping is decreased throughout the study period.
  - If HR > 1, the time rate of stopping is increased throughout the referred period

Example: Roadside safety analysis

Variable	Overturns			Fixed-object crashes		
	Coefficient estimate	p-value	Hazard ratio	Coefficient estimate	p-value	Hazard ratio
Clear weather	−0.148	0.031	0.862	−0.165	< 0.001	0.848
Daylight	0.195	< 0.001	1.215	–	–	–
Wet	0.146	0.110	1.157	–	–	–
Rural	–	–	–	−0.252	< 0.001	0.777
Two-way	–	–	–	0.179	< 0.001	1.196
Speed limit	−0.013	< 0.001	0.987	−0.017	< 0.001	0.983
AADT(/10000)	–	–	–	0.029	< 0.001	1.029
Shoulder width	–	–	–	−0.015	< 0.001	0.985
Tree	–	–	–	0.698	< 0.001	2.009
Non-breakaway pole	–	–	–	0.927	0.001	2.527
Breakaway pole	–	–	–	−1.589	0.001	0.204
Sign non-breakaway	–	–	–	0.427	< 0.001	1.532
Guardrail	–	–	–	0.437	< 0.001	1.548
Bridge rail	–	–	–	0.450	0.010	1.568
Curb/median	–	–	–	−0.530	0.003	0.588
Ditch	–	–	–	0.361	0.001	1.414
Front of the vehicle	–	–	–	0.291	< 0.001	1.338
Airbag deploy	0.104	0.058	1.110	0.146	< 0.001	1.158
Normal condition	0.226	0.001	1.254	0.387	< 0.001	1.472
Driver PDO	0.510	< 0.001	1.666	0.272	< 0.001	1.312
Ejection	–	–	–	0.201	0.002	1.222
Male	−0.087	0.150	0.917	−0.117	< 0.001	0.890
Variance of log-normal random effects	0.215	< 0.001		0.482	< 0.001	
Likelihood ratio test statistics	177.4			1656.9		
Sample size	2009			17545		

Cox mixed-effects model estimation results of distance traveled by an errant vehicle.

Source: Roque and Jalayer (2018)

Example: Roadside safety analysis

- The hazard ratio is the ratio of the hazard for a unit change in the covariate
  - HR = 1.157 for wet road surface vs. other conditions (overturns model)
  - This indicates that there is a 16% increase in the risk associated with stopping after adjusting for the other explanatory variables in the model, resulting in a decrease in the expected distance traveled.
- Hazard ratio assumed constant over time
  - At any time point, the stopping hazard for wet road surface is 1.157 times the hazard for other road surface conditions

Exercise 1: work-to-home departure delay

- A survey of 204 Seattle-area commuters was conducted to examine the duration of time that commuters delay their work-to-home trips in an effort to avoid peak period traffic congestion. Of the 204 commuters surveyed, 96 indicated that they sometimes delayed their work-to-home trip to avoid traffic congestion. These commuters provided their average time delay—thus each commuter has a completed delay duration so that neither left nor right censoring is present in the data.
  - Plot the Kaplan-Meier estimate of the duration of time that commuters delay their work-to-home trips
  - Determine the significant factors that affect the duration of commuters' delay using a Cox model.
  - Examine the work-to-home departure delay using exponential, Weibull, and log-logistic proportional-hazards models.

Exercise 1: work-to-home departure delay

Variable No.	Variable Description
1	Minutes delayed to avoid congestion
2	Primary activity performed while delaying: 1 if perform additional work, 2 if engage in nonwork activities, or 3 if do both
3	Number of times delayed in the past week to avoid congestion
4	Mode of transportation used on work-to-home commute: 1 if by single occupancy vehicle, 2 if by carpool, 3 if by vanpool, 4 if by bus, 5 if by other
5	Primary route to work in Seattle area: 1 if Interstate 90, 2 if Interstate 5, 3 if State Route 520, 4 if Interstate 405, 5 if other
6	In the respondent's opinion, is the home-to-work trip traffic congested: 1 if yes, 0 if no
7	Commuter age in years: 1 if under 25, 2 if 26–30, 3 if 31–35, 4 if 36–40, 5 if 41–45, 6 if 46–50, 7 if over 50
8	Respondent's gender 1 if female, 0 if male
9	Number of cars in household
10	Number of children in household
11	Annual household income (US dollars per year): 1 if less than 20,000, 2 if 20,000–29,999, 3 if 30,000–39,999, 4 if 40,000–49,999, 5 if 50,000–59,999, 6 if over 60,000
12	Respondent has flexible work hours? 1 if yes, 0 if no
13	Distance from work to home (in kilometers)
14	Respondent faces level of service D or worse on work-to-home commute? 1 if yes, 0 if no
15	Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time
16	Population of work zone
17	Retail employment in work zone
18	Service employment in work zone
19	Size of work zone (in hectares)

Source: Washington *et al.* (2011)

Exercise 1: work-to-home departure delay

- **Install and load packages**

```
install.packages("survival")
install.packages(“coxme")
install.packages("survminer")
library(survival)
library(coxme)
library(survminer)
```

- **Read and attach data**

```
data.delay <-
read.table(file="C:\\Users\\Carlos\\OneDrive\\Cursos\\Exercise.txt",head
er=T)
attach(data.delay)
head(data.delay,5)
```

Exercise 1: work-to-home departure delay

- **Renaming variables**

```
data.delay["minutes"] <- NA
data.delay$minutes <-  data.delay$X1
data.delay["number_of_times"] <- NA
data.delay$number_of_times <-  data.delay$X3
```

- **Sort the data by time**

```
data.delay <- data.delay[order(data.delay$minutes),]
print(data.delay)
```

- **Create graph**

```
with(data.delay, plot(minutes, type="h"))
```

Exercise 1: work-to-home departure delay

- **Create the life table survival object for data.delay**

```
# The functions survfit() and Surv() create a life table survival object.
data.delay2 <-subset(data.delay, minutes>0)
data.delay.survfit = survfit(Surv(minutes) ~ 1, data= data.delay2)
summary(data.delay.survfit)
```

- **Plot the Kaplan-Meier curve**

```
plot(data.delay.survfit, xlab = "Time (minutes)", ylab="Survival
probability", conf.int=TRUE)
ggsurvplot(data.delay.survfit, xlab = "Time (minutes)", xlim =
range(0:250) , conf.int = TRUE, color = "red", ggtheme =
theme_minimal())
```

Exercise 1: work-to-home departure delay

- **Cox Proportional Hazard Model Estimates of the Duration of Commuter Work-To-Home Delay to Avoid Congestion**

  result.cox <- coxph(Surv(minutes) ~ gender + rate_of_travel + distance + population, data= data.delay2)

  summary(result.cox)

Exercise 1: work-to-home departure delay

Testing proportional Hazards assumption

- Include an interaction between the covariate and a function of time (or distance). Log time often used but could be any function. If significant then assumption violated
- Test the proportional hazards assumption on the basis of partial residuals. Type of residual known as Schoenfeld residuals

```
test.ph <- cox.zph(result.cox)
```

- For each covariate, the function cox.zph() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole.

```
plot(test.ph)
ggcoxzph(test.ph)
```

- In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption.

Exercise 1: work-to-home departure delay

- **Plot the baseline survival function**

  ggsurvplot(survfit(result.cox), color = "#2E9FDF", ggtheme = theme_minimal())

- **Plot cumulative hazard function**

  ggsurvplot(survfit(result.cox), conf.int = TRUE, palette = c("#FF9E29", "#86AA00"), risk.table = TRUE, risk.table.col = "strata", fun = "event")

- **Log-likelihood**

  #Initial log-likelihood

  result.cox $loglik[1]

  #Final log-likelihood

  result.cox $loglik[2]

- **McFadden Pseudo-R2**

  Pseudo.R2 <- (1- (result.cox $loglik[2]/ result.cox $loglik[1]))

Exercise 1: work-to-home departure delay

- **Parametric Model Estimates of the Duration of Commuter Work-To-Home Delay to Avoid Congestion**

  # The argument dist has several options to describe the parametric model used ("weibull", "exponential", "gaussian", "logistic", "lognormal", or "loglogistic")

  result.expon <- survreg(Surv(minutes)~ gender + rate_of_travel + distance + population, data= data.delay2, dist="exponential")

  result.weib <- survreg(Surv(minutes)~ gender + rate_of_travel + distance + population, data= data.delay2, dist="weibull")

  result.loglog <- survreg(Surv(minutes)~ gender + rate_of_travel + distance + population, data= data.delay2, dist="loglogistic")

Exercise 2

- What else can we do with this dataset?

Variable No.	Variable Description
1	Minutes delayed to avoid congestion
2	Primary activity performed while delaying: 1 if perform additional work, 2 if engage in nonwork activities, or 3 if do both
3	Number of times delayed in the past week to avoid congestion
4	Mode of transportation used on work-to-home commute: 1 if by single occupancy vehicle, 2 if by carpool, 3 if by vanpool, 4 if by bus, 5 if by other
5	Primary route to work in Seattle area: 1 if Interstate 90, 2 if Interstate 5, 3 if State Route 520, 4 if Interstate 405, 5 if other
6	In the respondent's opinion, is the home-to-work trip traffic congested: 1 if yes, 0 if no
7	Commuter age in years: 1 if under 25, 2 if 26–30, 3 if 31–35, 4 if 36–40, 5 if 41–45, 6 if 46–50, 7 if over 50
8	Respondent's gender 1 if female, 0 if male
9	Number of cars in household
10	Number of children in household
11	Annual household income (US dollars per year): 1 if less than 20,000, 2 if 20,000–29,999, 3 if 30,000–39,999, 4 if 40,000–49,999, 5 if 50,000–59,999, 6 if over 60,000
12	Respondent has flexible work hours? 1 if yes, 0 if no
13	Distance from work to home (in kilometers)
14	Respondent faces level of service D or worse on work-to-home commute? 1 if yes, 0 if no
15	Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time
16	Population of work zone
17	Retail employment in work zone
18	Service employment in work zone
19	Size of work zone (in hectares)

Bibliography

- https://www.rstudio.com/resources/cheatsheets/
- Jenkins, S.P., 2005. Survival Analysis. https://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf
- **Moore D, F., 2016. Applied Survival Analysis Using R. Springer.**
- R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
- Roque, C. and Jalayer, M. 2018. Improving roadside design policies for safety enhancement using hazard-based duration modeling, Accident Analysis & Prevention, Volume 120, 2018, Pages 165-173.
- **Washington, S., Karlaftis, M., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, second edition. Chapman and Hall/CRC.**