Introduction to Survival Analysis

Statistical Physics Approaches to Systems Biology, Havana, Feb 2019

ACC Coolen
King's College London and Saddle Point Science

1982:
Commodore 64

next generation data

previous generation analysis ...

1982:
Commodore 64

Regression Models and Life-Tables

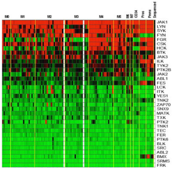
D. R. Cox

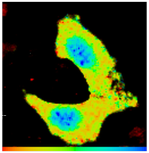
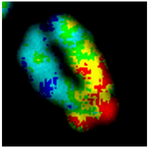
Journal of the Royal Statistical Society. Series B (Methodological), Volume 34, Issue 2 (1972), 187-220.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281972%2934%3A2%3C187%3ARMAL%3E2.0.CO%3B2-6

Biomedicine has changed drastically in recent decades

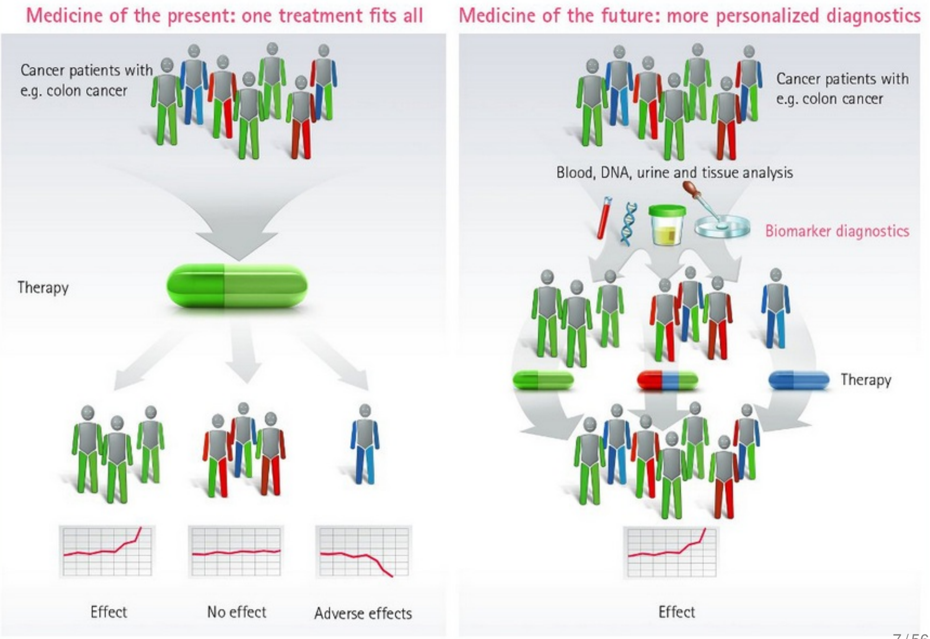
modern biomedical data



- *volume* of data ...
- *diversity* of data sources
  (clinical, genomic, biomarkers, health records, imaging, ...)
- *complexity* of experimental pipelines
  (confounders, batch effects, variability between centres, ...)
- *dimensionality* of measurements
  clinical images ($\sim 10^6$), transcriptome/proteome ($\sim 10^5$),
  DNA and methylation ($\sim 10^{10}$), ...

Personalized medicine: tailored treatments



generating 'big data' is not enough ...

- 'right drug, right dose, at the right time ...'

  need *predictive models* $p(y|\boldsymbol{z})$,

  $\boldsymbol{z}$: individual's makeup (DNA, gene expr, metabolism, environment, ...)
  y: response to treatment

- regression:
  find parameters $\boldsymbol{\theta}$ of model $p(y|\boldsymbol{z}, \boldsymbol{\theta})$ from historic data

  curse of dimensionality ...

  pre-genome medicine: $N \sim 10^3$ data points, $\dim(\boldsymbol{\theta}) \sim 10^2$
  post-genome medicine: $N \sim 10^4$ data points, $\dim(\boldsymbol{\theta}) \sim 10^{10}$

- simpler question: predict patient's *individual* risk
  (target aggressive treatments wisely)

  cancer, heart disease, diabetes, ...:
  relevant outcome is often a *duration* t,
  OS (overall survival), or PFS (progression-free survival)

  predictive model: $p(t|\boldsymbol{z}, \boldsymbol{\theta})$

Statistics in medicine:
tricky business ...

"I can prove it or disprove it! What do you want me to do?"

why is statistics tricky?

Monty Hall problem

‘Let’s Make a Deal’
(USA gameshow, 1963-1977)

standard quiz show,
winner has to choose prize at the end,
three doors: one with big prize, two with goats ...

..

- winner selects one door randomly
- Monty opens one door with a goat (two doors left ...)
- Monty gives winner the chance to change selection at last minute

would it matter?

The main pitfalls in statistics

- *accidental conditioning*

  (Monty Hall problem, share statistics,
  shop opening hours consultation, ...)

  extra knowledge:

  $\rightarrow$ *reduces possibilities*
  $\rightarrow$ *affects probabilities*

  $$\overbrace{P(A|B)}^{posterior} = \frac{P(A,B)}{P(B)} = \overbrace{P(A)}^{prior} \times \frac{P(B|A)}{P(B)}$$

- *often counterintuitive*

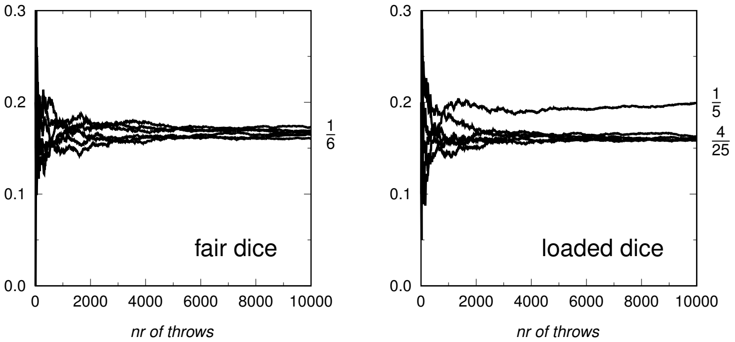
  (Monty Hall problem, gambling,
  human inability to generate random numbers, ...)

  I have just thrown 10 successive sixes!
  $Prob \approx 16.5.10^{-8}$

  how likely am I to throw yet another six?

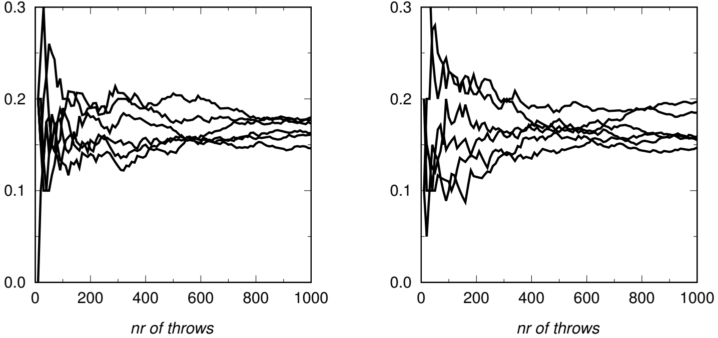
- *how many data do we need to be sure of something?*

  Is a genetic mutation harmless, or dangerous?
  Is a given dice fair, or loaded?



typical data sets in cancer research:
$n \approx 500$ patients, at 2K£ each ...

what can we say
after 500 throws?



- *'probability' can mean different things ...*

  our ignorance of

  (a) something *that cannot be known*
  (Russian roulette, we will spin the cylinder)

  (b) something *that is known, but not by us*
  (Russian roulette, cylinder has already been spun)

relevant in medicine?

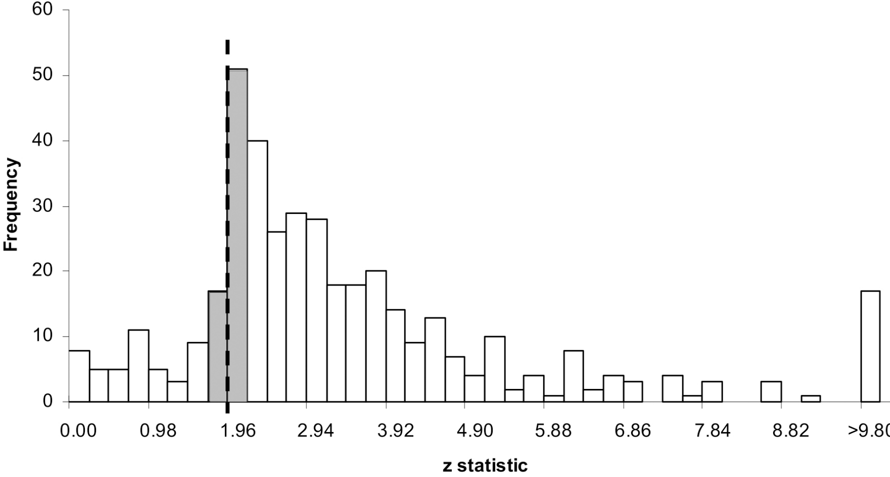
suppose we find survival function $S(t) = \mathrm{e}^{-t/\tau}$

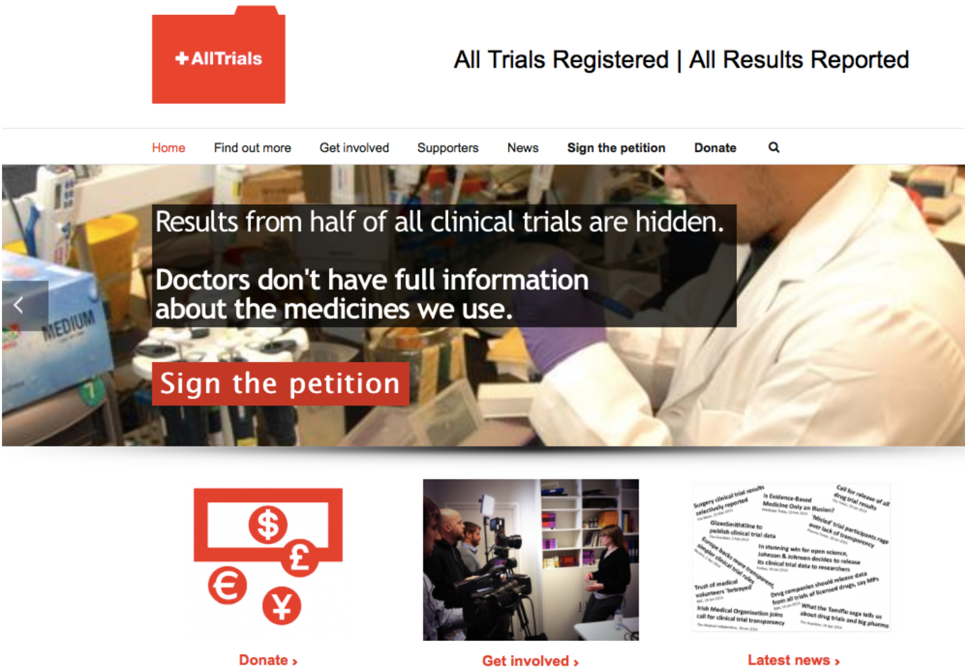
explanation I: homogeneous cohort, *random* death times,
each individual i has hazard rate $1/\tau$

explanation II: heterogeneous cohort, *deterministic* death times t_i,
distributed according to $p(t) = \tau^{-1}\mathrm{e}^{-t/\tau}$
(potential for stratification!)

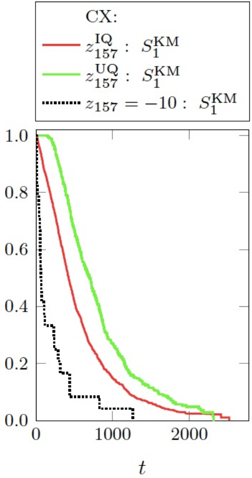
z-scores
reported in PLoS Medicine

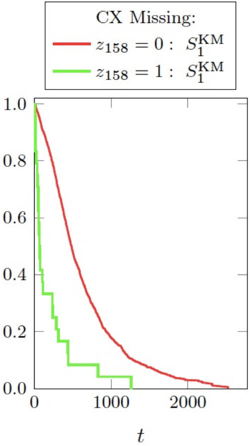
Selective reporting
(aka cheating)




AllTrials
All Trials Registered | All Results Reported
Home
Find out more
Get involved
Supporters
News
Sign the petition
Donate
Results from half of all clinical trials are hidden.
Doctors don't have full information about the medicines we use.
Sign the petition
Donate ›
Get involved ›
Latest news ›

- *missing values in data sets. ...*
  red herrings or white sharks?





always check for
informative missingness!

- *correlation/association is <u>not</u> the same as causality!*

  imagine Z
  is nr of hospital visits ...

  result: positive correlation
  between Z and risk!

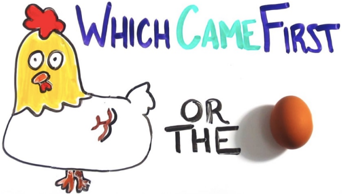
  $\beta > 0$, ergo hospital visits dangerous?

  or:

  $Z = 1, 0$: given strong chemotherapy yes/no
  but treatment not given if patient too weak ...

  result: positive association between Z and health!

  protective effect reported, *even if chemotherapy ineffective!*



two chemotherapies, A and B,
data on response rates from 880 patients

Q: which treatment is better?

	CHEMO A	CHEMO B
response rate	25% (76/300)	28% (162/580)

so treatment B is better,

now we zoom in ...

	CHEMO A	CHEMO B
medical centre 1	40% (40/100)	30% (150/500)
medical centre 2	18% (36/200)	15% (12/80)
response rate	25% (76/300)	28% (162/580)

still sure that B is better?

Simpson's paradox

Terminology and objectives

- *Data available* $\mathcal{D} = \{(\boldsymbol{z}_1, r_1, t_1), \ldots, (\boldsymbol{z}_N, r_N, t_N)\}$

  N samples/individuals $(\boldsymbol{z}_i, r_i, t_i)$,
  drawn independently from $p(t, r, \boldsymbol{z})$ (the population)

  the 'covariates':

  $\boldsymbol{z} \in \mathbb{R}^p$: *p characteristics, measured at* $t=0$

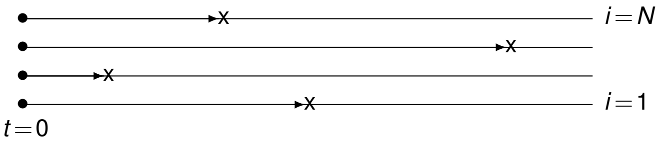
  uncontrolled : *e.g. gender, genome, ...*
  controlled : *e.g. medical treatment, ...*
  modifiable : *e.g. smoking, BMI, nutrition, ...*

  the 'clinical outcome':

  $t \in \mathbb{R}^+$: *failure time*
  e.g. death, onset/recurrence of disease, ...

  $r \in \{0, 1, \ldots, R\}$: *risk type that triggered failure*
  $r = 1 \ldots R$: *true risks/diseases*
  $r = 0$: *end of observation ('censoring')*

pat	BMI	SELENIUM	PHYS_ACT_LEIS	PHYS_ACT_WORK	Smoking	Time	PCcens
1	22.63671875	105	2	1	0	33.6892539357	0
2	34.21875	65	2	2	0	24.810403833	1
3	20.06640625	72	2	3	2	23.1047227926	2
4	28.3984375	81	2	0	2	33.6783025325	0
5	22.94921875	73	0	3	2	32.8843258042	0
6	20.59765625	73	2	0	2	23.2936344969	1
7	26.46875	70	2	3	2	27.9069130732	2
8	26.38671875	91	1	0	1	26.6119096509	2
9	24.296875	73	-10	0	0	26.379192334	1
10	30.01953125	84	1	1	2	26.8254620123	2
11	25.4296875	95	1	0	2	30.6557152635	2
13	23.19921875	67	3	1	0	33.2457221081	0
14	22.90625	82	2	2	0	33.6399726215	0
15	21.62890625	53	1	1	1	33.2457221081	0
16	23.046875	77	2	0	2	33.6125941136	0
17	21.70703125	76	1	2	2	27.8466803559	1
18	22.91796875	102	1	0	2	33.6125941136	0
19	24.5078125	57	2	0	2	25.8097193703	2
20	26.58984375	72	1	2	2	26.803559206	2
21	26.76953125	78	2	0	2	33.234770705	0
22	20.4296875	75	-10	1	1	33.5934291581	0
23	25.0078125	69	0	3	0	33.6125941136	0
24	24.296875	73	2	3	2	21.6098562628	1
25	23.65625	75	2	3	1	33.2320328542	0
26	25.9296875	90	1	1	2	33.5359342916	0
27	23.3671875	58	1	1	2	33.2320328542	0
28	30.08984375	77	-10	-10	2	32.0438056126	2
29	31.08984375	66	1	0	0	29.4893908282	1
30	27.13671875	82	1	-10	1	33.1526351814	0
31	19.828125	68	2	2	2	28.2600958248	2
32	27.30859375	97	2	3	1	33.5742642026	0
33	23.41796875	77	2	0	0	25.9219712526	1



- *Objective*

  find and quantify patterns that relate
  covariates $\boldsymbol{z}$ to clinical outcomes (r, t), in order to:

  - *predict clinical outcome for individuals*
  - *discover disease mechanisms*
  - *design interventions (modifiable covariates)*

- *Complications*

  - ‘noise’ caused by censoring
  - we only know the earliest event
    (different risks prevent each other from happening)
  - correlations between risks
  - heterogeneity in patient cohorts
  - overfitting danger, when p is large relative to N

Survival probability and cause-specific hazard rates

- *Joint event time statistics*

  imaginary situation: all events can be observed,
  t_r: time at which event r occurs,
  event time distributiion: $P(t_0, \ldots, t_R)$

  normalisation: $$\int_0^\infty \ldots \int_0^\infty \mathrm{d}t_0 \ldots \mathrm{d}t_R \; P(t_0, \ldots, t_R) = 1$$

- *Integrated event time distribution*

  $$S(t_0, \ldots, t_R) \;\; = \;\; \int_{t_0}^\infty \ldots \int_{t_R}^\infty \mathrm{d}s_0 \ldots \mathrm{d}s_R \; P(s_0, \ldots, s_R)$$

  meaning: probability that event 0 occurs later than t_0,
  and event 1 occurs later than t_1, *and* ...

  $$S(0, \ldots, 0) \;\; = \;\; \int_0^\infty \ldots \int_0^\infty \mathrm{d}s_0 \ldots \mathrm{d}s_R \; P(s_0, \ldots, s_R) = 1$$

- *Survival function $S(t)$*

  probability that *all* events happen later than time t:

  $$S(t) \;\; = \;\; \int_t^\infty \ldots \int_t^\infty \mathrm{d}s_0 \ldots \mathrm{d}s_R \; P(s_0, \ldots, s_R) \;\; = \;\; S(t, t, \ldots, t)$$

- *Cause-specific hazard rates $h_r(t)$*

  how do individual risks impact on survival?

$$h_r(t) \quad = \quad -\Big[\frac{\partial}{\partial t_r}\log S(t_0,\ldots,t_R)\Big]_{t_k=t \text{ for all } k}$$

work out, using $\frac{\mathrm{d}}{\mathrm{d}z}\theta(z)=\delta(z)$:

$$\begin{aligned}
h_r(t) &= \left[\frac{\frac{\partial}{\partial t_r}\int_0^\infty\!\ldots\int_0^\infty \mathrm{d}s_0\ldots\mathrm{d}s_R\; P(s_0,\ldots,s_R)\prod_k\theta(s_k-t_k)}{S(t_0,\ldots,t_R)}\right]_{t_k=t\;\forall k}\\
&= \left[\frac{\int_0^\infty\!\ldots\int_0^\infty \mathrm{d}s_0\ldots\mathrm{d}s_R\; P(s_0,\ldots,s_R)\delta(s_r-t_r)\prod_{k\neq r}\theta(s_k-t_k)}{S(t_0,\ldots,t_R)}\right]_{t_k=t\;\forall k}\\
&= \frac{\int_t^\infty\!\ldots\int_t^\infty\Big(\prod_{r\neq\mu}^R \mathrm{d}s_r\Big)P(s_0,\ldots,s_{\mu-1},t,s_{\mu+1},\ldots,s_R)}{S(t)}
\end{aligned}$$

$h_r(t)\mathrm{d}t$: probability that event r happens in time interval $[t,t+\mathrm{d}t)$,
given that no event has happened yet prior to t

$$h_r(t)\mathrm{d}t \quad = \quad \text{Prob}\Big(t_r\in[t,t+\mathrm{d}t)\Big|\text{ no events yet at time } t\Big) \qquad (\mathrm{d}t\downarrow 0)$$

Data likelihood in terms of cause specific hazard rates

most of the relevant quantities in survival analysis
can be written in terms of the cause specific hazard rates

- *Survival function*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\log S(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\log S(t,t,\ldots,t)\\ &= \sum_{r=0}^{R}\Big[\frac{\partial}{\partial t_r}\log S(t_0,\ldots,t_R)\Big]_{t_r=t\ \forall r}\\ &= -\sum_{r=0}^{R}h_r(t)\end{aligned}$$

Hence, using $S(0)=1$,

$$\log S(t) = \log S(0) - \int_0^t \mathrm{d}s \sum_{r=0}^{R} h_r(s) = -\sum_{r=0}^{R}\int_0^t \mathrm{d}s\ h_r(s)$$

result:

$$S(t) = \mathrm{e}^{-\sum_{r=0}^{R}\int_0^t \mathrm{d}s\ h_r(s)}$$

- *Data likelihood*

$p(t,r)\mathrm{d}t$: likelihood to observe *first* event being of type r, <u>and</u> occurring in time interval $[t, t+\mathrm{d}t)$ (with $\mathrm{d}t \downarrow 0$)

To observe the above, the following statements must be true:

time of the event is in $[t, t+\mathrm{d}t)$,
type of the event is r,
no events occurred prior to t

$$\theta(t_r-t)\theta(t+\mathrm{d}t-t_r)\prod_{k\neq r}\theta(t_k-t) = 1$$

hence

$$\begin{aligned} p(t,r) &= \lim_{\mathrm{d}t\downarrow 0}\frac{1}{\mathrm{d}t}\,\mathrm{Prob}\Big(\theta(t_r-t)\theta(t+\mathrm{d}t-t_r)\prod_{k\neq r}\theta(t_k-t) = 1\Big) \\ &= \lim_{\mathrm{d}t\downarrow 0}\frac{1}{\mathrm{d}t}\int_0^\infty\!\!\ldots\int_0^\infty \mathrm{d}t_0\ldots t_R\; P(t_1,\ldots,t_R)\theta(t_r-t)\theta(t+\mathrm{d}t-t_r)\prod_{k\neq r}\theta(t_k-t) \\ &= \int_0^\infty\!\!\ldots\int_0^\infty \mathrm{d}t_0\ldots t_R\; P(t_1,\ldots,t_R)\lim_{\epsilon\downarrow 0} h_\epsilon(t_r-t)\prod_{k\neq r}\theta(t_k-t) \end{aligned}$$

$$h_\epsilon(z) = \epsilon^{-1}\theta(z)\theta(\epsilon-z) = \begin{cases} \epsilon^{-1} & \text{for } z\in[0,\epsilon] \\ 0 & \text{elsewhere} \end{cases}$$

note: $\lim_{\epsilon\downarrow 0} h_\epsilon(z) = \delta(z)$, so

$$
\begin{aligned}
p(t,r) &= \int_0^\infty \ldots \int_0^\infty \mathrm{d}t_0 \ldots t_R\, P(t_1,\ldots,t_R) \lim_{\epsilon\downarrow 0} h_\epsilon(t_r - t) \prod_{k\neq r} \theta(t_k - t) \\
&= \int_0^\infty \ldots \int_0^\infty \mathrm{d}t_0 \ldots t_R\, P(t_1,\ldots,t_R)\delta(t_r - t) \prod_{k\neq r} \theta(t_k - t) \\
&= h_t(t) S(t) \;=\; h_t(t)\mathrm{e}^{-\sum_{r'=0}^R \int_0^t \mathrm{d}s\; h_{r'}(s)}
\end{aligned}
$$

- *Further relation*

$$
\begin{aligned}
p(t) &= \sum_{r=0}^R p(t,r) \;=\; \Big(\sum_{r=0}^R h_t(t)\Big)\mathrm{e}^{-\sum_{r'=0}^R \int_0^t \mathrm{d}s\; h_{r'}(s)} \\
&= -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-\sum_{r'=0}^R \int_0^t \mathrm{d}s\; h_{r'}(s)} \;=\; -\frac{\mathrm{d}}{\mathrm{d}t} S(t)
\end{aligned}
$$

- *Cause-specific hazard rates in terms of data probabilities*

$$
S(t) \;=\; S(0) + \int_0^t \mathrm{d}t'\, \frac{\mathrm{d}}{\mathrm{d}t'} S(t') \;=\; 1 - \int_0^t \mathrm{d}t'\, p(t') \;=\; \int_t^\infty \mathrm{d}s\, p(s)
$$

substitute into formula for $p(t,r)$:

$$
h_r(t) = \frac{p(t,r)}{\sum_{r'=0}^R \int_t^\infty \mathrm{d}s\, p(s,r')}
$$

Pitfalls and misconceptions

cause specific hazard rates can be tricky ...

$$S(t) = \prod_r \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}$$

- *Survival function formula factorizes over risks, does this imply that the risks are uncorrelated?*

  No. All risks $k \neq r$ can contribute to each $h_r(t)$ via the conditioning, i.e. the likelihood that nothing has happened yet prior to t. Risks may well interact strongly with each other, but we can no longer see this after we have calculated the rates $\{h_r(t)\}$ and forget about the times $(t_0, \ldots, t_R)$.

- *Do we get the survival function for the situation where risk μ is disabled (e.g. a disease removed from the world) by setting $h_\mu(t)$ to zero?*

  $$S(t) \to \mathrm{e}^{-\sum_{r\neq\mu} \int_0^t \mathrm{d}s\ h_r(s)}$$

  No. We would have $h_\mu(t) = 0$ for all t, but that is not all. Disabling risk μ can change also *all* hazard rates $h_r(t)$ with $r \neq \mu$, due to correlations among the different risks in combination with the conditioning.

Special cases

- *Time-independent hazard rates*

  $h_r(t) = h_r$: $\quad S(t) = \mathrm{e}^{-t\sum_{r=0}^{R} h_r} \quad p(t,r) = h_r \mathrm{e}^{-t\sum_{r'=0}^{R} h_{r'}}$

- *A single risk, $R{=}1$*

  One risk,
  hazard rate $h(t)$: $\quad S(t) = \mathrm{e}^{-\int_0^t \mathrm{d}s\; h(s)} \quad p(t) = h(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\; h(s)}$

- *Most probable event time distribution for $R{=}1$*

  Suppose we know only average event time $\langle t\rangle = \int_0^\infty \mathrm{d}t\; tp(t)$,
  most probable $p(t)$: maximize Shannon entropy
  $H = -\int_0^\infty \mathrm{d}t\; p(t)\log p(t)$, subject to $\int_0^\infty \mathrm{d}t\; p(t){=}1$ and $\int_0^\infty \mathrm{d}t\; tp(t){=}\langle t\rangle$

  Lagrange's method:

  $$\frac{\delta}{\delta p(t)}\int_0^\infty \mathrm{d}s\; p(s)\log p(s) = \frac{\delta}{\delta p(t)}\Big\{\lambda_0\int_0^\infty \mathrm{d}s\; p(s) + \lambda_1\int_0^\infty \mathrm{d}s\; p(s)s\Big\}$$

  $$1+\log p(t) = \lambda_0+\lambda_1 t \quad \text{so} \quad p(t) = \mathrm{e}^{\lambda_0-1+\lambda_1 t}$$

  use constraints: $\quad p(t) = \langle t\rangle^{-1}\mathrm{e}^{-t/\langle t\rangle}$

Independently distributed event times

$$P(t_0,\ldots,t_R) = \prod_{r=0}^{R} P_r(t_r)$$

- *Integrated event time distr*

$$S(t_0,\ldots,t_R) = \int_{t_0}^{\infty}\!\!\ldots\int_{t_R}^{\infty} \mathrm{d}s_0\ldots\mathrm{d}s_R \prod_{r=0}^{R} P_r(t_r) = \prod_{r=0}^{R} S_r(t_r)$$

$$S_r(t) = \int_t^{\infty}\mathrm{d}s\; P_r(s)$$

- *Cause specific hazard rates*

$$h_r(t) = -\Big[\frac{\partial}{\partial t_r}\sum_{r'=0}^{R}\log S_{r'}(t_{r'})\Big]_{t_k=t \;\; for \;\; all \;\; k} = -\frac{\mathrm{d}}{\mathrm{d}t}\log S_r(t)$$

hence

$$S_r(t) = \mathrm{e}^{-\int_0^t \mathrm{d}s\; h_r(s)}$$

joint event time distr now follows from the cause-specific hazard rates

$$P_r(t) = -\frac{\mathrm{d}}{\mathrm{d}t}S_r(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-\int_0^t \mathrm{d}s\; h_r(s)} = h_t(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\; h_r(s)}$$

$$P(t_0,\ldots,t_R) = \prod_{r=0}^{R}\Big[h_t(t_r)\mathrm{e}^{-\int_0^{t_r}\mathrm{d}s\; h_r(s)}\Big]$$

The identifiability problem (Tsiatis)

- *Observable from data*: $p(t,r)$,
  equivalently: $\{h_0(t),\ldots,h_R(t)\}$, since

$$p(t,r)=h_r(t)\mathrm{e}^{-\sum_{r'=0}^{R}\int_0^t \mathrm{d}s\; h_{r'}(s)},\qquad h_r(t)=\frac{p(t,r)}{\sum_{r'=0}^{R}\int_t^\infty \mathrm{d}s\; p(s,r')}$$

- For any $\{h_0(t),\ldots,h_R(t)\}$, even those corresponding to *statistically dependent event times*, there exists a distribution for *independent* event times that will give exactly the same cause-specifc hazard rates, namely

$$P(t_0,\ldots,t_R)=\prod_{r=0}^{R}\prod_{r=0}^{R}\Big[h_t(t_r)\mathrm{e}^{-\int_0^{t_r}\mathrm{d}s\; h_r(s)}\Big]$$

  Hence, survival data alone do not generally permit us to identify the underlying joint distribution of event times – in particular, we cannot infer whether or not the event times of the different risks are independent.

  a serious problem ...

- *The Bayesian view* on the identifiability problem

  – multiple hypotheses H may explain our data
  – but not all are equally probable ...
  – calculate each $Prob(H|\mathcal{D})$ from Bayes' formula

- *Illustration*

  true data:

$$p(t_2) = ae^{-at_2}, \quad \begin{cases} \textit{with prob } \epsilon: & t_1 = t_2 + \tau \\ \textit{with prob } 1-\epsilon: & \textit{draw } t_1 \textit{ from } p(t_1) = be^{-bt_1} \end{cases}$$

  explanation assuming risk independence:

$$p(t_2) = ae^{-at_2}, \qquad p(t_1) = \underbrace{-\Big(\epsilon+(1-\epsilon)e^{-bt_1}\Big)\log\Big(\epsilon+(1-\epsilon)e^{-bt_1}\Big)}_{\textit{with prob } \epsilon:\ \textit{event 1 } \underline{\textit{never}} \textit{ happens}}$$

implausible if e.g. risk 2 is cancer, risk 1 is death ...

Proportional hazards regression (Cox) definitions and assumptions

- *Survival analysis with covariates*

  Add $\boldsymbol{z}$ as conditions to definitions and identities

$$S(t) \to S(t|\boldsymbol{z}), \quad h_r(t) \to h_r(t|\boldsymbol{z}), \quad p(t,r) \to p(t,r|\boldsymbol{z})$$

$$S(t|\boldsymbol{z}) = \int_t^\infty \ldots \int_t^\infty \mathrm{d}s_0 \ldots \mathrm{d}s_R \; P(s_0,\ldots,s_R|\boldsymbol{z})$$

$$S(t|\boldsymbol{z}) = \mathrm{e}^{-\sum_{r=0}^R \int_0^t \mathrm{d}s \; h_r(s|\boldsymbol{Z})}, \quad p(t,r|\boldsymbol{z}) = h_r(t|\boldsymbol{z})\mathrm{e}^{-\sum_{r'=0}^R \int_0^t \mathrm{d}s \; h_{r'}(s|\boldsymbol{Z})}$$

- *Cox model*

  Parametrized form for the hazard rates:

$$h_r(t|\boldsymbol{Z}) = \lambda_r(t)\mathrm{e}^{\boldsymbol{\beta}^r\cdot\boldsymbol{z}}, \qquad \boldsymbol{\beta}^r = (\beta_1^r,\ldots,\beta_p^r), \quad \boldsymbol{\beta}^r\cdot\boldsymbol{z} = \sum_{\mu=1}^p \beta_\mu^r z_\mu$$

  $\lambda_r(t)$: 'base hazard rate' of risk r
  (covariate-independent contribution to risk)

  $\boldsymbol{\beta}^r$: 'association parameters' of risk r
  (impact of covariate values on risk)

- *Main features of Cox's choice*
  - *'Proportional hazards'*

    due to exponential form, effect of each covariate is multiplicative:

$$h_r(t) = \underbrace{\lambda_r(t)}_{base\ hazard\ rate} \times \underbrace{e^{\beta_1^r z_1} \times \ldots \times e^{\beta_p^r z_p}}_{'proportional\ hazards'}$$

  - *Effects of covariates on risk independent of time*
  - *There exists a hyper-plane in covariate space that separates high risk individuals from low risk individuals*

$$\begin{aligned} 'high\ risk\ individuals' : \quad & \beta_1^r z_1 + \ldots + \beta_p^r z_p \quad large \\ 'low\ risk\ individuals' \quad & \beta_1^r z_1 + \ldots + \beta_p^r z_p \quad small \end{aligned}$$

  - *One can quantify risk impact of each individual covariate* μ *in a single time-independent number: the 'hazard ratio'*

$$HR_\mu^r = \frac{h_r(t|\boldsymbol{z})|_{z_\mu=1}}{h_r(t|\boldsymbol{z})|_{z_\mu=0}} = \frac{\lambda_r(t) e^{\beta_\mu^r . 1 + \sum_{\nu \neq \mu} \beta_\nu^r z_\nu}}{\lambda_r(t) e^{\beta_\mu^r . 0 + \sum_{\nu \neq \mu} \beta_\nu^r z_\nu}} = e^{\beta_\mu^r}$$

    If no impact on risk: $\beta_\mu^r = 0, \ \mathrm{HR}_\mu^r = 1.$

Detour: parameter estimation from data

given data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$,
and a model $p(\boldsymbol{x}|\boldsymbol{\theta})$ to explain these data,
what can we say about the parameters $\boldsymbol{\theta}$?

- *Bayesian parameter inference*

  assume the $\{\boldsymbol{x}_i\}$ were indeed drawn
  randomly & independently from $p(\boldsymbol{x}|\boldsymbol{\theta})$,

  $$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta})$$

  Bayes' identity:

  $$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta}'\ p(\mathcal{D}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}, \qquad p(\boldsymbol{\theta}):\ \textit{prior}$$

- *Simplifications*

  $$\textit{MAP}: \qquad \boldsymbol{\theta}^\star = \mathrm{argmax}\ p(\boldsymbol{\theta}|\mathcal{D}) = \mathrm{argmin}\Big[\overbrace{-\log p(\mathcal{D}|\boldsymbol{\theta})}^{\textit{minus log–likelihood}}\ \overbrace{-\log p(\boldsymbol{\theta})}^{\textit{regularizer}}\Big]$$

  $$\textit{ML}: \qquad \boldsymbol{\theta}^\star = \mathrm{argmin}\Big[-\log p(\mathcal{D}|\boldsymbol{\theta})\Big] \qquad \textit{i.e. } p(\boldsymbol{\theta}) = \textit{constant}$$

- *Maximimum Likelihood (ML) regression*

$$\boldsymbol{\theta}^\star = \mathrm{argmin}\big[-\log p(\mathcal{D}|\boldsymbol{\theta})\big]$$

define empirical data distribution

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i)$$

note:

$$
\begin{aligned}
-\frac{1}{N}\log p(\mathcal{D}|\boldsymbol{\theta}) &= -\frac{1}{N}\log\prod_{i=1}^{N}p(\boldsymbol{x}_i|\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\log p(\boldsymbol{x}_i|\boldsymbol{\theta}) \\
&= -\int \mathrm{d}\boldsymbol{x}\;\hat{p}(\boldsymbol{x})\log p(\boldsymbol{x}|\boldsymbol{\theta}) \\
&= \int \mathrm{d}\boldsymbol{x}\;\hat{p}(\boldsymbol{x})\log\left[\frac{\hat{p}(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{\theta})}\right] - \int \mathrm{d}\boldsymbol{x}\;\hat{p}(\boldsymbol{x})\log\hat{p}(\boldsymbol{x}) \\
&= \underbrace{D(\hat{p}||p_{\boldsymbol{\theta}})}_{KL\ distance} + \underbrace{H[\hat{p}]}_{Shannon\ entropy}
\end{aligned}
$$

hence: ML finds the parameter vector $\boldsymbol{\theta}$ that minimizes the KL distance between $\hat{p}(\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{\theta})$

- *Beyond most probable parameters: error bars*

  return to full posterior distribution

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathrm{e}^{-\Omega(\boldsymbol{\theta},\mathcal{D})}}{\int \mathrm{d}\boldsymbol{\theta}'\ \mathrm{e}^{-\Omega(\boldsymbol{\theta}',\mathcal{D})}}, \qquad \Omega(\boldsymbol{\theta},\mathcal{D}) = -\log p(\mathcal{D}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})$$

  expand Ω around minimum $\boldsymbol{\theta}^\star$:

$$\Omega(\boldsymbol{\theta},\mathcal{D}) \quad = \quad \Omega(\boldsymbol{\theta}^\star,\mathcal{D}) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\cdot\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) + \mathcal{O}(|(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)|^3)$$

  truncate after quadratic term:

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx \left[\frac{\det\boldsymbol{A}}{(2\pi)^N}\right]^{\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\cdot\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)}, \qquad \langle(\theta_\mu-\theta_\mu^\star)(\theta_\nu-\theta_\nu^\star)\rangle = (\boldsymbol{A}^{-1})_{\mu\nu}$$

  hence, error bars for
  MAP/ML estimators:

$$\theta_\mu = \theta_\mu^\star \pm (\boldsymbol{A}^{-1})_{\mu\mu}, \qquad A_{\mu\nu} = \frac{\partial^2}{\partial\theta_\mu\partial\theta_\nu}\Big[-\log p(\mathcal{D}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\Big]_{\boldsymbol{\theta}^\star}$$

ML parameters of Cox's model

$$\boldsymbol{\theta} = \{\lambda_r, \boldsymbol{\beta}^r\} :$$

$$p(t, r|\boldsymbol{z}, \boldsymbol{\theta}) = h_r(t|\boldsymbol{z}, \boldsymbol{\theta}) \mathrm{e}^{-\sum_{r'=0}^{R} \int_0^t \mathrm{d}s\ h_{r'}(s|\boldsymbol{z},\boldsymbol{\theta})}, \qquad h_r(t|\boldsymbol{z}, \boldsymbol{\theta}) = \lambda_r(t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}}$$

- *ML inference*

$$\begin{aligned}
\boldsymbol{\theta}^\star &= \mathrm{argmin}_{\boldsymbol{\theta}} \Big[-\log p(\mathcal{D}|\boldsymbol{\theta}) \Big] \\
&= \mathrm{argmin}_{\boldsymbol{\theta}} \Big[-\sum_{i=1}^{N} \log p(t_i, r_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \Big] \\
&= \mathrm{argmin}_{\boldsymbol{\theta}} \Big[-\sum_{i=1}^{N} \log h_{r_i}(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}) + \sum_{i=1}^{N} \sum_{r=0}^{R} \int_0^{t_i} \mathrm{d}t\ \lambda_r(t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i} \Big] \\
&= \mathrm{argmin}_{\boldsymbol{\theta}} \Big[-\sum_{i=1}^{N} \log \lambda_{r_i}(t_i) - \sum_{i=1}^{N} \boldsymbol{\beta}^{r_i} \cdot \boldsymbol{z}_i + \sum_{i=1}^{N} \sum_{r=0}^{R} \int_0^{t_i} \mathrm{d}t\ \lambda_r(t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i} \Big] \\
&= \mathrm{argmin}_{\boldsymbol{\theta}} \sum_{r=0}^{R} \Big[-\sum_{i=1}^{N} \delta_{r,r_i} \log \lambda_r(t_i) - \sum_{i=1}^{N} \delta_{r,r_i} \boldsymbol{\beta}^r \cdot \boldsymbol{z}_i + \sum_{i=1}^{N} \int_0^{t_i} \mathrm{d}t\ \lambda_r(t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i} \Big]
\end{aligned}$$

To minimize

$$\Psi[\{\lambda_r, \boldsymbol{\beta}^r\}] \quad = \quad \sum_{r=0}^{R} \Big[-\int \mathrm{d}t \; \log \lambda_r(t) \sum_{i=1}^{N} \delta_{r,r_i} \delta(t-t_i) - \boldsymbol{\beta}^r \cdot \sum_{i=1}^{N} \delta_{r,r_i} \boldsymbol{z}_i$$
$$+ \int \mathrm{d}t \; \lambda_r(t) \sum_{i=1}^{N} \theta(t_i - t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i} \Big]$$

- *Minimize over functions $\lambda_r(t)$ first*

$$\frac{\delta}{\delta \lambda_r(t)} \Bigg[\sum_{i=1}^{N} \int \mathrm{d}t \; \theta(t_i - t) \; \lambda_r(t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i} - \int \mathrm{d}t \; \log \lambda_r(t) \sum_{i=1}^{N} \delta_{r,r_i} \delta(t-t_i) \Bigg] = 0$$

(functional differentiation)

$$\sum_{i=1}^{N} \theta(t_i - t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i} - \frac{1}{\lambda_r(t)} \sum_{i=1}^{N} \delta_{r,r_i} \delta(t-t_i) = 0$$

$$\lambda_r(t) = \frac{\sum_{i=1}^{N} \delta_{r,r_i} \delta(t-t_i)}{\sum_{i=1}^{N} \theta(t_i - t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z}_i}} \qquad \textit{Breslow's estimator}$$

- *Insert into Ψ, then minimize over $\{\boldsymbol{\beta}^r\}$*

- *To minimize*

$$\Psi[\{\boldsymbol{\beta}^r\}] = \sum_{r=0}^{R}\Big[-\int \mathrm{d}t\ \log\Big(\frac{\sum_{i=1}^{N}\delta_{r,r_i}\delta(t-t_i)}{\sum_{i=1}^{N}\theta(t_i-t)\mathrm{e}^{\boldsymbol{\beta}^r\cdot\boldsymbol{z}_i}}\Big)\sum_{j=1}^{N}\delta_{r,r_j}\delta(t-t_j) - \boldsymbol{\beta}^r\cdot\sum_{i=1}^{N}\delta_{r,r_i}\boldsymbol{z}_i + \int \mathrm{d}t\ \sum_{i=1}^{N}\delta_{r,r_i}\delta(t-t_i)\Big]$$

$$= \sum_{r=0}^{R}\Big[\int \mathrm{d}t\ \log\Big(\sum_{i=1}^{N}\theta(t_i-t)\mathrm{e}^{\boldsymbol{\beta}^r\cdot\boldsymbol{z}_i}\Big)\sum_{j=1}^{N}\delta_{r,r_j}\delta(t-t_j) - \boldsymbol{\beta}^r\cdot\sum_{i=1}^{N}\delta_{r,r_i}\boldsymbol{z}_i\Big] + \textit{ terms independent of } \{\boldsymbol{\beta}^r\}$$

$$= \sum_{r=0}^{R}\Psi_r(\boldsymbol{\beta}^r) + \textit{ terms independent of } \{\boldsymbol{\beta}^r\}$$

with

$$\Psi_r(\boldsymbol{\beta}) = \sum_{j=1}^{N}\delta_{r,r_j}\log\Big(\sum_{i=1}^{N}\theta(t_i-t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}\Big) - \boldsymbol{\beta}\cdot\sum_{i=1}^{N}\delta_{r,r_i}\boldsymbol{z}_i$$

hence

$$\boldsymbol{\beta}^{r\star} = \mathrm{argmin}_{\boldsymbol{\beta}}\Psi_r(\boldsymbol{\beta})$$

- *Each risk r*, find minima of Ψ_r:

$$\frac{\partial}{\partial \beta_\mu}\Psi_r(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_\mu}\Big[\sum_{j=1}^N \delta_{r,r_j} \log\Big(\sum_{i=1}^N \theta(t_i - t_j) \mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}\Big) - \boldsymbol{\beta}\cdot\sum_{j=1}^N \delta_{r,r_j}\boldsymbol{z}_j\Big]$$

$$= \sum_{j=1}^N \delta_{r,r_j} \frac{\sum_{i=1}^N z_{i\mu}\theta(t_i - t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}{\sum_{i=1}^N \theta(t_i - t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}} - \sum_{j=1}^N \delta_{r,r_j} z_{j\mu}$$

so $\boldsymbol{\beta}^{r\star}$ is solution of:

$$\sum_{j=1}^N \delta_{r,r_j}\Big[\frac{\sum_{i=1}^N z_{i\mu}\theta(t_i - t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}{\sum_{i=1}^N \theta(t_i - t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}} - z_{j\mu}\Big] = 0$$

p coupled nonlinear equations, for each risk

Final protocol:

1. Solve $\{\beta^{r\star}\}$ from above eqns (numerically)
2. Calculate $\{\lambda^r(t)\}$ (from Breslow's formula)
3. Predict outcomes via $p(r, t|\boldsymbol{z})$ for new patients (using Cox's model, with ML parameters)

Uniqueness, error bars, and and *p*-values

Curvature of $\Psi_r(\beta)$:

$$\frac{\partial^2\Psi_r(\boldsymbol{\beta})}{\partial\beta_\mu\beta_\nu} = \frac{\partial}{\partial\beta_\nu}\Big[\sum_{j=1}^N \delta_{r,r_j}\frac{\sum_{i=1}^N z_{i\mu}\theta(t_i-t_j)e^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}{\sum_{i=1}^N \theta(t_i-t_j)e^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}} - \sum_{j=1}^N \delta_{r,r_j} z_{j\mu}\Big]$$

$$= \sum_{j=1}^N \delta_{r,r_j}\Big[\langle z_\mu z_\nu\rangle_j - \langle z_\mu\rangle_j\langle z_\nu\rangle_j\Big]$$

with

$$\langle f(\boldsymbol{z})\rangle_j = \sum_{i=1}^N w(i|j)f(\boldsymbol{z}_i), \qquad w(i|j) = \frac{\theta(t_i-t_j)e^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}{\sum_{i=1}^N \theta(t_i-t_j)e^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}$$

properties, consequences:

- *Convexity*

  curvature matrix is positive definite, i.e. $\Psi_r(\beta)$ convex, since

$$\forall \boldsymbol{x}\in\mathbb{R}^p: \quad \sum_{\mu,\nu=1}^p x_\mu x_\nu \frac{\partial^2\Psi_r(\boldsymbol{\beta})}{\partial\beta_\mu\beta_\nu} = \sum_{j=1}^N \delta_{r,r_j}\Big[\langle(\boldsymbol{x}\cdot\boldsymbol{z})^2\rangle_j - \langle\boldsymbol{x}\cdot\boldsymbol{z}\rangle_j^2\Big]$$

$$= \sum_{j=1}^N \delta_{r,r_j}\Big[\langle(\boldsymbol{x}\cdot\boldsymbol{z} - \langle\boldsymbol{x}\cdot\boldsymbol{z}\rangle_j)^2\rangle_j\Big] \geq 0$$

- *Uniqueness of $\boldsymbol{\beta}^{r\star}$*

  since $\Psi_r(\beta)$ convex, can have only one minimum

- *Error bars for association parameters*

  Neglect fluctuations in $\{\lambda_r(t)\}$, focus on $\Delta\beta_\mu^{r\star}$:

$$A(r)_{\mu\nu} = \sum_{j=1}^{N} \delta_{r,r_j}\Big[\langle z_\mu z_\nu\rangle_j - \langle z_\mu\rangle_j\langle z_\nu\rangle_j\Big]$$

$$\langle f(\boldsymbol{z})\rangle_j = \sum_{i=1}^{N} w(i|j) f(\boldsymbol{z}_i), \qquad w(i|j) = \frac{\theta(t_i - t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}{\sum_{i=1}^{N}\theta(t_i - t_j)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}}$$

  Then

$$\beta_\mu^r = \beta_\mu^{r\star} \pm \sigma_\mu^r, \qquad \sigma_\mu^r = (\boldsymbol{A}(r)^{-1})_{\mu\mu}$$

- *p-values of inferred* $\beta_\mu^\star$

  definition: the probability to find a value $\beta_\mu^\star$ (or one further away from zero) due to fluctuations, when the true value is zero

  approx: assume above error bar is correct, disregard correlations,

$$p\text{-}value = \mathrm{Prob}\Big(|\beta_\mu| \geq |\beta_\mu^\star|\Big) = 1 - \frac{2}{\sigma_\mu\sqrt{2\pi}}\int_0^{|\beta_\mu^\star|} \mathrm{d}\beta\ \mathrm{e}^{-\frac{1}{2}\beta^2/\sigma_\mu^2}$$

$$= 1 - \mathrm{Erf}\Big(|\beta_\mu^\star|/\sigma_\mu\sqrt{2}\Big), \qquad \beta_\mu^\star/\sigma_\mu : \ \ z\text{-}score$$

Explanation for Simson's paradox

	CHEMO A	CHEMO B
medical centre 1	40% (40/100)	30% (150/500)
medical centre 2	18% (36/200)	15% (12/80)
response rate	25% (76/300)	28% (162/580)

$$P(response|chemo) = \sum_{centres} P(response|chemo, centre) P(centre|chemo)$$

$$P(resp|A) = \frac{40}{100}\frac{100}{300} + \frac{36}{200}\frac{200}{300} = 25\%$$

$$P(resp|B) = \frac{30}{100}\frac{500}{580} + \frac{15}{100}\frac{80}{580} = 28\%$$

if chemo choice indep of centre:

$$P(resp|A) = \frac{40}{100}\frac{1}{2} + \frac{36}{200}\frac{1}{2} = 29\%$$

$$P(resp|B) = \frac{30}{100}\frac{1}{2} + \frac{15}{100}\frac{1}{2} = 22.5\%$$