Medical Stats Notes

Autumn Mapes

Fall 2025

1 Part 1: Causal Inference

Causation versus association

Causation is the language of intervention. Association is the language of probability.

Card ID: 1779140587297

Def: Confounding

When some event is a common cause of two variables. The two variables may appear correlated, but don't cause each other.

Card ID: 1779140587298

Def: Collider Bias / Berkson's Paradox

Celebrity example: You may notice that ugly celebrities tend to be talented, and that untalented celebrities tend to be attractive. But this is only because they wouldn't be a celebrity if they weren't either.

When two different events cause the same outcome, and that outcome makes observation easier, then those two events may seem erroneously negatively correlated.

Card ID: 1779140587300

Def: Reverse Causation

If symptoms are consistently treated, treatment may appear to 'cause' the illness.

Card ID: 1779140587302

How can we measure causality?

---

1. Perform controlled experiments
2. Perform randomized experiments
3. Account for confounding in analyses

Use instrumental variables

4. Assume that some part of the data-generating process behaves like randomization (eg mendelian randomization)

This is called a natural experiment

Card ID: 1779140587303

todo assumptions of causal inference?

Def: Negative Controls

To check your work, check causality of obviously uncorrelated phenomena

Card ID: 1779140587305

Def: Difference-in-Differences

Sudden divergence of trendlines in two populations: there must be some reason for the divergence!

Card ID: 1779140587307

Def: Regression Continuity

Check that measured data that should be smooth is actually smooth. Jump points are suspicious.

Card ID: 1779140587308

Def: Immortal Time Bias

Must consider the age requirement of some property: nobel prize winners are older than non-nobel prize winners.

Card ID: 1779140587310

Def: Intention-to-treat

Who will be included in the study must be set-in-stone before time zero (can't disclude people on the fly)

Card ID: 1779140587312

Def: Case-control studies

Case studies paired with a control group. Not optimal, but sometimes necessary for rare diseases.

Card ID: 1779140587314

Def: Cohort Studies

Look at a (snapshot) population and observe what effects it over time.

Card ID: 1779140587315

Def: Rubin's Dictum

For objective causal inference, design trumps analysis.

Card ID: 1779140587317

2 Part 2: Models for Non-transmissible Disease

2.1 Continuous-time Markov models

Def: Transition Rate, and when it is a Markov process and time homogeneous

$$q_{rs}(t;\mathcal{F}_t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(X(t+\delta t) = s \,|\, X(t) = r,\, \mathcal{F}_t)}{\delta t}$$

$X(t)$ represents the state the individual is in at time t, $\mathcal{F}_t$ is all information about the process prior to t.

It's a Markov process when q_{rs} has no dependence on $\mathcal{F}_t$, and it's time-homogeneous when q_{rs} has no dependence on time (time probabilities over an interval only depend on the duration of the interval).

Card ID: 1779140587319

Def: Transition Intensity Matrix Q

$$Q_{r,s} := q_{r,s} \text{ for } r \neq s$$

$$Q_{r,r} = -\sum_{s \neq r} q_{rs}$$

So rows of Q sum to zero.

Card ID: 1779140587321

Def: Sojourn Time

Distribution of time spent in state r before moving
Has an exponential distribution with rate

$$\lambda = \sum_{s,\, s \neq r} q_{rs} = -q_{rr}$$

So the average time in a state before switching is given by

$$\frac{1}{\lambda} = \frac{1}{\sum_{s,\, s \neq r} q_{rs}} = \frac{1}{-q_{rr}}$$

Card ID: 1779140587323

Probability that next state is s given current state is r in a continuous Markov process

---

$$\frac{q_{rs}}{\sum_{j,\, j \neq r} q_{rj}} = \frac{-q_{rs}}{q_{rr}}$$

Card ID: 1779140587325

Def: Transition Probability Matrix $P(t)$

$$P_{rs}(t) = \mathbb{P}(\text{State } s \text{ at time } t \,|\, \text{State } r \text{ at time } 0)$$

Solution to the forward Kolmogorov equations:

$$\frac{dP(t)}{dt} = P(t)Q(t)$$

Card ID: 1779140587327

How to solve for the transition probability matrix given $P(0) = I$, Q is time-homogeneous

---

$$P(t) = \exp(tQ) = \sum_{n=0}^{\infty} \frac{t_n}{n!} Q^n$$

Use an eigendecomposition for $Q = UDU$,

$$= \exp(tQ) = U \exp(Dt) U$$

Card ID: 1779140587329

Def: Expected time matrix $T(t)$ (in a continuous time Markov process)

Expected total length of time spent in state s, starting from state r:

$$T_{rs}(t) = \mathbb{E}\left[\int_0^t I_{X(u)=s} du\right] = \int_0^t P_{rs}(u) du$$

Card ID: 1779140587331

Expected number of visits to a state (in a continuous-time Markov process)

$$\sum_{i \neq s} \int_0^t P_{ri}(u) q_{is} du = \sum_{i \neq s} T_{ri}(t) q_{is}$$

Card ID: 1779140587333

How to find the probability of ever visiting a state (in a continuous-time Markov process)

Create a new problem by zeroing out the sth row of Q (a trap state), then just compute the new $P^*(t) = \exp(tQ^*)$.

Card ID: 1779140587335

Def: Likelihood contribution of discrete data in a continuous-time Markov process

Let a be x_{i0} (state before), and let b be x_{i1} (state after), then

$$\mathbb{P}(X_{i1} = b \,|\, a) = P_{ab}(t_{i1} - t_{i0} \,|\, \theta)$$

Take the product of these all to get the full likelihood.

Card ID: 1779140587337

Def: Likelihood contribution of a known time interruption (like death) in a continuous time Markov process

Problem: we don't know what state the individual was in prior to death. So we have to allow for every possibility:

$$\sum_{s \neq d} p_{rs}(t_{i,n_i} - t_{i,n_{i-1}}) q_{sd}$$

where d is the death state, and r is the last known pre-death state $x_{i,n_{i-1}}$.

Card ID: 1779140587338

Def: Likelihood contribution of a censored individual in a continuous time Markov process

Problem: We don't know what disease state the individual is in at the end of the time period, only that they never died (or else we would have known)

$$\sum_{s \neq d} P_{r,s}(t_{i,n_i} - t_{i,n_{i-1}}) = 1 - P_{r,d}(t_{i,n_i} - t_{i,n_{i-1}})$$

where d is the death state, and r is the last known pre-death state $x_{i,n_i - 1}$.

Card ID: 1779140587340

todo modeling misclassified observations, semi-markov models

2.2 Infectious Disease Modeling

Def: Assumptions made in the SIR model

$X_i(t)$ maps the ith individual to the state they are in at time t.
$S(t), I(t), R(t)$ are the number of people susceptible, infected, and recovered, respectively. ie,

$$S(t) = \sum_{i=1}^{N} I_{X_i(t)=S}$$

N is the total population, assume it is fixed, so that for all t,

$$N = S(t) + I(t) + R(t)$$

The rate of moving from S to I is denoted $\lambda(t)$, is usually equal to $\beta I(t)$.
The rate of moving from I to R is denoted $\gamma(t)$, usually just a constant γ

Card ID: 1779308511090

Def: Gillespie Algorithm

Continuous time discrete individual algorithm:
Two events:
someone is infected $((s, i) \mapsto (s-1, i+1))$
or someone recovers $((s, i) \mapsto (s, i-1))$.
Draw next time: $\sim \exp(\beta is + \gamma i)$,
At that next time, pick one of the two events (like a binomial): infected w.p. $\frac{\beta is}{\beta is + \gamma i}$

Card ID: 1779308511092

Def: Chain-Binomial Model

Discrete time discrete individual algorithm, sort of like a probabilistic ODE:
Take time steps of size $\delta : [t, t+\delta)$
$S_{t+\delta} = S_t - B_t$
$I_{t+\delta} = I_t + B_t - C_t$
$R_{t+\delta} = R_t + C_t$
where B_t, C_t are binomial RVs:

$$B_t \sim \text{Binom}(S_t, 1 - \exp(-\beta I_t \delta))$$

$$C_t \sim \text{Binom}(I_t, 1 - \exp(-\gamma\delta))$$

(where the exponential comes from ≥ 1 event in a Poisson RV)
We can simplify B_t if we assume $\beta\delta$ is small:

$$B_t \sim \text{Binom}(S_t, \beta I_t \delta)$$

Card ID: 1779308511094

Def: Reed-Frost Model

Given state (S_t, I_t), individuals have a latent (infected) time of 1, and at the end of their latent period, they try to infect every susceptible individual with probability p. So

$$I_{t+1} \sim \text{Binomial}(S_t, 1 - (1-p)^{I_t})$$

Useful for small populations like households.

Card ID: 1779308511095

Def: Greenwood Formulation

Chain binomial model where instead of the infection rate being proportional to the number of infected individuals, it's just p if any infected individual exists:

$$I_{t+1} \sim \text{Binomial}(S_t, p)$$

Card ID: 1779308511098

Def: Continuous Deterministic SIR Model

Assume we start with N susceptible individuals and 1 infected individual. System of ODEs:
$\frac{d}{dt}S(t) = -\lambda(t)S(t)$
$\frac{d}{dt}I(t) = \lambda(t)S(t) - \gamma I(t)$
$\frac{d}{dt}R(t) = \gamma I(t)$

Card ID: 1779308511099

Def: Threshold result for the SIR ODE system

We need $\frac{dI}{dt}(0) > 0$, so we must have $S(0) = N > \gamma/\beta$

Card ID: 1779308511100

Def: The reproductive number R_0

We need $\frac{dI}{dt}(0) > 0$, so we must have $S(0) = N > \gamma/\beta$

Define $R_0 = N\frac{\beta}{\gamma}$: number of secondary infections caused by one infected individual in a fully susceptible population (epidemic will only take off if at time zero, $R_0 > 1$.

Related also to $R_e(t)$, just the R_0 with the current $S(t)$ in place of N.

Card ID: 1779308511102

Def: Prevalence of an infection

$$\pi(t) = I(t)/N$$

Card ID: 1779308511104

todo environmental stochasticity, exponential growth rates, death data

2.3 Backcalculation for Infectious Disease Modeling

Def: Delayed Incubation Period Equation

$$\mu(t) = \int_0^t h(u) f(t-u) du$$

where $\mu(t)$ is the rate of newly observed cases at t, $h(t)$ is the rate of infections at t, and f is a probability distribution for the incubation length of the disease. Note this is a convolution!

Note this simplifies to the following in discrete time:

$$\mu_k = \sum_{i=1}^{k} h_i f_{k-i}$$

Card ID: 1779308511105

todo fitting to likelihood, time since infection models,

3 Part 3: Adaptive Trial Design

TODO

4 Part 4-6: Analysis of Survival Data

4.1 Introduction and Kaplan-Meier

Def: Censored Random Variable

$$X := \min(T, C)$$

$$V := \begin{cases} 1 & T \leq C \\ 0 & T > C \end{cases}$$

Where T is the time to event and C is the time to censoring. V is called the visibility.

Card ID: 1769639660334

Def: Survivor Function

Also its two properties

$$F(t) = \mathbb{P}(T > t) = \int_t^\infty f(t)\,dt$$

1. Decreasing in t
2. When continuous, $F(0) = 1$

NOTE: Our professor denotes the survivor function as $F(t)$, but the usual convention is to write $S(t)$.

Card ID: 1769639660338

TODO: Kaplan-Meier estimate form 1! Do we really need to know it?

Def: Risk Set

$$R = \{j : x_j \geq a_j\}$$

Note $r_j = |R|$
Note that this includes everyone that hasn't been censored yet right before t_j

Card ID: 1780146566287

Def: Event Set / Death Set

$$D = \{i : x_i = a_i \wedge v_i = 1\}$$

Note $d_j = |D|$ The number of individuals with (uncensored) events exactly at t_i

Card ID: 1780146566288

Def: Kaplan-Meier Estimate, Form 2

$$\hat{F}(t) = \prod_{j:a_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

with $d_j = |\{i : x_i = a_i \wedge v_i = 1\}|$ (non-censured death times at event time a_i)
and $r_j = |\{i : x_i \geq a_i\}|$ (all events at and after $t = a_i$; number at risk)

Note this is based on the erroneous assumptions that $a_i ... a_m$ are fixed constants and that T is discrete only with events at a_j. Though it ends up being computationally equivalent to method 1!
$\frac{d_j}{r_j}$ represents the probability of any event occurring at a_i (of the number of people at risk, how many were known to have perished in the time interval?)

Card ID: 1769686190875

Def: The hypothesis tested in the Log Rank Test

Null hypothesis: $F_j^{(0)} = F_j^{(1)}$ for all $j \in [m]$
Alternate hypothesis: $F_j^{(0)} \neq F_j^{(1)}$ for any $j \in [m]$

where $F_j^{(k)}$ represents the survivor function over $(a_j, a_{j+1}]$ for the kth group in $k \in \{0, 1\}$. (assume the a_j are all the event times of the two groups put together)

Card ID: 1769705013138

Aside: The Hypergeometric Distribution

Def: The Hypergeometric Distribution

The probability of k successes given n draws, where we draw without replacement from N objects with K that we want.

$$\text{Hypergeometric}(N, K, n)$$

(probability distribution over k)

Card ID: 1775594408080

Expectation of the hypergeometric distribution

$$\mathbb{E}(\text{Hypergeometric}(N, K, n)) = n\frac{K}{N}$$

Card ID: 1775594408081

Def: The Log-Rank Test

$$\frac{Z}{S} \underset{approx}{\sim} \mathrm{N}(0,1)$$

with

$$Z = \sum_j Z_j = \sum_j d_j^{(0)} - d_j \frac{r_j^{(0)}}{r_j}$$

Z is the number of observed (uncensored) events minus the number of expected events were H_0 true.

NONEXAMINABLE:

$$S^2 = \sum_j \frac{d_j(r_j - d_j) r_j^{(0)} r_j^{(1)}}{r_j^2 (r_j - 1)}$$

Card ID: 1769705013142

What's a good rule of thumb for when the Log-Rank test will be powerful?

The curves are well-seperated and don't cross

Card ID: 1769705013145

Def: Relative Risk

$$\begin{aligned} RR &= \frac{\text{sum of observed / sum of expected (in group 0)}}{\text{sum of observed / sum of expected (in group 1)}} \\ &= \frac{\sum_j d_j^{(0)} / \sum_j d_j \frac{r_j^{(0)}}{r_j}}{\sum_j d_j^{(1)} / \sum_j d_j \frac{r_j^{(1)}}{r_j}} \end{aligned}$$

Events are $RR\times$ more likely in group 0 than group 1.

Card ID: 1769705013148

4.2 Likelihood-Based Methods

Def: Accelerated-Life Families

A family of distributions generated by $F(\lambda t)$ for some scaling $\lambda > 0$.

Card ID: 1777034779211

Proportional-Hazards Families

A family of distributions generated by $F(t)^k$ for some shape parameter $k > 0$.

Card ID: 1777034779213

Def: Survivor Function of the Weibull Distribution

$$F(t) = e^{-(\theta t)^k}$$

where $\theta > 0$ is the shape parameter and $\theta > 0$ is the scale parameter.
Note that this reduces to the exponential RV when $k = 1$.
Weibull distributions are both an accelerated life and proportional hazards family!

Card ID: 1777034779215

Idea: Represent λ and k as functions of the explanatory variables, then optimize.

State the contribution to the likelihood for an observed event versus a censored event.

When $v_i = 1$ (uncensored), the contribution is the density $f(x_i, \theta)$.
When $v_i = 0$ (censored), the contribution is $\mathbb{P}(x_i < T_i) = F(x_i, \theta)$ (where F is the survivor function).
So the total likelihood is

$$L(\theta) = \prod_{v_i=1} f(x_i, \theta) \prod_{v_i=0} F(x_i, \theta)$$

Card ID: 1777034779217

PDF of the exponential distribution

$$f(t; \theta) = \theta e^{-\theta t}$$

given $\theta > 0$.

Card ID: 1777034779219

Survivor Function of the exponential distribution

$$F(t; \theta) = e^{-\theta t}$$

given $\theta > 0$.

Card ID: 1777034779221

Requirements for a nonparametric survival function $\tilde{F}(t)$

1. Bounded between zero and one: $0 \leq \tilde{F}(t) \leq 1$
2. Decreasing: where $a \leq b$, $\tilde{F}(a) \geq \tilde{F}(b)$

Card ID: 1777034779223

Contributions to a non-parametric survivor function likelihood

Where uncensored, we have:

$$\mathbb{P}(T = x_i) = \mathbb{P}(T \geq x_i) - \mathbb{P}(T > x_i) = F(x_i^-) - F(x_i) = F_i^- - F_i$$

where the minus subscript represents the limit from the left.

Where censored, we have:

$$\mathbb{P}(T > x_i) = F(x_i) = F_i$$

So for each event the likelihood contribution is

$$L(\theta) = \prod_{v_i=1} (F_i^- - F_i) \prod_{v_i=0} F_i$$

Card ID: 1777034779225

How can we simplify a non-parametric survivor function likelihood?

We can reduce the number of parameters by our assumptions:

$$F_1^- = 1$$

$$F_n = 0$$

Also the function should only change between F_i^- and F_i terms! (we're trying to build a step function; piecewise linear)
We should end up with a polynomial we can maximimize easily.

Card ID: 1777034779227

Non-parametric likelihood contributions of left-censored observations and interval-censored observations

Left-censored: $1 - F(x_i) = 1 - F_i$
Interval-censored on an interval (a, b): $F(a) - F(b)$

Card ID: 1777034779229

MLE for the log-likelihood for the exponential survivor function

$$\hat{\theta} = \frac{V_+}{X_+}$$

where V_+ is the total number of (non-censored) events, and X_+ is the total time spent at risk.

Card ID: 1780667438414

4.3 Hazard and Deriving Nelson-Aalen

Def: Hazard Function (raw form)

$$h(t) = \lim_{\Delta \to 0} \frac{\mathbb{P}(t < T < t + \Delta \,|\, T > t)}{\Delta}$$

You can simplify this to the form generally used using Bayes' Theorem and recognizing the definition of derivative:

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

Card ID: 1777034779231

Def: Hazard Function (practical form)

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

where $F(t)$ is the survivor function.

Card ID: 1777034779233

Def: Integrated Hazard

$$H(t) = \int_0^t h(t')dt'$$

Models the accumulation of hazard over time.

Card ID: 1777034779234

Survivor function as a function of integrated hazard

$$F(t) = e^{-H(t)}$$

or equivalently

$$H(t) = -\log(F(t))$$

Card ID: 1777034779235

Shift from talking about hazard as a property of a population to thinking about hazard as a property of an individual

Def: Counting Process

A function $N : \mathbb{R} \to \mathbb{N}$ that counts how many times an event has occurred. It always fulfills:

$$N(0) = 0$$

Always increasing
Always right-continuous (increments at an event time, not right after it)

Card ID: 1777034779236

Survival Analysis as a Counting Process

In survival analysis, we only care about the first time an event happens, so $N(t) \in \{0, 1\}$.

Putting it in our usual notation, we have:

$$N(t) = I\{X \leq t, V = 1\}$$

Card ID: 1777034779237

Def: History (of a stochastic process)

$\mathcal{H}_t$ denotes knowledge of everything that has happened in a stochastic process up to and including t

$\mathcal{H}_{t-}$ denotes knowledge up to but not including t.

For our purposes, $\mathcal{H}_{t-}$ is completely captured by membership to the risk set at t.

Card ID: 1777034779239

What do we mean by $dN(t)$?

It's called Stieltjes Notation:

$$dN(t) = N(t + dt) - N(t)$$

It can only take the values 0 or 1.

Card ID: 1777034779241

Def: Intensity $\lambda(t)$

$$\lambda_i(t) = Y_i(t)h_i(t)$$

where $Y_i(t) = I\{X_i \geq t\}$ (1 when person i is still in the risk set).
Let Λ be the integrated intensity, so equivalently we have

$$d\Lambda_i(t) = \sum_i Y_i(t)dH_i(t)$$

Define $\Lambda_+ = \sum_i \Lambda_i$ and $Y_+ = \sum_i Y_i$, and assume $H_i(t) = H(t)$ for all i (could also assume proportionality instead)

$$d\Lambda_+(t) = Y_+(t)dH(t)$$

Card ID: 1777034779243

Show that $dN(t) - d\Lambda(t)$ is a martingale

Work from the definition of the derivative of the integrated intensity:

$$\begin{aligned}\mathbb{P}(dN(t) = 1 \,|\, \mathcal{H}_{t-}) &= d\Lambda(t) \\ \mathbb{E}[dN(t) \,|\, \mathcal{H}_{t-}] &= d\Lambda(t) \\ \mathbb{E}[dN(t) - d\Lambda(t) \,|\, \mathcal{H}_{t-}] &= 0\end{aligned}$$

Card ID: 1777034779245

Given the per-individual martingale identity

$$\mathbb{E}[dN(t) - d\Lambda(t) \,|\, \mathcal{H}_{t-}] = 0,$$

show that the population derivative of integrated intensity equals the derivative of the population counting process at all t.

Define the population counting process and integrated hazard:

$$N_+(t) = \sum_i N_i(t)$$

$$\Lambda_+(t) = \sum_i \Lambda_i(t)$$

By our assumption, we have:

$$\begin{aligned}\mathbb{E}[dN(t)] - \mathbb{E}[d\Lambda(t) \,|\, \mathcal{H}_{t-}] &= 0 \\ \mathbb{E}[dN(t)] &= \mathbb{E}[d\Lambda(t) \,|\, \mathcal{H}_{t-}]\end{aligned}$$

So by the method of moments, we have:

$$dN_+(t) = d\Lambda_+(t)$$

Card ID: 1777034779247

Use $dN_+(t) = d\Lambda_+(t)$ and $d\Lambda_+(t) = Y_+(t)dH(t)$ to build the Nelson-Aalen Estimator (find an estimator $d\hat{H}(t)$)

Substituting in and isolating $d\hat{H}(t)$, we get:

$$d\hat{H}(t) = \frac{dN_+(s)}{Y_+(s)}$$

Integrate both sides:

$$\int_0^t d\hat{H}(s)ds = \hat{H}(t) = \int_0^t \frac{dN_+(s)}{Y_+(s)}ds$$

But $N_+(t)$ is just a step function that jumps at each failure time, so this integral is just a sum of 'snapshots' at each $t_j \leq t$:

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{1}{Y_+(t_j)}$$

Sum of snapshots of hazard increments.

Card ID: 1777034779249

Def: Nelson-Aalen Estimator

Given a set of events $\{a_1, ..., a_j, ..., a_n\}$ with no ties, we can estimate the cumulative hazard $H(t)$ as

$$\hat{H}(t) = \sum_{a_j \leq t} \frac{1}{Y_+(a_j)}$$

where $Y_+(t)$ is the number of people at risk at t (this is exactly the r_j used in Kaplan-Meier).

Note that we can estimate the survivor function using Nelson-Aalen:

$$\hat{F}(t) = \exp(-\hat{H}(t))$$

Card ID: 1777034779251

4.4 Modeling Hazard

Def: Semiparametric Modeling

Model the bits we care about explaining (β) with a parametric method, and model the bits we don't care so much about (ψ) with a nonparametric method.

Card ID: 1777034779253

Def: Proportional Hazards Modeling

Semiparametric method for modeling hazard. Assume that everyone has hazards in the same proportional hazards family:

$$h_2(t) = Kh_1(t),\ K > 0$$

Express K with the explainable parameters β, and express $h_0(t)$ with the nonparametric ψ:

$$h(t, z^i, \beta, \psi) = \phi(z^i, \beta)h_0(t, \psi)$$

Note that this is a strong assumption!

Card ID: 1777034779255

Def: Partial Likelihood

An expression of likelihood that neglects some of the data in order to solve for only some of the model parameters.

In Cox regression, it was shown that this is mathematically valid!

Card ID: 1777034779257

Def: (Partial) Likelihood in Proportional Hazards Modeling

$$L(\beta) = \prod_{i\,:\,v_i=1} \frac{\phi(z^i, \beta)}{\sum_{i' \in R_i} \phi(z^{i'}, \beta)}$$

where R_i is the risk set at t_i. Note that $h_0(x_i, \psi)$ gets cancelled out in the fraction.

Card ID: 1777034779259

Def: Cox Regression

Proportional hazards modeling with the following form for $\phi(z, \beta)$:

$$\phi(z, \beta) = e^{\beta^T z}$$

This yields

$$L(\beta) = \prod_{i\,:\,v_i=1} \frac{e^{\beta^T z^i}}{\sum_{i' \in R_i} e^{\beta^T z^{i'}}}$$

It's easy to take the log-likelihood then:

$$\log L(\beta) = \sum_i v_i \left(\beta^T z^i - \log \sum_{i' \in R_i} e^{\beta^T z^{i'}} \right)$$

And a derivative with respect to β:

$$\frac{d}{d\beta} \log L(\beta) = \sum_i v_i \left(z^i - \sum_{i' \in R_i} \frac{z^{i'} e^{\beta^T z^{i'}}}{e^{\beta^T z^{i'}}} \right)$$

Use Newton-Raphson from here.

Card ID: 1777034779262

The two kinds of ties in a Nelson-Aalen model, and how we can account of them

---

1. Tie by lack of precision
Suppose in the dataset ordered 1,2,3,4, 2 and 3 tied. Then we can account for this by just adding the two possibilities:

$$... \left(\frac{\phi_2}{\phi_2 + \phi_3 + \phi_4} \frac{\phi_3}{\phi_3 + \phi_4} + \frac{\phi_3}{\phi_2 + \phi_3 + \phi_4} \frac{\phi_2}{\phi_2 + \phi_4} \right) ...$$

We can approximate this with

$$... \left(\frac{\phi_2 \phi_3}{(\phi_2 + \phi_3 + \phi_4)(\frac{1}{2}\phi_2 + \frac{1}{2}\phi_3 + \phi_4)} \right) ...$$

2. Genuine Tie
Same example, use the following:

$$... \left(\frac{\phi_2 \phi_3}{\phi_2\phi_3 + \phi_2\phi_4 + \phi_3\phi_4} \right) ...$$

Card ID: 1778015739854

Def: Estimating Baseline Hazard

$$\hat{H}_0(t) = \sum_{i: x_i \leq t,\, v_i = 1} \frac{1}{\sum_{i' \in R_i} \hat{\phi}(z'_i, \beta)}$$

Card ID: 1778106432332

Def: Many Baseline Hazards Modeling

Suppose we had a treatment trail where individuals in the US and the UK are separately given the same medication. We expect the treatment effect to be the same, but the baseline risk to be different.

To model this: we say there are L strata, with the function $q(i)$ mapping the ith individual to their strata group. Then just use the custom risk set:

$$R_i := \{i' : x_{i'} \geq x_i \wedge q(i') = q(i)\}$$

For an event occurring to individual i, we only consider other individuals in the same strata (with known events occurring after that individual's events).

Estimate β by taking the product of each stratum's partial likelihood. Estimate the baseline hazard separately.

Card ID: 1778106432335

Def: Matched Pair Analysis Setup

Smallest possible stratification. We have two treatments $k \in \{0, 1\}$ and L pairs where one is exposed to treatment, and the other is not. Using our partial hazard split per individual:

$$h_{k,l}(t) = \phi_k h_l(t)$$

Clearly we can't estimate each individual $h_l(t)$, but we can estimate the ϕ_k shared among each pair.

Assume we have the baseline effect $\phi_0 = 1$, and the treatment effect $\phi_1 = e^{\beta}$ (or combined $\phi_k = e^{k\beta}$. β is called the log hazard ratio between treatments.

Card ID: 1778106432337

Def: Matched Pair Analysis Likelihood Contributions

Ignore pairs with ties: nothing useful we can learn from a tie. There's only two pieces of information we need from each pair l:

$$k(l) = I\{X_{1,l} < X_{0,l}\}$$

Just the index of whichever treatment fails first.

$$v(l) = V_{k(l),l}$$

0 if it was actually a censoring, not a failure; 1 if not.
This yields the full likelihood contribution:

$$\left(\frac{e^{k(l)\beta}}{1+e^{\beta}}\right)^{v(l)}$$

This ends up just being equivalent to a binomial likelihood!

Note that there's no likelihood contribution when:
- there's a tie
- one individual is censored before the other has an event
- both individuals are censored

Card ID: 1778106432338

When we graph log(survival) and log(log(survival)) over time, what are we graphing?

---

log(survival): Integrated hazard, slope is hazard
log(log(survival)): We'll see a constant vertical difference between two curves if their hazards are proportional.

Card ID: 1778106432340

Def: Probability Integral Transform for Hazard

Take the random variable T with integrated hazard $H(t)$. Then the random variable $U := H(T)$ is a $Exp(1)$ distribution.
Proof:

$$\begin{aligned}\mathbb{P}(u > U) &= \mathbb{P}(u > H(T)) \\ &= \mathbb{P}(H^{-1}(u) > T) \\ &= F(H^{-1}(T)) \\ &= e^{-H(H^{-1}(u))} \\ &= e^{-u}\end{aligned}$$

We can use this to quantify how accurate our models are!

Card ID: 1778106432342

Def: Cox-Snell Residual

Fit the model to construct $y_i = \hat{H}_i(x_i)$. By the PIT for Hazard, we expect $y_i \sim Exp(1)$.

Problem: we still may have censored individuals. So we can use a method like Kaplan-Meier to instead build a nonparametric representation of $\hat{F}_{y_i}$.

Graphically, $\hat{F}_y(y)$ should look like a straight line against y, with a slope of -1.

Card ID: 1778106432343

Trick for accounting for smaller-than-expected y_i mean in computing the Cox-Snell Residual

---

Add one to the y_is for each censored individual:

$$y_i' := y_i + (1 - v_i)$$

This comes from the memoryless property of the exponential:

$$\mathbb{E}[Y|Y + c] = c + 1$$

This strategy is called mean imputation.

Card ID: 1778106432344

Def: Martingale Residual

$$y_i'' = 1 - y_i' = v_i - \hat{H}_i(x_i)$$

We know this must have mean zero: $\mathbb{E}[y_i''] = 0$

Card ID: 1778106432346

How can we use Martingale plots as a means of designing an explanatory model?

---

1. Fit a model without using explanatory variables (z).
2. Plot the Martingale residuals (y_i'') against z: where the graph is above 0, we see more events than expected. For instance if you see a log plot, that implies you should probably use Cox regression.
3. Fit a fixed dataset and see if the means have been corrected towards zero.

Card ID: 1778106432347

4.5 Frailty

Def: Frailty

Each individual has some frailty u that describes how susceptible they are to risk. We denote the population risk $\bar{F}(t)$:

$$\bar{F} = \mathbb{E}_u[F(t \mid U = u)]$$

Note however that this doesn't imply that the population hazard $\bar{h}(t)$ is $\mathbb{E}_u[h(t \mid u)]$:

$$\bar{h}(t) = \frac{\bar{f}(t \mid u)}{\bar{F}(t \mid u)}$$

Card ID: 1778592619399

Def: Population Hazard under Proportional Frailty

$$h(t \mid U = u) = uh_0(t)$$

Card ID: 1778592619401

Def: Gamma Distribution PDF

$$Gamma(x; p, \lambda) = \frac{\lambda^p x^{p-1} e^{-\lambda x}}{\Gamma(p)}$$

p is called the shape or index parameter, λ is the rate parameter.

Card ID: 1778592619403

Mean of the Gamma Distribution

$$\frac{p}{\lambda}$$

Card ID: 1778592619405

Variance of the Gamma Distribution

$$\frac{p}{\lambda^2}$$

Card ID: 1778592619406

Def: Laplace Transform of the Gamma Distribution

With $Gamma(x; p, \lambda)$:

$$\left(\frac{p}{1+p}\right)^{\lambda}$$

Card ID: 1778592619408

Def: Population Survivor under Gamma-distributed Frailty

Assume the density of u is given by $g(u)$, with $\mathbb{E}[g(u)] = 1$. Then we have:

$$\mathbb{E}_u[F(t)] = \mathbb{E}_u[e^{-H(t\,|\,u)}] = \int_0^\infty \exp(-uH_0(t))g(u)du$$

Recognizing the Laplace transform and plugging in $u \sim Gamma(p, \lambda)$:

$$\bar{F}(t) = \left(\frac{p}{p + H_0(t)}\right)^\lambda$$

Usually we want the expected value of $g(u)$ to be one, so we normally use $p = \lambda$, denoted ψ.

Card ID: 1778592619410

Def: Population Hazard under Gamma-distributed Frailty

$$\bar{h}(t) = \frac{ph_0(t)}{p + H_0(t)}$$

A lower $p = \lambda = \psi$ means a higher variance among individuals: they either die off quickly or stick around.

Card ID: 1778592619412

Def: Cox Regression with Frailty

$$h_z(t\,|\,U = u) = ue^{\beta z}h_0(t)$$

You can use this to substitute into population hazard:

$$\bar{h}(t) = \frac{pe^{\beta t}h_0(t)}{p + e^{\beta z}H_0(t)}$$

Note then the population hazard ratio will start at e^β at $t = 0$, but then converge to 1 as $t \to \infty$.

Card ID: 1778592619414

What is the central problem with frailty in practice? How can we resolve this?

There's not enough information in (X_i, V_i) pairs alone to assess g and H_0 separately. Three resolutions:

1. Be aware that we can't tell apart a homogeneous population with decreasing hazard from a population varying in frailty but with constant hazard over time.

2. Measure as many explanatory variables as possible

3. Use a model where frailty goes into the error term (accelerated frailty):

$$F(t \mid u) = F_0(ue^{\beta z}t)$$

Card ID: 1778592619416

4.6 Competing Hazards

The two approaches to competing risks modeling

---

1. Two different times to events, T_A and T_b, we only observe one.

2. One time-to-event variable T, and at event time, choose between event type A or B.

Card ID: 1778592619417

What do $F(t)$, $\tilde{f}_A(t)$, and $G_A(t)$ represent in competing hazards modeling, and how do we compute them?

$F(t)$ is just the probability of any event before t, so

$$F(t) = \exp(-H_A(t) - H_B(t))$$

$\tilde{f}_A(t)$ is the event density at t (not a proper distribution):

$$\tilde{f}_A(t) = h_A(t)F(t)$$

$G_A(t)$ is the probability of a the specific event A occurring before or at t, so

$$G_A(t) = \int_0^t \tilde{f}_A(t')dt'$$

Note we can also compute $\tilde{F}_A(t) = \exp(-H_A(t))$, the survivor function for A if the event B entirely disappeared.

Card ID: 1778592619419

Def: Aalen-Johansen Estimator

$$\hat{G}_A(t) = \sum_{j:a_j \leq t} \left[\prod_{j'=1}^{j'=j-1} (1 - \Delta\hat{H}_{j'}) \right] \Delta\hat{H}_j^A$$

where

$$\Delta\hat{H}_j = \frac{1}{r_j}$$

$$\Delta\hat{H}_j^A = \frac{1}{r_j} \mathbb{1}_{E_j=A}$$

Card ID: 1778592619420

4.7 Net Survival

Net Survival Setup

Want to consider excess deaths (E) versus the chance of a background event that has nothing to do with the disease we're studying (B).

$$h^i(t) = h^i_B(t) + h^i_E(t)$$

We want an unbiased estimator for

$$d\bar{H}_E(t) = \frac{\sum_i F^i_E(t) dH^i_E(t)}{\sum_i F^i_E(t)}$$

Incremental excess integrated population hazard

(Note that we assume $h^i_B(t)$ is known from government data)

Card ID: 1778592619422

Def: Net Survival Integral

$$F_{net} = \exp \int_0^t d\bar{H}_E(s) ds$$

What fraction of patients would survive to time t if it were the only possible cause of death?

Card ID: 1779140587369

Def: Pohar-Perme Estimator

$$d\tilde{H}_E(t) = \frac{\sum_i dN_i(t)/F^i_B(t)}{\sum_i Y_i(t)/F^i_B(t)} - \frac{\sum_i Y_i(t) dH^i_B(t)/F^i_B(t)}{\sum_i Y_i(t)/F^i_B(t)}$$

Card ID: 1778592619423

4.8 Period Survival

How can we make up for left-truncated data?

Divide by $F(S_i)$ in the likelihood contributions (if the individual had had the event before S_i, they would not have been included in the dataset)

Card ID: 1779140587371

Accounting for piecewise hazard rates

Divide time into intervals; use left-truncation on all but the first region and right-censor any observation after the cutoff time.

$$\hat{\theta}_{interval} = \frac{\text{number of events in interval}}{\text{sum of all time at risk in interval}}$$

Card ID: 1779140587373

Def: Best way to account for left-truncated data in practice

Kaplan-meier accounts for this already; just make sure that left-truncated individuals aren't considered as part of the risk set until their introduction.

This gives the same estimator as the likelihood and piecewise approaches.

Card ID: 1779140587375

Def: Period Survival

We want to characterize survival rates in some restricted period, say 2024. Denote the interval of interest $(a, b]$.

Convert calendar time (time u, y_i diagnosis date, z_i event observation date, w_i event or censored) into individual's time (s_i truncation time, x_i event/censoring time, v_i event or censored)

Only include individuals that spend some amount of time in the interval pre-event. Left-truncate time before interval cutoff; censor events that occur after the time cutoff.

Card ID: 1779140587377