Department of Statistics, Stanford University

"Regression models and life tables" (D.R. Cox)

Joshua Kazdan & Ivy Zhang

March 17, 2024

The problem

You want to buy a 10-year life insurance plan from Farmer's with a payout of \$2M. Farmer's needs to compute the probability that you will die within 10 years so that they can price your insurance and maintain a positive expected value (with accrued interest from your payments).

1. What is the probability that you will die within 10 years?
2. How does this probability change when we factor in your specific medical profile/risk-taking tendencies?

In 1950, how would Famer's have computed the probability that you'd survive another 10 years?

Life tables calculate the probability of survival to time t by the proportion of the initial cohort that survived to time t. But life tables hide the difficulties of computing these probabilities.

1. Censored data
2. Extrapolation beyond the age of the oldest surviving member

Exact age	Male			Female		
	Death probability [a]	Number of lives [b]	Life expectancy	Death probability [a]	Number of lives [b]	Life expectancy
0	0.005837	100,000	74.12	0.004907	100,000	79.78
1	0.000410	99,416	73.55	0.000316	99,509	79.17
2	0.000254	99,376	72.58	0.000196	99,478	78.19
3	0.000207	99,350	71.60	0.000160	99,458	77.21
4	0.000167	99,330	70.62	0.000129	99,442	76.22
5	0.000141	99,313	69.63	0.000109	99,430	75.23
6	0.000123	99,299	68.64	0.000100	99,419	74.24
7	0.000113	99,287	67.65	0.000096	99,409	73.25
8	0.000108	99,276	66.65	0.000092	99,399	72.25
9	0.000114	99,265	65.66	0.000089	99,390	71.26
10	0.000127	99,254	64.67	0.000092	99,381	70.27
11	0.000146	99,241	63.68	0.000104	99,372	69.27
12	0.000174	99,227	62.69	0.000123	99,362	68.28
13	0.000228	99,209	61.70	0.000145	99,349	67.29
14	0.000312	99,187	60.71	0.000173	99,335	66.30
15	0.000435	99,156	59.73	0.000210	99,318	65.31

Naive estimates

$S(t)$ is the probability of surviving up until time t. $T_i = \min(\tilde{T}_i, C_i)$ (i.e. the time of either death or censoring, whichever comes first).

1. $\hat{S}_1(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(T_i > t)$: biased downwards because there's a smaller risk of censoring for lower times.
2. $\hat{S}_2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(T_i > t, \Delta_i = 1)$ (i.e. only look at uncensored data): biased because dividing by n but only considering uncensored data.
3. $\hat{S}_3(t) = \frac{1}{\sum_j \Delta_j}\sum_{i=1}^{n} \mathbf{1}(T_i > t, \Delta_i = 1)$: doesn't take into account the time dependency of the censoring process.

Bias of $\hat{S}_3$

1. Let F be the CDF of the failure time and G be the CDF of the censoring time.
2. $\mathbb{P}[T > t, \Delta = 1] = S(t) - \int_t^\infty G(u)dF(u)$.
3. $\mathbb{P}[T > t|\Delta = 1] = \frac{S(t) - \int_t^\infty G(u)dF(u)}{1 - \int_0^\infty G(u)dF(u)}$.
4. Subtract $S(t)$ from both sides:

$$\mathbb{P}[T > t|\Delta = 1] - S(t) = \frac{\int_0^t G(u)dF(u) - F(t)\int_0^\infty G(u)dF(u)}{1 - \int_0^\infty G(u)dF(u)}. \quad (1)$$

5. (1) is 0 at $t = 0$ and ∞. Analyzing the derivative shows that it is decreasing up to some t_* and then increasing between t_* and ∞.

Unbiased estimation of survival curves

Kaplan-Meier Curves:

1. $$p_t = \mathbb{P}(\text{survive to time } t+1 | \text{survived to time } t) = 1 - \frac{\text{uncensored people who died in the interval } [t, t+1)}{\text{uncensored people who were alive at time } t}$$
2. $\mathbb{P}(\text{survive to time } t) = \prod_{s=1}^{t-1} p_t$.

Unbiased estimation of survival curves

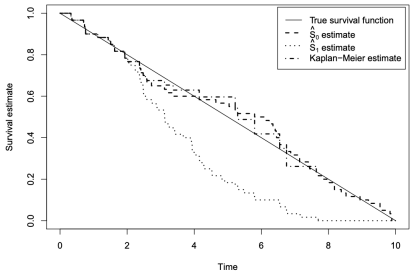
1. Even though it's difficult to come up with an unbiased estimate of survival, it's not difficult to find an unbiased estimate of the hazard rate

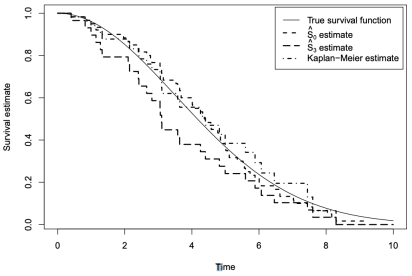
$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t]}{\Delta t}.$$

2. In continuous time,

$$\lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t + \Delta t, \Delta = 1 | T \geq t]}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq \tilde{T} < t + \Delta t | \tilde{T} \geq t]}{\Delta t}.$$

Unbiased estiamtion

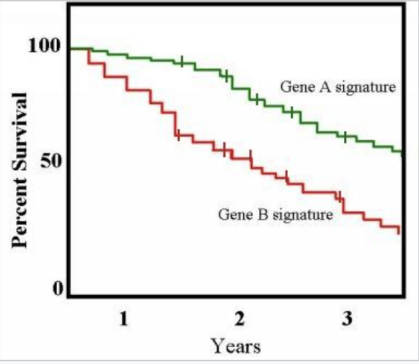




Unbiased estimation of survival curves

Kaplan Meier curves

1. Fix the issue of censoring
2. Still can't project probabilities beyond the time of death for the last living person in the cohort.
3. Can't consider other factors (i.e. heart disease, riding a motorcycle, etc.).



An example of a Kaplan–Meier plot for two conditions associated with patient survival.

Hypothesis testing

Suppose that we wish to determine the effect of age on the risk of dying from breast cancer using the following data:

TABLE 1. BREAST CANCER DATA

Patient no.	Age (yr)	Age group	Observation time (days)	Censored?
1	68	2	413	No
2	54	1	701	Yes
3	72	2	1075	No
4	60	1	1735	No
5	70	2	1801	No
6	50	1	2989	No
7	73	2	3044	No
8	71	2	3351	No
9	69	2	5551	No
10	61	2	6277	No
11	32	1	7293	Yes
12	46	1	7352	Yes
13	55	1	7434	Yes

Hypothesis testing

Prior to the Cox model, the standard method would have been through a log-rank test.

1. Split into two groups greater than 60 and less than 60.
2. Suppose that the means are the same and the expected numbers of deaths are $E_{>60}$ and $E_{<60}$. The observed deaths are $O_{>60}$ and $O_{<60}$.
3. The statistic of interest is $(O_{>60} - E_{>60})^2 + (O_{<60} - E_{<60})^2 = 6.92$.
4. The variance of this statistic is 2.17.
5. The statistic $L = 6.92/2.17 = 3.19$, and a χ^2 table will give a p-value of 0.07

Hypothesis testing

The problems with a test like this:

1. It ignores the exact ages of death by splitting into just two groups of < 60 and > 60.
2. It is unlikely that there is a turning point for breast cancer at age 60: more likely the probability of dying from breast cancer depends continuously on age.

Hazards

1. For an individual with characteristics given by z, we want to assess the impact of z on risk of death
2. Cox modelled hazards by

$$\lambda(t; z) = \exp(z\beta)\lambda_0(t)$$

   with parameters β.
3. So the probability of survival is

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

Proportional Hazards

1. By Cox's parameterization of a hazard, the effect of t is independent of z.
2. Conditional on one death occurring at time t, the probability that patient i died is

$$\frac{\exp(z_{(i)}\beta)}{\sum_{j \text{ still alive}} \exp(z_{(j)}\beta)}.$$

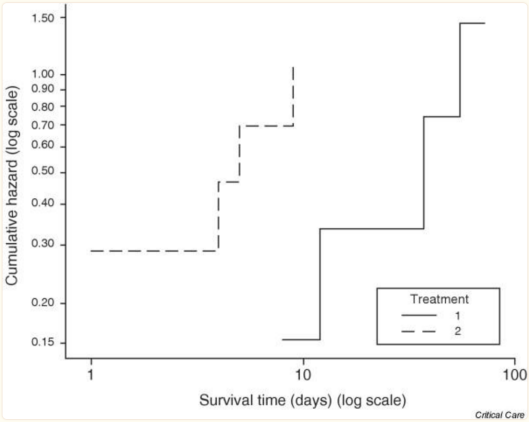
3. This observation gives rise to a partial likelihood

$$\sum_{i=1}^{k} z_{(i)}\beta - \sum_{i=1}^{k} \log\left(\sum_{j \in R(t_i)} \exp(z_{(j)}\beta)\right)$$

4. Can find β by maximizing the partial likelihood.
5. Note that censored data doesn't impact the model– it just doesn't show up in the denominator when calculating likelihood.

Key assumption

The effect on the probability of death is constant throughout time because the difference in the log cumulative hazard functions remains roughly constant.



Hypothesis testing

The CoxPH model can be used to estimate the probability that a given variable has an effect on mortality.

1. In the breast cancer example, to estimate the effect of being 60 versus 70 on dying from breast cancer, one can compute $\beta_{\text{age}} = 0.0096$ with a significance of 0.01.
2. For binary variables (i.e. smokes/doesn't smoke), the test resulting from the CoxPH model agrees exactly with the result of a log-rank test.

"The applications are more likely to be in industrial reliability studies and medical statistics than in actuarial science."

1. The model is a powerful tool for assessing the effect of a given factor on the time of failure.
2. The model doesn't actually predict the time or probability of failure, only relative probabilities. Therefore, it's not necessarily useful for trying to assess the probability of death within a certain time period.

Back to insurance

1. Insurance doesn't care whether eating deep-fried Twinkies makes you more likely to die; they only care how much extra they should charge you for each Twinkie that you eat.
2. Since the reliability of Cox models for probability estimation depends on the choice of $\lambda_0(t)$, they can't always be used to set premiums. However, they might be useful for scaling premiums.

Back to insurance

1. Let P be the standard premium that someone who doesn't eat Twinkies has to pay.
2. Under the assumptions of the Cox model, the premium for someone who does eat Twinkies should be

$$\exp(\beta_{\text{twinkies}})P.$$

Summarizing Advantages and Limitations

Advantages

1. Semiparametric method
2. Computationally quick
3. Incorporates censoring

Summarizing Advantages and Limitations

Advantages

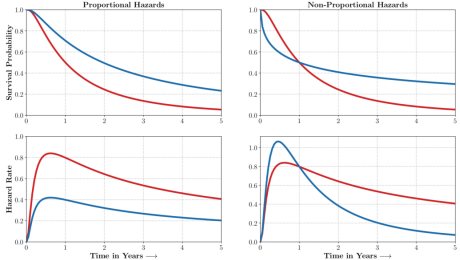
1. Semiparametric method
2. Computationally quick
3. Incorporates censoring

Limitations

1. Proportional hazards assumption
2. Incorporating time-dependent covariates loses some beauty
3. Only one failure event

Proportional hazards assumption

Hazard at time t of a person with covariate X = Constant(t) × Constant(x)



Examples of violation*:

- Resistance to therapy: **converging hazards** for treatment vs control groups
- Surgery in oncology has higher initial risk, but better long-term prognosis while radiation hazards start out lower, but grow over time: **crossing hazards**
- Effects of a variable growing more pronounced over time: **diverging hazards**

*** 110/318 2019 PubMed database TJA studies and 11/58 oncology RCT with PH violations**

Alternatives

Model set-up

1. Stratification: by group or by covariate
2. Separate models for disjoint time periods
3. Time dependent covariates (Cox)

Other models

1. Weighted Cox
2. Parametric models: Accelerated Failure Time
3. Nonparametric model: Random Survival Forest; Deep-Learning Based Techniques

Alternatives

Model set-up

1. **Stratification: by group or by covariate**
2. Separate models for disjoint time periods
3. **Time dependent covariates** (Cox)

Other models

1. Extensions: Weighted Cox
2. Parametric models: Accelerated Failure Time
3. Nonparametric model: **Random Survival Forest**; Deep-Learning Based Techniques

Stratified Cox Model

Consider the following example:

$$W: \text{ treatment group } = \begin{cases} 0 \text{ control} \\ 1 \text{ treatment} \end{cases} \qquad X: \text{ sex } = \begin{cases} 0 \text{ female} \\ 1 \text{ male} \end{cases}$$

The usual Cox model of $h(t|W,X) = h_0(t)e^{\beta_1 W + \beta_2 X}$ assume proportional hazards ie

$$\text{Treatment Hazard Ratio} = \frac{h(t|W=1,X)}{h(t|W=0,X)} = e^{\beta_1}$$

$$\text{Sex Hazard Ratio} = \frac{h(t|W,X=1)}{h(t|W,X=1)} = e^{\beta_2}$$

What if the HR for male and female are not proportional in reality?

Stratified Cox Model

Instead, we can model it as:

$$h(t|W, X) = h_X(t)e^{\beta W}$$

"Treatment has the same e^{β} HR within each level of X but the underlying hazards for female and male are arbitrary"

Let $L^{(x)}(\beta)$ be the usual partial likelihood of the of the group $X = x$. Then this model has the following simple overall partial likelihood:

$$L(\beta) = \prod_{x=1}^{|X|} L^{(x)}(\beta)$$

Extended Cox Model (Time Dependent Covariates)

Let T: time until death by heart attack

$$W: \text{ treatment group } = \begin{cases} 0 \text{ control} \\ 1 \text{ treatment} \end{cases} \qquad X(t): \ \# \text{ of times R crashes}$$

has model $h(t|W, X(t)) = h_0(t)e^{\beta_1 W + \beta_2 X(t)}$

1. Everyone has a hazard function that begins at $h_0(t)$
2. Allows for hazard to vary with time
3. Estimation just as in fixed-covariate case

Extended Cox Model (Time Dependent Covariates)

Proceed with caution!

1. Make sure not to use data from the future
   - Ex: a measurement that comes every 3 months can be used to model 4 month survival but not 2 month survival
2. Loss of the interpretation of: "individual i has a predicted estimated time to event of x"
   - Might not know what the future value of a quantity is
3. Use of some covariates may mean we cannot estimate the survival curve
   - When the existence of some covariate implies event not occurred
   - Ex: if we have a measurement of blood pressure, then they didn't die yet

Alternatives

Model set-up

1. **Stratification: by group or by covariate**
2. Separate models for disjoint time periods
3. **Time dependent covariates** (Cox)

Other models

1. Extensions: Weighted Cox
2. Parametric models: Accelerated Failure Time
3. Nonparametric model: **Random Survival Forest**; Deep-Learning Based Techniques

Random Survival Forest

1. Variation of RF, incorporating survival time and censoring information into the splitting criterion
2. Just like RF, constructed from ensemble of binary decision trees
   - Each tree is built from a random bootstrap sample through a node splitting process
   - At each split, a number of covariates are considered for split candidates and one is chosen based on which maximizes the difference between the number of people who have reached an event vs those who have not at the time of the daughter nodes, similar to the impurity measure used in RF
   - Iterative split until all the uncensored events have happened
3. The hazard function estimates, used for predicting survival at a point t, from each tree are calculated and then averaged over the whole forest to get the final estimate

RSF vs Cox-PH: Nonlinear + interaction effects under PH

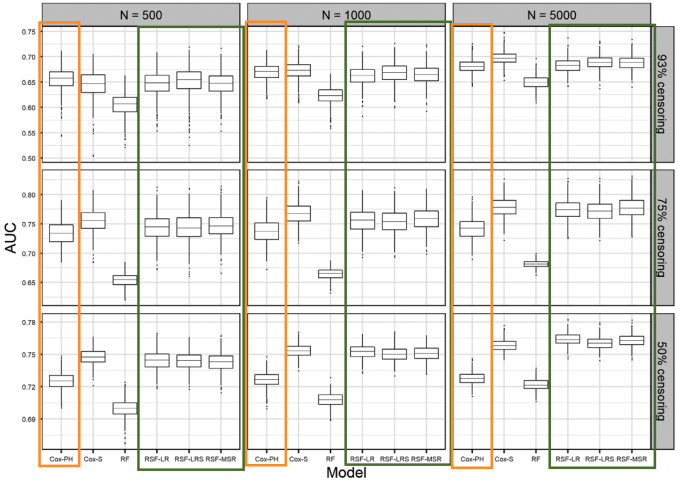


FIGURE 3 Predictive performance in terms of time-dependent AUC under the scenario with both nonlinear and interaction effects and increasing hazard function when $N = 500, 1000, 5000$ and censoring rates 93%, 75%, and 50%

Considerations

1. No parameter tuning
2. Simulated data assumes proportional hazards
3. Another simulation study showed superior results in RSF

RSF vs Cox, not PH

For mortality of patients with hemorrhagic stroke

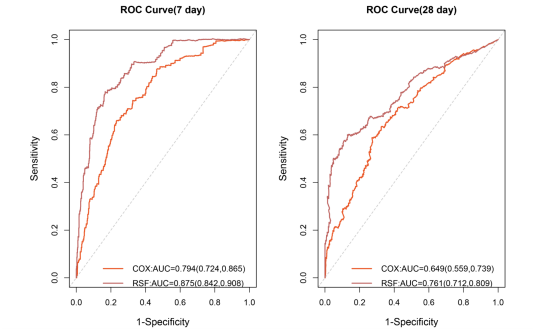


Fig. 4 ROC curve for predicting 7-day and 28-day hospital mortality

For High-Grade Glioma after Proton and Carbon Ion Radiotherapy

1. Cox C-index: 62.9 %
2. RSF C-index: 61.1%

Summarizing Advantages and Limitations

Advantages

1. Semiparametric method
2. Computationally quick
3. Incorporates censoring

Limitations

1. Proportional hazards assumption ✓
2. Incorporating time-dependent covariates loses some beauty ✓
3. Only one failure event

Competing Risks (Prentice et al. 1978)

Imagine we are oncologists who wants to understand not just time until death of a patient but time until death or time until onset of second primary cancer.

One way to extend the Cox model to this scenario is to model *cause-specific hazard*:

$$h_j(t|X) = h_{0j}(t)\exp(X'\beta_j)$$

1. Each cause has a different baseline hazard and HR
2. Like in the stratified case, the overall partial likelihood is the product of the j different causes' likelihoods
3. Interpretation of HR: cause-specific hazard had the competing risks not occurred