Descriptive Statistics
Meet Machine Learning

Jon A. Wellner

University of Washington, Seattle

Breslow Lecture, Seattle, November 12, 2020

Breslow Lecture

Alternative Titles:

- (Nearly) 50 years of the Cox model.
- Survival Analysis: Inference versus prediction?

Part I: Parameters defined by models (& estimators)

- A. Kullback - Leibler divergence and projections
- B. Questions / properties
- C. Four examples

Part II: Extensions or enlargements of the Cox model.

- Parametric relative risk
- Semiparametric relative risk
- Nonparametric relative risk.
- Nonparametric

Part III. Descriptive statistics (parameters) by design

- Bickel and Lehmann:
  Descriptive Statistics for Nonparametric Models,
  I (1975), II (1976), III (1976), IV - (1979).
- Buja, Brown, . . .: Models as Approximations I & II:
  Statistical Science (2019) with discussions;
  A Model-Free theory of parametric regression
- Followup(s) in survival analysis?
- Estimators that can be "plugged in"?
  Bickel and Ritov (2003).

Part I: Parameters defined by projection on models

- Model misspecification: the basics
- Probability Model:

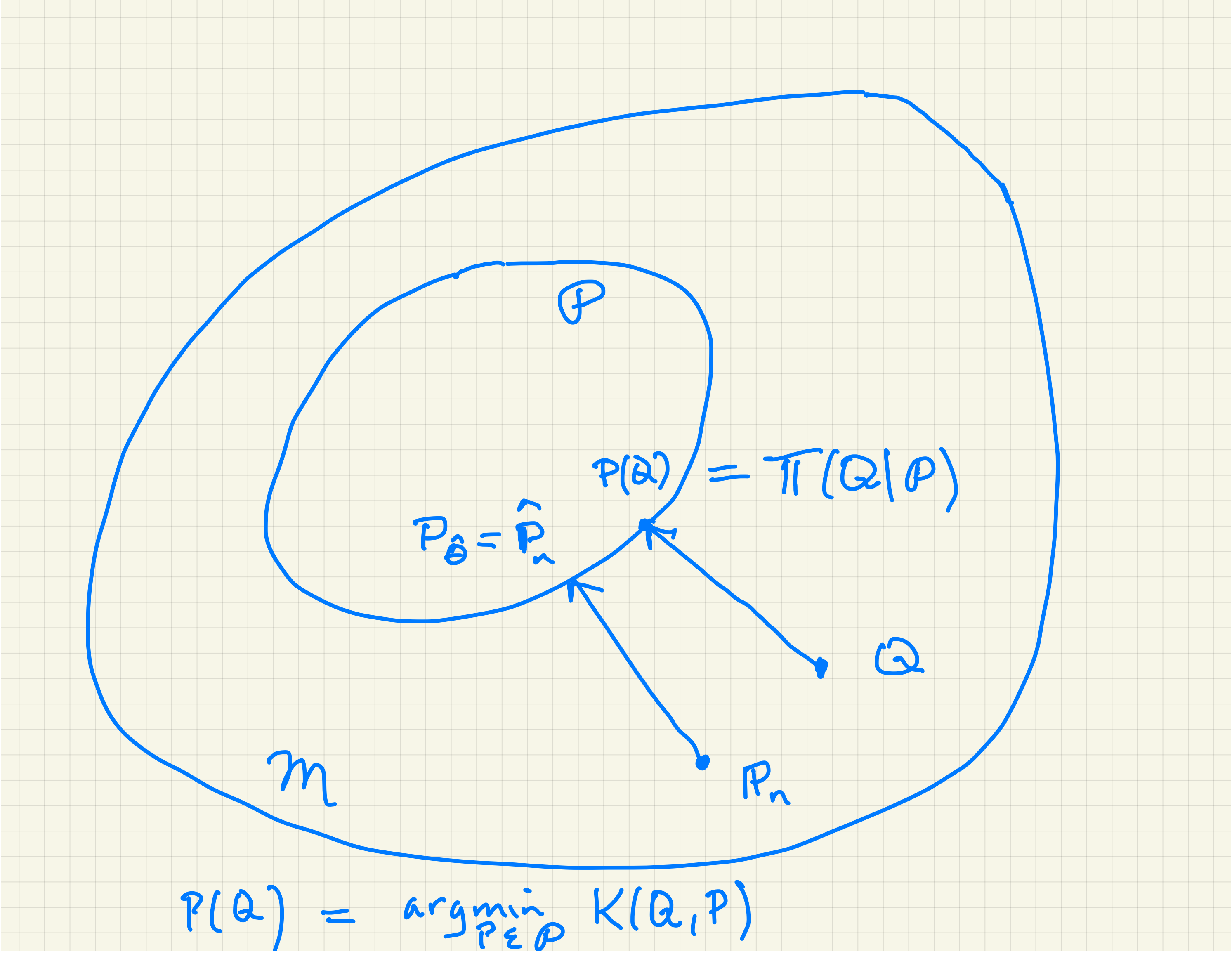
$$\mathcal{P} \subset \mathcal{M} = \text{all probability measures on } \ (\mathcal{X}, \mathcal{A}).$$

  Often $\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$ for some "parameter space" Θ.

- "true" $P =$ the P that generated the data $\equiv Q$.
- Let $K(Q,P)$ be the Kullback-Leibler divergence between $Q \in \mathcal{M}$ and $P \in \mathcal{P}$:

$$K(Q,P) = E_Q \text{log} \frac{dQ}{dP} = E_Q \text{log} \frac{q}{p}$$

  where $q = dQ/d\mu$, $p = dP/d\mu$ where μ dominates Q and P.


$\mathcal{P}$
$P(Q) = \Pi(Q \mid \mathcal{P})$
$P_{\hat\theta} = \hat{P}_n$
Q
$\mathbb{P}_n$
$\mathcal{M}$

$$P(Q) = \operatorname*{argmin}_{P \in \mathcal{P}} K(Q, P)$$

Let

$$\begin{aligned} P(Q) &\equiv \Pi(Q|\mathcal{P}) \equiv \text{argmin}\{K(Q,P): \ P \in \mathcal{P}\}, \\ P_\theta(Q) &\equiv \Pi(Q|\mathcal{P}) \quad \text{when} \quad \mathcal{P} = \{P_\theta : \theta \in \Theta\}. \end{aligned}$$

This is called the Kullback-Leibler (or maximum likelihood) projection of Q onto $\mathcal{P}$, since at least heuristically,

$$\widehat{\theta}_n = \text{argmin}_{\theta\in\Theta} K(\mathbb{P}_n, P_\theta)$$

where $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure of $X_1, \ldots, X_n$. This is the MLE considered as an "M- estimator". Note that

$$K(Q,P) = E_Q \text{log} \frac{dQ}{dP} = E_Q \text{log} \frac{dQ}{d\mu} - E_Q \text{log} \frac{dP}{d\mu},$$

so maximizing $E_Q \text{log} dP/d\mu$ over $\mathcal{P}$ minimizes $K(Q,P)$ over $P \in \mathcal{P}$.

Assuming that P_θ has density p_θ (w.r.t. μ, $\theta \in \Theta \subset \mathbb{R}^d$) is suitably differentiable w.r.t, θ, then we typically find that $\widehat{\theta}_n$ is characterized via a system of score equations:

$$0 = \mathbb{P}_n \dot{\ell}_\theta(X) = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_\theta(X_i).$$

where

$$\dot{\ell}_\theta(x) = \nabla_\theta \text{log} p_\theta(x)$$

is the vector of scores. The corresponding (population) equations characterizing $\Pi(Q|\mathcal{P})$ and $\theta(Q) = \text{argmin}_{\theta\in\Theta} K(Q, P_\theta)$ are simply

$$0 = Q\dot{\ell}_\theta(X).$$

This is the "$Z-$ estimator" view of the MLE and its corresponding population version.

History: Building on Wald (1949)

- Peter Huber (1967)
- H. Akaike (1973)
- Halbert White (1982)
- Aad van der Vaart (1995)
- Valentin Patilea (2001) .
- R.H. Berk (1966, 1970)
- van der Vaart & Kleijn (2006,2012)

White (1982) writes:

> *If one does not assume that the probability model is correctly specified, it is natural to ask what happens to the properties of the maximum likelihood estimator. Does it still converge to some limit asymptotically, and does this limit have any meaning?* . . .

Questions: $\Pi(Q|\mathcal{P})$ and $\theta(Q)$.

- 0. Do they exist for all $Q \in \mathcal{M}$ large? Or for $\mathcal{M}$ large enough to include $\mathbb{P}_n$? Are they unique?
- 1. Is $\widehat{P}_n \equiv \Pi(\mathbb{P}_n|\mathcal{P})$ computable?
- 2. Is the map $Q \mapsto \Pi(Q|\mathcal{P})$ continuous on $\mathcal{M}$?
- 3. Is the map $Q \mapsto \Pi(Q|\mathcal{P})$ Lipschitz? Differentiable?
- 4. Are $\theta(Q)$ and/or $P(Q) = \Pi(Q|\mathcal{P})$ interpretable?

Four Examples:

- **Example 1.** $\mathcal{P}$: Exponential distributions on $\mathbb{R}^+$. $\Theta = \mathbb{R}^+$.
  $\mathcal{M} = \{Q \text{ on } \mathbb{R}^+, Q(X) = \int_0^\infty x dQ(x) < \infty\}$.

- **Example 2.** $\mathcal{P}$: Weibull distributions on $\mathbb{R}^+$. $\Theta = \mathbb{R}^+ \times \mathbb{R}^+$.
  $\mathcal{M} = \{Q \text{ on } \mathbb{R}^+, Q(X \log X) < \infty\}$.

- **Example 3.** $\mathcal{P}$: log-concave distributions on $\mathbb{R}^d$: $dP/d\mu = p(x) = e^{\theta(x)}$ where θ is concave:
  $\Theta = \{\theta : \mathbb{R}^d \to \mathbb{R}, \theta \text{ concave with } \int e^\theta d\mu = 1\}$.
  $\mathcal{M} = \{Q \text{ on } \mathbb{R}^d : Q(\|X\|) < \infty; Q(H) < 1 \text{ for all hyperplanes } H\}$.

- **Example 4.** $\mathcal{P}$: The (classical) Cox model for right - censored survival data with covariates, (T, Δ, Z) where $T = \min\{X, Y\}$, $\Delta = 1_{[X \le Y]}$, X and Y are conditionally independent given Z.

Example 1. Here $p_\theta(x) = \theta^{-1}\text{exp}(-x/\theta)$, so

$$\dot{\ell}_\theta(x) = -\frac{1}{\theta} + \frac{x}{\theta^2} = \theta^{-2}(x - \theta),$$

so the solution of $0 = Q\dot{\ell}_\theta(X) = \theta^{-2}(Q(X) - \theta)$ is given by $\theta(Q) = E_Q X = Q(X)$. Here the projection map $\theta(Q)$ is simply the mean of Q, an easily interpretable parameter. This is connected with the fact that the model $\mathcal{P}$ is an *exponential family*.

Example 2. $\mathcal{P}$ = Weibull: Here $\theta = (\alpha, \beta)$, the density is $p_{\alpha,\beta}(x) = (\beta/\alpha)(x/\alpha)^{\beta-1}\text{exp}(-(x/\alpha)^{\beta})$, and the resulting score functions are

$$\begin{aligned} \dot{\ell}_{\alpha}(x) &= \frac{\beta}{\alpha}\left\{\left(\frac{x}{\alpha}\right)^{\beta} - 1\right\} \\ \dot{\ell}_{\beta}(x) &= \frac{1}{\beta}\left\{1 - \log\left(\frac{x}{\alpha}\right)^{\beta}\left(\left(\frac{x}{\alpha}\right)^{\beta} - 1\right)\right\}. \end{aligned}$$

So the projection $P_{\theta(Q)}$ is given by $\theta(Q) = (\alpha(Q), \beta(Q))$ solving

$$\underline{0} = Q(\underline{\dot{\ell}}_{\theta}) = \begin{pmatrix} Q\dot{\ell}_{\alpha} \\ Q\dot{\ell}_{\beta} \end{pmatrix}.$$

Each fixed $\beta > 0$ the first equation can be solved explicitly: $\alpha_{\beta}(Q) = \{Q(X^{\beta})\}^{1/\beta}$. Substitution of this into the second equation (just as in the derivation of the profile likelihood estimator) shows that

$$K(Q, P_{\alpha_{\beta},\beta}) = -\text{log}\beta - (\beta - 1)Q(\text{log}X) + Q(X^{\beta}) + \text{a constant}$$

which has a unique maximizer $P_{\alpha(Q),\beta(Q)} = P_{\theta(Q)}$ if Q is not degenerate.

I do not know of a natural interpretation of this projection -- other than the fact that it minimizes the Kullback-Leibler divergence between Q and the Weibull family $\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$. Note that this $\mathcal{P}$ is *not* an exponential family.

Example 3. $\mathcal{P} = \{P_\theta : dP_\theta/d\mu = \; p_\theta = e^\theta$ on $\mathbb{R}^d, \; \theta$ concave$\}$.

$\mathcal{M} = \{Q$ on $\mathbb{R}^d : \; Q(\|X\|) < \infty, \; Q(X \in H) < 1$ for all hyperplanes $H\}$.

Let

$$\begin{aligned} H^2(p_1, p_2) &= 2^{-1} \int_{\mathbb{R}^d} (\sqrt{p_1} - \sqrt{p_2})^2 d\mu, \\ d_W(Q_1, Q_2) &= \text{inf} \Big\{ E_J \|X - Y\| : \; (X, Y) \sim J \text{ on } \mathbb{R}^d \times \mathbb{R}^d, \\ &\qquad\qquad\qquad X \sim Q_1, \; Y \sim Q_2 \} \\ &\equiv \text{Wasserstein}_1 \text{ distance between } Q_1, \; Q_2. \end{aligned}$$

Theorem: (Dümbgen, Samworth, Schumacher, 2011): $P_{\theta(Q)} = \Pi(Q|\mathcal{P})$ exists and is unique for all $Q \in \mathcal{M}$. Furthermore $Q \mapsto \Pi(Q|\mathcal{P})$ is continuous - - w.r.t. H on $\mathcal{P}$ and d_W on $\mathcal{M}$.

Corollary: If $X_1, \ldots, X_n$ are i.i.d. $Q \in \mathcal{M}$, then

$$H(P_{\theta(\mathbb{P}_n)}, P_{\theta(Q)}) \to_{a.s.} 0.$$

Example 4. Right-censored survival data; continued:

Primer: notation and basic facts, hazards and hazard rates

- Suppose $T \sim F$ on $\mathbb{R}^+ = [0, \infty)$: $F(t) = P(T \le t)$.
- If F has density $f(t) = F'(t) = (d/dt)F(t)$, then T has hazard rate function

$$\lambda(t) \equiv \frac{f(t)}{1 - F(t)}.$$

- The cumulative hazard function $\Lambda(t)$ is given by

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1 - F(s)} ds = \int_0^t \frac{1}{1 - F(s-)} dF(s) \\ &= -\log(1 - F(t)) \quad \text{for continuous } \ F. \end{aligned}$$

- Thus for F continuous,

$$\exp(-\Lambda(t)) = 1 - F(t) = \text{the survival function.}$$

More generally, for an arbitrary d.f. F,

$$
\begin{aligned}
1 - F(t) &= \prod_{s \le t} (1 - \Delta\Lambda(s)) \exp(-\Lambda_c(t)) \\
&= \textstyle\prod_{s \le t} (1 - d\Lambda(s)) = \text{the ``product integral''}
\end{aligned}
$$

where $\Lambda_c(t) \equiv \Lambda(t) - \sum_{s \le t} \Delta\Lambda(s)$, and $\Delta\Lambda(s) \equiv \Lambda(s) - \Lambda(s-)$.

Example 4, Cox model; with right-censored survival data, continued

- The underlying random variables and assumptions:

  for $1 \leq i \leq n$,

  - $X_i \sim F(x|Z_i)$ for $x \in \mathbb{R}^+$, $Z_i \in \mathbb{R}^p$. (the survival times)
  - $Y_i \sim G(x|Z_i)$ for $x \in \mathbb{R}^+$, $Z_i \in \mathbb{R}^p$. (the censoring times)
  - X_i, Y_i conditionally independent given Z_i.
  - $Z_i \sim H$ on $\mathbb{R}^p$.

- The observed data:
  - $T_i = \min\{X_i, Y_i\}$, $\Delta_i = 1\{X_i \leq Y_i\}$, and Z_i.

The “classical” Cox model (1972): parametric relative risk:

- The model:

$$\lambda(t|Z) = \lambda_0(t)\text{exp}(\beta^T Z)$$

  where

  - λ_0 is an unknown (baseline) hazard (rate) function,
  - $\beta \in \mathbb{R}^p$ is a vector of unknown regression parameters.

- The Cox partial likelihood estimator $\widehat{\beta}$ of β: maximize the Partial Likelihood $PL(\beta)$ defined by

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta^T Z_i}}{\sum_{j \in \mathcal{R}_i} e^{\beta^T Z_j}} \right\}^{\Delta_i}$$

  where $\mathcal{R}_i \equiv \{j : \ T_j \geq T_i\}$, $i = 1, \ldots, n$. Then

$$\widehat{\beta} = \text{argmax}_\beta PL(\beta). \tag{1}$$

- The Breslow (Aalen) estimator of Λ_0: $\widehat{\Lambda}_0(t;\widehat{\beta})$ where

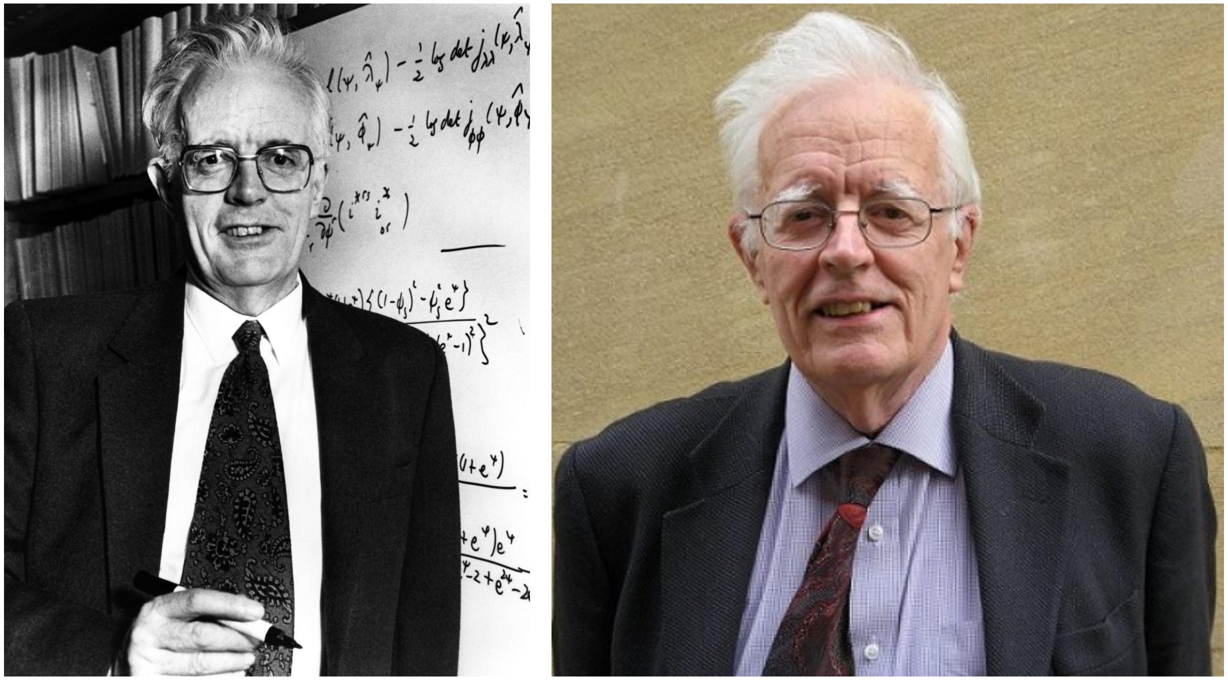
$$\widehat{\Lambda}_0(t;\beta) \;=\; \sum_{i=1}^{n} \frac{\Delta_i 1_{[T_i \le t]}}{\sum_{j\in\mathcal{R}_i} e^{\beta^T Z_j}} = \text{argmax}_{\lambda_0} L(\beta,\lambda_0) \qquad (2)$$

  where $L(\beta,\lambda_0)$ is an appropriate "full likelihood" for (β,Λ).

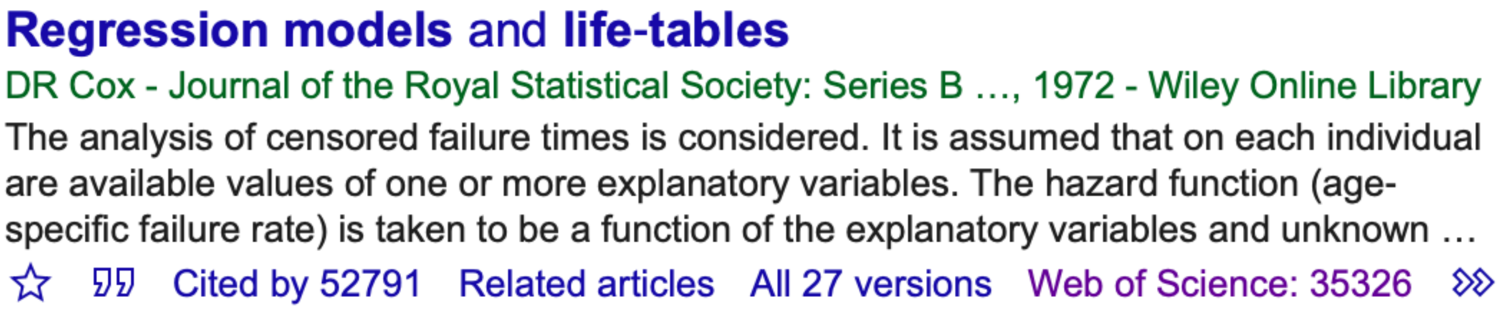
- This leads to the corresponding estimator of the conditional survival function at time t and covariate value $Z = z$:

$$\widehat{S}(t|z) = \exp\left(-\int_0^t e^{\widehat{\beta}^T z} d\widehat{\Lambda}_0(s)\right) = \exp\left(-e^{\widehat{\beta}^T z}\widehat{\Lambda}_0(t)\right).$$

  where $\widehat{S}(t|z) = 1 - \widehat{F}(t|z)$.



Sir David Cox: 1980 and 2003

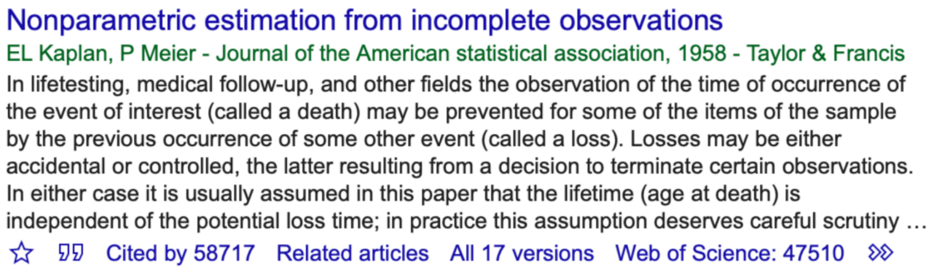


Regression models and **life-tables**

DR Cox - Journal of the Royal Statistical Society: Series B …, 1972 - Wiley Online Library

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown …

☆ ❞ Cited by 52791 Related articles All 27 versions Web of Science: 35326 »



Nonparametric estimation from incomplete observations

EL Kaplan, P Meier - Journal of the American statistical association, 1958 - Taylor & Francis

In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a death) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a loss). Losses may be either accidental or controlled, the latter resulting from a decision to terminate certain observations. In either case it is usually assumed in this paper that the lifetime (age at death) is independent of the potential loss time; in practice this assumption deserves careful scrutiny …

☆ ❞ Cited by 58717 Related articles All 17 versions Web of Science: 47510 »

- **Classical Cox estimators when the model holds**
  Write $X = (T, \Delta, Z)$ for the observed data with $n = 1$.
  Let $\mathbb{N}(t) \equiv \Delta 1_{[T \le t]}$, $Y(t) \equiv 1_{[T \ge t]}$, and define

$$M(t) \equiv M_{\beta,\Lambda}(t) \equiv \mathbb{N}(t) - \int_0^t e^{\beta^T Z} Y(s) d\Lambda(s).$$

  Then $M_0 \equiv M_{\beta_0,\Lambda_0}(t)$ is a martingale under $P_0 \in \mathcal{P}_{Cox}$.

- **Score function for β and score operator for Λ:**

$$\begin{aligned} \dot{\ell}_{\beta,\Lambda}(X) &= \int_0^\tau Z dM_{\beta,\Lambda}(s), \\ (B_{\beta,\Lambda} h)(X) &= \int_0^\tau h(s) dM_{\beta,\Lambda}(s), \quad \text{for} \quad h \in \mathcal{H} \equiv BV[0,\tau], \\ B_0^* \dot{\ell}_0 &= P_0(Z e^{\beta^T Z} Y), \\ B_0^* B_0 h &= h P_0(Z e^{\beta^T Z} Y). \end{aligned}$$

  Thus

$$(B_0^* B_0)^{-1} h = \frac{h}{P_0(e^{\beta^T Z} Y)}.$$

Now set

$$m(t) \equiv \frac{P_0(Ze^{\beta_0^T Z} Y(t))}{P_0(e^{\beta_0^T Z} Y(t))} = E(Z|T = t, \Delta = 1).$$

Then the efficient score for β, efficient information for β, and efficient score operator for Λ are given by

$$\begin{aligned} \ell_0^* &= \left[I - B_0(B_0^* B_0)^{-1} B_0^*\right]\dot{\ell}_0 \\ &\equiv \int_0^\tau (Z - m) dM_0, \\ \tilde{I}_0 &= = P_0\left(e^{\beta_0^T Z} \int_0^\tau [Z - m]^{\otimes} Y d\Lambda_0\right), \quad \tilde{\ell}_0 = \tilde{I}_0^{-1} \ell_0^* \\ Ah &= \int_0^\tau \frac{h}{P_0(e^{\beta_0^T Z} Y)} dM_0 - P_0\left(\int_0^\tau \frac{h}{P_0(e^{\beta_0^T Z} Y(t))} dM_0\right) \tilde{\ell}_0 \end{aligned}$$

These lead, via the infinite-dimensional $Z-$ theorem of van der Vaart (1995) or Bickel et al. (1993) to the following expansions:

$$\begin{aligned} \sqrt{n}(\widehat{\beta}_n - \beta_0) &= \mathbb{G}_n(\tilde{\ell}_0) + o_p(1), \\ \sqrt{n}(\widehat{\Lambda}_n - \Lambda_0)(h) &= \mathbb{G}_n(Ah) + o_p(1), \end{aligned}$$

where $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P_0)$. This yield asymptotic normality and efficiency **when the model holds**; i.e. when $P_0 \in \mathcal{P}_{Cox}$.

What happens when (T_i, Δ_i, Z_i) are i.i.d. $Q \notin \mathcal{P}_{Cox}$?

The estimating equations for β and $\Lambda \equiv \Lambda_0$ can now be written as

$$\mathbb{P}_n \int Z dM_{\beta,\Lambda} = \mathbb{P}_n(\Delta Z - \Lambda(Ze^{\beta^T}Y) = 0, \qquad \text{and}$$
$$\mathbb{P}_n \int h dM_{\beta,\Lambda} = \mathbb{P}_n \left(\Delta h(T) - \Lambda(hZe^{\beta^T Z}Y)\right) = 0,$$

for all h in the unit ball of $BV[0,\tau]$, where $\Lambda(h) \equiv \int_0^\tau h d\Lambda$.

Now suppose that $(\widehat{\beta}_n, \widehat{\Lambda}_n)$ solve the estimating equations in the last display, but that Q does not satisfy the Cox model. The infinite-dimensional system of equations can again be reduced to a finite dimensional system as follows: Taking $h(s) = 1_{[s\le t]}/\mathbb{P}_n e^{\beta^T Z} Y(s)$ in the score equation for Λ yields

$$\widehat{\Lambda}(\beta)(t) = \mathbb{P}_n \left(\frac{\Delta 1_{[T\le t]}}{(\mathbb{P}_n e^{\beta^T Z} Y)(T)}\right).$$

Using this $\widehat{\Lambda}(\beta)$ in the first equation above shows that the resulting $\widehat{\beta}_n$ solves

$$\mathbb{P}_n \Delta \left(Z - \frac{\mathbb{P}_n Z e^{\beta^T Z} Y}{\mathbb{P}_n e^{\beta^T Z} Y}(T) \right) = 0.$$

Letting $n \to \infty$ (so $\mathbb{P}_n \to_d Q$), we see that the limiting versions of the score equations are given by

$$\begin{aligned} \Psi_{1:\beta,\Lambda} &= Q\Delta \left(Z - \Lambda \left(Q Z e^{\beta^T Z} Y \right) \right) = 0, \\ \Psi_{2:\beta,\Lambda} &= Q\Delta h(T) - \Lambda \left(h Q (e^{\beta^T Z} Y) \right) = 0, \qquad h \in \mathcal{H}. \end{aligned}$$

Choosing $h = Q(Z e^{\beta^T Z} Y)/Q(e^{\beta^T Z} Y)$ and subtracting we see that $\beta_0 \equiv \beta_0(Q)$ solves

$$Q\Delta \left(Z - \frac{Q Z e^{\beta^T Z} Y}{Q e^{\beta^T Z} Y}(T) \right) = 0. \tag{3}$$

See Struthers and Kalbfleisch (1986) and Sasieni (1992) for a careful treatment of existence and uniqueness. By convexity arguments Struthers and Kalbfleisch show that $\widehat{\beta}_n \to_p \beta_0 = \beta_0(Q)$. It then follows that

$$\begin{aligned}\widehat{\Lambda}_n(t) &= \mathbb{P}_n\left(\frac{\Delta 1_{[T\le t]}}{\mathbb{P}_n(e^{\widehat{\beta}^T Z}Y)(T)}\right)\\ &\to_p Q\left(\frac{\Delta 1_{[T\le t]}}{Q(e^{\beta_0^T Z}Y)(T)}\right) \equiv \Lambda_0(t).\end{aligned}$$

Is $\beta_0 \equiv \beta_0(Q)$ an interpretable parameter when $Q \notin \mathcal{P}_{Cox}$?

As Sasieni (1992) notes, if $Q \in \mathcal{P}_{Cox}$, then

$$\frac{E(Ze^{\beta_0^T Z})}{Ee^{\beta_0^T Z}} = E\{Z|T = t, \Delta = 1\},$$

and hence $\beta_0(Q) = \beta_0$ is a solution of (3).

But what happens when:

(a) Hazards are not proportional?

(b) A relevant covariate is (incorrectly) left out?

(c) The correct model involves additive hazards?

(d) Is $Q \mapsto \beta_0(Q)$ continuous?

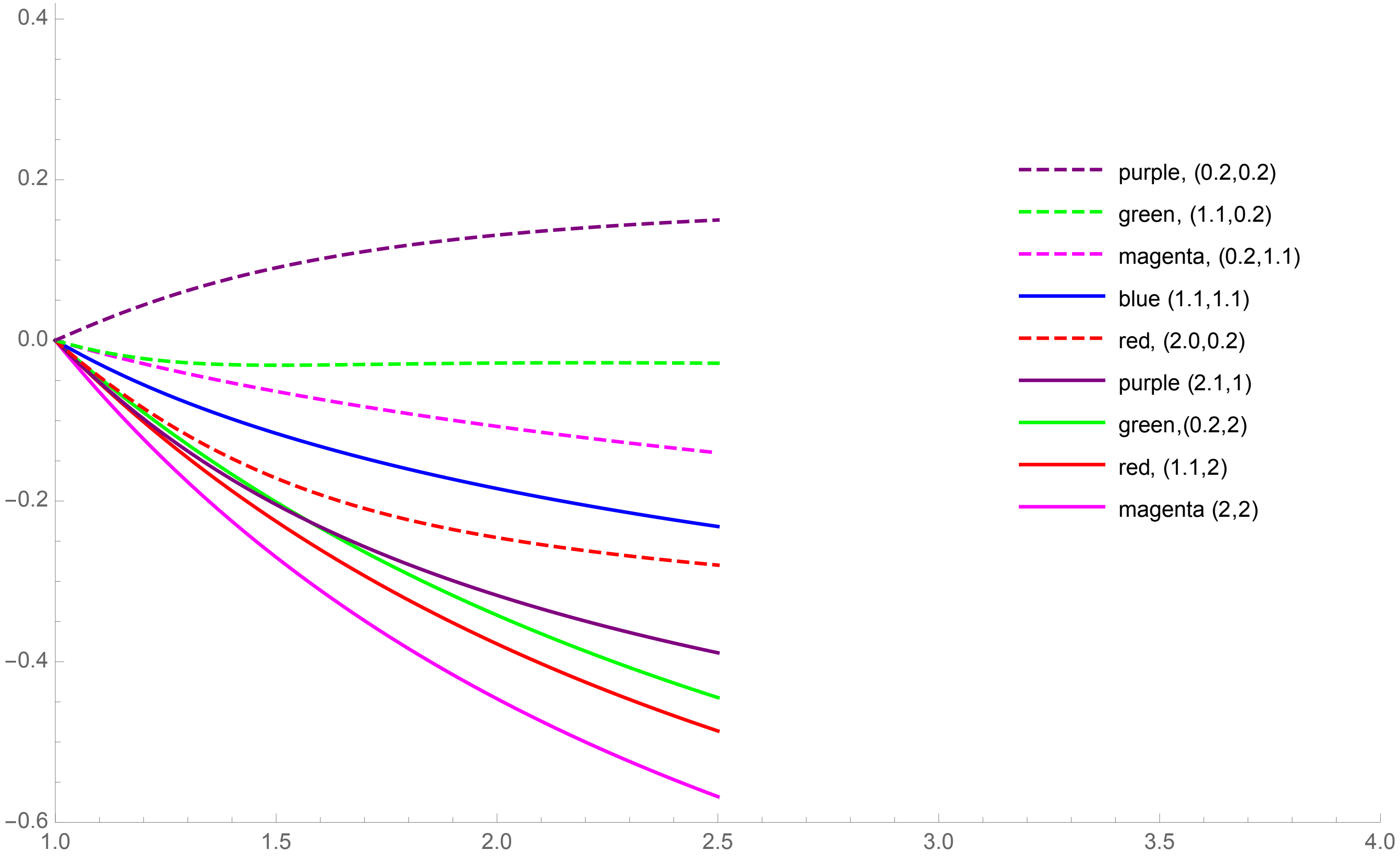
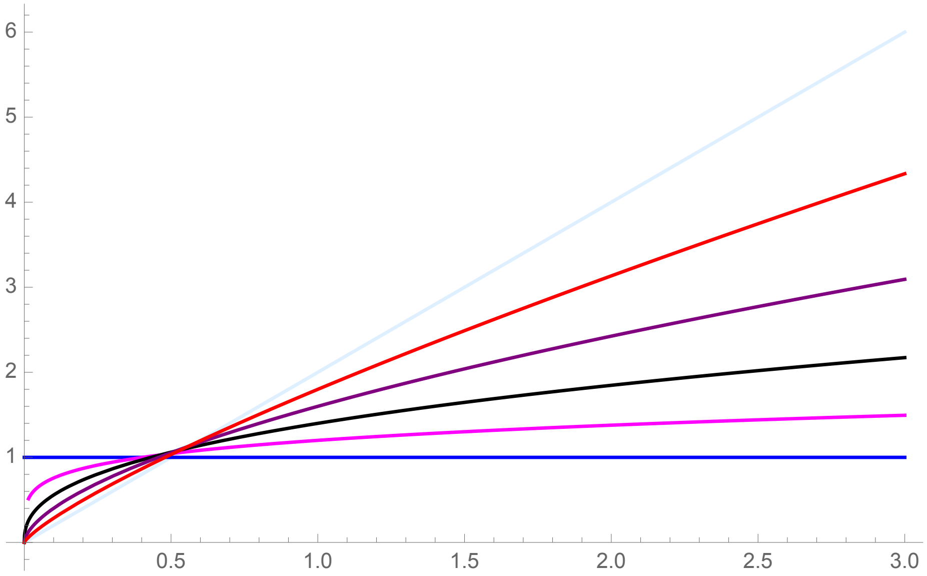
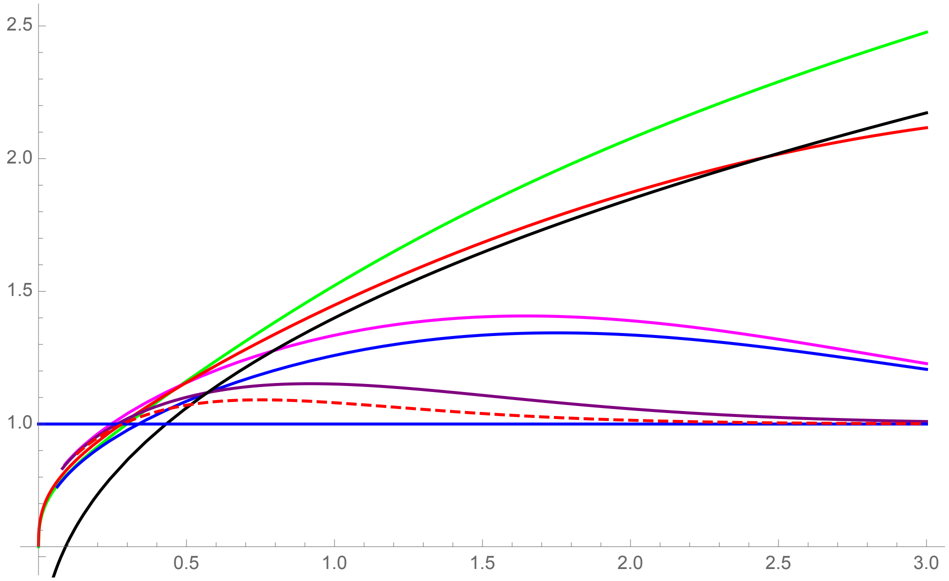


Figure 2: $\beta(Q)$ as a function of α; $\lambda_1(t) = \alpha t^{\alpha-1}$, $\lambda_0(t) = 1$
$C(t|z = 0, 1) = \exp(\gamma_0), \exp(\gamma_1)$.
Whitney, Shojaie, and Carone (2019).


6
5
4
3
2
1
0.5
1.0
1.5
2.0
2.5
3.0

,


2.5
2.0
1.5
1.0
0.5
1.0
1.5
2.0
2.5
3.0

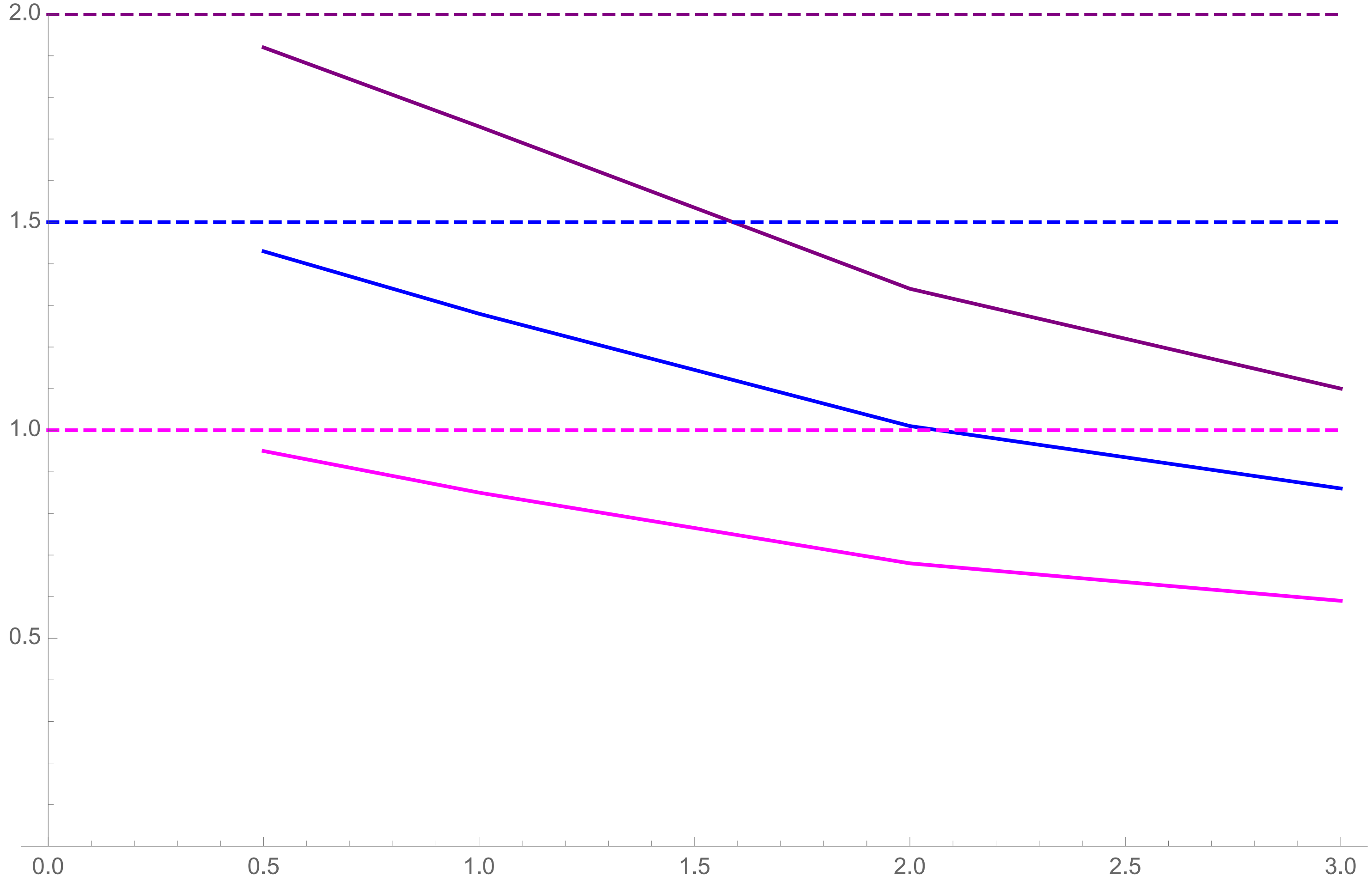


Figure 3: $\alpha_1(Q)$ as a function of α_2, the regression coefficient of the missing covariate for three levels of α_1. Z_1, Z_2 indep. Bernoulli(1/2) (From Struthers and Kalbfleisch (1986).)

Part II. Extensions of the Cox model: beyond parametric relative risk:

Parametric relative risk:

$$\begin{aligned}\lambda(t|Z) &= \lambda_0(t)\text{exp}(\beta^T Z)\\ \lambda(t|Z) &= \lambda_0(t)r(\beta^T Z), \quad r \text{ known; Prentice and Self (1983)}\end{aligned}$$

- Precursors with parametric λ_0: Feigl and Zelen (1965) Prentice (1973);
- Efficiency of the the partial likelihood estimators: Efron (1977), Begun, Hall, Huang and W (1983)
- Robustness: Sasieni (1993a,1993b); Bednarski (1993).
- Martingale theory: P.K. Andersen - R. Gill (1982), Odd Aalen

Semiparametric relative risk:

$$\lambda(t|Z) = \lambda_0(t)\text{exp}(\beta^T Z_1 + \eta(Z_2))$$

where $Z = (Z_1, Z_2) \in \mathbb{R}^p = \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$;

$$Z_\eta : \mathbb{R}^{p_2} \to \mathbb{R}.$$

- Sasieni (1992), SJS; Jian Huang (1999), AOS,
- J. Cai, J. Fan, J. Jiang, and H. Zhou (2007), Biometrika
- R. Tibshirani (1996, 1997): $Z \in \mathbb{R}^p$ with $p > n$. Lasso!

- Pang Du, Shuangge Ma, and Hua Liang (2010), AOS.

Nonparametric relative risk:

$$\begin{aligned} \lambda(t|Z) &= \lambda_0(t)\text{exp}(\eta(Z)); \quad \eta : \mathbb{R}^p \to \mathbb{R}, \quad \text{or} \\ \lambda(t|Z) &= \lambda_0(t) r(Z); \quad r : \mathbb{R}^p \to \mathbb{R} \end{aligned}$$

where λ_0, η, and r are unknown functions.

- Hastie and Tibshirani (1987), Biometrics
- Fan, Gijbels, and King (1997), AOS
- Huang, J.Z., Kooperberg, Stone and Truong (2000)
- LeBlanc and Crowley (1993, 1995, 1999) (JASA, JASA,CJS)
- Gentleman and Crowley (1991), Biometrics
- $\cdots$

Nonparametric: Beyond relative risk

$$\lambda(t|Z) \geq 0 \text{ an ``arbitrary'' hazard rate function;}$$

$$X_i, \; Y_i \text{ conditionally independent given } Z_i.$$

and (reminder):

$$T_i \equiv \min\{X_i, Y_i\}, \quad \Delta_i \equiv 1\{X_i \leq Y_i\}.$$

First attempts:

- Beran (1981): unpubl. UC Berkeley, Tech Report
- Dabrowska (1987). Kernel methods
- Zucker and Karr (1990). Time dependent β's
- O'Sullivan (1988), (1993). Spline methods
- $\cdots$

Nonparametric approaches, more recent work:

- Random survival forests:
  - Ishwaran, Kogalur, Blackstone, and Lauer; (2008), AOAS
- Neural nets:
  - Faraggi and Simon (1995). Statistics in Medicine.
  - Fotso (2018). arXiv
  - Katzman, Shaham, Cloninger, Bates, and Jiang (2018). BMC Medical Research Methodology.
  - Kvamme, Borgan, Scheel (2019), JMLR
    Non-linear Cox & non-proportional Cox
  - Tarkan and Simon (2020). arXiv

Main focus so far:
• Computation via tree-based methods and/or stochastic gradient descent
• empirical performance measures based on prediction of survival.

Progress on theory front? without survival complications: Consistency

- Random forests: Bernoulli random forest framework (involves additional randomization). (2018) A novel consistent random forest framework: Bernoulli random forests. Wang, Xia, Tan, Wu, and Zhu (IEEE transactions on neural networks and learning systems).

- Neural networks:
  Schmidt-Hieber, Johannes.
  Nonparametric regression using deep neural networks with ReLU activation function.
  Ann. Statist. 48 (2020), no.4 1875 - 1897.

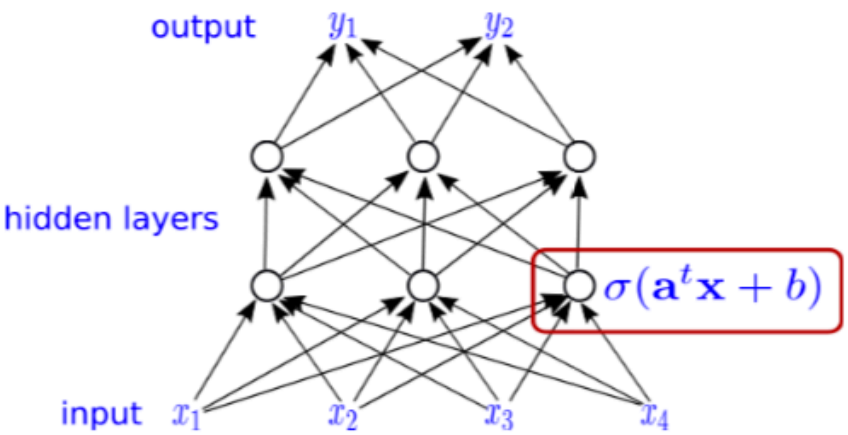


Figure 4: Directed graph representation of a network with two hidden layers, $L = 2$, and $p = (4, 3, 3, 2)$.

$$\sigma_v \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} \sigma(y_1 - v_1) \\ \vdots \\ \sigma(y_1 - v_r) \end{pmatrix}$$

Part III: Descriptive statistics (parameters) by design

Descriptive statistics for nonparametric models:

- I: Introduction (1975).
- II: Location (1975).
- III: Dispersion (1976). Birnbaum (1948)
- IV: Spread (1979).

Question 1. What properties do we want for a measure of location?
Question 2. Are the properties resulting from Q1 logically independent?
Question 3. Do there exist measures satisfying the properties from our answer to Q1?

Bickel & Lehmann, II: Suppose $X \sim F$ on $\mathbb{R}$.
Let $\mathcal{F}$ be the collection of all distribution functions F on $\mathbb{R}$. Then $\mu : \mathcal{F} \to \mathbb{R}$ is a *measure of location* if it satisfies the following:

- If $X <_s Y$ then $\mu(X) \le \mu(Y)$.
- If $a > 0$ and $b \in \mathbb{R}$, then $\mu(aX + b) = aX + b$.
- $\mu(-X) = -\mu(X)$.

Here are some consequences of this definition:
(a) If F is symmetric about θ, then $\mu(F) = \theta$.
(b) If $P(X = c) = 1$, then $\mu(X) = c$.
(c) If $P(a \le X \le b) = 1$, then $a \le \mu(X) \le b$.
(d) If $P(X \ge 0) = 1$, then $\mu(X) \ge 0$.

Examples of measures of location: (carefully chosen) L, M, and R estimators.

- $\mu(F) = \int_0^1 F^{-1}(t)dK(t)$ where K is any d.f. on $(0,1)$ which is symmetric about 1/2.
- $\mu(F) = \text{argmin}_{\theta\in\mathbb{R}} \int \rho(x-\theta)dF(x)$ where ρ is positive, even, convex and twice differentiable with derivative ψ.
- $\mu(F) =$ median of the distribution of $(X_1+X_2)/2$ with X_1, X_2 i.i.d. F; i.e. the solution θ of $\int F(2\theta - x)dF(x) = 1/2$.

Coefficients in linear regression as functionals?

- Models as approximations II: a model-free theory of parametric regression. (Perhaps regard this as "Bickel & Lehmann V"?)
- Buja, Brown, Kuchibhotla, Berk, George, and Zhao (2019).
- *Statistical Science* **34**, 545 - 565. (with discussion).

$$\theta(P) = \theta(P_{Y|\underline{X}} \otimes P_{\underline{X}})$$

Definition: The regression functional $\theta(P)$ is well-specified for $P_{Y|\underline{X}}$ if

$$\theta(P_{Y|\underline{X}} \otimes P_{\underline{X}}) = \theta(P_{Y|\underline{X}} \otimes P_{\underline{X}'})$$

for all (acceptable) regressor distributions $P_{\underline{X}}$ and $P_{\underline{X}'}$.

Examples from survival analysis?

- Sasieni , P. (1996) Proportional excess hazards. *Biometrika.*

  $$\lambda(t; X, Z) = \alpha(t|Z) + \lambda_0(t)\text{exp}(\beta^T Z)$$

  where α is *known*.

- Sasieni, P. and Brentnall (2017). On standardized relative survival. *Biometrics*.

- Wanted: A functional R of two conditional survival functions and a covariate distribution that is a function of time t only (i.e. it is not a function of covariates Z) which describes the ratio of survival functions: for example, R might be the net survival

  $$R(S^C, S^P, H)(t) = E_H \left\{ \frac{S^C(t|Z)}{S^P(t|Z)} \right\},$$

(where C denotes the cohort of interest and P stands for the general population from which the cohort was derived) or it could be the relative survival

$$R(S^C, S^P, H)(t) = \frac{E_H S^C(t|Z)}{E_H S^P(t|Z)}.$$

The authors write:

> If the purpose is to recreate the ratio of survival functions when they are independent of covariates, then this should be a requirement:
>
> $$R(S^C, S^P, H)(t) = S^C(t)/S^P(t)$$
>
> whenever $S^C(t|z) = S^C(t)$ and $S^P(t|Z) = S^P(t)$ for all Z.
>
> . . ..

Sasieni and Brentnall (2017) go on to develop five different criteria that such a measure R should satisfy (their A1 - A5, page 474), and then identify candidates for such a functional R which satisfy at least the first three of their requirements or criteria. They credit Bickel and Lehmann (1975) for at least part of their approach:

> "Our argument mirrors Bickel and Lehmann (1975) who showed that although a trimmed mean is not an unbiased estimate of the mean of an asymmetric distribution, it has a place as a measure of central location of a distribution, and may be better for this than the mean in many situations."

Problems and Questions:

- Q1: Is there an appropriate notion of "well-specified regression functionals" for semiparametric models?
- Q2: Other desirable properties in the survival context? transitivity? collapsibility?
- Q3: What properties do we want a multivariate *measure of location* to satisfy?
- Q4: Which model-based functionals yield interpretable K-L projections?

- Q5: What properties do we want a multivariate depth functional to satisfy?
- Q6: What properties do we want a multivariate clustering functional to satisfy?
- Q7: Once we (I?!) understand some of the machine - learning algorithms better, can we achieve both improved prediction and reliable inference? (Can the machine-learning methods be “plugged in” in the sense of Bickel and Ritov (2003)?)

MANY THANKS!

Two papers with application to breast cancer:

- S. Gore, S. J. Pocock, and G. Kerr (1984).
  Regression models and non-proportional hazards in the analysis of breast cancer.
  J.R.S.S. **33**, 176 - 195.
  $n = 3922$; $p = 14$.

- Weina Zhang, W. and Yilun Zhang (2020).
  Integrated survival analysis of mRNA and microRNA signature of patients with breast cancer based on Cox model.
  J. Computational Biology **27**, 1486 - 1494.
  $n = 626$; number censored $= 560$; $p \geq 77,697$.