Marginal estimates

Terry Therneau

7 October 2024

1 Introduction

"Statistics is the art of clever averaging."

My musings on marginal estimates have had three stages, widely separated in time.

- A Cox model fit with multiple covariates leads to a flock of predicted survival curves; confusing to the eye. How can we create an "adjusted survival curve" for one factor controlling for others. Vignette: Adjusted Survival Curves.
- A user query of "why doesn't coxph provide type III tests?". This led to an unfortunately long side excursion to determine what exactly a type III test is. As it turns out, it is a rather odd test of equality for a rather odd marginal estimand. Vignette: Population Contrasts.
- Extending this to multistate models
  - Which estimands make sense in this context.
  - Sensible variance estimates and tests for these.

2 Adjusted survival curves

As a running example consider the data set of free light chain, which is based on a cross sectional random sample of all the residents of Olmsted County Minnesota aged 50 or more, in a short calendar window. (The parent study has nearly complete coverage, but not all samples had sufficient remaining sera for the later FLC assay.) Figure 1 shows the overall curves.

```
Call:
coxph(formula = Surv(years, death) ~ group, data = flc2)

                 coef exp(coef) se(coef)     z      p
groupMed FLC  0.75680   2.13145  0.05186 14.59 <2e-16
groupHigh FLC 1.65226   5.21874  0.05518 29.94 <2e-16

Likelihood ratio test=772.2  on 2 df, p=< 2.2e-16
```

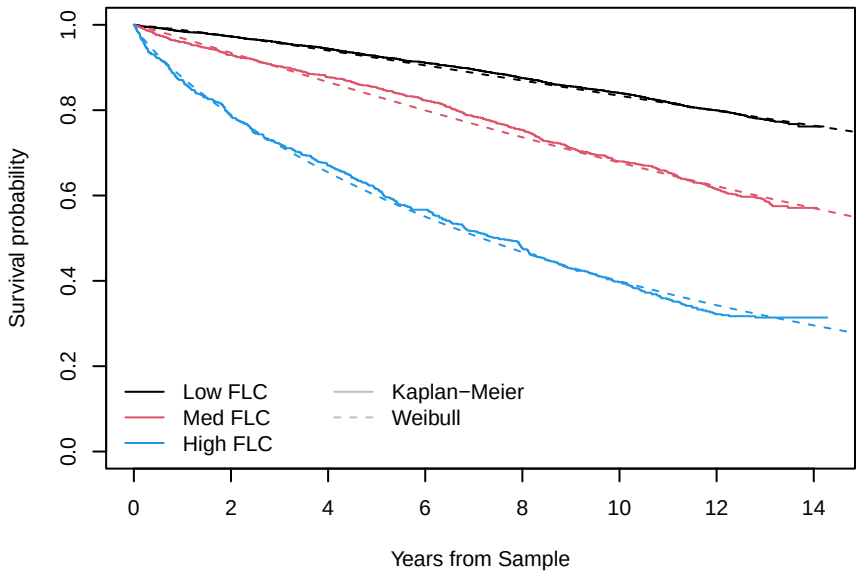


Figure 1: Kaplan-Meier curves from the free light chain (FLC) study, separated by subjects' level of FLC. Overlaid are predictions from a fitted Weibull model.

```
> boxplot(age ~ strata(sex, group), flc2, xaxt='none', xlab='')
> axis(1, 1:6, c("F low", "F med", "F high", "M low", "M med", "M high"))
```

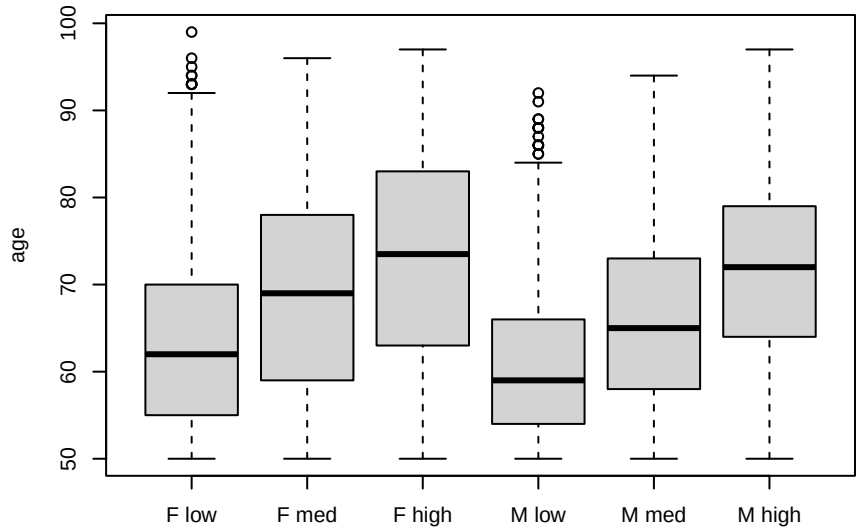


Figure 2: Age distribution for FLC by sex groups.

```
n= 7840, number of events= 2146
Call:
coxph(formula = Surv(years, death) ~ group + age + sex, data = flc2)

                  coef exp(coef) se(coef)      z        p
groupMed FLC  0.297187  1.346067 0.053013  5.606 2.07e-08
groupHigh FLC 0.867334  2.380556 0.058484 14.830  < 2e-16
age           0.102845  1.108319 0.002379 43.237  < 2e-16
sexM          0.331862  1.393560 0.044314  7.489 6.94e-14

Likelihood ratio test=2780  on 4 df, p=< 2.2e-16
n= 7840, number of events= 2146
Call:
coxph(formula = Surv(years, death) ~ group * (age + sex), data = flc2)

                       coef exp(coef)  se(coef)      z        p
groupMed FLC        1.071068  2.918495  0.428299  2.501   0.0124
groupHigh FLC       2.824229 16.847951  0.456765  6.183 6.28e-10
age                 0.110790  1.117161  0.003168 34.970  < 2e-16
sexM                0.319165  1.375978  0.062332  5.120 3.05e-07
groupMed FLC:age   -0.010851  0.989207  0.005597 -1.939   0.0525
groupHigh FLC:age  -0.026307  0.974036  0.005869 -4.482 7.39e-06
groupMed FLC:sexM   0.012078  1.012151  0.106609  0.113   0.9098
groupHigh FLC:sexM  0.055080  1.056625  0.111812  0.493   0.6223

Likelihood ratio test=2802  on 8 df, p=< 2.2e-16
n= 7840, number of events= 2146


> dummy <- expand.grid(group= levels(flc2$group), sex= levels(flc2$sex),
                       age= quantile(flc2$age, c(.2, .4,.6, .8)))
> csurv2 <- survfit(cfit2, newdata=dummy)
> dim(csurv2)
data
  24
```

Predicted curves from the Cox model can be obtained for any fixed values of the covariates. The problem is *what* curves to show, as there are so many possibilities. For patient counseling the answer is quite simple: show the curve applicable to that patient. For an expository paper it is not as clear.

Vignette:

- Balance then model
  - Balance: IPW using logistic regression
  - Model: weighted KM

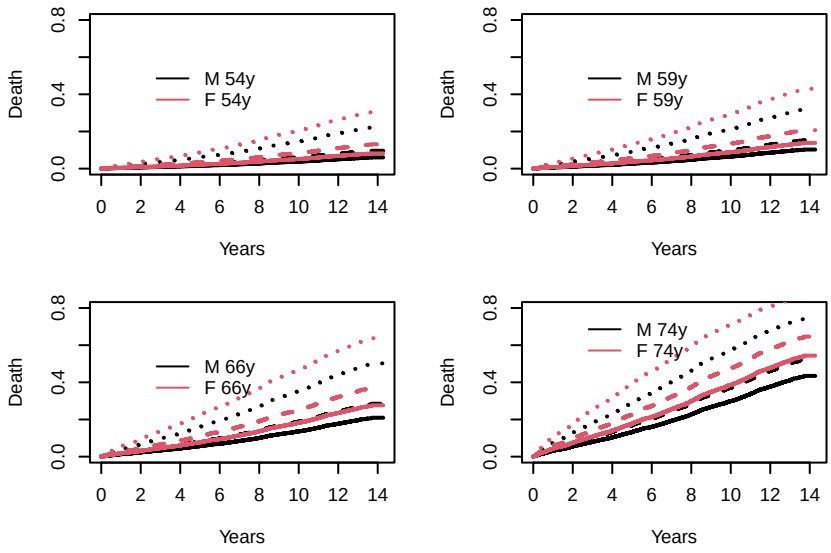


Figure 3: Predicted survival for the free light chain data, from a Cox model with group *(age + sex), at the 20, 40, 60 and 80th percentile of enrollment age.

- Model then balance
  - Model: Stratified KM or predicted coxph curves
  - Balance: weighted average of curves

There is a comment that balanced data leads to curves with standard errors, but averaged Cox model curves does not.

3 Marginal summaries

> "Comparative experiments are mandatory in order to not view coincidences as cause-effect relationships. ... The comparative experiment requires, to be of some value, to be run in the same time and on as similar as possible patients, else the physician walks at random and becomes the sport of illusions." C.Bernard, Introduction à L'Etude de la Médicine Expérimantale, 1866

Statisticians and their clients have always been fond of single number summaries for a data set, perhaps too much so. Consider the hypothetical data shown in figure 4 comparing treatments A and B with age as a confounder. The prediction comes from a linear model with a `trt * age` term. What is a succinct but useful summary of the difference between treatment arms A and B?

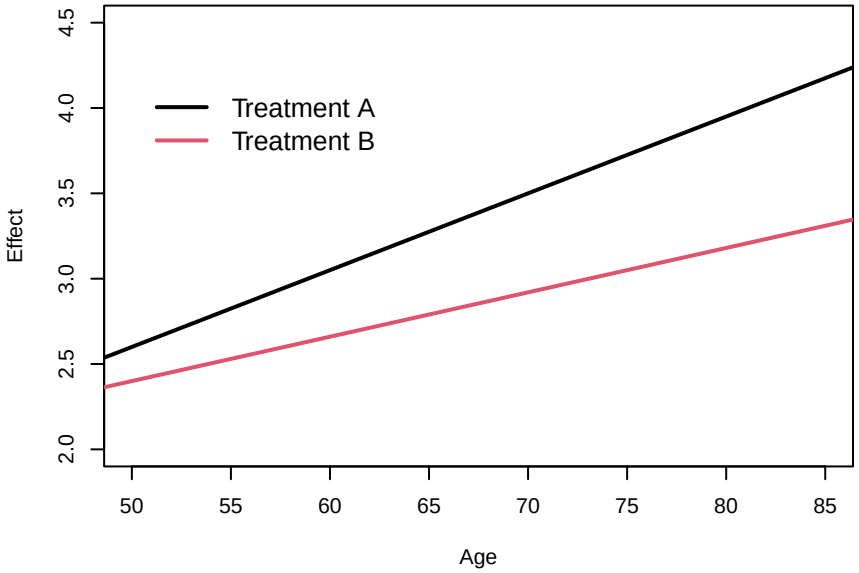


Figure 4: Treatment effects for a hypothetical study.

One approach is to select a fixed *population* for the age distribution, and then compute the mean effect over that population. The two "population" means are

$$m_A = E_F(\hat{y}|trt = A, age)$$
$$m_B = E_F(\hat{y}|trt = B, age)$$

where F is a chosen distribution over the ages. Which F to use depends entirely on the question we want to answer. For instance, perhaps we want to predict the average treatment effect in nursing home patients?

More generally, let $\hat{y}$ be some prediction from a model and divide the covariates as $X=$ the one(s) for which we want a marginal prediction and Z all the others. For a given value of X define a population marginal mean (PMM) as

$$\text{pmm}(X = x) = E_F(\hat{y}|X = x, Z) \tag{1}$$

where F is a chosen distribution for Z.

This simple definition includes two key factors that are in the DNA of our profession: balance on confounders and take averages. It will thus be no surprise that this basic idea has been re-invented multiple times, with multipe names, in multiple contexts.

Here are four key questions, listed in (my) order of importance.

1. The choice of $\hat{y}$
    - $\hat{\eta}$: common (emmeans, LSM), easy variance
    - GLM: type="response" predictions make more sense

- Survival: HR, $p(t)$, $E[N(t)]$, sojourn(t)

2. The choice of population
   (a) External (census totals, SMR, indirect and direct adjustment)
   (b) Factorial or Yates
   (c) Dataset
   (d) LSM
3. Statistical properties of the PMM estimates, including tests for difference
4. Compuational methods

In survival analysis, attention has often focused on hazard ratios (HR), which have a long reign as the primary and often only Cox model summary that is reported. However, there are several shortcomings to HR, particularly HR in the absence of anything else. Both the causal modeling and estimand efforts of recent years have added a welcome refocus in this regard.

The use of an external population has a long history in the study of comparitive death rates across populations. A direct adjusted estimate of deaths, for instance, will use the estimated survival at a fixed time t for each age and sex (z), weighted by a frequency of that group from a national census, e.g., the United States 2020 population.

$$E(d) = \sum \hat{S}(t; z) f(z)$$

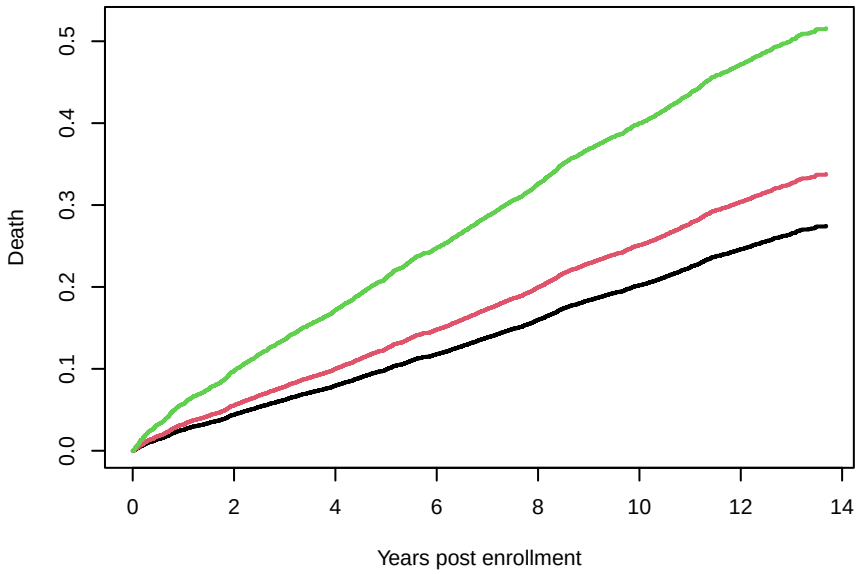
See Keiding and Clayton [**?**] for a review of such methods, over 100+ years and multiple literatures. Another example is the comparison of result from two studies using either direct or indirect standardization. Let A be the study in hand and B an external reference. Direct standardization uses the population of A and predicted values from B, indirect uses predicted values from B and the population from A.

3.1 Yates function

The Population Effects vignette talks about many of these issues. Re-reading it now, there are many opportunities for revision.

```
> yfit1 <- yates(cfit2, ~group, test="pairwise")
> yfit1
    group     pmm     std        test  chisq df        Pr
  Low FLC 0.46829 0.21126      1 vs 2  25.97  1 3.463e-07
  Med FLC 0.84749 0.24285      1 vs 3 178.50  1   < 1e-08
 High FLC 1.62683 0.24716      2 vs 3  58.83  1   < 1e-08
>
> yfit2 <- yates(cfit2, ~group, test="pairwise", predict="survival", nsim=10)
> yfit2
    group     pmm     std        test chisq df        Pr
  Low FLC 11.8115 0.22734      1 vs 2 31.93  1 1.602e-08
  Med FLC 11.3487 0.27927      1 vs 3 44.75  1   < 1e-08
```

```
 High FLC  9.8882 0.44548      2 vs 3 34.30  1   < 1e-08
> plot(yfit2$summary, fun="event", col=1:3, lwd=2,
      xlab="Years post enrollment", ylab = "Death")
```



The yates() function attempts to look at this. Good: it uses the observed data as the default population, and can accept a data set as the population. Bad: it defaults to the HR as the measure to summarize.

SAS: The least squares means of SAS GLM use a mixture population: factorial for the class variables and data for the remainder. (I can think of no situations in which this population makes sense as an estimand.) The GLM test for equality of LSM is based on a clever but obscure algorithm.

- Assume that X is in standard order: intercept, then main effects, then 2 way interactions, etc.
- Let $L'L = X'X$ where L is lower triangular. Then $L\beta$ are constrasts corresponding to sequential SS (type I).
- Let Z be a subset of rows of X that correspond to a balanced design and $L'L = Z'Z$ where L is lower triangular. The rows of L correspond to tests for the Yates contrasts. These are the "type III" tests for those contrasts.
- If Z is not full rank this is not reliable.

All other routines in SAS that report type III results are focused on $\eta = X\beta$. The phreg routine does not use this algorithm for the tests; the one it employs creates invalid results unless categoricals use the sum contrasts ("effect" in SAS).

The popular emmeans (expected marginal means) library in R also focuses entirely on η. It is much more focused on the estimands, not just the tests, and is more careful in its computations.

4 Next steps

1. What is an interesting estimand? One answer is "one that can be verified":

- The prediction $\hat{y}$ for a single subject can assessed, at least in principle.
- The prediction for a group = mean(prediction of the subjects)

The odds ratio and hazard ratio fail. P(event), $p(t)$, $E[N(t)]$, and sojourn are tenable.

2. How do we assess uncerainty? Propose the IJ. How good is it? How good are tests of equality?

$$\frac{\partial \Lambda_C(t;z)}{\partial w_i} = \int_0^t \frac{dM_{ijk}(s)}{\sum_i w_i Y_{ij}(s) r_{ijk}(s)} - \int_0^t Y_{ij}(s) D_{i(jk)} (\overline{x}(t) - z)_{(jk)} d\hat{\lambda}_{jk}(s)$$

The first term is identical to the AJ influence, so I have efficient code. The second term just needs to be added.

3. What single number summary of a curve is most useful?

- $p(\tau)$
- expected number of visits by τ (lifetime risk)
- sojourn time in a state

4. How does this relate to estimands and or causal modeling?

5 Laments

"mean subject"

A recurring, erroneous approach is to deal with the multiplicity by using the mean for each of the confounding covariates, leading to a single curve for each FLC group. Unfortunately, many packages produce $\hat{S}(t, \overline{x})$ as the default "predicted survival" if no covariate set is specified by the user. Most simply the issue is that

$$E[S(t;x)] \neq S(t; E[x]) \,;$$

expected values cannot be moved inside a nonlinear function, and the "mean covariate" curve corresponds neither to any subject nor to a population. With respect to the first, in the FLC dataset the 0/1 dummy variable for sex has a mean value of 0.45. Who exactly does this represent?

Treating the $\overline{x}$ curve as a population curve is as large an error, though more subtle. Figure 5 reprises an example from Therneau and Grambsch [2]. Consider a set of grandfathers and grandsons at a baseball game, with mean ages of 60 and 10, respectively. The predicted survival

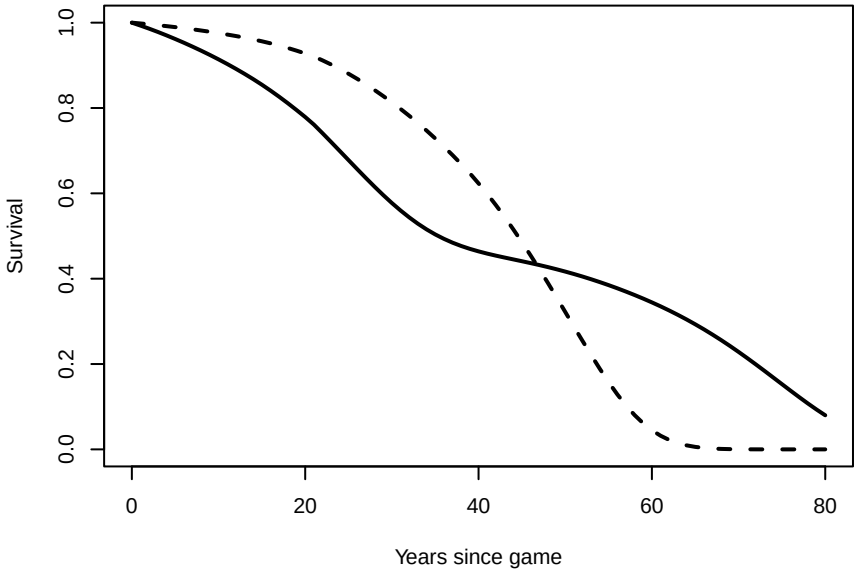


Figure 5: Predicted survival for a fictional cohort of 10 and 60 year olds (solid line); the dashed line shows the predicted survival for a cohort of 35 year olds.

$E[S(t;x)]$ for this cohort dips twice as first the grandfathers and then later the grandsons reach old age. It looks nothing like the predicted survival $S(t;E[x])$ of a 35 year old.

This error has been recurrently highlighted in the literature, but is still widespread [1]. One of its more common variants is to fit and plot a stratified Cox model:

```
> cfitc1 <- coxph(Surv(futime, death) ~ strata(group) + age + sex,
                  data=flc2)
> sfitc1 <- survfit(cfitc1)
```

This is a very fast and easy way to produce a set of three curves, one for each stratum. But, as just stated, these are curves for some hypothetical subject of the mean study age (64.3) and indeterminate sex. A Cox model that treats the FLC group as a covariate imposes an additional constraint of proportional hazards across the 3 FLC groups and is even less satisfactory.

5.1 Inverse probability of censoring (IPC) weights

There is a close relationship between the redistribute-to-the-right (RTTR) algorithm and inverse probability of censoring weights (IPC) 1/G(t). The latter are often created by invoking the Kaplan-Meier function with a reversed status of 0=event and 1=censored. That is, compute $G(t)$ as the KM for *censoring* rather than death, and use weights of $1/G$ for the events and 0 for censored observations. One immediate advantage of this approach is that it gives an avenue to address informative censoring. For instance, say that for a particular study censoring was related to treatment arm. One can fit per-treatment estimates of G, and reweight each event by

the appropriate G for that event. An extension of this approach using a multivariable model for $G(t)$ plays an important role in marginal-structural models.

There is a technical detail, often overlooked, which is necessary to make this approach equivalent to the RTTR. If there are any tied censoring and event times, e.g., times 2 and 5 in the small example above, all survival methods assume that the censoring occurs after the event. This is a reflection of how data is gathered; if subject Smith is observed to still be alive on day 108, and Jones perishes on day 108, we nevertheless know that Smith dies after Jones, even though the recorded time values are the same. When the status variable is switched from 0/1 to 1/0 in order to compute G, it is necessary to preserve this ordering. A second hazard occurs when a subject's observations have been split into multiple rows, e.g., for a time-dependent covariate; the simple algorithm can mistakenly treat all of the intermediate rows as a censoring event.

It is also important to note that $G(t)$, like the at-risk indicator $Y_i(T)$, should be left continuous, whereas the ordinary survival curve is right continuous, weights should be based on $1/G(t+)$. If all these details are carried out correctly, then the sum of weights before and after IPC reweighting will be the identical. Much statistical software is not cognizant of these issues, however, and uses the naive "recode censoring" algorithm. Fortunately, for many datasets the total number of such ties is small and the subsequent error often ignorable. (The censor-after-death argument breaks down when time has been coarsened into intervals. For example, some datasets in the R survival package have time rounded to months, in order to preserve subject anonymity. We do not know the relative ordering of subjects within the same month.)

If the dataset included delayed entry for some subjects, then the exact equivalence between the RTTR, Kaplan-Meier estimate, and IPC weights breaks down. As well, the best approach to IPC weighting is no longer completely clear. One approach is to give a weight of $G(t_{i0}+)/G(t_i+)$, where t_{i0} is the entry time for subject i. (If a subject is broken into multiple intervals, this is not the starting time of the most recent interval, but of the first interval). If a give subject has a hole in their risk period, e.g., intervals of (0, 10) and (20, 50), then we have no suggestions for a proper IPC weight.

References

[1] W. A. Ghali, H. Quan, R. Brant, G. van Melle, C.m. Norris, P.D. Faris, P.D. Galbraith, M.L. Knudtson, and APPROACH investigators. Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA*, 286:1494–1497, 2001.

[2] N. Keiding and D. Clayton. Standardization and control for confounding in observational studies: a historical perspective. *Stat Science*, 29:529–558, 2014.

[3] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.