Survival Analysis

3. Cox Extensions. Flexible and Discrete Models

Germán Rodríguez

Princeton University

February 19, 2018

Baseline Cumulative Hazard

We can also define an estimate of the baseline cumulative hazard that extends the Nelson-Aalen estimate.

This is in fact easier to derive because it simply equates the observed and expected failures at each distinct failure time, yielding

$$\hat{\Lambda}_0(t) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j \in R_i} e^{x_j' \hat{\beta}}}$$

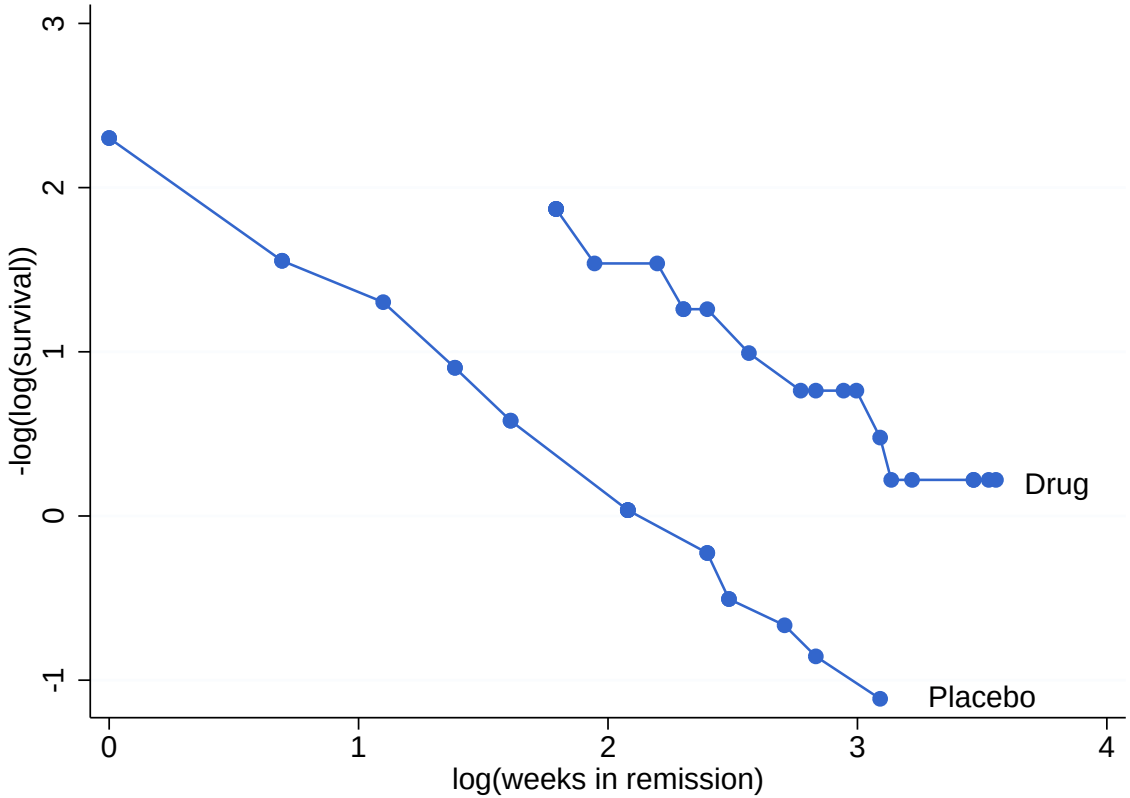
where the sum in the denominator is over the risk set at t_i.

If there are no covariates this estimator reduces to the ordinary Nelson-Aalen, just like the baseline survival reduces to Kaplan-Meier.

The hazard itself can be estimated by differencing the cumulative hazard, but is very "spiky" and usually requires smoothing.

The Log-log Plot

A simpler way to check proportionality of hazards with two or more groups is to plot $-\log(-\log \hat{S}(t))$ versus $\log t$ using separate Kaplan-Meier estimates.



If the assumption is tenable the lines should be parallel, as is clearly the case for the Gehan data.

Interactions With Time

Another way to check proportionality of hazards is to add interactions with time. In his original paper Cox allows the treatment effect to vary linearly with time, effectively fitting the model

$$\lambda(t|x) = \lambda_0(t)e^{\beta x+\gamma xt}$$

The log of the hazard ratio is β at time zero and increases γ per unit of time.

The hazard ratio itself is e^{β} at the origin and is multiplied by e^{γ} for each unit of time.

A test of $H_0 : \gamma = 0$ using a likelihood ratio, Wald, or score statistic checks proportionality of hazards against a linear trend in the log-hazard over time.

Software Notes

In order to include interactions with time in R we need to split the data using the powerful `survSplit()` function. In the computing logs I show how to split the Gehan data at each failure point and then add an interaction with time.

In Stata we can do the same thing with `stsplit`, which has the option `at(failures)` to split at each failure time. However, `stcox` can also fit time interactions without splitting the data: the option `tvc` defines a variable to be interacted with time, and `texp` defines the expression to be used, typically time itself.

Either way, we find that the estimated hazard ratio is 4.86 at remission and declines 0.1% per week. The trend is not significant, so we have no evidence against the proportional hazards assumption.

Time Expressions: Indicators

Another way to test for interactions is to allow different effects of a covariate before and after a set time, say 10 weeks.

In the computing logs I do this in R and Stata by splitting the observations at 10 weeks. In Stata one can avoid splitting the data by using `tvc` with $t > 10$ as `texp`.

We find that the hazard ratio is 3.70 in the first 10 weeks and 83% higher afterwards, but the change is not significant;

Once again we find no evidence against the proportionality assumption.

Time-varying Covariates

The main application of episode splitting, however, is to handle time-varying covariates.

Consider the more general model

$$\lambda(t|x(t)) = \lambda_0(t)e^{x(t)'\beta}$$

where $x(t)$ is the vector of covariates at time t.

For example $x(t)$ could be smoking status at age t in a study of adult mortality. A long-time smoker who enters the study at age t_0, quits at age $t_1 > t_0$ and remains a non-smoker until last seen alive at age t would be split into two records: $(t_0, t_1]$ with smoking status 1 and $(t_1, t]$ with smoking status 0.

Don't confuse time-varying covariates with time-dependent effects. Of course a covariate may change *and* have different effects over time.

Splitting and Standard Errors

At this point you may be worried that splitting adds observations and could affect standard errors, but this is not the case because the likelihood doesn't change!

- This is true of the parametric likelihood; a failure is counted just once, while the integral of the hazard from t_0 to t can be split into two (or more) segments
- It is also true of the partial likelihood, where each observation contributes to the risk set at each failure time while appearing in the numerator just once, no matter how we split the data

If looking at the likelihoods doesn't convince you, try fitting a model, splitting the data, and fitting the same model again. You'll get the same estimates and standard errors! Really.

There's no need to cluster the standard errors; if you do, all you get is a robust estimate.

AIDS Survival in Australia

Venables and Ripley have an interesting dataset on AIDS survival in Australia, included as `Aids2` in R's `MASS` library. A Stata version of the data is available in the course website as `aids2`.

The variables include date of diagnosis, date of death or censoring, and status, coded "D" for died. The predictors are age, sex, state and mode of transmission. The dates are coded as days since 1/1/1960.

There are 29 cases with the same date of diagnosis and death. These are cases diagnosed after death. VR add 0.9 days to all dates of death so they occur after other events the same day.

An important factor affecting survival was expected to be the widespread availability of zidovudine (AZT) from mid 1987. Create a time-varying covariate `azt` coded zero before July 1, 1987 and one thereafter. Note that the split is on a calendar date, not survival time.

Residuals in Cox Models

Residuals play an important role in model checking. Censoring, however, means we can't use ordinary residuals. We will review the most useful alternatives available for Cox models:

- Martingale residuals, which are useful to identify unusual observations and to determine suitable functional forms for continuous predictors
- Schoenfeld residuals, which can be used to check the proportional hazards assumptions, both globally and variable by variable

We will skip two other residuals which are less useful: deviance residuals, a transformed version of martingale residuals, and Cox-Snell residuals.

Martingale residuals are motivated by the theory of counting process. We will introduce some basic concepts, but one could skip the technicalities and jump to the definition.

Counting Processes and Martingales

Instead of focusing on the waiting time T_i consider a function $N_i(t)$ that counts events over time. With single events $N_i(t)$ is zero until individual i experiences an event and then it is one.

To keep track of exposure let $Y_i(t)$ take the value one while individual i is at risk and zero afterwards. Finally, let $\lambda_i(t)$ denote the hazard for individual i, which in turn follows a Cox model, so $\lambda_i(t) = \lambda_0(t)e^{x_i'\beta}$. The product $\lambda_i(t)Y_i(t)$ is called the *intensity*.

The stochastic process

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(u)Y_i(u)du$$

is a *martingale*, a process without drift where given two times $t_1 < t_2$ the $E[M_i(t_2)]$ given the history of the process until t_1 is simply $M_i(t_1)$. Martingale increments have mean zero and are uncorrelated. The integral is called a compensator.

Martingale Residuals

Martingales play a central role in establishing the asymptotic properties of Kaplan-Meier estimators, Mantel-Haenszel tests, and Cox partial likelihood estimators.

The martingale residual for each observation is defined as

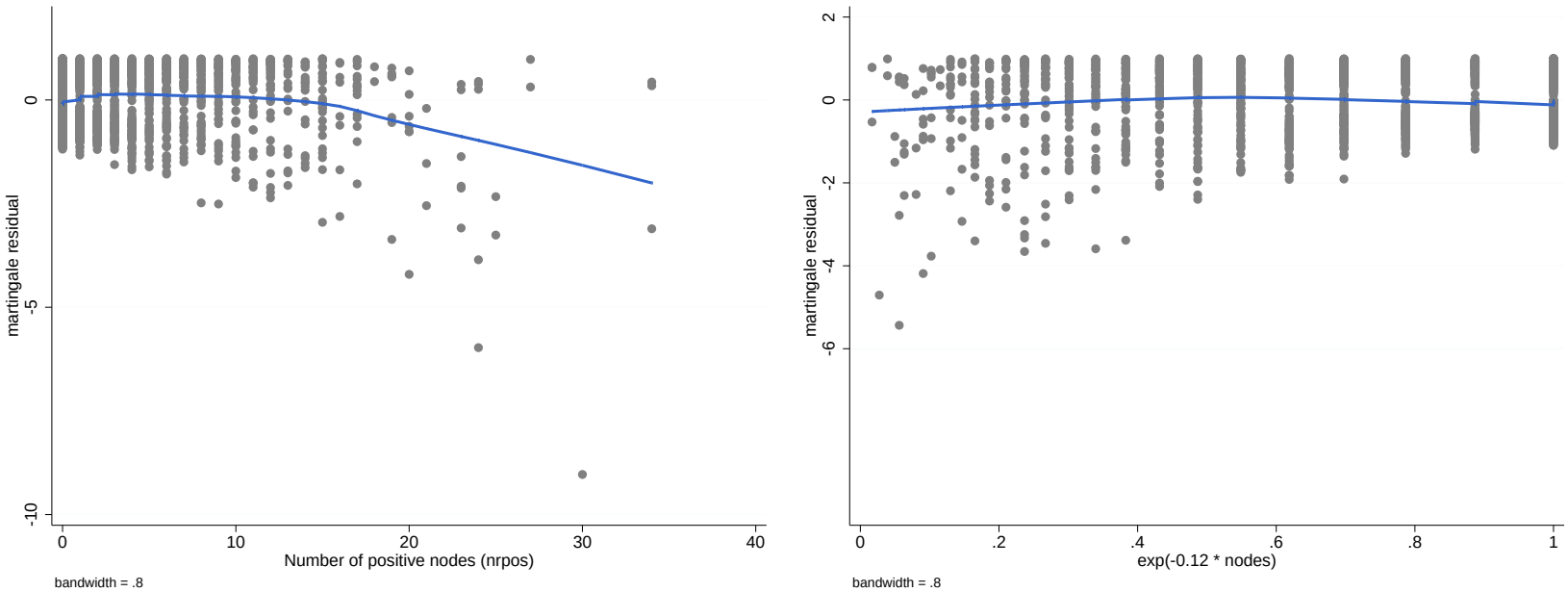
$$\hat{M}_i = d_i - e^{x_i'\beta}\hat{\Lambda}_0(t_i)$$

and may be interpreted as the difference between observed and expected failures over $(0, t_i)$. The range is $(-\infty, 1)$.

Fleming and Harrington showed in 1991 that if the model is correctly specified a plot of $\hat{M}_i$ against each continuous predictor should be linear, and otherwise the plot may help identify the transformation needed.

Breast Cancer in The Netherlands

Royston and Lambert illustrate the use of martingale residuals in an analysis of breast cancer in Rotterdam.



They fit a model using the number of nodes along with other predictors. The martingale residuals on the left show trend. They exponentiate the number of nodes (and take log of another predictor, not shown here). The new residuals on the right are flatter. Differences are clearer if you plot just the smooth.

Schoenfeld Residuals

The Schoenfeld residual for an observation that fails at t_i, assuming no ties, is simply the score

$$r_i = x_i - \frac{\sum_{j \in R_i} x_j e^{x_j' \beta}}{\sum_{j \in R_i} e^{x_j' \beta}}$$

the difference between the values of the covariates for the failure and the risk-weighted average of the covariates over the risk set.

Schoenfeld residuals are defined only for failures, not for censored observations, and each failure has a residual for each predictor.

Grambsch and Therneau showed in 1993 that if the coefficient of a covariate actually varies over time, say it is $\beta_k(t)$ rather than just β_k, the Schoenfeld residual can be scaled so that

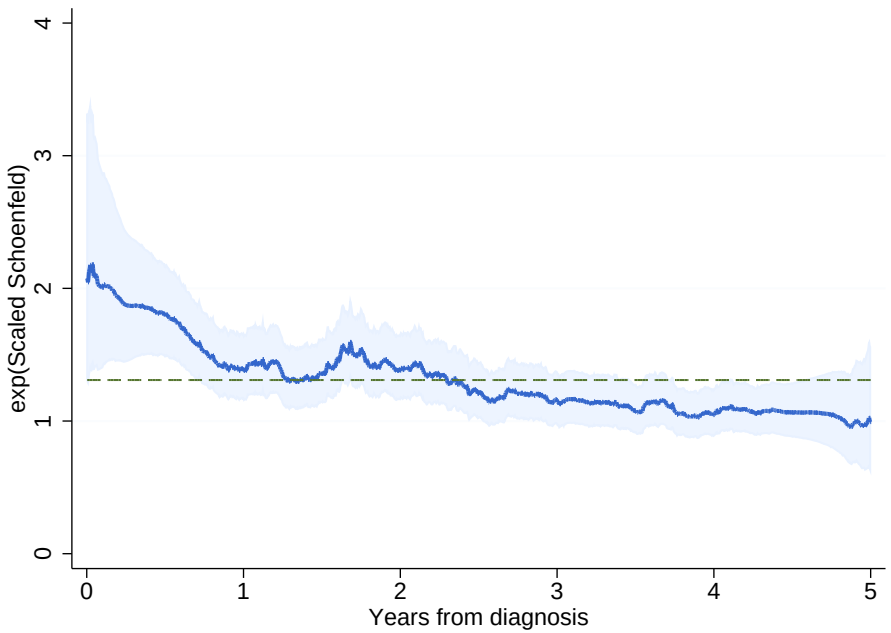
$$E(r_{ik}^* + \beta_k) = \beta_k(t)$$

so a plot of the scaled residuals against time helps identify how the relative risk varies over time.

Breast Cancer in England

Royston and Lambert also have data on breast cancer in England, and find a hazard ratio of 1.31 between the most and least deprived quintiles of women.

Here's a plot of the smoothed scaled Schoenfeld residuals and 95% confidence bands on the smooth, exponentiated to reflect hazard ratios



Clearly the hazard ratio is much higher immediately after diagnosis and declines over time, crossing the dashed line representing proportional hazards. What would you do in light of this result?

Schoenfeld Residuals for Recidivism

In the computing logs I fit a Cox model to the recidivism data, and check proportionality of hazards using Schoenfeld residuals.

The global χ^2 of 12.76 on 9 d.f. shows no evidence against the assumption of proportional hazards.

The only variable that might deserve closer scrutiny is time served, which had the largest chi-squared statistic, 3.59 on one d.f., although it doesn't reach the conventional five-percent level.

A plot of the residuals for this variable against time shows no evidence of time dependence. Please see the website for details.

Piecewise Exponential Regression

We consider models that assume a parametric form, so we can easily estimate the hazard or survival probabilities, yet are flexible.

One of my favorites is the piecewise exponential model, where the baseline hazard is assumed constant in well-chosen intervals, defined by cutpoints

$$0 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = \infty$$

so the baseline hazard at any time is one of k values

$$\lambda_0(t) = \lambda_{0i}, \quad \text{when} \quad t \in (\tau_{i-1}, \tau_i]$$

The model may be fit easily by splitting the data at the cutpoints τ_1 to τ_{k-1} and then fitting an exponential survival model with the interval treated as a factor.

Poisson Regression and Cox

Interestingly, the piece-wise exponential model may also be fit by treating the failure indicators as if they were independent Poisson outcomes.

Specifically, if d_{ij} is a failure indicator and t_{ij} is the exposure time for individual i in interval j then we "pretend" that

$$d_{ij} \sim P(\mu_{ij}) \quad \text{where} \quad \mu_{ij} = \lambda_{0j} t_{ij} e^{x'_{ij}\beta}$$

so $\log t_{ij}$ enters the model as an offset. This trick is useful because we can fit multilevel PWE models!

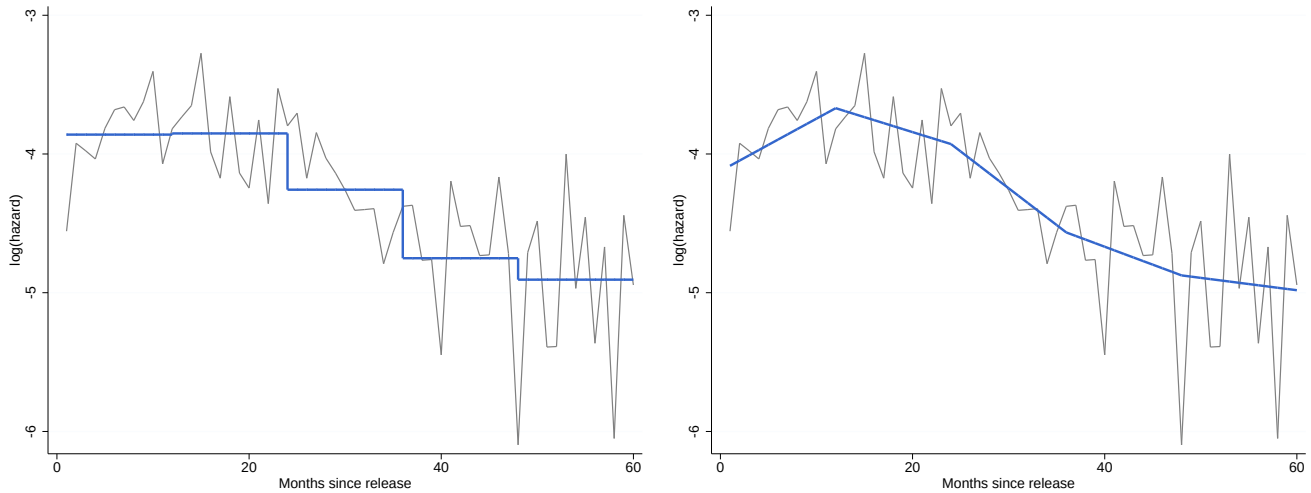
If we assume that the hazard is constant between the observed distinct failure times and fit a PWE model we get *exactly* the same result as with Cox's partial likelihood, provided there are no ties or we use Breslow's approximation.

In other words a PWE model can get arbitrarily close to a Cox model by using more detailed time intervals.

Piecewise Gompertz Models

Instead of assuming that the hazard is constant in each interval we could assume that the log hazard is linear on time in each interval but with possibly different slopes.

The figure below shows PWE and PWG log-hazards with annual intervals superimposed on the Cox estimates for the recidivism data



The software package *aML* implements this method. It also allows for interval censoring rather than just right-censoring.

Regression Splines

More generally, we could model the log of the hazard using a spline. A spline is a piecewise polynomial defined over a series of knots $\xi_1 < \ldots \xi_k$ such that the pieces join smoothly at each knot.

Cubic splines are particularly useful, and can be defined as

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \gamma_k (x - \xi_j)_+^3$$

where $(x - \xi_j)_+^3$ is zero when $x < \xi_j$ and $(x - \xi_j)^3$ otherwise. Because the spline is linear on the β and γ parameters it can be fit by regression for given knots. (With many knots a numerically more stable basis such as B-splines is advisable.)

A cubic spline is *natural* if it is linear outside the range of the knots. This requires $\beta_2 = \beta_3 = 0$ and two constraints on the γ's: $\sum \gamma_j = 0$ and $\sum \gamma \xi_j = 0$. Usually we add knots at the min and max, so we save only two parameters.

Smoothing Splines

Consider a general scatterplot smoothing problem, where we have data on n pairs (x_i, y_i) and want to estimate the relationship using a smooth function $y = s(x)$, by minimizing the criterion

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int [s''(x)]^2 dx$$

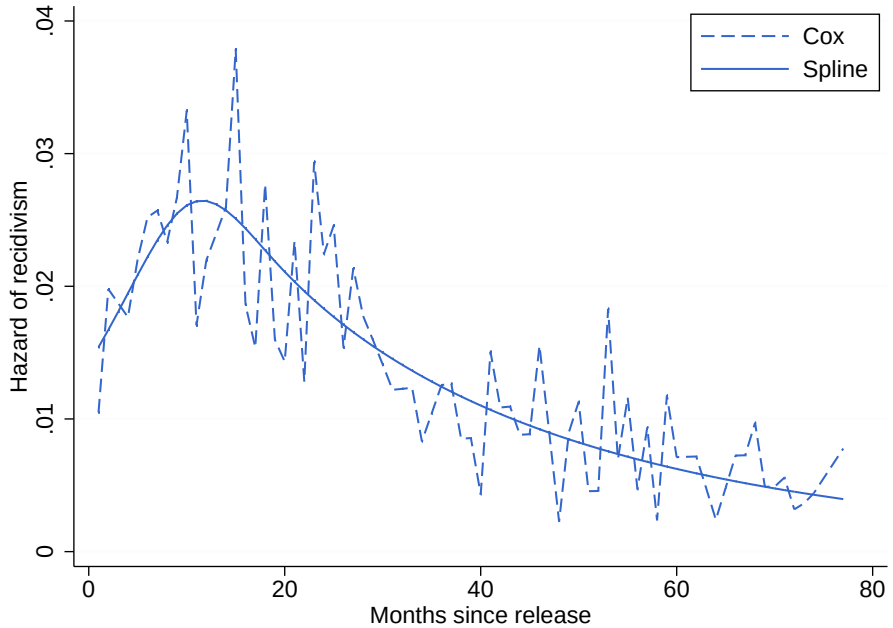
The first term is an ordinary sum of squares which captures lack of fit. The second term is a roughness penalty based on the second derivative of the smooth function. The parameter λ controls the trade off between fit and roughness.

At $\lambda = 0$ you get a perfect fit interpolating the data, which are usually rough. As $\lambda \to \infty$ you approach the ordinary least squares fit, which is perfectly smooth but may not fit well.

Minimizing this criterion for fixed λ over the space of all twice differentiable functions yields as unique solution a natural cubic spline with knots at all data points!

The Hazard of Recidivism

Splines are easy to fit if you split the data into small intervals of equal width and model the hazard at the midpoint using a regression spline. Here are some results for the recidivism data:



I split the data by month and fitted a natural cubic spline with internal knots at the quartiles of failures (10, 19, 34). The estimates of the parameters are almost identical to Cox's, but the baseline hazard is smooth.

Royston-Parmar Models

We now consider the Royston-Parmar (2002) family of models based on transformations of the survival function. Start from a standard proportional hazards model and take log-log to obtain

$$\log(-\log(S(t|x))) = \log(-\log(S_0(t))) + x'\beta$$

Starting from a proportional odds model and taking logits we get

$$\text{logit}(S(t|x)) = \text{logit}(S_0(t)) + x'\beta$$

A generalization uses the Aranda-Ordaz family of links

$$g(S_0(t)) = \log(\frac{S_0(t)^{-\theta} - 1}{\theta})$$

which includes the logit when $\theta = 1$ and approaches the log-log as $\theta \to 0$. Interpretation is difficult in the general case.

The family is completed with the probit link to include all standard links for binary data.

Royston-Parmar (continued)

What about the baseline survival? They model it using a natural cubic spline on log-time with df-1 internal knots (at quantiles). With one df (no knots) the spline is linear and the probit, logit and c-log-log links lead to log-normal, log-logistic and Weibull models.

The method is implemented in Stata's `stpm2` and R's `flexsurv`.

For the recidivism data I fitted Royston-Parmar models using the probit, logit, and c-log-log scales. I also let the θ parameter free. In all cases I used three df, leading to internal knots at the terciles.

Model	logL
Probit	−1570.07
PH	−1577.67
PO	−1568.88
θ	−1566.66

The estimated value of θ is 2.14. The evidence suggests that proportional odds fit better than proportional hazards. AIC would accept freeing θ because it reduces the deviance by 4.44, but the parameters are not directly interpretable.

Discrete Survival

Consider now the discrete case, where the event of interest can only occur at times $t_1 < t_2 < \ldots < t_m$, usually the integers 0,1,2,... My canonical example is waiting time to conception measured in menstrual cycles.

The discrete survival function or probability of surviving up to t_i is

$$S_i = \Pr\{T > t_i\}, \quad i = 1, \ldots, m$$

The discrete density function or probability of failing at t_i is

$$f_i = \Pr\{T = t_i\}, \quad i = 1, \ldots, m$$

Finally the discrete-time hazard or conditional probability of failure at t_i conditional on survival to that point is

$$\lambda_i = \Pr\{T = t_i | T \geq t_i\} = \frac{f_i}{S_{i-1}}, \quad i = 1, \ldots, m$$

Note: These are the definitions in K-P. Others define the survival using $T \geq t$ so that $\lambda_i = f_i/S_i$. Both conventions are used, so watch out.

The Logistic Model

Cox proposed a discrete-time proportional hazards model where

$$\text{logit}(\lambda_i(x)) = \text{logit}(\lambda_{i0}) + x'\beta$$

In this model the conditional odds of surviving (or failing) the i-th discrete time are proportional to some baseline odds.

Cox then proposed fitting this model using the partial likelihood, so β is estimated but λ_{i0} is not.

Allison wrote a very popular paper in 1982 proposing to fit this model using logistic regression with a separate parameter for each failure time.

To fit the model you split the data at the discrete failure times and treat the resulting records as independent Bernoulli observations. The proof follows the same lines as the equivalence between PWE and Poisson regression.

The Complementary Log-Log Model

An alternative discrete-time model uses the complementary log-log transformation

$$\log(-\log(\lambda_i(x))) = \log(-\log(\lambda_{i0})) + x'\beta$$

This model results from grouping data from a continuous-time proportional hazards model, as we noted in the GLM course.

To see this point write $S(t|x) = S_0(t)^{e^{x'\beta}}$ as in continuous time and note that $S_0(t) = \prod_{i:t_i \leq t}(1-\lambda_{i0})$ with grouped data to obtain

$$\lambda_i(x) = 1 - (1-\lambda_{i0}))^{e^{x'\beta}}$$

a relationship that is linearized by c-log-log.

Kalbfleish and Prentice (2002, p. 47) note that this is the uniquely appropriate model for grouped continuous-time data.

Discrete Models and Partial Exposure

Care must be exercised with partial exposure if you are using a discrete model with grouped continuous-time data.

Consider two contraceptors, one who is lost to follow up at 21 months and one who discontinues at 15 months, but you group by year. It is then very common to turn these two cases into four records as shown on the right. What's wrong with this setup?

Id	Year	Fail
1	1	0
1	2	0
2	1	0
2	2	1

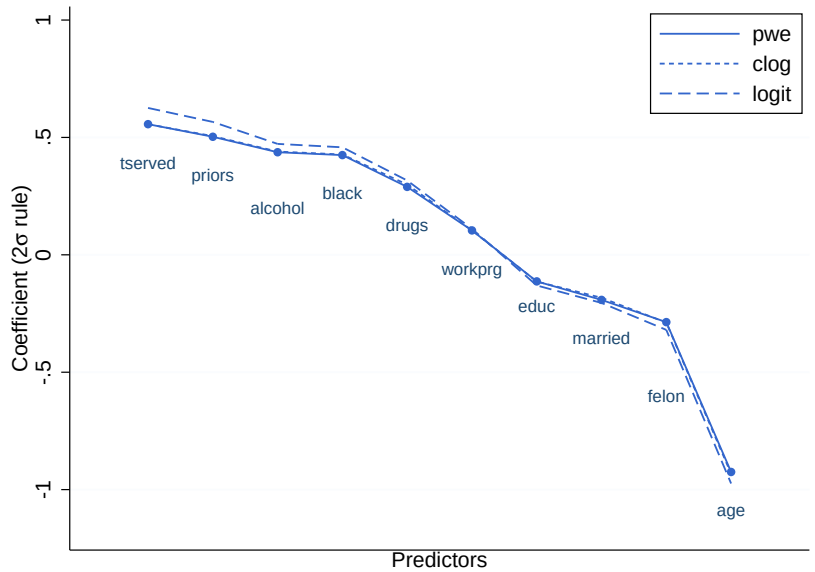
We don't really know if the first woman survived the second year of use. The second record should be deleted, effectively censoring the case at the end of the first year, see "reduced sample" in Cox and Oakes.

Less obviously, it is not clear that the failure should count, because we may not know if the second woman would have been observed throughout the second year had she not discontinued. Why is this a problem? The first woman could have failed before she was lost to follow up!

The Recidivism Data

The recidivism data are well-suited for discrete analysis because the data were collected retrospectively and everyone is potentially exposed for a full five years with no censoring. We focus on years one to five.

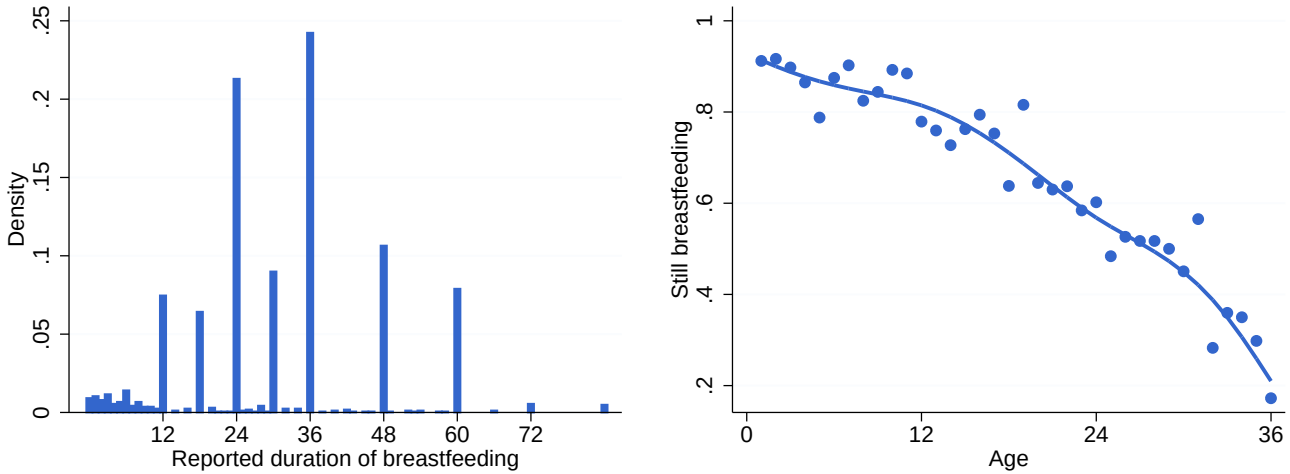
In the computing logs I compare three models, continuous PWE, and discrete c-log-log and logit. Here's a graphic summary of coefs:



The results of PWE and c-log-log are indistinguishable, while logit is a bit different, reflecting odds ratios rather than relative risks. The annual hazard is just 8%, so odds and hazards are not too different.

Current Status Survival

Retrospective reports of breastfeeding duration typically show substantial heaping at multiples of 12, as in Bangladesh in 1976



The figure on the right ignores reported duration and simply shows the proportion still breastfeeding by current age of child, together with a spline with knots at 12 and 24. There is little evidence of heaping.

All observations here are censored. If a child has been weaned the duration is less than current age and is left censored. If a child is still breastfeeding the duration is at least the current age and is right censored. Yet we can estimate the survival curve!