Wypasek Data Science, Inc.

# Commentary and Supplemental Documentation of the Multinomial Proportional Hazards Model Methodology Used by Christian Wypasek

Version 1.0.0

Author(s): Christian Wypasek

**Abstract**

This article with editorial and commentary has been written by Christian Wypasek (the author or CW) and should be treated as such, having editorial and commentary. Some aspects of the commentary were motivated by recent discussions on alternate methods for fitting competing risk models. Hopefully, the more interesting portions come from the just the supplemental documentation of the author's process for fitting and implementing these models.

# 1 Disclaimer

To the author's knowledge, everything detailed here is in the public domain and has been presented at public forums. The author's personal methodology has been used and refined at multiple firms and does not represent any restricted IP. This article and other details can also be found at: https://github.com/wdatasci/WDS-ModelSpec

# 2 Background discussion, the part with commentary

For this section, the author will dispense with the third person language.

The method I have used since late 1999 or early 2000 is different than another which has now been implemented in multiple frameworks such as SAS and R. For lack of free time, I will let readers look up references. (Or, if there is something publishable herein, perhaps an old colleague can rope in some some poor grad student who needs a paper. If interested, please give me a shout.) However, as a single reference to direct some discussion points towards, consider a SAS proceedings paper:

https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2159-2018.pdf

I have never read "Fine and Gray" (FG) or any of the references mentioned in the SAS paper. The simple reason not showing an interest before is that I solved the problem back in 1999, came up with a fitting and implementation style which suited my purposes better than anything available, and alternatives to this day are still lacking.

Why comment on it now? Because someone recently showed me what an FG method was doing. I was floored at how grossly inappropriate it would be for consumer credit modeling.

One of the beautifully simple aspects of Sir David Cox's proportional hazards model is the partial likelihood function with a Nelson-Aalen-estimator-like effect where it is fundamental to treat the hazard estimate as the ratio of the subject weight at events to the total weight at risk. For the sample unit hazard, this can be intuitively referred to as *the number of events over the number able.*

In consumer credit modeling, possibly with a long span of data crossing multiple market environments (in other words, having significant time-varying covariates, such as with mortgages), once a subject has an exit, such as prepayment or default, the data stream ends. Unless one has already built the "calculator" for forecasting out the process (which may also include other more complex exogenous variables), to include data from exited subjects in the denominator is akin-to using made-up-data (FG-mud). [1]

Furthermore, the authors of the SAS article referenced above make a claring mistake that represents the problem with a simplistic cause-specific fitting of competing risks. Another way of interpreting the *cause-specific* term is that the competing risk hazards are effectively *non-concurrently fit but concurrently evaluated* (NCF-CE). For my purposes, the root cause of misunderstanding is the statement:

[1] Oddly enough, for years, I had an ongoing argument about validating hazard models with a few colleagues who did not have mathematical sciences backgrounds. The concept of comparing the empirical hazard against the at-risk average of model estimates was mis-termed by a colleague as somehow *with replacement.* For context, this was around 2009/10 when modeling the LoanPerformance dataset for subprime mortgages. This dataset had approximately 2M loans and eventually grew to over 1B time series observations, all actual, uncensored observations, originated over 10+ years.

A basic reality in industry is that one has to pick one's battles, it can be difficult to argue against a simple statement like *Why can't you just run out the model and compare the average cumulatives against the empirical?* when the counter party does not have an understanding of Nelson-Aalen, Kaplan-Meier, censoring, hazards, compensators, mixtures of distributions, etc. Another colleague commented, *you're just never going to give in on this one*, to which the only response was *I can't, otherwise some aspects of my dissertation and the mathematics are wrong.*

If the hazards do not validate, the cumulatives should not also. On the other hand, some explainable events or outliers in the hazards cause persistent forward deviations in cumulatives. Further confounding of the issue comes from time varying covariates, heavy censoring, and mixtures of distributions.

A few years later, an old deal was revisited and re-evaluated with updated data. That particular deal's analysis was fit on a rep-line basis without significant exogenous time varying effects. I was able to show a tight comparison between cumulative empiricals and model estimates. The second colleague finally understood when I pointed to the model tail extending past the end of empiricals and could comment, *You know what is different here? There is no actual data here, but I also had a "calculator" to forecast all of the variables necessary.*

Before anyone might pish-off this footnote with *well, at our shop, we were more sophisticated and actually did run 1000s of simulations on the whole LoanPerformance dataset for a full 10 years*, in 2009, as part of a small team, I was solely responsible for the model prep, building, implementation, integration into cash flows, and then producing CPRs/CDRs which were fed through the Intex C-api. Having a big team, resources, and too much free time, would have been great. Instead, we got the job done, made the firm money, and moved on to more interesting projects.

*Because the K models are mutually exclusive, they can be analyzed independently.*

Mutual exclusivity does not imply independence.[2] In fact, mutually exclusive screams dependency. In this way, the fundamental premise of non-informative sensoring is ignored in both NCF-CE and FG-mud.

That being said, in consumer credit modeling, there may be cases (with prime mortgages as an example), where one might argue that prepayment and default are independent. In these cases, default is usually the more rare event and often due to an unrelated event for the consumer. From personal experience of modeling collateral on the order of $Ts, once one gets outside of prime credit, competing risks are generally not independent.

It is fairly common to talk with modellers who fit risks separately and then used combined systems, as in NCF-CE, only to find that adjustments (corrections) have to be made to the final system. That was one of the motivating factors for the method discussed in this article/commentary/editorial: In a shop where one colleague developed a first-serious-delinquency model and another fit a prepayment model, I developed the roll-to-loss model and got stuck implementing them all. That is when the misspecifications showed up. The first-serious-delinquency model looked like it validated well, but when carried out, it was a full distribution function, over-estimating ultimate defaults, lacking the the sub-distribution effect associated with prepayments. To my colleague's credit, he subsequently did come up with a really interesting accelerated life type of model for a subdistribution function, where the terminal value was the expected default level. The prepay model was, well, just wrong.

One can see the motivation in something like FG to create a hazard which results in a sub-distribution, but instead of mud, how about just doing it right in the first place? Which brings me back to the method I have used for 20+ years.

Late in 1999, I thought I came up with an innovative way of fitting competing risks via an extension of Cox's Proportional Hazards model. I had considered writing an academic publication, but decided against it.

*Why?* Because Sir David Cox had already proposed it back in the 1980s.

For this reason, I believe the correct term for the model described here is ”Multinomial Proportional Hazards (or Cox)” model or possibly ”Marked Proportional Hazards (or Cox)” model. If one wants an acronym, how about MPH/Cox?

A continual learning process requires oneself to periodically review what might be new, different, and potentially helpful. In terms of competing risk models, with a working methodology in hand, I have generally ended up dismissing the few that have come up. An R-based FG-mud function was rejected, before even knowing about the mud part, because it failed a quick check against some of my basic requirements:

- Does it fit multinomial or competing risk hazards models in a logical manner?
  - This includes appropriate handling of non-informative sensoring.
  - This includes some concept of the change of measure between outcomes.
- There has to be stratification handling.
- The variable specification for each risk type should not be required to be the same. For example, the set of covariates required to model the hazards for each competing risk should not have to be the same and the coefficient values should not have to be the same.
- Baselines in PH/Cox models are afterthoughts. However, the usual estimates lack reasonable smoothness. If the method doesn't discuss this, then the authors might never have done anything in consumer credit.

One can fit MPH/Cox models satisfying these requirements either by:

- Writing it yourself: Which back in 1999/2000, I first did with a combination of MATLAB and SAS code, using the equations detailed below.
- Tricking a normal PH/Cox method, like SAS's PHREG.[3] I figured this one out when a colleague promoted the concept that a SAS-only technique would be better in some way. Once one thinks about it, it is actually quite easy (details below), but data expensive. Someone may have published an article in the 1990s.

[2] In stochastic processes, independence implies orthogonality, but the converse is not generally true.

[3] I have always pronounced SAS's PHReg as P-H-Reg, since it stands for Proportional-Hazards-Regression. After one unfortunate event, I try not to snicker at those who pronounce it *FREG.* As an alternative, I try to stick with the southern-ism, *oh, bless its heart.*

Most of my original write-up is included here, but I have also decided to include refinements made along the way. Hopefully, after reading, one will understand the significant differences between MPH/Cox, NCF-CE, and FG-mud. If not, please keep re-reading.

## 3 The Racing Game, MTL and LTM

Obviously, in order to understand the cash flows around a consumer product such as mortgage, both prepayment and default need to be well understood. Whereas default is (hopefully) a small proportion of any overall pool with a small impact on prepayment, the prepayment rate constantly changes the dynamic of default propensity of the surviving pool. The marginal effect of late cycle default on default timing estimates becomes inflated. In the prime world, this may potentially be due to non-informative censoring witnessed by evidence that prepayment speed tends to be negatively correlated to default risk.

*Author's note:* For those not familiar with the term *mark*, it is used here in the sense of marked point processes. In these processes, an arrival brings a *mark*, or some associated value. Cox may also have used that term in reference to competing risks. Use of the term, Life-Then-Mark, was on purpose, to avoid using the term proportional hazards, since one colleague would otherwise just reject the idea.

A few of the techniques used in the past could be described as:

1. **A Racing Game: Competing lives with fastest receiving the mark**

   The intuition is that often the marginal default timing is inflated and similarly the prepayment timing may be slightly inflated. Compensate these two effects by "racing" the two estimated distributions, i.e., Let $X$ be the time to default and let $Y$ be the time to prepayment and let $S^d(t)$ and $S^p(t)$ be their respective survival functions. Then our estimate of $C^d(t)$ becomes:

   $$C^d(t) = \mathrm{P}\left(X \leq t, X < Y\right) = \int_0^t -\mathrm{P}\left[X \leq y | Y = y\right] dS^p(y) + \mathrm{P}\left(X \leq t, Y > t\right),$$

   which under the assumption of independence of $X$ and $Y$ (at least given the common covariate picture) becomes:

   $$C^d(t) = \int_0^t -\mathrm{P}\left(X \leq y\right) dS^p(y) + (1 - S^d(t))S^p(t).$$

   The challenge remains the overestimation of the default rate and the solution technique is to allow this transformation of $S^d$ and $S^p$ to damp out the biasing of their estimates. Even though there is a damping out of the biasing of $C^d(t)$ especially at $t = \infty$ by correcting for prepayment, both prepay and default curves may be faster than intended. Furthermore, the effectiveness may depend on the relative speeds and $C^d(\infty)$.

   *Author's original note:* The more I know and the more I have seen through examples, the more I realize that there are still too many effects confounding the overall timing with this technique. At GE, this seemed intuitive and seemed to work, but through work with MTL and LTM I think we now know why there are problems and how to correct them.

   *Author's new note:* This is effectively NCF-CE.

2. **Mark then life (MTL)**

   The intuition is that the ultimate default propensity exists prior to life. In this scenario, to find $C^p(t)$, the cumulative prepayment probability, one must first find $C^d(\infty)$, the ultimate default probability, and then find $C^p(t)$ via the product of $1 - C^d(\infty)$ and the conditional probability of prepayment by time $t$ given the loan will never go into default.

   The challenge remains the overestimation of the default rate and the solution technique is to treat prepayments as non-defaults throughout the study.

   *Author's new note:* One advantage here is that final outcomes are stable with regard to changes in environmental hazard covariates.

3. **Life then mark (LTM)**

   The intuition is the loan has a "productive life" the end of which carries a mark of prepayment or default. Let $S(t)$ denote the total survival function and let $h(t), h^d(t), h^p(t)$ represent the hazard rates for the total survival, default and prepayment respectively. It follows that $S(t)$ satisfies the differential equation:

   $$\frac{dS(t)}{dt} = -h(t)S(t) = -(h^d(t) + h^p(t))S(t).$$

   Furthermore,

   $$\frac{dC^d(t)}{dt} = h^d(t)S(t) \text{ and } \frac{dC^p(t)}{dt} = h^p(t)S(t).$$

   For conditional models, as in the case where a loan has not gone to default or prepayment after some time $t_1$, the differential equation can just be solved forward with $S(t_1) = 1$. This will also incorporate only the exit patterns which exist after time $t_1$.

## 3.1 LTM, the fast way

Before discussing the main MPH/Cox methodology of this article, there is another *fast* and relatively easy way to fit LTM scenarios.

Note that the estimation of $S(t)$ and $h(t)$ can be performed knowing that the censoring is truely non-informative. Under some absolute continuity assumptions, let $\nu(t) = \frac{h^d(t)}{h^p(t)}$, the instantaneous odds of a default at time $t$. The result becomes:

$$\frac{dC^d(t)}{dt} = \frac{\nu(t)}{1 + \nu(t)} h(t)S(t).$$

The derivative $\nu$ can be estimated via a separate but interesting (logistic) model for the odds of a default given that the non-productive time occurs at time $t$. Here, $t$ is treated as just another variable, but its inclusion redirects the exit patterns, i.e., early exits may have a higher default rate followed by a prepay exodus until burnout followed by steady state leaving rates.

This method may be relatively quick and easy, but the estimation of $S(t)$ may be confounded by the fact that it is a mixture of both default and prepayment timing distributions. Any variables used to fit $S(t)$ are equally applied to both parts of the mixture. The longer process for fitting both the derivative and the total survival proceeds below.

# 4 First, Some Light Background Discussion

*Author's notes:* Please forgive the following fault: I tend to move rather freely between discrete time and continuous time. Motivations tend to be in continuous time, then applied discrete time with a subjectively small $\Delta t$. This should not be much of a problem in a non-Brownian motion stochastic calculus. (Ok, I have found places like integrating out the baselines that need to be done in a discretized manner and not just rectangular on the theoretical, otherwise the discussion still flows.)

For a nicely behaved lifetime, $X$, with distribution function, $F$, and density, $f$, the hazard rate is defined as:

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

$F$ can then be written:

$$F(t) = 1 - e^{-\int_0^t h(u)du}.$$

Conditional survival distributions are taken as:

$$\mathrm{P}\left(X \leq t | X > c\right) = 1 - e^{-\int_c^t h(u)du}.$$

And for small increments, $\Delta t$,

$$\mathrm{P}\left(X \leq t + \Delta t | X > t\right) \approx h(t)\Delta t.$$

In proportional hazards models, for a given covariate picture, $Z$, the forms for $h$ in continuous and discrete times are taken as:

$$h(t) = g(\beta Z)h_0(t) \qquad \text{and} \qquad \frac{h(t)\Delta t}{1 - h(t)\Delta t} = \frac{h_0(t)\Delta t}{1 - h_0(t)\Delta t} g(\beta Z).$$

Here, $h_0$ is a baseline hazards rate that is not parameterized but can be estimated from the data, and $g$ is a function, usually $g(x) = e^x$ to ensure non-negativity. Essentially, given a baseline rate, a proportional hazards is a small increment proportional odds model (notice the similarities to $\frac{\nu(t)}{1+\nu(t)}h(t)$ above).

Fitting a proportional hazard rate model derives from Poisson processes. Suppose a collection of Poisson processes, $N_t^1, \ldots, N_t^n$, with intensities, $\lambda_1, \ldots, \lambda_n$, is aggregated, i.e., $N_t = \sum_{i=1}^n N_t^i$. Furthermore, suppose an event for $N$ occurs at time $t$, then the probability that the event occurs due to Poisson process $i$ is given by:

$$\frac{\lambda_i}{\sum_{j=1}^n \lambda_j}.$$

Given a collection of covariate pictures, $Z_1, \ldots, Z_N$, and event times $T_1, \ldots, T_N$, (assume times are ordered, $T_i < T_j$ when $i < j$), the likelihood for subject $i$ is:

$$\frac{h_i(t)}{\sum_{j=i}^N h_j(t)} = \frac{e^{\beta Z_i} h_0(t)}{\sum_{j=i}^N e^{\beta Z_j} h_0(t)} = \frac{e^{\beta Z_i}}{\sum_{j=i}^N e^{\beta Z_j}},$$

i.e., the baseline cancels out and the likelihood is the proportional factor over the sum of factors for all subjects able to fail at time $T_i$. In general, this is only considered a partial likelihood for the dynamics surrounding $h_0(t)$ are left un-specified, hence it is also a semi-parametric model. This interpretation also facilitates right censoring in which case the denominator above includes all subjects able to fail at time $T_i$, but only subjects that actually fail contribute to the overall likelihood of the model/dataset.

The nature of the likelihood function is inherently linked to the Kaplan-Meier estimator, a variant of which is used to estimate $S_0(t) = e^{-\int_0^t h(u)du}$. Once $S_0(t)$ is estimated from the data, $S_i(t) = [S_0(t)]^{e^{\beta Z_i}}$.

# 5 Working Marked Point Processes into the Mix

## 5.1 Fitting the long technique

Instead of playing a racing game with composite distributions as above which introduced scores of censoring/numerical/technique issues, let us look to the loan level. Instead of treating subject $i$ as a single counting process stream as described in the previous section, treat each subject as a composite stream itself, i.e., $N_t^i = N_t^{i,d} + N_t^{i,p}$ (even though we generally care only about the first event epoch). Note that this is exactly a 3-state case of a semi-Markov renewal model which is like a continuous time Markov chain except sojourn times are general distributions. In other words, this type of procedure could be used for fitting timed-multinomials, possibly conditional on a given state with the multinomial outcome as the transistion into another state.[4]

As before, assume that event times are ordered, $T_1, \ldots, T_N$, with a 1-1 mapping to subjects and let $M_1, \ldots, M_N$, be the respective marks which may be $d, p$ or $e$ for default, prepayment or end of study respectively. For our purposes end of study times are the only purposefull but truely non-informative censoring times for we will assume there are no other stopping times. As before censored observations do not have unique contributions to

[4] Again, leaving to the reader to look it up, or remember from graduate classes long ago: A discrete time Markov chain makes a Markov transition at each epoch. A continuous time Markov chain has the same concepts in continuous time, but has the effect of Markov state transitions with exponential sojourns. A renewal process is a counting process with i.i.d. interarrival (or sojourn) times. A Markov renewal process has Markov process driving the inter-renewal times with a state transition at renewals, and a semi-Markov renewal has the persistent concept of *state* between transitions.

likelihood. Let $h_i^d(t) = e^{\beta^d Z_i} h_0^d(t), h_i^p(t) = e^{\beta^p Z_i} h_0^p(t)$ be the default and prepayment hazards for subject $i$. For a non-censored observation, $i$, the contribution to likelihood is:

$$\frac{1_{\{M_i=d\}} e^{\beta^d Z_i} h_0^d(T_i) + 1_{\{M_i=p\}} e^{\beta^p Z_i} h_0^p(T_i)}{\sum_{j=i}^N e^{\beta^d Z_j} h_0^d(T_i) + e^{\beta^p Z_j} h_0^p(T_i)}.$$

Here, $1_{\{A\}}$ is the indicator of a set, i.e., $1_{\{A\}} = 1$ if $A$ occurs and 0 otherwise.

Of course, it would be nice if we could *cancel* out the baseline hazards, but they are not shared in common any longer.[5] However, dividing numerator and denominator by $h_0^d(T_i)$ yields a contribution to likelihood of :

$$\frac{1_{\{M_i=d\}} e^{\beta^d Z_i} + 1_{\{M_i=p\}} e^{\beta^p Z_i} \frac{h_0^p(T_i)}{h_0^d(T_i)}}{\sum_{j=i}^N e^{\beta^d Z_j} + e^{\beta^p Z_j} \frac{h_0^p(T_i)}{h_0^d(T_i)}}.$$

Here, $\frac{h_0^p(t)}{h_0^d(t)}$ is the instantaneous odds ratio of prepayments to defaults at time $t$ as before in LTM. Now, take a standard logistic form of $\frac{h_0^p(t)}{h_0^d(t)} = e^{\phi(t)}$. Note that this ratio (actually a Radon-Nikodym derivative between the two measures and therefore induces some absolute continuity assumptions between the default and prepayment support) can be found from the definition of $\nu$ in the LTM discussion above. $\nu(t)$ can be estimated from a separate odds model of the form $\nu(t) = e^{\hat{\beta} Z_i + \phi(t)}$ fit on loans with non-censored observations. The term $\hat{\beta} Z_i$ becomes a controlling factor to find the average effect of time alone. In a general sense, there may be terms of $Z_i$ which are correlated to time and could be included above, but this would become more computationally intensive and we will not consider that immmediately. To see this complexity, notice that we are dividing by $h_0^d(T_i)$ at the time of the numerators event, interactions with time would require the denominators covariates to be interacted with the numerators time for each $T_i$.

The process for fitting the marked point process becomes as follows:

- One could first fit the derivative effect, $\phi(t)$ via a logistic model, $\nu(t) = e^{\hat{\beta} Z_i + \phi(t)}$. This is optional, since it can also be fit concurrently as detailed in the equations for Newton-Raphson fitting detailed in a later section.
- Maximize the following likelihood function:

$$\prod_{M_i \neq e} \frac{1_{\{M_i=d\}} e^{\beta^d Z_i} + 1_{\{M_i=p\}} e^{\beta^p Z_i} e^{\phi(T_i)}}{\sum_{j=i}^N e^{\beta^d Z_j} + \left(\sum_{j=i}^N e^{\beta^p Z_j}\right) e^{\phi(T_i)}}.$$

Of course, it looks like $\hat{\beta} = \beta^p - \beta^d$, however, these sub-models can be entirely specified as to what is appropriate for prepayment or what is appropriate for default. Furthermore, these coefficients are fit in a non-informative censoring environment.

*Author's note:* Notice the fundamental difference between the partial likelihood function above, in a proper form above (and proposed by Cox), vs FG-mud. The additional terms in the denominator come from the competing risks and this dampens the hazard appropriately using just at-risk subjects, not using data from formerly at-risk subjects.

## 5.2 How to use a normal PH/Cox method for fitting a MPH/Cox model

*Author's note:* The fundamental concept of each single subject being replaced with a competing *set*, motivates how to fit MPH/Cox via a normal PH/Cox methodology without sacrificing anything ( in other words, in total contrast to NCF-CE ). It is actually so simple, it begs credulity. The only down-side is that it blows up the sample size. In software such as SAS (not necessarily done in memory), it can be handled without too much additional complexity.

The basic format of a PH/Cox (without time varying covariates is):

[5] *Author's note:* At this point, when presenting, I usually just say: What the hell, just do it anyways. In generally, for more stability in the Radon-Nikodym derivative (RND), the reference event should be the more frequent event at any given time. Or, in the even more general case, the reference measure should be a common effect and both/all outcomes might have RND's to change measure to the common.

$$\begin{pmatrix} obs_1 \\ obs_2 \\ \vdots \\ obs_N \end{pmatrix} = \begin{pmatrix} t_{obs_1} & Z_{obs_1} & e_{obs_1} \in \{0,1\} \\ t_{obs_2} & Z_{obs_2} & e_{obs_2} \in \{0,1\} \\ \vdots & & \\ t_{obs_N} & Z_{obs_N} & e_{obs_N} \in \{0,1\} \end{pmatrix}$$

Here, for each observation, $objs_i \in \{obs_1, \ldots, obs_N\}$, there is an associated time point, $t_{obs_i}$, and covariate picture, $Z_{obs_i}$, and a response $e_{obs_i}$. For simplicity in this discussion, assume that $e_{obs_i} = 1$ at time $t_{obs_i}$ is the event of a target event and that $e_{obs_i} = 0$ at time $t_{obs_i}$ is a censoring event.

In a marked or competing risk scenario, this is extended to:

$$\begin{pmatrix} obs_1 \\ obs_2 \\ \vdots \\ obs_N \end{pmatrix} = \begin{pmatrix} t_{obs_1} & Z_{obs_1} & e_{obs_1} \in \{0,\ldots,K\} \\ t_{obs_2} & Z_{obs_2} & e_{obs_2} \in \{0,\ldots,K\} \\ \vdots & & \\ t_{obs_N} & Z_{obs_N} & e_{obs_N} \in \{0,\ldots,K\} \end{pmatrix}$$

where there are $1, \ldots, K$ events of interest, all others (including censoring) are mapped to to event class 0.

Upon close inspection to the likelihood functions described below and their relationship to a normal one-event class PH/Cox model, it is clear that for $K$ response classes, there are $K$ factors to the baseline being fit concurrently, $e^{\beta^i Z}$, $i = 1, \ldots, K$.[6] The coefficients can be estimated with total consistency to the full MPH/Cox method by *stacking* the model fitting dataset via the following:

- Assume that responses are ordered and numbered, 1, ..., $K$.
- Without loss of generality, assume that the reference outcome is the first response.
- Recall that each singular subject, $i$, has $K$ competing risks (i.e., $e_{obs_i} \in \{0, \ldots, K\}$), for which the occurrence of any one, $e_{obs_i} = k > 0$, truncates any future behavior for the $k^{th}$ and all other response classes.
- Each observation, $obs_i, i \in \{1, \ldots, N\}$, is replaced with $K$ observations, $obs_{i,k}, k = 1, \ldots, K$:
  - Accordingly, $e_{obs_i} = 0$ indicates a censoring after time $t_{obs_i}$. In other words, through and including time $t_{obj_i}$, no marks have been observed, and the subject survives.
  - For each of the $K$ duplications, there is only one associated response. Whereas, in the non-duplicative datset, $e_{obs_i}$ can itself be represented by a vector valued indicator function (where only the $k^{th}$ position is non-zero and equals 1 if and only if $e_{obs_i} = k$), the multinomial or mark reponse will be replaced with the single corresponding stacked response. In other words, the response varaible for the $k^{th}$ duplication equals 1 if and only if $e_{obs_i} = k$.
- In addition to *row* duplication, there is column duplication in the following sense:
  - Assume the total covariate space is considered to be $\{Z_j\}_{\{j \in 1,\ldots,N\}}$.
  - Also, assume that the covariates can be grouped into $K$ sets, $Z^1, \ldots, Z^K$,[7] associated with each of the $K$ reponses, respectively.
    - ∗ Note: The $Z^1, \ldots, Z^K$ sets of covariates do not have to be mutually exclusive (and probably are not).
  - For each of the $K$ row duplications, there is also a copy of each of the $Z^1, \ldots, Z^K$ sets of covariates.
  - For the row duplication associated with reponse, $k \in \{1, \ldots, K\}$, only the covariate set, $Z_k$, is non-zero, each of the other covariate sets is overriden to 0.

[6] This includes any $\phi(t)$ RND factors.

[7] *Author's note:* There have been some interesting observations from personal experience, such as, variables in common to more than one outcome may end up having the same, implying that the variable also has a *duration* effect, i.e., works across outcomes in the same way. Or, if coefficents for outcomes have contrasting signs, there might be a particularly strong variable associated with positive change in one response with a negatively correlated response in another.

Therefore, if one is modeling, say 5 reponses, and with 10 covariates, one ends up with 5X rows and 5X covariates. The multivariate reponse is replaced with a bivariate reponse, associated with the row-duplicate. Futhermore, the 5X rows and 5X covariates, is actually quite sparse and representable by a block diagonal structure.

*Author's note:* This may seem awkward, but still easily handlable in something like SAS (which has been effectively file-based since the mainframe/COBOL days). This is still a hell of a lot better than making up data.

## 5.3 Second-to-last thoughts

When using MPH/Cox models, there is also one last implied assumption (if not stated elsewhere): There are unobservable processes going on in a subject's *life.* The assumption on everything in discussion here is: Once all of the covariates are controlled for, and only after that point, for that one subject, the competing risks become independent.

In other words, the covariates represent all of the information that, once controlled for, provides the context for treating the responses as independent (and self-censoring in a combined context).

## 5.4 Last thoughts before getting lost in formulas

PH/Cox models are thought of as *robust* estimation techniques. One of the basic concepts of these models is that the nuisance parameters do not confound the fitting of the pertinent covariates. In fact, any effect in common to all subjects cancels out. It follows that no energy is spent fitting the baseline in the first step.

Since the baseline could cause major shifts in behavior, its inclusion could take predictive power away from other variables. This is also a criticism of $\phi(t)$ above, so care needs to be taken in its artificial treatment and which response is the reference.

Something that might be new in this discussion is the introduction of *conal ordering.* One can think of *number of events over number able* as being stratified to a cohort, the *able.* The *able* is usually determined by ordered survival age where the number *able* includes all subjects that have survived past the reference time.

In a 2-dimensional plane, conal ordering perserves the transitive ordering process. How can we use this to our advantage? By cohort conal ordering in age and calendar time, the net effect is that both age-related and calendar-related effects are cancelled.

One way to continue this *robust* theme is to structure a series of conditional models. This is the sequence of models the author generally uses:

- Static: Fit a MPH/Cox model with a minimal number of time varying covariates and conal ordering. Any time varying covariates and $\phi(t)$ are treated as controlling variables and discarded.
- Time varying conal ordered (TVC): A model for the time varying covariates under conal ordering. The Static model is used as an *offset* or a covariate with pre-determined coefficient of 1. Any controlling time varying covariates, such as $\phi(t)$ are discarded.
- Time varying (TV): A model for time varying covariates in the usual sense, survival ordering by subject age, where both Static and TVC are used as offsets.
- Baseline: As described below, the baseline is fit via a weighted artificial martingale approach.

Using an artificial treatment approach discussed below, all of the above end up being communicated as linear components via a markup language. If one throws in a deterministic scorecard for model applicibility (such as a scorecard that returns 0 if the overall model is applicable, negative otherwise), one has the basis for specification of an entire suite of models.

In contrast to an after-the-fact markup, such as PMML, one might as well start with a markup for specification, fit the models and then return coefficients to the markup. Afterwhich, all encapsulated information is immediately available for implementation. Implementation is free and easily productionizable. The markup could be consumed by a model delivery engine, or via processes such as XSLT, and used to write out pedantic (and fast) implementations in any language. Please see WDSModelSpec for details.

## 5.5 Estimating the baseline, the easy one first

### 5.5.1 A fast way that can be smooth

Before hitting the usual suspect in PH/Cox models for the baseline, let us consider something that is more elegant and easy to fit.

The basic concept is: Even for a self-censoring counting process, i.e., one and done, the compensated process is still a martingale:

$$E[N_t - \int_0^t \lambda(s)ds] = 0$$

And with lack of any notational rigor, we also have:

$$E[dN_t - \lambda(t)] = E[dN_t - e^{\beta Z} h_0(t)] = 0$$

As a matter of practice, covariates are always pre-processed with some artificial treatment (see WDSModelSpec). This includes time, $t$. Therefore, the baseline $h_0(t)$ is modeled as

$$h_0(t) = \beta^{h_0} W(t)$$

where $W(t)$ is a vector of artificials of $t$. The only criterion is that the artificial treatment (or any linear combination thereof) does not yield negative values for $t \geq 0$. Again, see WDSModelSpec for detail, but the artificials could be something like *Hats* (BZ1) for piece-wise linear and continuous or BZ treatements with higher orders.

In this way, $\beta^{h_0}$ can be fit through linear regression of

$$N_t = \beta^{h_0}[\int_0^t e^{\beta Z} W(s)ds]$$

and/or

$$dN_t = \beta^{h_0}[e^{\beta Z} W(t)]$$

where $e^{\beta Z} W(t)$ is the vector weighted artificials and $[\int_0^t e^{\beta Z} W(s)ds]$ is the vector of integrated weighted artificials.

In the long form for MPH/Cox, the Radon-Nikodym derivative for the change in measure between baselines has already been incorporated into the model form for one of the competing risks:

$$h(t) = h^d(t) + h^p(t) = e^{\beta^d Z} h_0^d(t) + e^{\beta^p Z_i} h_0^p(t) = (e^{\beta^d Z} + e^{\beta^p Z_i + \phi(t)}) h_0(t)$$

Therefore, the common baseline can be fit with a method analogous to above where $N_t$ is the total exit counting process and the baseline artificials are weighted by $e^{\beta^d Z} + e^{\beta^p Z_i + \phi(t)}$.

Relatively short and simple.

### 5.5.2 The usual suspect

Most references on baselines either give a formula or some glancing idea at how it was derived, but to understand its extension, it helps to see the derivation. Even though it is commonly said that the baseline hazard rate is estimated empirically from the data, it is actually a maximum likelihood estimate. Assume that the event times $T_1, \ldots, T_N$ are ordered as above, but let us temporarily assume that there are no censored observations. The baseline will be estimated in a multiplicative manner, and let $\alpha_n$ be the factor:

$$\alpha_n = \frac{S_0(T_n)}{S_0(T_{n-1})}.$$

It follows that $S(T_n)$ has the form:

$$S(T_n) = \left[\prod_{1\le k\le n} \alpha_k\right]^{e^{\beta Z_n}}.$$

The probability mass function then has the form:

$$pmf_n = (1-\alpha_n^{e^{\beta Z_n}}) \prod_{1\le k<n} \alpha_k^{e^{\beta Z_n}}.$$

Given $\beta$, the total likelihood of the dataset with parameters $\alpha_k$ is given by the product of corresponding terms of the above form, the log-likelihood is:

$$\log \prod_{1\le n\le N} pmf_n = \sum_{1\le n\le N} \log(1-\alpha_n^{e^{\beta Z_n}}) + \sum_{1\le n\le N}\sum_{1\le k<n} e^{\beta Z_n}\log(\alpha_k).$$

To maximize this log-likelihood, let us look to the point where the derivatives of the above quantity with respect to each $\alpha_n$ equate to zero. These derivatives are given by:

$$\frac{\delta LL}{\delta \alpha_n} = \frac{1}{1-\alpha_n^{e^{\beta Z_n}}}\left[-e^{\beta Z_n}\alpha_n^{e^{\beta Z_n}-1}\right] + \frac{1}{\alpha_n}\sum_{k>n} e^{\beta Z_k}.$$

Equating to 0 for each $1\le n\le N$, and solving for the term, $\alpha_n^{e^{\beta Z_n}}$, one obtains:

$$\alpha_n^{e^{\beta Z_n}} = \frac{e^{-\beta Z_n}\sum_{k>n} e^{\beta Z_k}}{1+e^{-\beta Z_n}\sum_{k>n} e^{\beta Z_k}}.$$

A quick gyration yields:

$$\alpha_n = \left[1-\frac{e^{\beta Z_n}}{\sum_{k\ge n} e^{\beta Z_k}}\right]^{e^{-\beta Z_n}}.$$

This solution holds when each subject with an observed event as a unique event time. The equation is adapted in the case of ties and the technique is to solve a system of equations. If this is not easily observable, consider the following: Suppose there is a multiplicity of events at time $n$, say $l_n$. If every $Z_i$ with $t_i = t_n$ were identical, adding a factor of $l_n$ to the derivatives above, setting equal to 0 and solving would be elementary. However, the summation of $\alpha_n^{e^{\beta Z_i}}$ terms complicates this.

Extending this concept to censoring is simple, only non-censored observations contribute a term to the likelihood. Extending this to marked point processes follows exactly the same concept, if a loan prepays, its default stream is censored and only the prepayment term contributes to likelihood. There will also be switching from continuous to discrete time and back again to obtain the proceeding algorithm. **If one would want to save oneself considerable grief, with these insights one could skip to the final answer of this section and understand the result.** Otherwise,....

In continuous time, the total survival function for subject $n$ is given by:

$$S^n(t) = 1 - e^{-\int_0^t \left(e^{\beta^d Z_n}+e^{\beta^p Z_n}e^{\phi(u)}\right)h_0(u)du}.$$

For defaults, the cumulative distribution satisfies the following differential equation:

$$\frac{dC^d(t)}{dt} = e^{\beta^d Z_n} h_0(t) S^n(t),$$

and for prepayments:

$$\frac{dC^p(t)}{dt} = e^{\beta^p Z_n} e^{\phi(t)} h_0(t) S^n(t),$$

Given the mark of subject $n$, $M_n$, and taking the time at $t_n$, let $\Delta t_n = t_n - t_{n-1}$ and let us rewrite these two densities in a way amenable to the probability mass function above:

$$pmf_n \approx \left[1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}e^{\phi(t_n)}\right] h_0(t_n) e^{-\int_{t_{n-1}}^{t_n}\left(e^{\beta^d Z_n}+e^{\beta^p Z_n}e^{\phi(u)}\right)h_0(u)du} S^n(t_{n-1})\Delta t_n.$$

To make this equation a little less unwieldy, let us temporarily use the notation, $a_n = e^{\beta^d Z_n}$, $b_n = e^{\beta^p Z_n}$, and $c(t) = e^{\phi(t)}$, rewriting the equation as,

$$pmf_n \approx \left[1_{\{M_n=d\}}a_n + 1_{\{M_n=p\}}b_n c(t_n)\right] h_0(t_n) e^{-\int_{t_{n-1}}^{t_n}(a_n+b_n c(t_n))h_0(u)du} S^n(t_{n-1})\Delta t_n.$$

Taking $c(t) = t_n$ and $h_0(t) = h_0(t_n)$ over $(t_{n-1}, t_n]$, and recalling that $\lambda\Delta t \approx 1 - e^{-\lambda\Delta t}$ for an exponential, simplifies the expression to

$$\begin{aligned}
pmf_n &\approx \left[1_{\{M_n=d\}}a_n + 1_{\{M_n=p\}}b_n c(t_n)\right] h_0(t_n) e^{-(a_n+b_n c(t_n))h_0(t_n)\Delta t_n} S^n(t_{n-1})\Delta t_n. \\
&\approx \Big[1_{\{M_n=d\}}a_n h_0(t_n)\Delta t_n e^{-a_n h_0(t_n)\Delta t_n} e^{-b_n c(t_n) h_0(t_n)\Delta t_n} \\
&\qquad\qquad + 1_{\{M_n=p\}} b_n c(t_n) h_0(t_n)\Delta t_n e^{-b_n c(t_n) h_0(t_n)\Delta t_n} e^{-a_n h_0(t_n)\Delta t_n}\Big] S^n(t_{n-1}). \\
&\approx \Big[1_{\{M_n=d\}}(1 - e^{-a_n h_0(t_n)\Delta t_n}) e^{-b_n c(t_n) h_0(t_n)\Delta t_n} \\
&\qquad\qquad + 1_{\{M_n=p\}}(1 - e^{-b_n c(t_n) h_0(t_n)\Delta t_n}) e^{-a_n h_0(t_n)\Delta t_n}\Big] S^n(t_{n-1}).
\end{aligned}$$

Taking $\alpha_n = e^{-h_0(t_n)\Delta t_n}$, yields:

$$\begin{aligned}
pmf_n &\approx \left[1_{\{M_n=d\}}(1-\alpha_n^{a_n})\alpha_n^{b_n} + 1_{\{M_n=p\}}(1-\alpha_n^{b_n c(t_n)})\alpha_n^{a_n}\right] S^n(t_{n-1}). \\
&\approx \left[1_{\{M_n=d\}}(1-\alpha_n^{a_n})\alpha_n^{b_n} + 1_{\{M_n=p\}}(1-\alpha_n^{b_n c(t_n)})\alpha_n^{a_n}\right] \prod_{1\le k<n} \alpha_n^{a_n+b_n c(t_n)}.
\end{aligned}$$

Following the technique above for a vanilla baseline calculation, taking the derivative of the log-likelihood yields:

$$\begin{aligned}
\frac{\delta LL}{\delta\alpha_n} &= 1_{\{M_n=d\}}\left[\frac{1}{1-\alpha_n^{a_n}}(-a_n\alpha_n^{a_n-1}) + \frac{b_n}{\alpha_n}\right] \\
&\quad + 1_{\{M_n=p\}}\left[\frac{1}{1-\alpha_n^{b_n c(t_n)}}(-b_n c(t_n)\alpha^{b_n c(t_n)-1}) + \frac{a_n}{\alpha_n}\right] \\
&\quad + \frac{1}{\alpha_n}\sum_{k>n}(a_k + b_k c(t_n)).
\end{aligned}$$

It may take a moment of thought to convince oneself, but within the last summation for the $k^{th}$ term, $c(t)$ must be evaluated at $t_n$. Equating each partial derivative to zero and using a similar gyration as above, one obtains:

$$\begin{aligned}
\alpha_n &= \left[1 - \frac{1_{\{M_n=d\}}a_n + 1_{\{M_n=p\}}b_n c(t_n)}{\sum_{k\ge n} a_k + b_k c(t_n)}\right]^{\frac{1}{1_{\{M_n=d\}}a_n + 1_{\{M_n=p\}}b_n c(t_n)}} \\
&= \left[1 - \frac{1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}(t_n)}{\sum_{k\ge n} e^{\beta^d Z_k} + e^{\beta^p Z_k}e^{\phi(t_n)}}\right]^{\frac{1}{1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}e^{\phi(t_n)}}}.
\end{aligned}$$

In the case of tied event times, the system of equations generated by setting derivatives to 0 is solved.

## 6 Further extensions and equations for Newton-Raphson fitting

Once the format for two outcomes is established, extension to a greater number of outcomes is direct. For example, if the mark set is $m_0, m_1, \ldots, m_l$, (with $m_0$ as the censored mark), the key to this extension is a family of Radon-Nikodym derivatives, $\phi_2(t), \ldots, \phi_l(t)$, between the multiple hazard functions and one pre-specified reference outcome. These could of course be fit using a common reference generalized logit model. The quanitities above of the form $1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}e^{\phi(t_n)}$ would be replaced by

$$1_{\{M_n=m_1\}}e^{\beta^1 Z_n} + \sum_{k=2}^{l} 1_{\{M_n=m_k\}}e^{\beta^k Z_n}e^{\phi_k(t_n)}.$$

If we use a linear form for $\phi_k(t) = \phi_k t$, (now treating $\phi_k$ and $t$ as vector valued), the $\phi^k$ could also be fit simulataneously to the $\beta^k$. This would alleviate some of the censored data issues around the preconditioning model. Let us consider the following form for the likelihood function:

$$L = \prod_{M_i \neq e} \frac{\sum_{k=1}^{l} 1_{\{M_n = m_k\}} e^{\beta^k Z_i} e^{\phi_k T_{i,i}} \eta(i,i)}{\sum_{j=i}^{N} \left( \sum_{k=1}^{l} e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \right) \eta(i,j)}.$$

In case you missed it, there is a $\phi_1$. Let us generalize $T_j$ to $T_{j,i}$, where $T_{j,i}$ represents a vector of time varying covariates for observation $j$ taken at the time of observation $i$. By appropriate selection of variables, one can use time varying data such as incentive or home price index even for the reference class. There is an open question of identifiability. For time in the usual sense, this issue can be resolved by coding the problem in a fully flexible manner where one can totally specify which variables get used for each model, i.e., time in the usual sense will not be specified for the reference class. Recall that one of the reasons that time is a special case is that we are using the Radon-Nikodym derivatives to eliminate *as much as is common to all members of the cohort at that event epoch.* Therefore any variable that has the same impact for all observations must be treated as a time-like variable. Observation unique or at least some-what unique variables can be specified as time-varying.

Let us also consider the strata functional $\eta(i,j)$. It will generally be assumed that $\eta(i,i) = 1$ but it is included for completeness. In the usual sense of *strata* in a Cox model, $\eta$ will be either 0 or 1 depending on whether or not $i$ and $j$ are within the same strata. Let us generalize this to the sense of a potentially *soft* strata case where $\eta$ is a neighbor defining metric. It can also be a catch all term for a weighting scheme. Clearly bayesian notions will surround the parameter definitions within $\eta$ and the area will need some more thought.

Admittedly, it was first noticed during early modeling experiments, but there is a natural interaction between the concepts of the previous two paragraphs. A hard strata structure is used whereever one works under the assumption of a fixed proportionality function for each covariate, regardless of the baseline. However, the strata structure and functional, $\eta(i,j)$, are used to guarantee that only common baselines are canceled out. It is therefore only natural to discover that if different baselines are reasonable for different cohorts, then the Radon-Nikodym derivatives, $\phi^k$, are also likely to change by strata. In our generalized covariate, $T$, we may want to consider some elements treated as *time-like* variables in the usual sense which probably need to be interacted with strata, or maybe not, in order to fit a fixed proportionality function across cohorts. Once again, a fully flexible piece of software for fitting will cure the problem.

For model fitting, we will employ Newton's method, for a quick reminder: Let our vector of parameters be $\vec{\beta}$. Since the likelihood $L$ is the product of probabilities (usually small) and these often have a multiplicative or exponential form themselves, we will maximize the log-likelihood, $LL$, for an additive form. The candidate for a maximum must satisfy $\nabla LL(\vec{\beta}) = 0$. If the Fisher information matrix, $-\mathrm{E}\left[\nabla^2 LL(\vec{\beta})\right]$, is nicely behaved, i.e., nothing like quasi-complete separation in logistic regression, the following algorithm should converge to a maximum: Given an initial estimate, $\vec{\beta}_0$, iterate until desired convergence through $\vec{\beta}_{n+1} = \vec{\beta}_n + \Delta\vec{\beta}_n$ where

$$\nabla^2 LL(\vec{\beta}_n)\Delta\vec{\beta}_n = -\nabla LL\vec{\beta}_n.$$

For testing variable significance, we will use the inverse of the Fisher information matrix as a variance estimate of the parameter estimates,

$$cov(\vec{\beta}) = (-\nabla^2 LL(\vec{\beta}))^{-1}.$$

Thus, for a Newton-Raphson technique the following quantities are required:

$$\frac{\delta LL}{\delta \beta_a^{k_1}} = \sum_{M_i \neq m_0} \left[ 1_{\{M_n = m_{k_1}\}} Z_{i,a} \eta(i,i) - \frac{\sum_{j=i}^N Z_{j,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j)} \right]$$

$$\frac{\delta LL}{\delta \phi_a^{k_1}} = \sum_{M_i \neq m_0} \left[ 1_{\{M_n = m_{k_1}\}} T_{i,i,a} \eta(i,i) - \frac{\sum_{j=i}^N T_{j,i,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j)} \right]$$

$$\frac{\delta^2 LL}{\delta \beta_a^{k_1} \delta \beta_b^{k_1}} = \sum_{M_i \neq m_0} \left[ - \frac{\sum_{j=i}^N Z_{j,a} Z_{j,b} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j)} + \frac{\sum_{j=i}^N Z_{j,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \sum_{j=i}^N Z_{j,b} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\left( \sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j) \right)^2} \right]$$

$$\frac{\delta^2 LL}{\delta \beta_a^{k_1} \delta \phi_b^{k_1}} = \sum_{M_i \neq m_0} \left[ - \frac{\sum_{j=i}^N Z_{j,a} T_{j,i,b} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j)} + \frac{\sum_{j=i}^N Z_{j,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j) \sum_{j=i}^N T_{j,i,b} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\left( \sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j) \right)^2} \right]$$

$$\frac{\delta^2 LL}{\delta \phi_a^{k_1} \delta \phi_b^{k_1}} = \sum_{M_i \neq m_0} \left[ - \frac{\sum_{j=i}^N T_{j,i,a} T_{j,i,b} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j)} + \frac{\sum_{j=i}^N T_{j,i,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j) \sum_{j=i}^N T_{j,i,b} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j)}{\left( \sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j) \right)^2} \right]$$

$$\frac{\delta^2 LL}{\delta \beta_a^{k_1} \delta \beta_b^{k_2}} = \sum_{M_i \neq m_0} \left[ \frac{\sum_{j=i}^N Z_{j,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j) \sum_{j=i}^N Z_{j,b} e^{\beta^{k_2} Z_j} e^{\phi_{k_2} T_{j,i}} \eta(i,j)}{\left( \sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j) \right)^2} \right]$$

$$\frac{\delta^2 LL}{\delta \beta_a^{k_1} \delta \phi_b^{k_2}} = \sum_{M_i \neq m_0} \left[ \frac{\sum_{j=i}^N Z_{j,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j) \sum_{j=i}^N T_{j,i,b} e^{\beta^{k_2} Z_j} e^{\phi_{k_2} T_{j,i}} \eta(i,j)}{\left( \sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j) \right)^2} \right]$$

$$\frac{\delta^2 LL}{\delta \phi_a^{k_1} \delta \phi_b^{k_2}} = \sum_{M_i \neq m_0} \left[ \frac{\sum_{j=i}^N T_{j,i,a} e^{\beta^{k_1} Z_j} e^{\phi_{k_1} T_{j,i}} \eta(i,j) \sum_{j=i}^N T_{j,i,b} e^{\beta^{k_2} Z_j} e^{\phi_{k_2} T_{j,i}} \eta(i,j)}{\left( \sum_{j=i}^N \sum_{k=1}^l e^{\beta^k Z_j} e^{\phi_k T_{j,i}} \eta(i,j) \right)^2} \right]$$

## 6.1 Old notes on the longer form for the baseline

In the case of no ties and regarding the $LL$ for the baseline and talking the product below only over non-censored observations, we have

$$LL = \log \prod_{1 \leq n \leq N} pmf_n = \sum_{1 \leq n \leq N} \left[ \sum_{k=1}^l 1_{\{M_n = m_k\}} \log(1 - \alpha_n^{e^{\beta^k Z_n} e^{\phi^k T_{n,n}}}) + \sum_{1 \leq j < n} \left( \sum_{k=1}^l e^{\beta^k Z_n} e^{\phi^k T_{n,j}} \right) \log(\alpha_j) \right].$$

Recall that the parameters, $\alpha_k$, are the factors of the baseline survival. For the two outcome case, we had the following derivative (in a slightly simpler form than presented previously):

$$\frac{\delta LL}{\delta \alpha_n} = \frac{1}{\alpha_n} \left( -1_{\{M_n = d\}} \frac{a_n}{1 - \alpha_n^{a_n}} - 1_{\{M_n = p\}} \frac{b_n c(t_n)}{1 - \alpha_n^{b_n c(t_n)}} + \sum_{k \geq n} (a_k + b_k c(t_n)) \right).$$

In the case of ties and applying the multinomial form above, we obtain,

$$\frac{\delta LL}{\delta \alpha_n} = \frac{1}{\alpha_n} \left( - \sum_{\{t_g = t_n \cap M_g \neq m_0\}} \sum_{k=1}^l 1_{\{M_g = m_k\}} \frac{e^{\beta^k Z_g} e^{\phi_k T_{g,n}}}{1 - \alpha_n^{e^{\beta^k Z_g} e^{\phi_k T_{g,n}}}} + \sum_{t_g \geq t_n} \sum_{k=0}^l e^{\beta^k Z_g} e^{\phi_k T_{g,n}} \right).$$

With some abuse of notation, the above assumes only one $\alpha_n$ for each unique time $t_n$. The second derivative proceeds directly,

$$\frac{\delta^2 LL}{\delta\alpha_n^2} \;=\; \frac{1}{\alpha_n}\left(-\frac{\delta LL}{\delta\alpha_n} - \sum_{\{t_g=t_n\cap M_g\neq m_0\}}\sum_{k=1}^{l} 1_{\{M_g=m_k\}}\frac{\left(e^{\beta^k Z_g}e^{\phi_k T_{g,n}}\right)^2\alpha_n^{e^{\beta^k Z_g}e^{\phi_k T_{g,n}}-1}}{\left(1-\alpha_n^{e^{\beta^k Z_g}e^{\phi_k T_{g,n}}}\right)^2}\right).$$

It follows that the second derivative is negative whenever the first derivative is 0 and at least a local maximum criterion is satisfied.

The other more pressing reason to consider the the first and second derivatives is a parameterization of the baseline. Now, it is clear that anything where the first derivative above is non-zero is sub-optimal, however, a parameterization cleans up the baseline and reduces overfitting noise. Since for each $\alpha_n$ there is a unique $t_n$, suppose that we have the form $\alpha_n = \theta(t_n)$ and that $\theta(t)$ has a parameter $\theta_p$. Note that it is required that $0 \leq \theta(t) \leq 1$. We can then use the following equations for Newton-Raphson fitting:

$$\begin{aligned}
\frac{\delta LL}{\delta\theta_p} &= \sum_n \frac{\delta LL}{\delta\alpha_n}\frac{\delta\alpha_n}{\delta\theta_p}\\
\frac{\delta^2 LL}{\delta\theta_p^2} &= \sum_n \left[\frac{\delta LL^2}{\delta\alpha_n^2}\left(\frac{\delta\alpha_n}{\delta\theta_p}\right)^2 + \frac{\delta LL}{\delta\alpha_n}\frac{\delta^2\alpha_n}{\delta\theta_p^2}\right]\\
\frac{\delta^2 LL}{\delta\theta_p\delta\theta_q} &= \sum_n \left[\frac{\delta LL^2}{\delta\alpha_n^2}\frac{\delta\alpha_n}{\delta\theta_p}\frac{\delta\alpha_n}{\delta\theta_q} + \frac{\delta LL}{\delta\alpha_n}\frac{\delta^2\alpha_n}{\delta\theta_p\delta\theta_q}\right]
\end{aligned}$$

# 7 Old References

1. *Applied Survival Analysis, Regression Modeling of Time to Event Data*, David W. Hosmer, Jr. and Stanly Lemeshow, Wiley Series in Probability and Statistics, 1999.

2. Cinlar