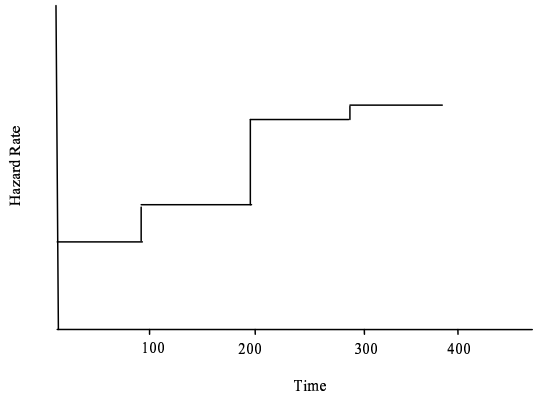
Semi-Parametric Duration Models: The Cox Model

There are basically two semi-parametric alternatives to the parametric models that we examined earlier: (i) the piecewise-constant exponential (PCE) model and (ii) the Cox model.

1 The Piecewise-Constant Exponential (PCE) model

The piecewise-constant exponential model is a semi-parametric continuous time duration model.[1] The model is semi-parametric in the sense that the shape of the hazard rate with respect to time is not specified *a priori* but is determined by the data. The basic idea underlying the PCE model is that the duration time can be divided up into discrete units in each of which the hazard rate is assumed to be constant across time. In other words, the hazard rate is allowed to differ in different time periods but is assumed to be constant within any given time period. The advantage of the PCE approach is that the overall shape of the hazard rate does not have to be imposed in advance as with the parametric models. One use of the PCE model is to see how the hazard rate varies with time and to use this information to choose an appropriate parametric model. For example, if the PCE reveals that the hazard rate increases monotonically with time as in Figure 1, we could safely adopt the Weibull parametric model. In this sense, the PCE model is an alternative method for adjudicating between different parametric models.

Figure 1: The Piecewise-Constant Exponential Model



One of the key issues with the PCE model involves determining the appropriate number of time intervals to be used. The number of time intervals is something that must be determined by the analyst. Although any number of time periods can be chosen, it is important to recognize that there is always a tradeoff to be made. If one chooses a large number of time periods, then we will get a better approximation of the unknown baseline hazard but we will have to estimate a larger number of coefficients and this may cause problems. Alternatively, if one chooses a small number of time periods, then there will be fewer estimation problems but the approximation of the baseline hazard will be worse. A key requirement when choosing the number of time periods is that there should be units that fail within each of the different time intervals. If this is not the case, then one will not obtain sensible estimates.

[1] The discussion of the PCE model is based largely on Jenkins (2008).

1.1 The Setup

The PCE model is a special case of models that employ time-varying covariates. This is because the PCE requires us to split up single-spell duration data in the same way that we had to when we wanted to incorporate time-varying covariates. The PCE model is also a proportional hazard (PH) model as its basic hazard rate can be specified in the following way: $h(t, X) = h_o(t)e^{X\beta}$.[2] The main difference is that the baseline hazard rate is allowed to vary in different time periods. Thus, the hazard rate in the PCE model is specified as:

$$h(t, X_t) = \begin{cases} \overline{h}_o(t_1)e^{X_1\beta} & t \in (0, \tau_1) \\ \overline{h}_o(t_2)e^{X_2\beta} & t \in (\tau_1, \tau_2) \\ \vdots & \vdots \\ \overline{h}_o(t_K)e^{X_K\beta} & t \in (\tau_{K-1}, \tau_K) \end{cases}$$

The baseline hazard, $\overline{h}_o(t)$ is constant with each of the K time periods but may differ between them. Covariates may be fixed or, if time-varying, constant within each time period.

It is possible to rewrite this expression as:

$$h(t, X_t) = \begin{cases} \exp[\log(\overline{h}_o(t_1)) + X_1\beta] & t \in (0, \tau_1) \\ \exp[\log(\overline{h}_o(t_2)) + X_2\beta] & t \in (\tau_1, \tau_2) \\ \vdots & \vdots \\ \exp[\log(\overline{h}_o(t_K)) + X_K\beta] & t \in (\tau_{K-1}, \tau_K) \end{cases}$$

or, more simply, as:

$$h(t, X_t) = \begin{cases} \exp(\tilde{\lambda}_1) & t \in (0, \tau_1) \\ \exp(\tilde{\lambda}_2) & t \in (\tau_1, \tau_2) \\ \vdots & \vdots \\ \exp(\tilde{\lambda}_K) & t \in (\tau_{K-1}, \tau_K) \end{cases}$$

where $\tilde{\lambda}_1 = \log(\overline{h}_o(t_1)) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$. As you can see, the constant time period-specific hazard rates are equivalent to having time period specific intercept terms in the overall hazard. This is the key to estimating the PCE model. In effect, we can estimate the PCE model by creating a series of dichotomous variables that refer to each time period and including them in our model. The estimated coefficients on these dichotomous variables then indicate the baseline hazard in each time period. Obviously, it is not possible to include all of the time period dummies as well as a constant in the model. As a result, one either includes all of the time period dummies but omit the constant or one includes the intercept and all but one of the time dummies.

1.2 Estimating the PCE Model in STATA

There is no canned command to estimate the PCE model in STATA. However, it is relatively straightforward to get STATA to estimate it. The first thing you must do is STSET the data. Unlike some of our earlier examples, you will need to use the ID() option in the STSET command.

```
stset duration, failure(event) id(cabinetcode);
```

[2] As with the standard exponential model, the PCE can also be specified as an AFT model and estimated by using the TIME option in STATA.

You will then need to split your data into different time periods. You can do this with the STSPLIT command. I now demonstrate how to do this using data on government duration from the Constitutional Change and Parliamentary Democracy project. Government duration is measured in days with the maximum duration equal to 1936 days. To split the data into ten time periods of 200 days each, we type:

```
stsplit time, at (200 (200) 1936);
```

This gives us data that is split into the following intervals: [0, 200), [200, 400), [400, 600), ..., [1800, ∞).[3] We then need to transform this categorical variable into a series of dichotomous variables indicating one of the ten time periods. There are a number of ways of doing this.

Approach 1:

You could type the following:

```
tab time, gen(t);
```

This will create ten dummy variables, t1, t2, t3, ..., t10. You can then estimate the PCE model by including all ten of these dummy variables and no constant:

```
streg mwc t1 t2 t3 t4 t5 t6 t7 t8 t9 t10, dist(exp) nohr
           noconstant;
```

or nine of these dummy variables and a constant:

```
streg mwc t1 t2 t3 t4 t5 t6 t7 t8 t9 , dist(exp) nohr;
```

Approach 2:

You could also type:

```
forvalues k=1/9{
    gen in_‘k’ = ((‘k’-1)*200 <=time) & (time < ‘k’*200)
    }
```

This will create nine dichotomous variables, in_1, in_2, ..., in_9, where the dummy variable for the tenth time period is omitted. You would then include all nine dummy variables and the intercept in your model:

```
streg mwc in_* , dist(exp) nohr;
```

Approach 3:

You could also use STATA's XI command and estimate the model in one step:

```
xi, prefix(in_) : streg mwc i.time, dist(exp) nohr
```

In this last approach, we include nine time period dummies and the constant, with the omitted dummy variable being the one for the first time period.

[3] Obviously, you do not have to split your spells into intervals of the same length. For example, you can type: STSPLIT TIME, AT(350, 450, 900, 1400) if you want the cutpoints to be at 350, 450, 900, and 1400 days.

The coefficients on the time dummies from these different approaches will vary depending on whether you decide to include a constant or not, and if you include a constant, on the particular time dummy that is omitted. However, the coefficients on your covariates will always be the same. In other words, all the alternatives shown above will lead to the same coefficient estimate for our one covariate, MWC.

The interpretation of the results from the PCE model is essentially the same as the interpretation of the results from a standard parametric model. For example, you can use factor and percent changes in the baseline hazard (survival time) by exponentiating the coefficients on the covariates of interest if you have used a PH specification (AFT specification).

In this standard setup of the PCE model, we have assumed that the hazard rate varies across time periods but that the effect of the covariates is the same. However, we can easily allow the effect of our covariates to also vary from time period to time period by including interactions between our time period dummies and our covariates of interest. For example, we might estimate the following model:

```
streg mwc t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
            mwc_t1 mwc_t2 mwc_t3 mwc_t4 mwc_t5 mwc_t6
            mwc_t7 mwc_t8 mwc_t9 mwc_t10,
            dist(exp) nohr noconstant;
```

2 The Cox Model

The Cox model is another semi-parametric model. However, it is more general than the PCE model because it allows us to estimate the slope parameters in the β vector *irrespective* of what the baseline hazard looks like. In other words, the Cox model makes no assumptions about the distribution of the survival times.

The Cox model is a proportional hazards model. Thus, its basic specification can be written as:

$$h(t) = h_0(t)e^{X\beta} \tag{1}$$

The key to being able to estimate the slope parameters in the β vector without having to make any assumptions about the functional form of the baseline hazard is the use of partial likelihood methods. Partial likelihood, as we will see, works by focusing on the *ordering* of events.[4]

2.1 Estimating the Cox Model

For each uncensored observation, we know $T_i =$ duration and $C_i =$ censoring variable. If there are no tied failure times (units don't fail at exactly the same time), then for a given set of data there are N distinct event times called t_j. If an event occurred at a particular time t_j, we might want to know, given that there was an event, what the probability it was that it was observation k (with covariates X_k) that failed i.e. P(Observation k had an event at $t_j \mid$ one event occurred at t_j). We can write this conditional probability as:

$$\frac{P(\text{Observation } k \text{ has an event at } t_j)}{P(\text{ One event at } t_j)} \tag{2}$$

The numerator is just the hazard for individual k at time t_j, while the denominator is the sum of all hazards at t_j for all individuals who were at risk at t_j. Thus, we can write this as:

$$\frac{h_k(t_j)}{\sum_{l \in R_j} h_l(t_j)} \tag{3}$$

[4]What follows is based on lecture notes from Chris Zorn.

where R_j denotes the risk set at t_j. By substituting in the hazard function for the Cox model we have:

$$\frac{h_0(t_j)e^{X_k\beta}}{\sum_{l\in R_j} h_0(t_j)e^{X_l\beta}} = \frac{e^{X_k\beta}}{\sum_{l\in R_j} e^{X_l\beta}} \tag{4}$$

where the equality holds because the baseline hazards cancel out. This is the nice bit of the Cox model. Because the baseline hazards drop out, we do not need to make any assumptions about it. Note at this point that because the baseline hazard drops out, the Cox model does not estimate a constant. This is because the constant is essentially absorbed into the baseline hazard just as we saw before with proportional hazard parametric models i.e. $h_0(t)e^{\beta_0}$. Each observed event in our sample contributes one term like the one shown in Eq. (4) to the partial likelihood for the sample. It is referred to as a partial likelihood function because the Cox model is only using 'part' of the available data, i.e., $h_0(t)$ is not estimated. The partial likelihood can for all intents and purposes be treated like the likelihood. The partial likelihood for the sample is:

$$\begin{aligned} PL &= \prod_{j=1}^{N} \frac{e^{X_j\beta}}{\sum_{l\in R_j} e^{X_l\beta}} \\ &= \prod_{j=1}^{N} P_j \end{aligned} \tag{5}$$

where j denotes the N distinct event times, X_j denotes the covariate vector for the unit that actually experienced the event at t_j, and P_j is the conditional probability, that of those units at risk at t_j, it was observation j that experienced the event.

As before, we prefer to work with the log partial likelihood to get rid of the product term. So we have:

$$\ln PL = \sum_{j=1}^{N} \left\{ X_j\beta - \ln \left[\sum_{l\in R_j} e^{X_l\beta} \right] \right\} \tag{6}$$

We then maximize Eq. (6) with respect to β. It should be noted again that the partial likelihood does not take account of the actual duration each observation lasts; all that matters is the order in which observations fail. Censored observations enter the calculations only because they determine the size of the risk set. This is exactly the same idea as when we discussed how the Kaplan-Meier survival function was calculated.

2.2 An Example

Table 1: Sample Data

Subject	t	x
1	2	4
2	3	1
3	6	3
4	12	2

There are four failure times in these data: 2, 3, 6, and 12. As we have mentioned before, the values of the failure times do not matter; all that matters is the order in which observations fail. There are four distinct failure times and so we need to generate four distinct risk pools:

1. Time 2:
   Risk group: 1, 2, 3, 4
   Subject #1 is observed to fail.

2. Time 3:
   Risk group: 2, 3, 4
   Subject #2 is observed to fail.

3. Time6:
   Risk group: 3, 4
   Subject #3 is observed to fail.

4. Time 12:
   Risk group: 4
   Subject #4 is observed to fail.

As noted above, we assume that one observation fails at each failure time. As a result, we need to calculate the conditional probability of failure for the observation that is actually observed to fail. Thus, the partial likelihood function is:

$$PL = P_1 P_2 P_3 P_4 \tag{7}$$

where P_i, $i = 1, \ldots, 4$ indicates a conditional probability for each failure time. We now need to calculate these conditional probabilities. The last one is the easiest - $P_4 = 1$. In other words, the probability that observation 4 fails at $t = 12$ is 1. What about P_3?

$$P_3 = \frac{e^{X_3\beta}}{e^{X_3\beta} + e^{X_4\beta}} \tag{8}$$

We also have:

$$P_2 = \frac{e^{X_2\beta}}{e^{X_2\beta} + e^{X_3\beta} + e^{X_4\beta}}$$

$$P_1 = \frac{e^{X_1\beta}}{e^{X_1\beta} + e^{X_2\beta} + e^{X_3\beta} + e^{X_4\beta}} \tag{9}$$

Thus, we have:

$$\begin{aligned} PL &= \frac{e^{X_1\beta}}{e^{X_1\beta} + e^{X_2\beta} + e^{X_3\beta} + e^{X_4\beta}} \times \frac{e^{X_2\beta}}{e^{X_2\beta} + e^{X_3\beta} + e^{X_4\beta}} \times \frac{e^{X_3\beta}}{e^{X_3\beta} + e^{X_4\beta}} \times 1 \\ &= \prod_{j=1}^{4} \frac{e^{X_j\beta}}{\sum_{l \in R_j} e^{X_l\beta}} \end{aligned} \tag{10}$$

2.3 Tied Data

This all seems relatively straightforward. However, what happens when we have ties in the data i.e. what happens when multiple observations fail at some t_j? The basic problem tied events pose for the partial likelihood function is in the determination of the composition of the risk set at each failure time and the sequencing of event occurrences. In order to estimate the parameters of the Cox model with tied data, it becomes necessary to *approximate* the partial likelihood function. There are several ways to do this.

2.4 Breslow Method

Call d_j the number of events occurring at t_j, and D_j the set of d_j observations which have events at t_j. The logic is that since it is not possible to determine the order of occurrence in tied events, then we might want to assume that the size of the risk set is the same regardless of which event occurred first. In other words, we group all the tied events together. This means that we need to modify the numerator of Eq. (5) to include the covariates from all the observations that had an event at t_j and modify the denominator to account for the multiple possible orderings. After doing this we have:

$$PL = \prod_{j=1}^{N} \frac{e^{\left[(\sum_{q \in D_j} X_q)\beta\right]}}{\left[\sum_{l \in R_j} e^{X_l \beta}\right]^{d_j}} \tag{11}$$

An example might help to illustrate what is going on. Say we have four observations with respective failure times 5, 5, 8, 14. Two of the cases have tied failure times. Since we cannot determine which of the two cases failed first, the Breslow method would approximate the partial likelihood by assuming that the two cases failed from the risk set containing all four cases. To ease presentation, let say that $\psi_i = e^{X_i \beta}$. Let P_{12} be the probability that case one fails before case two and let P_{21} be the probability that case two fails before case one. The Breslow method will say that:

$$\begin{aligned} P_{12} &= \frac{\psi_1}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \times \frac{\psi_2}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \\ &= \frac{\psi_1 \psi_2}{(\psi_1 + \psi_2 + \psi_3 + \psi_4)^2} \end{aligned} \tag{12}$$

and

$$\begin{aligned} P_{21} &= \frac{\psi_2}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \times \frac{\psi_1}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \\ &= \frac{\psi_2 \psi_1}{(\psi_1 + \psi_2 + \psi_3 + \psi_4)^2} \end{aligned} \tag{13}$$

The contribution to the likelihood of the two observations that failed at $t = 5$ is obtained as:

$$P_{12} + P_{21} = \frac{2\psi_1 \psi_2}{(\psi_1 + \psi_2 + \psi_3 + \psi_4)^2} \tag{14}$$

The Breslow method is normally the default method in most statistical packages but is technically the least accurate. It works reasonably well when the number of failure times is small relative to the size of the risk group itself.

2.5 Efron Method

A modification of the Breslow method leads to the Efron method. The Efron method takes account of how the risk set changes depending on the sequencing of tied events. In effect, it adjusts the risk sets using probability weights. To illustrate how this method works consider again the little example from above. The Efron method would say that there are two possible sequences by which case 1 and case 2 could have failed. Case 1 may have failed first and so we have:

$$\frac{\psi_1}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \times \frac{\psi_2}{\psi_2 + \psi_3 + \psi_4} \tag{15}$$

or case 2 may have failed first and so we have:

$$\frac{\psi_2}{\psi_1+\psi_2+\psi_3+\psi_4} \times \frac{\psi_1}{\psi_1+\psi_3+\psi_4} \tag{16}$$

As you can see, the composition of the second risk set changes depending on the possible sequencing of events. Because case one is equally likely to fail first as case 2, the appearance of case 1 or case 2 in the second risk set is equally likely. This means that the probability of the second risk set would be $\frac{1}{2}(\psi_1+\psi_2)+\psi_3+\psi_4$. Thus, the Efron method leads to:

$$P_{12} = \frac{\psi_1}{\psi_1+\psi_2+\psi_3+\psi_4} \times \frac{\psi_2}{\frac{1}{2}(\psi_1+\psi_2)+\psi_3+\psi_4} \tag{17}$$

$$P_{21} = \frac{\psi_2}{\psi_1+\psi_2+\psi_3+\psi_4} \times \frac{\psi_1}{\frac{1}{2}(\psi_1+\psi_2)+\psi_3+\psi_4} \tag{18}$$

and so the contribution of these two cases to the partial likelihood is:

$$P_{12}+P_{21} = \frac{2\psi_1\psi_2}{(\psi_1+\psi_2+\psi_3+\psi_4)[\frac{1}{2}(\psi_1+\psi_2)+\psi_3+\psi_4]} \tag{19}$$

The Efron method is more accurate than the Breslow method.

2.6 Average Likelihood or Exact Partial Likelihood Method (exactm)

We can go a step further than the Efron method and take account of the fact that, if there are d events at t_j, then there are $d!$ possible orderings of those events. Thus, if there are four cases that all failed at t_j, then this method would take account of the 24 (4!) possible orderings of the event times in its approximation of the partial likelihood. This is the most accurate method. Continuing our example from before we have

$$P_{12} = \frac{\psi_1}{\psi_1+\psi_2+\psi_3+\psi_4} \times \frac{\psi_2}{\psi_2+\psi_3+\psi_4} \tag{20}$$

or case 2 may have failed first and so we have:

$$P_{21} = \frac{\psi_2}{\psi_1+\psi_2+\psi_3+\psi_4} \times \frac{\psi_1}{\psi_1+\psi_3+\psi_4} \tag{21}$$

and so the contribution to the likelihood is exactly $P_{12}+P_{21}$. This method is also called the marginal calculation, the exact-marginal calculation, or the continuous-time calculation and is the most accurate of the methods.

2.7 Exact Discrete Partial Likelihood Method (exactp)

All of the previous methods assume that data is generated from a continuous-time process. As a result, when two or more events occur simultaneously it is important to account for the possible sequencing of the events. In other words, the continuous-time assumption implies that the events do not really occur at the same time, just that our way of measuring time is not accurate enough to spot the different sequencing. The exact discrete method is different in that it assumes that events really do occur at the same time. This is essentially a multinomial problem. Given that two failures are to occur at the same time among cases 1, 2, 3, and 4 in our running example, the possibilities are that:

1. 1 and 2 fail
2. 1 and 3 fail
3. 1 and 4 fail
4. 2 and 3 fail
5. 2 and 4 fail
6. 3 and 4 fail

The conditional probability that cases 1 and 2 fail is:

$$P_{12} = \frac{\psi_1\psi_2}{\psi_1\psi_2 + \psi_1\psi_3 + \psi_1\psi_4 + \psi_2\psi_3 + \psi_2\psi_4 + \psi_3\psi_4} \tag{22}$$

This is the contribution of these two cases to the partial likelihood. Note that this calculation is also the calculation that conditional logit uses to calculate probabilities when conditioning on more than one event taking place. Here the observations are grouped together by the time period at which they are at risk of experiencing an event. In other words, the probability is conditional on the risk set at each discrete time period. This method is also known as the partial calculation, the exact-partial calculation, the discrete-time calculation, and the conditional logistic calculation.

So, which method should one use? Certainly, the exact partial likelihood and the exact discrete partial likelihood are more accurate. Choosing between these will depend on how you think about the tied failures. Do they arise from imprecise measurement in which case you might want to go with the exact partial likelihood, or do they arise from a discrete time model in which case you might want to go with the exact discrete partial likelihood. Zorn notes that he always uses the Efron method at a minimum.

To use the Breslow method (STATA's default) for dealing with ties, you would type:

- stcox X, nohr breslow

If you wanted to use something other than the BRESLOW, then get rid of BRESLOW and add either EFRON, EXACTP, or EXACTM.

2.8 Interpretation of Results

As I noted earlier, the Cox model uses a proportional hazards (PH) specification. As a result, the output can be interpreted in exactly the same way as you would interpret results from a parametric PH model.

1. You can interpret the sign and statistical significance of the coefficients.

   The sign of the coefficient indicates how a covariate affects the hazard rate. Thus, a positive coefficient increases the hazard rate and, therefore, reduces the expected duration. A negative coefficient decreases the hazard rate and, therefore, increases the expected duration. The statistical significance of the coefficient indicates whether these changes in the expected duration will be statistically significant or not.

2. You can calculate factor and percentage changes in the hazard ratio.

   You can exponentiate the coefficients to obtain hazard ratios. You can then use these hazard ratios to calculate the factor change or percentage change in the baseline hazard associated with a one unit increase in a covariate.

2.9 Baseline Hazard and Survivor Functions

Although the benefit of the Cox model is that we do not need to specify a particular hazard function, it turns out that we can still get estimates of the baseline hazard, the integrated hazard, and the survivor functions from the Cox model. To see how this is possible mathematically, see Box-Steffensmeier and Jones (2004, 64). I will not go through the math here. However, I will give you an intuition about how this works. Remember that:

$$h(t) = h_0(t)e^{X\beta} \tag{23}$$

We would like to get an estimate of the baseline hazard $h_0(t)$. How do we get this? Essentially, we estimate our model to obtain the coefficients in the β vector and then we estimate the components that make up $h_0(t)$.

Recall the following result from an earlier set of notes:

$$H(t) = \ln[-S(t)] \tag{24}$$

From this, we have:

$$\begin{aligned}
S(t) &= \exp[-H(t)] \\
&= \exp\left[-\int_0^t h_\tau d\tau\right] \\
&= \exp\left[-\exp(X_i\beta)\int_0^t h_0(\tau)d\tau\right] \\
&= \left[\exp\left(-\int_0^t h_0(\tau)d\tau\right)\right]^{\exp(X_i\beta)} \\
&= [\exp(-H_0(\tau))]^{\exp(X_i\beta)} \\
&= [S_0(t)]^{\exp(X_i\beta)}
\end{aligned} \tag{25}$$

where $S_0(t)$ is the baseline survivor function.

It turns out that Eq. (25) is very useful. After we have estimated our model, we know $e^{X_i\beta}$. We can then use the data to estimate the survivor function, $S(t)$ - this was essentially the KM plot we saw earlier. Thus, from the data we know $S(t)$ and $e^{X_i\beta}$. From here, we can get an estimate of the baseline survivor function, $S_0(t)$.

From the fact that we know $S(t)$ from the data, we also know $H(t)$. This is because $S(t) = e^{-H(t)}$. From here, we know that:

$$\begin{aligned}
H(t) &= \int_0^t h(\tau)d\tau \\
&= e^{X_i\beta}\int_0^t h_0(\tau)d\tau \\
&= e^{X_i\beta}H_0(t)
\end{aligned} \tag{26}$$

Since we know $H(t)$ because we know $S(t)$, and we know $e^{X_i\beta}$, we can estimate $H_0(t)$ and so get an estimate of the baseline cumulative hazard. Note also, that in deriving $H_0(t)$, we had the equation such that:

$$H(t) = e^{X_i\beta}\int_0^t h_0(\tau)d\tau \tag{27}$$

Since we know $H(t)$ and $e^{X_i\beta}$, we can get an estimate of $h_0(t)$. Thus, the basic story is that because we know $e^{X_i\beta}$ once we estimate the model and because we know S(t) (from the Kaplan-Meier estimate), we can get estimates of $H_0(t)$, $S_0(t)$, and $h_0(t)$.

One question that people sometimes have is whether the baseline hazard function from the Cox model is the same as the baseline hazard function from the Kaplan-Meier estimate we saw earlier. The confusion arises because we interpret the baseline hazard in the Cox model (and other PH models) as the hazard rate when all the covariates are 0 and because the Kaplan-Meier hazard rate is when we do not condition on any covariates. It would seem at first sight that these would be the same things. However, this is not the case, precisely because the baseline hazard from the Cox model uses the results from the Cox model (the βs) to get the estimate of the baseline hazard. In other words, the baseline hazard from the Cox model is *conditional* on the covariates. If we add covariates or remove covariates from the Cox model, the estimated baseline hazard will change. The Kaplan-Meier estimate of the baseline hazard on the other hand is not conditional on any covariates.[5]

To get the baseline integrated hazard in STATA, we type:

```
stcox X, nohr basechazard(cumhazard);

graph twoway connected cumhazard duration, c(J) sort
```

Recall that the baseline integrated hazard is for the situation where all of the covariates are set to 0. You may want to graph the cumulative hazard for scenarios where the covariates are set at more substantively meaningful values (or you might want to compare two scenarios). You can do this by remembering that the integrated or cumulative hazard $H(t)$ is given by:

$$H(t, X) = e^{X\beta}H_0(t) \tag{28}$$

As a result, you just choose some values for the covariates and calculate X_BETAHAT. Then you would type something like the following:

```
stcox X, nohr basechazard(cumhazard);

gen cumhazard2 = cumhazard*exp(x_betahat);

graph twoway connected cumhazard cumhazard2 duration, c(J J) sort;
```

To get the baseline survivor function, we simply type:

```
stcox X, nohr basesurv(surv);

graph twoway connected surv duration, c(J J) sort;
```

Again, the baseline survivor function is for the situation where all of the covariates are set to 0. You may want to graph the survivor function for scenarios when the covariates are set at more substantively meaningful values. You can do this by remembering that the survivor function $S(t)$ is given by:

$$S(t|x) = S_0(t)^{e^{x_i\beta}} \tag{29}$$

Again, you would choose some values for your covariates, calculate X_BETAHAT, and then type the following:

[5] If we estimate the Cox model with no covariates, then we would have the same baseline hazard as the KM estimate.

```
stcox X, nohr basesurv(surv);

gen e_xbetahat = exp(x_betahat);

gen surv2= surv^e_xbetahat;

graph twoway connected surv surv2 duration, c(J J) sort;
```

To get the baseline hazard function, we type:

```
stcox X, nohr basehc(baslinehazard);

graph twoway connected baslinehazard duration, c(J) sort;
```

You'll see that the baseline hazard is very jerky unlike the baseline hazards of the parametric models from earlier.[6] This is because the baseline hazard can only be estimated at the time in which failures are recorded. The baseline hazard is assumed constant between time periods and so we get the appearance of step level changes.

Because we know that the baseline hazard is not really jerky like this, we might want to smooth the baseline hazard from the Cox model. This can be done by typing:

```
stcox X, nohr basehc(baslinehazard);

lowess baselinehazard duration, c(J) sort;
```

Again, the baseline hazard is for the situation where the covariates are all set at 0. You can look at the baseline hazard for other scenarios by recognizing that:

$$h(t, X) = h_0(t)e^{X\beta} \tag{30}$$

3 Diagnostics

After we have run whatever continuous time duration model we think appropriate, we should probably conduct some diagnostic tests.

3.1 Proportional Hazards Assumption

One of the assumptions underlying PH models such as the Cox and Weibull models is the proportional hazards assumption. Recall that this assumption is the idea that covariates will have a proportional and constant effect that is invariant to time. Non-proportional hazards can arise if some covariate only affects survival up until some time t or if the size of its effect changes over time. Non-proportional hazards can result in biased estimates, incorrect standard errors, and faulty inferences about the effect of our covariates.

Tests for non-proportionality fall into three camps:

1. Piecewise regression to detect changes in parameter values.
2. Residual-based tests.
3. Explicit tests of coefficients for interactions of covariates and time.

[6] This raises a potential issue with the Cox model. The baseline hazard is closely adapted to the observed data in the sample and this may lead to an 'overfitted' estimate. The result is that the estimate of the baseline hazard can be highly sensitive to outlying event times and may lead to poor out of sample predictions.

3.1.1 Piecewise Regression

To conduct these tests, one estimates separate event history regression models for observations whose survival times fall above or below some predetermined value and assess whether the estimated covariate effects are consistent across the two separate models. You can, of course, divide the data into more than two categories. Obviously, these tests are somewhat subjective and based on an arbitrary division of the data. Better tests for non-proportionality exist for the Cox model; however, piecewise regression is the best that you can do for testing non-proportionality in parametric PH models.

3.1.2 Residual-Based Tests: Schoenfeld Residuals

In OLS, a residual is simply the difference between the observed value of the dependent variable and its predicted value. Residuals are not so obvious in a duration model since the value of the dependent variable might be censored and the fitted model may not provide an estimate of the systematic component of the model due to the use of partial likelihood methods. However, there are a number of 'residuals' that can be obtained from a duration model. Schoenfeld residuals are particularly useful for testing the assumption of proportional hazards.[7]

Schoenfeld residuals can be thought of as the observed minus the expected values of the covariates at each failure time. If the residual exhibits a random i.e. unsystematic pattern at each failure time, then this suggests that the covariate effect is not changing with time i.e that the PH assumption fits. If it is systematic, it suggests that the covariate effect is changing with time. This suggests that one test would be to plot Schoenfeld residuals against time. If the PH assumption holds, then the slope of the Schoenfeld residuals should be zero. This is the basis for graphical tests of the PH assumption. However, a slope of zero is in the eye of the beholder and so we might prefer to conduct a statistical test. You can do this by typing:

```
stcox X, nohr efron schoenfeld(sc*) scaledsch(ssc*);

stphtest, plot (some variable)
```

This will produce a graph of the Schoenfeld residuals against time. This allows you to conduct an 'eyeball test': are the slopes flat with respect to time? If you then type:

```
stphtest, detail
```

you will get a better test based on the scaled Schoenfeld residuals. A table will be produced illustrating whether the individual covariates pass the PH assumption and whether the model as a whole (the global test) passes the assumption. The null hypothesis is that the PH assumption holds. Thus, p-values less than 0.05 or 0.10 suggest that the PH assumption is violated. If you find evidence of non-proportional hazards, then you need to alter your model.[8] The solution that you need to implement is to interact all of the variables that show signs of non-proportional hazards with some function of time (Box-Steffensmeier & Zorn 2001, 978). The most straightforward and common way to do this is to interact your covariate with the natural log of time.[9] The idea is to explicitly allow (model) the effect of your variable to vary across time. Thus, if you found that some variable X_1 failed to pass the test for proportional hazards, then you should include X_1 and $X_1 \times \ln(t)$ in the model.

[7] To see the mathematical derivation of the Schoenfeld residuals, see the notes from Jones.

[8] Note that you may find that you pass the global test, but that one or more of your individual covariates fail. This is still evidence of non-proportional hazards.

[9] An alternative is to interact your variable with time or time squared.

```
gen ln_time = ln(duration);

gen time_volatility=ln_time*volatility
```

Do NOT include $\ln(t)$ as a separate variable when running a continuous time duration model. Note that there is no way to know if this 'solution' really did solve the non-proportional hazards problem. This is because you cannot now use the Schoenfeld residuals test on this new model. In effect, you have to assume that this solves the problem.

3.1.3 Explicit tests of coefficients for interactions of covariates and time.

The third test follows from the second. Basically, you interact your variables with some function of time and then evaluate whether the coefficient on these interaction terms are significant or not. If they are, then the original model had non-proportional hazards. This approach is not entirely recommended because the correlation among the covariates that these separate interaction terms induce has been shown to affect the presence or absence of proportional hazards (Box-Steffensmeier & Zorn 2001, 978).

3.2 Model Fit: Cox-Snell Residuals

Another type of residual that can be obtained from duration models is called the Cox-Snell residual. These residuals can be derived from the Cox model and parametric models. Let's assume that we have a Cox model:

$$h(t) = h_0(t)e^{X\beta} \tag{31}$$

Given this, the estimates of the survival times from the posited model i.e. $\hat{S}(t)$ should be similar to the true value of $S(t)$. This is where the Cox-Snell residuals come in. The Cox-Snell residual is:

$$r_{CS} = \exp(X\hat{\beta})\hat{H}_0(t) \tag{32}$$

where $\hat{H}_0(t)$ is the estimated integrated baseline hazard (or cumulative hazard). If you remember, this is the Nelson-Aalen estimator shown at the beginning. It turns out that the Cox-Snell residuals can also be written as the log of the estimated survival time i.e.

$$r_{CS} = \ln\hat{S}(t) \tag{33}$$

If the Cox model fits the data, then the residuals should be distributed unit exponential i.e. they should behave as if they are from a unit exponential distribution.

So, how does this work? First, you compute the KM estimator on the Cox-Snell residuals. From these estimates, you compute the integrated hazard. Then plot the integrated hazard based on the residuals against the hazard rate estimates backed out of the Cox model. If the model 'holds', then the plot should have a 45-degree slope. To conduct the test, you need to type

```
stcox X, nohr efron mgale(martingale);

predict CoxSnell, csnell;
```

Now you need to re-STSET the data to treat the Cox-Snell residuals as 'the data'i.e. the time variable.

```
stset CoxSnell, fail(censoring variable)
```

Now generate the KM estimates for the "new data" and then generate the integrated hazard using the double option for increased computer precision.

```
sts generate km=s;

gen double H_cs=-log(km);
```

Now you need to plot this:

```
sort CoxSnell;

graph twoway line H_cs CoxSnell CoxSnell,  s(..);
```

3.3 Functional Form of Covariates

You might also wonder whether a covariate should be entered linearly, as a quadratic, or some other form in your model. You can evaluate these different functional forms using another type of residual – a Martingale residual.

There are two different approaches to using Martingale residuals to evaluate functional form. The first approach is to estimate a Cox model, compute the Martingale residuals, and then plot smoothed versions of these residuals against each of the covariates. If the residuals deviate from the zero line, then this might indicate incorrect functional form. To do this test, type:

```
stcox X, nohr efron mgale(mg);

ksm mg polar, lowess;
```

where the KSM command smoothes out the residuals.

The second approach is to assume that there are m covariates in the model and first estimate a model with $m-1$ covariates. From this model, compute the Martingale residuals. Now plot the smoothed residuals against the omitted variable. If the smoothed plot is linear, no transformation of the omitted variable is necessary. To do this test, type:

```
stcox X, nohr efron mgale(mg);

ksm mg polar, lowess;
```

4 Time Varying Covariates

In the model of government coalition data that we have been looking at, duration is a function of coalition characteristics that don't change over time. However, you might think that a government coalition fails partly as a function of the economy. Now we would have time varying covariates (TVCs). These are easiest to handle in terms of a Cox model.

The basic idea behind TVCs requires thinking in terms of a 'counting process' setup. In this setup, each record (line of data) gives the value of covariates that are constant between a beginning and ending time point, and whether an event (failure) has occurred by the ending time point or not. Note that this setup will be useful for thinking about discrete time duration models and allows us to keep "count" of the number of failures for multiple failure data - we'll see all of this later. Note that for relatively continuously varying measures, such as the economy, we might wish to take economic measures as constant over a year or so in

order to simplify the data handling (the counting process allows us to deal with monthly varying covariates or daily varying covariates). The point, though, is that covariates do have to be constant over some discrete time interval. Since our data automatically comes in discrete time periods, this is not much of a limitation. In Table 2, we show what time varying covariate data might look like using an example from the MIDs project.

Table 2: Example of Counting Process Data with a Yearly TVC

Year	Dyad I.D.	Interval (Start, Stop)	Indicator Variable (δ)	Economic Growth	Contiguity Status
1951	2020	(0, 1]	0	0.01	1
1952	2020	(1, 2]	0	0.03	1
1953	2020	(2, 3]	0	0.02	1
1954	2020	(3, 4]	0	0.01	1
⋮	⋮	⋮	⋮	⋮	⋮
1971	2020	(20, 21]	0	0.01	1
1972	2020	(21, 22]	0	0.01	1
1973	2020	(22, 23]	1	0.03	1
1974	2020	(23, 24]	0	0.01	1
⋮	⋮	⋮	⋮	⋮	⋮
1961	2041	(0, 1]	0	-0.05	0
1962	2041	(1, 2]	0	-0.01	0
1963	2041	(2, 3]	1	-0.01	0
1964	2041	(3, 4]	0	0.00	0

As you can see, the observed survival time for each observation is described by a triple (t_0, t, δ), where t_0 is the time at which observation i enters the risk set, t is the time at which the individual is observed as failing or surviving (being right-censored), and δ is an indicator variable indicating whether an observation fails ($\delta = 1$) or survives ($\delta = 0$). As Table 2 illustrates, each period has its own record of data because the TVC changes values each period.

In contrast to the data in Table 2, analysts might have TVCs that change intermittently across time. An example of what this sort of dataset might look like is shown in Table 3. The dataset records the duration in weeks that passes from the start of a campaign cycle until the emergence of a high quality challenger against the incumbent (Box-Steffensmeier 1996).

Table 3: Example of Duration Dataset with TVCs

Case I.D.	Weeks to Event	Southern District	Incumbent's Party	1988	War Chest in Millions	Censoring Indicator
100	26	0	0	0.62	0.003442	0
100	50	0	0	0.62	0.010986	1
201	26	1	0	0.59	0.142588	0
201	53	1	0	0.59	0.15857	0
201	65	1	0	0.59	0.202871	0
201	75	1	0	0.59	0.217207	0
516	26	0	1	0.79	0.167969	0
516	53	0	1	0.79	0.147037	0
516	65	0	1	0.79	0.164970	0
516	72	0	1	0.79	0.198608	1
706	26	0	0	0.66	0.139028	0
706	53	0	0	0.66	0.225633	0
706	65	0	0	0.66	0.225817	0
706	78	0	0	0.66	0.342801	0
706	83	0	0	0.66	0.262563	1

4.1 Endogeneity Issues with TVCs

As Beck points out in his notes, once we allow for time varying covariates, we have to be particularly careful of using endogenous or jointly causal covariates. His example is marriage duration and the use of the number of children as a TVC. We might expect that people decide to have children if they think that their marriage is stable. Thus, finding that children increase marriage duration might well be spurious. The key is to think about whether the TVC is under the control of the unit. Typically, the economy is a good exogenous measure (usually), but any strategic variable must be suspect. There are no statistical means to assess whether a TVC is suitable or not.

As Beck also notes, we have to be careful about what we mean by TVCs. His example is the duration of unemployment spells. Perhaps we have some states where people can have 26 weeks of benefits but in other states, they can have 39 weeks of benefits. Thus, we might think to include a TVC measuring the number of weeks left of benefits. However, this is not a TVC since the people know the rules throughout their unemployment spell. In effect, there is nothing that is changing across time in terms of the rules. An alternative that would be better is to include a dummy variable indicating which rule a person is under – this would be a non-time varying covariate.

References

Box-Steffensmeier, Janet. 1996. "A Dynamic Analysis of the Role of War Chests in Campaign Strategy." *American Journal of Political Science* 40:352–371.

Box-Steffensmeier, Janet & Bradford Jones. 2004. *Timing and Political Change: Event History Modeling in Political Science*. Ann Arbor: University of Michigan Press.

Box-Steffensmeier, Janet & Christopher Zorn. 2001. "Duration Models and Proportional Hazards in Political Science." *American Journal of Political Science* 45:972–988.

Jenkins, Stephen P. 2008. "Survival Analysis." Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester.