Chapter 5: Cox Proportional Hazards Model

A popular model used in survival analysis that can be used to assess the importance of various covariates in the survival times of individuals or objects through the hazard function. In addition, the quantitative impact of these variables on important lifetime variables of interest (such as median survival) can be described.

The heuristics behind the model:

Suppose we have 3 individuals, labeled A, B, and C, who are at risk for some particular event. Then we are told a week later that one of them had the event. If we know that each individual had the same risk for that event, what can we say about the probability that it was A (or B, or C)?

Knowing that each had the same risk tells us that each person is equally likely to have had the event, therefore:
P(A had the event) = 1 / 3

The same holds for both B and C, i.e. P(B) = P(C) = 1/3

Now suppose (instead) that we are told that the risk for individual A was twice as high as B and C. Now what can we say about the probability of it being A (or B, etc.)? One way to think of this problem is to label the number of “risk balls” that correspond to the risks of each individual. If A has twice as much risk, then it has twice as many balls.

If these balls were randomly mixed and blindly chosen, what is the probability that the ball is labeled ‘A’, ‘B’, or ‘C’? The answer to this question gives us the probability that A had the event.
We know that $P(A) = \frac{1}{2}$ whereas $P(B) = P(C) = \frac{1}{4}$.

Now suppose that the risk of an event is proportional to the person’s age. If we know the ages of these people, we can calculate the probability that it was A, B, or C. Suppose A was 47, B was 68, and C was 75. What are the various probabilities now?

\[P(A) = \frac{47}{47+68+75} = \frac{47}{190} = 0.247\]
\[P(B) = \frac{68}{190} = 0.358\]
\[P(C) = 0.395\]

In a sense, this is how the “pmf” for the proportional hazard model works. It assigns the probability that a particular individual has an event by dividing the risk (actually, the hazard) for that person by the sum of all the hazards for all of the people who are at risk.
That is, \[ P(A) = \frac{h(A)}{h(A) + h(B) + h(C)} \]

Of course, in statistics, the effects of the various covariates on the probability of individuals having an event are unknown because their associated parameters, \( \beta_i \), are unknown. In fact, the investigator is actually more interested in estimating the \( \beta_i \)'s for a given dataset. This estimation problem can use a likelihood function, and inference can rely on the asymptotic properties of the MLEs for the \( \beta_i \)'s. Before we move on to this problem, let's take a closer look at the proportional hazards assumption.

**The Basic Assumption**

\[ h_i(t) = h_0(t) \times \exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}\} \]

That is, the hazard for individual, i, is based on two parts. \( h_0(t) \) is called the **baseline hazard function**. It can be any complicated function of \( t \) as long as \( h_0(t) \geq 0 \). Fortunately, it doesn’t have to be specified.
The other part of the hazard function involves the covariates, which are in the exponent along with the unknown parameters. Note that this term does not involve a time variable. Therefore, the ratio of the hazards of two individuals does not depend on time, i.e. $h_0(t)$. This will simplify the problem of estimating the $\beta_i$'s.

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \exp\{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}\}}{h_0(t) \exp\{\beta_1 x_{j1} + \cdots + \beta_k x_{jk}\}} = \exp\{\beta_1 (x_{i1} - x_{j1}) + \cdots + \beta_k (x_{ik} - x_{jk})\}$$

Remember that the likelihood function is a function of the unknown parameters, given the outcomes of the random variables. If the events for the individuals are independent of each other, then the likelihood is a product of the likelihoods for all individuals in the sample: $PL = \prod_{i=1}^{n} L_i$
where PL stands for “partial likelihood” and $L_i$ is the likelihood for the $i^{th}$ event.

Assume for simplicity that there is only one covariate collected in an experiment, call it $x_i$, which is the covariate for the individual who had the $i^{th}$ event or censoring time (indicated by $\delta_i = 1$ if event and $\delta_i = 0$ if censored). If an event occurred at 5 months, the $L_i$ could be written as:

$$L_i = \left( \frac{h_0(5) \exp\{\beta x_i\}}{h_0(5) \exp\{\beta x_i\} + h_0(5) \exp\{\beta x_{(i+1)}\} + \cdots + h_0(5) \exp\{\beta x_n\}} \right)^{\delta_i}$$

Note that when $\delta_i = 0$, $L_i = 1$. More importantly, the $h_0(5)$ term factors out of the denominator and cancels the $h_0(5)$ term in the numerator. $L_i$ then simplifies to:
The hazards included in the denominator are only those individuals who are at risk at the $i^{th}$ event (or censoring) time. The entire likelihood function can be expressed very concisely as:

$$PL = \prod_{i=1}^{n} \left[ \frac{\exp \{ \beta x_i \} \sum_{j=1}^{n} Y_{ij} \exp \{ \beta x_j \}}{\sum_{j=1}^{n} Y_{ij} \exp \{ \beta x_j \}} \right]^{\delta_i}$$

where $Y_{ij}=1$ if $t_j \geq t_i$, and $Y_{ij}=0$ if $t_j < t_i$
Although the expressions of the likelihood can be written in short form mathematically, using them in practice can yield massive expressions. Remember this example dealt with only one covariate. Imagine the complexity of dealing with several or more covariates, and one quickly gains an appreciation of the task!

Fortunately computers are very good at solving these kinds of problems. In many cases, they can estimate the MLEs and their standard errors in seconds.

What is also nice about this method is the fact that the same asymptotic properties of MLEs in the last chapter also hold for this chapter.

Therefore, many inferential techniques learned in the previous chapter (LRT, Wald Chi Square) will work here as well.
The PH model is invoked in SAS with “Proc phreg” as the following example shows:

```
proc phreg data=asa.hmohiv;
model time*censor(0)=drug;
run;
```

This model includes the variable, drug, as a covariate.

Notice that SAS provides the log-likelihood statistic in –2ln(L) format.

The –2ln(L) statistics are provided both with and without the covariates.
Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; Chi Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>10.1973</td>
<td>1</td>
<td>0.0014</td>
</tr>
<tr>
<td>Score</td>
<td>10.7432</td>
<td>1</td>
<td>0.0010</td>
</tr>
<tr>
<td>Wald</td>
<td>10.3451</td>
<td>1</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Std Error</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi Sq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug</td>
<td>0.24226</td>
<td>1</td>
<td>10.3451</td>
<td>0.0013</td>
<td>2.180</td>
</tr>
</tbody>
</table>

The global test is analogous to the overall F-test in an ANOVA or linear regression. It is testing whether all of the covariates have no “influence” on survival time.

The MLE for the drug covariate is 0.779 with SE=0.242. Notice that this parameter is significantly positive, not negative as in the last chapter.
Remember that this parameter is estimating a quantity that is related to the hazard. A $\beta = 0$ would indicate that IV drug use has no association with survival time; a $\beta > 0$ would indicate that IV drug users have a higher hazard of death, and a $\beta < 0$ would indicate that IV drug users have a lower hazard of death. So, qualitatively speaking, the conclusions of this analysis is the same as the non-parametric analysis (e.g. log-rank) and the parametric analysis (using the log-normal distribution).

Within the Cox model, the best interpretation of $\beta$ for a 0-1 categorical variable is the hazard ratio. That is, $\exp\{\beta\}$ is the hazard ratio for being in the group where $x=1$ versus the group where $x=0$. The MLE of this hazard ratio is provided as $\exp\{0.77919\}=2.18$, which is listed in the standard output.
A 95% C.I. for $\beta$ is provided with:

$$0.779 \pm 1.96 \times 0.242 = [0.305, 1.25]$$

The 95% C.I. for the hazard ratio is:

$$[\exp\{0.305\}, \exp\{1.25\}] = [1.36, 3.49]$$

--------------------------------------------------------------------------

Although no examples were given in the notes for the last chapter, numerical covariates can be included into the model just as easily as categorical ones. The same holds for phreg.

```plaintext
proc phreg data=asa.hmohiv;
model time*censor(0)=drug age;
run;
```

$$\ln\{h_i(t)\} = \ln\{h_0(t)\} + \beta_1 x_{i1} + \beta_2 x_{i2}$$

\[ x_{i1} = 1 \text{ if IV drug user} \]
\[ x_{i1} = 0 \text{ otherwise} \]
\[ x_{i2} = \text{age} \]
**Model Fit Statistics**

<table>
<thead>
<tr>
<th>Without</th>
<th>With</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>Covariates</td>
</tr>
<tr>
<td>-2 LOG L</td>
<td>598.390</td>
</tr>
<tr>
<td>AIC</td>
<td>598.390</td>
</tr>
<tr>
<td>SBC</td>
<td>598.390</td>
</tr>
</tbody>
</table>

**Testing Global Null Hypothesis: BETA=0**

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>34.9819</td>
<td>2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Score</td>
<td>34.3038</td>
<td>2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Wald</td>
<td>32.4859</td>
<td>2</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

**Analysis of Maximum Likelihood Estimates**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug</td>
<td>0.25550</td>
<td>13.5662</td>
<td>0.0002</td>
<td>2.563</td>
</tr>
<tr>
<td>age</td>
<td>0.01849</td>
<td>24.5009</td>
<td>&lt; 0.001</td>
<td>1.096</td>
</tr>
</tbody>
</table>
From the results provided by SAS, we can see that age is also a significant prognostic factor for survival after including IV drug use into the model. Because the estimate for $\beta_2$ is positive, we see that the hazard of death increases with age. Specifically, the hazard of death increases by an estimated 

$$(\exp\{0.0915\}-1)*100\% = 9.5\%$$

for each additional year of age.

The Wald Chi-square statistic is 24.5 for this variable.

The LR Chi-square for testing the significance of age in the model is found with:

$$X^2 = -2[\text{LL}(\text{drug})] - \{-2[\text{LL}(\text{drug, age})]\}$$

$$X^2 = 588.198 - 563.408 = 24.79$$

For large samples, the Wald and LRT statistics should be close when testing for individual parameters.
Tied Data

The likelihood function shown on p7 of these notes assumes that all of the survival (or censor) times are distinct. In reality, this is probably true in most cases, but in practice, there are often datasets that have many tied survival times.

There needs to be a way of accounting for these tied data. Several formulas are available in SAS to help the analyst with this practical problem.

**Breslow’s approximation**: is the default option in SAS and works well when the ties are relatively few.

**Efron’s approximation**: tends to do much better and generally requires about the same amount of computational time as Breslow, so between the two methods, Efron’s is generally preferred.
The exact method: assumes the tied results are stemming from imprecise time measurements and calculates the likelihood using all of the possible orderings of the tied data.

The exact method will give the “best” estimates for the effects of the covariates, but the computational time can be long. For relatively small datasets, however, this increase in computing time is relatively trivial.

The discrete method uses a slightly different model than the other methods. Rather than assuming the time variable is continuous, as we have been for most of this course, this method assumes the time variable is inherently discrete.

A good example of this would be a change of the political party in the Whitehouse. By the very nature of our political system, this can occur only once every four years (except in extreme circumstances).
Estimates of the coefficients using the discrete method can be considerably different from the exact method. Both methods are technically “exact,” so these differences arise from different model assumptions. Therefore, the meaning of the coefficients depend on the model. In particular, the coefficients for the discrete method are related to the log odds rather than the log hazard.

For most applications, the discrete method does not make sense.

More details on the Exact Method:

Some general ideas are worth understanding. Earlier, we talked about 3 people labeled A, B, and C. Suppose both ‘A’ and ‘B’ had tied event times which occurred before ‘C’ had his event. How does the method compute the likelihood?
The exact method computes the “probability” that ‘A’ evented first followed by ‘B’ OR ‘B’ evented first followed by ‘A.’ Since both ‘A’ and ‘B’ had the same observed survival times, we assume that each sequence (permutation) is equally likely:

\[ P(A, B) = \left( \frac{h(A)}{h(A) + h(B) + h(C)} \right) \cdot \left( \frac{h(B)}{h(B) + h(C)} \right) \]

\[ P(B, A) = \left( \frac{h(B)}{h(A) + h(B) + h(C)} \right) \cdot \left( \frac{h(A)}{h(A) + h(C)} \right) \]

\[ L_1 = \left( \frac{h(A)}{h(A) + h(B) + h(C)} \right) \cdot \left( \frac{h(B)}{h(B) + h(C)} \right) + \left( \frac{h(B)}{h(A) + h(B) + h(C)} \right) \cdot \left( \frac{h(A)}{h(A) + h(C)} \right) \]
• The covariates would be related to the hazard function in the usual way, e.g. \( h(A) = \exp\{\beta_1 \times x_A\} \) where \( x_A \) is same variable of interest such as age. The likelihood function would relate the outcomes to the unknown parameters such as \( \beta_1 \).

• If there are a total of \( n_{Ti} \) tied event times for a particular time, \( t_i \), then the exact method must evaluate \( n_{Ti}! \) separate permutations of possible outcomes. Therefore, the likelihood function can become extremely complicated (very quickly) as \( n_{Ti} \) increases. In practical terms, this translates into a longer wait as the computer estimates the parameters, their standard errors, and so on.

• If there are censored times tied with the event times, they are included in the risk set(s) only and do not complicate the likelihood in any substantial way.
Using the methods in SAS

The specific method used to evaluate ties can be accessed with the “ties = option” in the model statement of the phreg procedure. The options are efron, exact, discrete, and breslow (default). For example:

```sas
proc phreg data=asa.hmohiv;
  model time*censor(0)=drug age /ties=efron;
run;
```

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; Chi Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>39.1316</td>
<td>2</td>
<td>&lt; 0001</td>
</tr>
<tr>
<td>Score</td>
<td>38.3936</td>
<td>2</td>
<td>&lt; 0001</td>
</tr>
<tr>
<td>Wald</td>
<td>36.1251</td>
<td>2</td>
<td>&lt; 0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi Sq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>drug</td>
<td>1</td>
<td>1.01670</td>
<td>0.25622</td>
<td>15.7459</td>
<td>&lt; 0001</td>
<td>2.764</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>0.09714</td>
<td>0.01864</td>
<td>27.1597</td>
<td>&lt; 0001</td>
<td>1.102</td>
</tr>
</tbody>
</table>
Number of ties in the HMOHIV Dataset:

```
proc freq data=asa.hmohiv;
  table time / nocum nopercent;
run;
```

<table>
<thead>
<tr>
<th>time</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
</tbody>
</table>

Partial listing of output for the time variable. Note that there are 17 individuals who had survival (or censoring) times equal to 1 month, 10 who had time equal to 2, and so on...

The number of ties in the dataset are not trivial, so the more precise methods should be examined.
Coefficients by method used to evaluate ties:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi Sq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breslow</td>
<td>drug</td>
<td>1</td>
<td>0.94108</td>
<td>0.25550</td>
<td>13.5662</td>
<td>0.0002</td>
<td>2.563</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>1</td>
<td>0.09151</td>
<td>0.01849</td>
<td>24.5009</td>
<td>&lt;.0001</td>
<td>1.096</td>
</tr>
<tr>
<td>Efron</td>
<td>drug</td>
<td>1</td>
<td>1.01670</td>
<td>0.25622</td>
<td>15.7459</td>
<td>&lt;.0001</td>
<td>2.764</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>1</td>
<td>0.09714</td>
<td>0.01864</td>
<td>27.1597</td>
<td>&lt;.0001</td>
<td>1.102</td>
</tr>
<tr>
<td>Exact</td>
<td>drug</td>
<td>1</td>
<td>1.02263</td>
<td>0.25716</td>
<td>15.8132</td>
<td>&lt;.0001</td>
<td>2.781</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>1</td>
<td>0.09768</td>
<td>0.01874</td>
<td>27.1731</td>
<td>&lt;.0001</td>
<td>1.103</td>
</tr>
<tr>
<td>Discrete</td>
<td>(Method technically inappropriate)</td>
<td>drug</td>
<td>1</td>
<td>1.07004</td>
<td>0.27438</td>
<td>15.2084</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>1</td>
<td>0.10315</td>
<td>0.02006</td>
<td>26.4449</td>
<td>&lt;.0001</td>
<td>1.109</td>
</tr>
</tbody>
</table>
**Time-Dependent Covariates:**

Time dependent covariate analysis with phreg can be a powerful tool in the analysis of survival data for several reasons:

- Sometimes a meaningful origin time is difficult to determine and people may modulate their risk over the course of the period of risk. A good example is establishing an observational study to determine the effectiveness of the flu vaccine in a particular year.

  - It is not immediately apparent what the origin time should be. You could say it is Sept. 15\textsuperscript{th}, Oct. 1\textsuperscript{st}, or Nov. 1\textsuperscript{st}. Each of these origin times will effect the parameter estimates.

  - Regardless of the origin time, people will become vaccinated over a period of time, typically from Oct. 1\textsuperscript{st} to Jan. 31\textsuperscript{st}. As they become vaccinated, their risk of getting the flu changes over the course of the season.
• In many cases, the explanatory variable is heavily correlated with the outcome or endpoint. In such instances, it is easy to confuse the “causation” of effect. What is worse, a naive statistical analysis will confirm an investigator’s hunch or hypothesis with significant results (e.g. small p-values).

• A good example is the Stanford Heart Transplant study. Primary interest focused on the effectiveness of transplants in reducing the risk of death. However, the waiting period to receive a transplant was highly variable, so it was difficult to determine if transplants actually reduced the risk of death or if people who lived longer were more likely to receive a transplant, and a transplant covariate ($X_T = \text{Yes}/\text{No}$) was simply a reflection of that fact.
• Finally, some variables that are important to the risk of an endpoint will vary in individuals over the course of a study and there is no way to control them (i.e. keep them constant) by controlling the people’s risk factors.

• A good example was a study of dietary and behavior factors that influenced the risk of a first attack of gout in men. Because the study lasted for many years (This was necessary because the risk was relatively small.), it was unreasonable to assume that diet would remain the same throughout the study period. People’s alcohol related drinking habits, meat intake, and dairy intake will change. If any of these factors are important to risk (hazard), then their hazards will change as well.
Typical model for time varying covariates:

\[ \ln \{h_i(t)\} = \log \{h_0(t)\} + \beta_1 X_{i1} + \beta_2 X_{i2}(t) \quad \text{Eq. L5.1} \]

where \( X_{i2}(t) \) is a covariate that depends on time.

For example, \( X_{i2}(t) = 1 \) if \( t > t_{\text{wait}} \) and \( X_{i2}(t) = 0 \) if \( t < t_{\text{wait}} \) in the Stanford transplant example. Therefore the person’s hazard changes at the time of the transplant by a factor of \( \beta_2 \).

----------------------------------------------------------------

Before we look at an analysis in SAS using time varying covariates, let's analyze the Stanford dataset using a covariate, \( X_{i2} = 1 \) if the person had a transplant, and \( X_{i2} = 0 \) otherwise. The origin time is the date of acceptance into the study, and the survival time is from the date of acceptance until death (or censoring).
SAS is programmed as follows:

```sas
proc phreg data=asa.stan;
model surv1*dead(0)=trans surg ageaccpt;
run;
```

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>trans</td>
<td>1</td>
<td>-1.70813</td>
<td>0.27860</td>
<td>37.5902</td>
<td>&lt; 0.0001</td>
<td>0.181</td>
</tr>
<tr>
<td>surg</td>
<td>1</td>
<td>-0.42130</td>
<td>0.37098</td>
<td>1.2896</td>
<td>0.2561</td>
<td>0.656</td>
</tr>
<tr>
<td>ageaccpt</td>
<td>1</td>
<td>0.05860</td>
<td>0.01505</td>
<td>15.1611</td>
<td>&lt; 0.0001</td>
<td>1.060</td>
</tr>
</tbody>
</table>

trans is the indicator variable just discussed on the last page
surg is an indicator that gives information about prior heart surgery before acceptance into this study
ageaccpt is the age of acceptance into the study
This analysis indicates that ‘trans’ and ‘ageaccpt’ are statistically significant variables that are important in determining the hazard of death.

Unfortunately, as we discussed earlier, patients may live longer because they received transplants, or they may have received transplants because they lived longer, or it could be a combination of the two. The analysis cannot answer this question for us. In order to determine that, another approach is needed such as using time varying covariates.

Without time varying covariates, the only types of questions that can be answered are ones that involve patient characteristics such as prior surgery, age, waiting time, and genetic compatibility on survival time after transplantation. However, the most important question (whether transplants are effective), cannot be answered!
For example, we can analyze the survival time for patients who had transplants after the date of transplant surgery:

```latex
\begin{verbatim}
proc phreg data=asa.stan;
  where trans=1;
model surv2*dead(0)=surg m1 m2 m3 ageaccpt wait dot;
run;
\end{verbatim}
```

Subsets the analysis to only those who had the transplants.

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi Sq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>surg</td>
<td>1</td>
<td>-0.77029</td>
<td>0.49718</td>
<td>2.4004</td>
<td>0.1213</td>
<td>0.463</td>
</tr>
<tr>
<td>m1</td>
<td>1</td>
<td>-0.24857</td>
<td>0.19437</td>
<td>1.6355</td>
<td>0.2009</td>
<td>0.780</td>
</tr>
<tr>
<td>m2</td>
<td>1</td>
<td>0.02958</td>
<td>0.44268</td>
<td>0.0045</td>
<td>0.9467</td>
<td>1.030</td>
</tr>
<tr>
<td>m3</td>
<td>1</td>
<td>0.64407</td>
<td>0.34276</td>
<td>3.5309</td>
<td>0.0602</td>
<td>1.904</td>
</tr>
<tr>
<td>ageaccpt</td>
<td>1</td>
<td>0.04927</td>
<td>0.02282</td>
<td>4.6619</td>
<td>0.0308</td>
<td>1.050</td>
</tr>
<tr>
<td>wait</td>
<td>1</td>
<td>-0.00183</td>
<td>0.00514</td>
<td>0.1276</td>
<td>0.7210</td>
<td>0.998</td>
</tr>
<tr>
<td>dot</td>
<td>1</td>
<td>-0.0001650</td>
<td>0.0002991</td>
<td>0.3044</td>
<td>0.5811</td>
<td>1.000</td>
</tr>
</tbody>
</table>
```

The variables are defined in the text on page 118-9 or in the appendix.
Now let's see how we can use SAS to create a time dependent covariate that changes for each individual from 0 to 1 at the time of the transplant:

```sas
proc phreg data=asa.stan;
model surv1*dead(0)=plant surg ageaccpt / ties=exact;
if wait>surv1 or wait= . then plant=0; else plant=1;
run;
```

![Analysis of Maximum Likelihood Estimates](image)

A new variable was created, called ‘plant’ which equals 0 if the waiting time is greater than the survival time or if the waiting time is missing. Otherwise plant = 1. Does this make intuitive sense?
The answer should be ‘no.’ It will take a little effort to understand what this conditional statement in phreg is doing. One temptation is to put this statement into a data-step where such assignments are more familiar. Let’s see what happens if we create another dataset with the variable ‘plant’ and if we get the same results:

```plaintext
data stan2;
set asa.stan;
if wait>surv1 or wait=. then plant=0; else plant=1;
run;

proc phreg data=stan2;
model surv1*dead(0)=plant surg ageaccpt / ties=exact;
run;
```

<table>
<thead>
<tr>
<th>Parameter Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>1</td>
<td>-1.71733</td>
<td>0.27855</td>
<td>38.0101</td>
<td>&lt;.0001</td>
<td>0.180</td>
</tr>
<tr>
<td>surg</td>
<td>1</td>
<td>-0.41897</td>
<td>0.37117</td>
<td>1.2742</td>
<td>0.2590</td>
<td>0.658</td>
</tr>
<tr>
<td>ageaccpt</td>
<td>1</td>
<td>0.05889</td>
<td>0.01505</td>
<td>15.3119</td>
<td>&lt;.0001</td>
<td>1.061</td>
</tr>
</tbody>
</table>
```
This analysis does not yield the same results as the one that had the conditional statement inside phreg. We see the analysis is essentially the same as the first one that used the ‘trans’ variable (differences in results are from the method used to handle ties). As a matter of fact, when we define ‘plant’ in the datastep, we create a replicate of ‘trans’

The point: The conditional statement in phreg is handled differently than in a datastep. It’s more complicated than simply evaluating whether surv1>wait for each patient.

```r
proc phreg data=asa.stan;
model surv1*dead(0)=plant surg ageaccpt / ties=exact;
if wait>surv1 or wait=. then plant=0; else plant=1;
run;
```

`surv1` is a variable that is evaluated at each event time. Therefore, for each event time, a comparison is made between that event time and each person’s waiting time, and based on those comparisons, the variable ‘plant’ is assigned as 0 or 1.
Therefore, for each person in the dataset, there are potentially many evaluations for the variable ‘plant.’ The variable ‘plant’ is 0 until ‘wait’ < ‘surv1’ which means the event time is now longer than the waiting time, and that means the person had the transplant (so his/her ‘plant’ variable is now appropriately equal to 1).

In essence, surv1 is basically the time variable, t, in Eq. L5.1

**Question:** So why are we using ‘surv1’ as a proxy for t?

**Answer:** Because the likelihood is evaluated at the event times.

It’s helpful to examine specific cases to see exactly how the computer determines the value of a time dependent covariate.
### Partial listing of data for the Stanford Heart Transplant study

<table>
<thead>
<tr>
<th>Obs</th>
<th>surv1</th>
<th>dead</th>
<th>wait</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>16</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>17</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>23</td>
<td>27</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>24</td>
<td>29</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>0</td>
<td>10000</td>
</tr>
<tr>
<td>26</td>
<td>31</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>27</td>
<td>34</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>28</td>
<td>35</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>29</td>
<td>36</td>
<td>1</td>
<td>10000</td>
</tr>
<tr>
<td>30</td>
<td>38</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>

On day 16, the 19th patient died, so the contribution to the partial likelihood is:

\[ L_{19} = \frac{\exp{\beta \cdot x_{19}(16)}}{\exp{\beta \cdot x_{19}(16)}} + \exp{\beta \cdot x_{20}(16)} + \cdots + \exp{\beta \cdot x_{103}(16)}} \]

\[ x_{19}(16) = 1, \quad x_{24}(16) = 1. \quad \text{But} \quad x_{23}(16) \quad \text{and} \quad x_{30}(16) = 0 \quad \text{because} \quad \text{wait} > \text{surv1} = 16. \]
On day 17, the 20\textsuperscript{th} patient dies, and the contribution to the partial likelihood is:

$$L_{20} = \frac{\exp{\beta \cdot x_{20}(17)}}{\exp{\beta \cdot x_{20}(17)}} + \exp{\beta \cdot x_{21}(17)} + \cdots + \exp{\beta \cdot x_{103}(17)}$$

In this instance, $x_{19}(17)$ is not included because the patient is no longer in the risk set. $x_{24}(17) = 1$ and $x_{23}(17) = 1$ because $\text{wait} \leq \text{surv1} = 17$, but $x_{30}(17) = 0$, still.

$x_{30}(t)$ does not change to 1 until the 28\textsuperscript{th} patient dies, who has a survival time of 35 days.

All of the other time dependent covariates listed on the last page remain at 0 throughout the entire calculation of the partial likelihood (for as long as they are in the risk set) because they never received a transplant.
Now that we understand how the computer handles time dependent covariates, let's take another look at the output:

```
Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>1</td>
<td>-0.04615</td>
<td>0.30276</td>
<td>0.0232</td>
<td>0.8788</td>
<td>0.955</td>
</tr>
<tr>
<td>surg</td>
<td>1</td>
<td>-0.77145</td>
<td>0.35961</td>
<td>4.6021</td>
<td>0.0319</td>
<td>0.462</td>
</tr>
<tr>
<td>ageaccpt</td>
<td>1</td>
<td>0.03109</td>
<td>0.01391</td>
<td>4.9952</td>
<td>0.0254</td>
<td>1.032</td>
</tr>
</tbody>
</table>
```

Notice that the ‘plant’ variable is now statistically insignificant. This indicates that the risk of death is not reduced after having a heart transplant. In addition, prior surgery became a significant factor in reducing the risk of death.

Using SAS arrays to handle time dependent covariates

The recidivism dataset:
The “recid” dataset looked at factors that could influence the time until a newly released prisoner was arrested again. Such factors included age, financial aid, race, and employment status.

The data on employment status were collected on a weekly basis for up to 1 year after release. Therefore, each person had 52 variables, called emp1 through emp52, which indicated whether the person was employed during that week (1 = Yes, 0 = No, . = arrested). These variables can be used to construct a time dependent covariate such as:

\[ X_{\text{emp}}(t) = 1 \text{ if person was employed at time } t \]

\[ X_{\text{emp}}(t) = 0 \text{ if person was not employed.} \]

SAS arrays are particularly useful in constructing a single variable from many. The code for this example follows:
PROC PHREG DATA=ASA.RECID;

MODEL WEEK*ARREST(0)=FIN AGE RACE WEXP MAR PARO
   PRIOR EMPLOYED / TIES=EFRON;

ARRAY EMP[*] EMP1-EMP52;  \[ Statement defines an array, called emp[ ], linked to emp1-52 \]

EMPLOYED=EMP[WEIGHT];  \{ Note that survival time is being used to evaluate the time dependent covariates at each event time, the same as before. \}

RUN;

ARRAYS USE BRACKETS!!!  [ ]

With arrays, variables can be accessed by their subscripts, e.g.

emp[1] is emp1
emp[3] is emp3
emp[20] is emp20
The output for the program is provided below:

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>fin</td>
<td>1</td>
<td>-0.35672</td>
<td>0.19113</td>
<td>3.4835</td>
<td>0.0620</td>
<td>0.700</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>-0.04633</td>
<td>0.02174</td>
<td>4.5442</td>
<td>0.0330</td>
<td>0.955</td>
</tr>
<tr>
<td>race</td>
<td>1</td>
<td>0.33867</td>
<td>0.30960</td>
<td>1.1966</td>
<td>0.2740</td>
<td>1.403</td>
</tr>
<tr>
<td>wexp</td>
<td>1</td>
<td>-0.02557</td>
<td>0.21142</td>
<td>0.0146</td>
<td>0.9037</td>
<td>0.975</td>
</tr>
<tr>
<td>mar</td>
<td>1</td>
<td>-0.29374</td>
<td>0.38303</td>
<td>0.5881</td>
<td>0.4432</td>
<td>0.745</td>
</tr>
<tr>
<td>paro</td>
<td>1</td>
<td>-0.06420</td>
<td>0.19468</td>
<td>0.1088</td>
<td>0.7416</td>
<td>0.938</td>
</tr>
<tr>
<td>prio</td>
<td>1</td>
<td>0.08515</td>
<td>0.02896</td>
<td>8.6455</td>
<td>0.0033</td>
<td>1.089</td>
</tr>
<tr>
<td>employed</td>
<td>1</td>
<td>-1.32823</td>
<td>0.25071</td>
<td>28.0679</td>
<td>&lt;0.001</td>
<td>0.265</td>
</tr>
</tbody>
</table>

The employed covariate is extremely significant. Even when using time dependent covariates, it is still easy to run into problems with *causal ordering*. In this case, it is possible that people may be unemployed because they were arrested, again getting statistically significant results that are misleading.
One way to prevent this error is to lag the covariate values. For example, we could look at employment status the week before the arrest. The code is modified simply as follows:

```sas
proc phreg data=asa.recid;
where week>1;
model week*arrest(0)=fin age race wexp mar paro
    prio employed / ties=efron;
array emp[*] emp1-emp52;
employed=emp[week-1];
run;
```

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variable</th>
<th>DF</th>
<th>Standard Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi Sq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>fin</td>
<td>fin</td>
<td>1</td>
<td>-0.35129</td>
<td>0.19181</td>
<td>3.3541</td>
<td>0.0670</td>
<td>0.704</td>
</tr>
<tr>
<td>age</td>
<td>age</td>
<td>1</td>
<td>-0.04977</td>
<td>0.02189</td>
<td>5.1697</td>
<td>0.0230</td>
<td>0.951</td>
</tr>
<tr>
<td>race</td>
<td>race</td>
<td>1</td>
<td>0.32149</td>
<td>0.30912</td>
<td>1.0816</td>
<td>0.2983</td>
<td>1.379</td>
</tr>
<tr>
<td>wexp</td>
<td>wexp</td>
<td>1</td>
<td>-0.04765</td>
<td>0.21323</td>
<td>0.0499</td>
<td>0.8232</td>
<td>0.953</td>
</tr>
<tr>
<td>mar</td>
<td>mar</td>
<td>1</td>
<td>-0.34477</td>
<td>0.38322</td>
<td>0.8094</td>
<td>0.3683</td>
<td>0.708</td>
</tr>
<tr>
<td>paro</td>
<td>paro</td>
<td>1</td>
<td>-0.04709</td>
<td>0.19630</td>
<td>0.0576</td>
<td>0.8104</td>
<td>0.954</td>
</tr>
<tr>
<td>prio</td>
<td>prio</td>
<td>1</td>
<td>0.09201</td>
<td>0.02880</td>
<td>10.2085</td>
<td>0.0014</td>
<td>1.096</td>
</tr>
<tr>
<td>employed</td>
<td>employed</td>
<td>1</td>
<td>-0.78689</td>
<td>0.21808</td>
<td>13.0195</td>
<td>0.0003</td>
<td>0.455</td>
</tr>
</tbody>
</table>
PH Survival Functions

When the PH model is true, the following relationship holds for an individual with the set of covariates, \( x_i \):

\[
S(t) = [S_0(t)]^{\exp\{x_i \beta\}},
\]

where \( S_0(t) \) is called the baseline survival function. That is, the survival function for the set of covariates that yield \( x_i \beta = 0 \) so that \( \exp\{x_i \beta\} = 1 \).

(Remember from linear algebra that \( x_i \beta = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \))

To see this, remember that:

\[
h_i(t) = h_0(t) * \exp\{x_i \beta\}, \text{ and}
\]

\[
H_i(t) = \int_0^t h_0(u) \cdot \exp\{x_i \beta\} du = H_0(t) \cdot \exp\{x_i \beta\}
\]
\[ S_i(t) = \exp\{-H_i(t)\} = \exp\{-H_0(t) \cdot \exp[x_i \beta]\} \]

Noting that,

\[ (x^a)^b = x^{a \cdot b} \]

We get,

\[ S_i(t) = (\exp\{-H_0(t)\})^{\exp\{x_i \beta\}} = S_0(t)^{\exp\{x_i \beta\}} \]
What we would like to do is estimate the baseline survival function, $S_0(t)$, so we could examine the effects of the various covariates on the survival curve, the hazard function, or make predictions for a particular individual with a given set of covariates, among other things.

Typically, $S_0(t)$ is estimated by assuming that the parameters, $\beta$, are known. In practice, they are assumed to be equal to their maximum likelihood estimators.

For simplicity, we will assume that each survival time (associated with events, not censoring) is distinct. That is, $t_1 < t_2 < \ldots < t_k$, where there are a total of $k$ events. The baseline conditional probability of surviving beyond $t_i$ given that a person survived to $t_i$ is estimated with,

$$\alpha_i = \frac{S_0(t_i)}{S_0(t_{i-1})}$$

(See KM estimation)
The conditional probability of surviving beyond $t_i$ given that a person survived to $t_i$ with covariates $x_i$ is similarly derived as:

$$\frac{S(t_i | \beta, x_i)}{S(t_{(i-1)} | \beta, x_i)} = \left\{ \frac{[S_0(t_i)]^{\exp\{x_i, \beta\}}}{[S_0(t_{(i-1)})]^{\exp\{x_i, \beta\}}} \right\} = \alpha_i^{\exp\{x_i, \beta\}}$$

To estimate the probability defined above using data provided in the dataset, we ask a related question: For an individual with covariates $x_i$, what is the probability of dying (or eventing) at $t_i$ given that he has made it to $t_i$?

We already asked this question when looking at the MLEs of the covariates in the Cox model.
It is the ratio involving hazards:

\[
\frac{\exp\{x_i \beta\}}{\sum_{j=i}^{k} \exp\{x_j \beta\}}
\]

And with the usual definition of hazards, this ratio simplifies:

\[
\alpha_i^{\exp\{x_i \beta\}} = 1 - \frac{\exp\{x_i \beta\}}{\sum_{j=i}^{k} \exp\{x_j \beta\}}
\]

Individuals in the risk set at time, \(t_i\)
The estimate of the baseline conditional probability is then given as:

\[\alpha_i = \left(1 - \frac{\exp\{x_i \beta\}}{\sum_{j=i}^{k} \exp\{x_j \beta\}}\right)\exp\{x_i \beta\}^{-1}\]

Substituting the MLE’s will give a good approximation of the probability of interest. Then the baseline survival probability is estimated by taking the product of the conditional survival probabilities, just like the KM estimate.
Mathematically, the estimate of the baseline survival function can be written as:

$$\hat{S}_0(t) = \prod_{i:t_i<t} \hat{\alpha}_i$$

When there are ties in survival times, evaluating the conditional probabilities is considerably more complex.

 Commands in SAS can be used to provide an estimate of the baseline survival function by using the “baseline” command. The default, however, is to evaluate the survival function when the covariates are evaluated at their respective means, i.e., $x = \bar{x}$
Example:

Output Dataset is named “aa”

```sas
proc phreg data=asa.hmohiv;
model time*censor(0)=drug age / ties=efron;
baseline out=aa survival=s logsurv=ls;
run;

data ab;
set aa;
ls=-ls;
run;

proc gplot data=ab;
symbol1 v=none i=join;
symbol2 v=none i=join;
plot (s ls)*time;
run;

proc print data=ab;
run;
```

An estimate of the survival for an “average” person is provided in “aa” under “s”

An estimate of the cumulative hazard for an “average” person is provided in “aa” under the variable name “ls”
Partial Listing of Dataset “ab”:

<table>
<thead>
<tr>
<th>Obs</th>
<th>drug</th>
<th>age</th>
<th>time</th>
<th>s</th>
<th>ls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.49</td>
<td>36.07</td>
<td>0</td>
<td>1.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>2</td>
<td>0.49</td>
<td>36.07</td>
<td>1</td>
<td>0.88452</td>
<td>0.12270</td>
</tr>
<tr>
<td>3</td>
<td>0.49</td>
<td>36.07</td>
<td>2</td>
<td>0.84069</td>
<td>0.17354</td>
</tr>
<tr>
<td>4</td>
<td>0.49</td>
<td>36.07</td>
<td>3</td>
<td>0.73796</td>
<td>0.30387</td>
</tr>
<tr>
<td>5</td>
<td>0.49</td>
<td>36.07</td>
<td>4</td>
<td>0.68897</td>
<td>0.37255</td>
</tr>
<tr>
<td>6</td>
<td>0.49</td>
<td>36.07</td>
<td>5</td>
<td>0.59206</td>
<td>0.52415</td>
</tr>
<tr>
<td>7</td>
<td>0.49</td>
<td>36.07</td>
<td>6</td>
<td>0.56224</td>
<td>0.57582</td>
</tr>
</tbody>
</table>

PH Estimate of $S(t)$

![Survivor Function Estimate](image-url)
Of course, it may be more interesting to compare the survival curves of drug users against those who don’t while controlling for the effects of age. To do this, we use the “strata” command (More will be discussed about this command later).

```plaintext
proc phreg data=asa.hmohiv;
model time*censor(0)=age / ties=efron;
baseline out=aa survival=s logsurv=ls;
strata drug;
run;
```

**S(t) by Drug**

**H(t) by Drug**
Technically, these curves are step functions, just like the KM curve, but it seems difficult to generate attractive step functions in “gplot” so the values were interpolated instead.

These survival and cumulative hazard curves look quite similar to the ones obtained through non-parametric (Chapter 3) and fully parametric methods (Chapter 4). As a matter of fact, the differences between the KM and the PH survival curves are only slight.

If you are interested in predicting the survival properties of an individual with a particular set of covariates, there will be a need to define a new dataset. Once the dataset is created, SAS will process the survival function (or transformations) for a person with these characteristics when the “covariates” option is used in the baseline statement.
**Example:** Predict the probability (and give 95% C.I.) of a drug free 35 y.o. individual and a 23 y.o. drug user of surviving beyond 8 months.

**SAS Code Follows:**

```sas
data vars;
input age drug;
datalines;
35 0
23 1;
run;

proc phreg data=asa.hmohiv;
model time*censor(0)=age drug / ties=efron;
baseline out=aa covariates=vars survival=s lower=lcl upper=ucl / nomean;
run;
```

Dataset “vars” created for two individuals: (1) aged 35 and drug free, and (2) aged 23 and a drug user.

Covariates option used with the dataset, “vars.”

The lower and upper confidence intervals are obtained with these key words. Their values are placed into the variables, lcl and ucl.

The default level of confidence is 95%. Use the “alpha=value” for different levels of confidence.
A partial listing of the output dataset, “aa” follows:

<table>
<thead>
<tr>
<th>Obs</th>
<th>time</th>
<th>s</th>
<th>lcl</th>
<th>ucl</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>6</td>
<td>0.72954</td>
<td>0.63473</td>
<td>0.83851</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>0.65874</td>
<td>0.55342</td>
<td>0.78410</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>0.60128</td>
<td>0.49018</td>
<td>0.73755</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>0.55670</td>
<td>0.44263</td>
<td>0.70016</td>
</tr>
<tr>
<td>35</td>
<td>6</td>
<td>0.76209</td>
<td>0.65194</td>
<td>0.89085</td>
</tr>
<tr>
<td>36</td>
<td>7</td>
<td>0.69792</td>
<td>0.57265</td>
<td>0.85060</td>
</tr>
<tr>
<td>37</td>
<td>8</td>
<td>0.64514</td>
<td>0.50816</td>
<td>0.81905</td>
</tr>
<tr>
<td>38</td>
<td>9</td>
<td>0.60371</td>
<td>0.45891</td>
<td>0.79420</td>
</tr>
<tr>
<td>39</td>
<td>10</td>
<td>0.56155</td>
<td>0.41030</td>
<td>0.76854</td>
</tr>
</tbody>
</table>

The 9\textsuperscript{th} obs. gives the relevant information about the drug free 35 y.o. whereas the 37\textsuperscript{th} obs. gives the relevant info about the 23 y.o. drug user.

We see that despite the fact that the 35 y.o. is drug free, the 23 y.o. drug user has a higher estimated survival probability at 8 mos. The effect of age is not insignificant.
This demonstrates one of the major advantages of using the Cox model over some of the non-parametric methods discussed earlier such as the log-rank test statistic. Cox modeling enables us to control for several (and perhaps many) covariates at the same time, so confusing results can sometimes be resolved.

For example, if most of the drug users were young and non-users were older, it is possible that the “detrimental effects” of drug use would be masked by the age factor, yielding an insignificant log-rank test (or worse, indicating that drug use is beneficial!). This did not happen with this particular dataset because the distribution of age in both groups were roughly the same.

Cox modeling cannot correct for problems were severe confounding exists, such as a case were all drug users < 25 y.o. and all drug free people > 35.
Residual Diagnostics, Influential Observations, and Remedial Measures

There are residuals available for model checking for the Cox-model. These residuals can be used to assess the model fit and outlying observations.

Cox-Snell Residual

\[ r_{Ci} = \exp \{ x_i \hat{\beta} \} \hat{H}_0(t_i) \]

Basically, the Cox-Snell residual is the the cumulative hazard function for the \( i^{th} \) individual (See eq. 3.21 Collett).

When the model is adequate, the distribution of these residuals is exponential. In practice, they are not very helpful in assessing the adequacy of the Cox-model.
Martingale Residuals: Are defined as follows...

\[ r_{Mi} = \delta_i - r_{Ci} \]

where \( \delta_i \) are the censor indicators (1 if evented, 0 if censored). One way to think of martingale residuals is the difference between the observed death (or censoring) and the “expected” number of deaths for an individual with covariates, \( x_i \), in the interval \((0, t_i)\), (when the specific Cox-model is evaluated).

**Example:** Suppose an individual dies at a time when the cumulative hazard predicts 2.1 deaths at the time of his death (for a person with his covariates). Then his martingale residual is \( r_{Mi} = 1 - 2.1 = -1.1 \)

People who “live” longer than expected have negative martingales, and people who die quicker than expected have positive martingale residuals.
When the Cox-model is true, martingales have a “reversed” exponential distribution on $(1, -\infty)$ for those that evented. For those that don’t event, the martingales occur on $(0, -\infty)$.

Because of the heavy skewness of the distribution for martingales, deviance residuals are often the preferred since they have a more symmetric distribution. They can be interpreted in a similar manner as the residuals in LS regression. They are defined as:

$$r_{Di} = \text{sgn}(r_{Mi})\left[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}\right]^{\frac{1}{2}}$$

Martingale and deviance residuals can be obtained in SAS using the “resmart” and “resdev” options in the output statement of the phreg procedure.

An example follows with the HMOHIV dataset:
```sas
proc phreg data=asa.hmohiv;
model time*censor(0)=drug age;
output out=resid resmart=mart resdev=dev;
run;

proc gplot data=resid;
symbol1 v=dot h=0.2;
plot (mart dev)*age;
run;
```

Often there is a pattern in the residuals that do not necessarily indicate any problems with the model (see p175 of Allison).
These residuals can be used to help assess the functional form of a potential covariate in the model. An example follows:

```sas
proc phreg data=asa.hmohiv;
model time*censor(0)=drug;
output out=resid resmart=mart resdev=dev
/ order=data;
run;

data resid2;
merge asa.hmohiv resid;
run;

proc gplot data=resid2;
symbol1 i=sm70s v=dot h=0.2;
plot (dev)*age;
run;
```

Notice that age is left out of the model.

The “order=data” option keeps the output data in the same order as the “asa.hmohiv” dataset. This is necessary for the “merge” statement in the following data step.

This interpolation method uses a smoothed spline to fit the data. The larger the number, the smoother the spline.

Because the data are typically so noisy, it is often difficult to see any relationship without graphical guidance, such as the use of splines.
The deviance residuals indicate that a linear term for age is adequate. In other cases, a quadratic term may be needed (e.g. age2 = age **2) or an indicator function (e.g. if age > 32, then Iage = 1; else Iage = 0;).
Schoenfeld Residuals

These residuals are defined, for covariate $x_k$, as the difference between the $i^{th}$ individual’s covariate, $x_{ik}$, who died at time $t_i$, and the “expected” value of $x_k$ for all people in the risk set at $t_i$.

$$r_{Si_k} = x_{ik} - \sum_{j=1}^{n_i} x_{jk} p_j$$

where $p_j$ is the probability that person $j$ dies at time $t_i$.

Schoenfeld residuals are mainly used to detect departures from the proportional hazards assumption. If there is a pattern in these residuals against survival time, the PH assumption is questionable. It can be shown that: $E[r_{Si_k}] \approx \beta_k(t_i) - \hat{\beta}_k$

where $\beta_k(t_i)$ is a time dependent coefficient (or, alternatively, that the corresponding covariate is time dependent).
An example of using Schoenfeld residuals in SAS is provided below:

```sas
proc phreg data=asa.stan;
model surv1*dead(0)=trans surg ageaccpt;
output out=resid ressch=schtrans schsurg schage;
run;

proc gplot data=resid;
symbol1 v = dot h = 0.2;
plot (schtrans schsurg schage)*surv1;
run;
```

Schoenfeld Residuals for ‘Trans’

Lack of residuals for long survival times for the ‘trans’ covariate when it is zero indicates a possible time dependent coefficient.

As it turned out, a time dependent covariate analysis was required for this dataset, and time dependent covariates violate the proportional hazards assumption.
These residuals show essentially no relationship, or at worst, a mild increase over time. If one believed that $\beta_{\text{age}}(t_i) \propto t$, we could change the model so that, $x_{\text{age}}(t) = \text{age} + t$. However, the results of this model would give identical results as the original (see p142). The pattern is probably meaningless as it results from only a few observations.
Influence diagnostics

Like linear regression diagnostics, the influence of each observation can be assessed through $\Delta \beta$’s and overall impact measures like Cook’s distance. That is, the model is estimated with the entire dataset, and then the effect of the $i^{th}$ observation on the estimates is assessed by fitting the model without the $i^{th}$ observation and comparing to the original results. If the differences are substantial, there is concern about influence.

**Delta-Beta:** For parameter $\beta_j$, the impact of observation ‘i’ on the estimate is assessed with:  
\[
\Delta \beta_{j(i)} = \hat{\beta}_j - \hat{\beta}_{j(i)}
\]

Most computer algorithms estimate this difference through approximations.
Likelihood Displacement (LD): Looks at the effect of removing the i\textsuperscript{th} observation from the dataset on the likelihood: 

$$LD_{(i)} = 2\left[\log[L(\hat{\beta})] - \log[L(\hat{\beta}_{(i)})]\right]$$

Often plots are made with these influence statistics against the rank of the survival time.

---

```plaintext
proc phreg data=asa.hmohiv;
model time*censor(0)=drug age;
output out=infl ld=ldstat dfbeta=ddrug dage;
run;

proc sort data=infl;
by time;
run;

data infl;
set infl;
rank=_N_;
run;

proc gplot data=infl;
symbol1 v=dot h=0.2 i=join;
plot (ldstat ddrug dage)*rank;
run;
```
Remedial Measures:

With potentially influential observations, accuracy of the statistics should be determined. Analysis of the data without the observations should be conducted to see if the conclusions of the study are dependent on them. If they are, further investigation with scientists who have an expertise in the area of study should be carried out. If this cannot resolve the problem, results of all analyses should be presented.
Interactions with Time:

When the covariate is mostly a nuisance variable (i.e. not of direct interest to the investigator) that is categorical, a strata command can be issued to the phreg procedure. This command will allow the model to incorporate different baseline hazard functions for each level (group) of the nuisance variable while still enabling estimation of the effects of more important factors such as treatment effects. The model would be written as:

\[
\ln \{h_{ij}(t)\} = \ln \{h_j(t)\} + \beta \cdot x_i
\]

where \(h_j(t)\) is the baseline hazard for the nuisance covariate at the \(j^{th}\) level.

Example: In the Stanford heart transplant example, there might be primary interest in evaluating the mismatch score on determining “good” candidates for transplants. Knowing that
“rejection” is potentially violating the PH assumption and could be associated with the mismatch score, there is some concern over the appropriateness of using a Cox model, so the analysis is stratified on the rejection factor:

```plaintext
proc phreg data=stan2;
model surv2*dead(0)= surg wtime m3;
strata reject;
run;
```

```
Analysis of Maximum Likelihood Estimates

Parameter    Standard     Hazard
Variable DF   Estimate     Error Chi-Square Pr > Chi Sq  Ratio

surg     1    -0.61211    0.48492     1.5934     0.2068    0.542
wtime    1     -0.0009428  0.00466     0.0409     0.8397    0.999
m3        1     0.43933    0.31155     1.9885     0.1585    1.552
```

The conclusions are essentially the same as in the unstratified analysis.
Alternatively, an explicit interaction term with time can be written into the model and handled analytically by a computer. For example:

\[
\ln \{ h(t) \} = \ln \{ h_0(t) \} + (\beta_1 + \beta_2 t) x = \alpha(t) + \beta_1 \cdot x + \beta_2 \cdot t \cdot x
\]

To handle such a problem in SAS, we would code the following:

```sas
proc phreg data=stan2;
model surv2*dead(0)= surg wtime m3 reject rejtime;
rejtime=reject*surv2;
run;
```

```
Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi Sq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>surg</td>
<td>1</td>
<td>-0.58073</td>
<td>0.48637</td>
<td>1.4256</td>
<td>0.2325</td>
<td>0.559</td>
</tr>
<tr>
<td>wtime</td>
<td>1</td>
<td>-0.0003638</td>
<td>0.00477</td>
<td>0.0058</td>
<td>0.9392</td>
<td>1.000</td>
</tr>
<tr>
<td>m3</td>
<td>1</td>
<td>0.43796</td>
<td>0.32315</td>
<td>1.8368</td>
<td>0.1753</td>
<td>1.550</td>
</tr>
<tr>
<td>reject</td>
<td>1</td>
<td>0.43454</td>
<td>0.44335</td>
<td>0.9607</td>
<td>0.3270</td>
<td>1.544</td>
</tr>
<tr>
<td>rejtime</td>
<td>1</td>
<td>0.00443</td>
<td>0.00260</td>
<td>2.9062</td>
<td>0.0882</td>
<td>1.004</td>
</tr>
</tbody>
</table>
```
The Wald statistics are not statistically significant for either reject or rejtime, however, to properly determine the significance of these two variables, it is better to perform a likelihood ratio test. Doing that, we notice that $X^2 = 283.06 - 264.67 = 18.39$, which has a $\chi^2(2 \text{ df})$ under the null. This is certainly significant.

Since the product of time and reject is borderline significant with the Wald test, it is reasonable to conclude that among those who are destined to reject, their risk of death increases over time.

The lack of statistical significance for the Wald test may be the result of a small sample size (i.e. insufficient power).

In any case, the mismatch score is insignificant when rejection is known. If the investigator was truly interested in the utility of m3, he/she may try logistic regression or SA with the endpoint as the time of organ rejection.