CPH 931 — Fall 2009 — Dr. Charnigo

Lecture 7

Recalling proportional hazards regression

Introduction. In CPH 930 you learned that the proportional hazards regression model is often used to relate the hazard function for a time-to-event response variable T to the values assumed by one or more explanatory variables $X_1, \ldots, X_k$.

The proportional hazards regression model has the form

$$\log[h_{x_1,\ldots,x_k}(t)] = \alpha(t) + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{1}$$

where

$$\begin{aligned} h_{x_1,\ldots,x_k}(t) &:= -\frac{d}{dt}\{\log P(T > t | X_1 = x_1, \ldots, X_k = x_k)\} \\ &= \text{limit}_{\Delta \to 0^+} \frac{P(t < T < t + \Delta | T > t, X_1 = x_1, \ldots, X_k = x_k)}{\Delta} \end{aligned} \tag{2}$$

is the hazard function at time t (> 0). Note that the intercept in (1) — but not the partial slope coefficients! — can vary with t.

Putting $h_0(t) := \exp[\alpha(t)]$, we can rewrite (1) as

$$h_{x_1,\ldots,x_k}(t) = h_0(t) \exp[\beta_1 x_1 + \cdots + \beta_k x_k]. \tag{3}$$

The proportional hazards regression model is so named because the t-dependence of the hazard function is proportional to $h_0(t)$ no matter the values of $X_1, \ldots, X_k$.

Consequently, the hazard ratio corresponding to a one-unit increase in X_1 (holding fixed $X_2, \ldots, X_k$) is

$$\frac{h_{x_1+1,\ldots,x_k}(t)}{h_{x_1,\ldots,x_k}(t)} = \frac{h_0(t) \exp[\beta_1(x_1+1) + \cdots + \beta_k x_k]}{h_0(t) \exp[\beta_1 x_1 + \cdots + \beta_k x_k]} = \exp[\beta_1]. \tag{4}$$

The hazard ratio $\exp[\beta_1]$ in (4) cannot depend on t because β_1 does not depend on t. Thus, even though the hazard function may depend on t, the proportional hazards regression model constrains hazard ratios to be independent of t.

Today we will discuss two problems left unresolved in CPH 930. First, how do we proceed if one or more of $X_1, \ldots, X_k$ can change values over time? Second, how do we proceed if we believe that constraining the hazard ratios to be independent of t is unrealistic?

Time-dependent explanatory variables

Motivating example. The file {BMTdata.xls}, obtained from {http://www.biostat.mcw.edu/homepgs/klein/bmt.html}, contains information from a study on bone marrow transplants for leukemia patients.

As Klein and Moeschberger explain in their book, the prognosis for a leukemia patient receiving a bone marrow transplant may depend not only on factors known at the time of transplantation but also on factors that emerge during the recovery process.

For simplicity we will not consider all of the variables in {BMTdata.xls}, but a few variables to which I will call your attention are as follows:

• The variable DiseaseGroup equals 1 for patients with acute lymphoblastic leukemia (ALL), 2 for patients with acute myeloctic leukemia deemed to be at low risk based on the number of relapses (AML Low Risk), and 3 for patients with acute myeloctic leukemia deemed to be at high risk (AML High Risk). Once inside SAS I also created Z_1 as an indicator for DiseaseGroup = 2 and Z_2 as an indicator for DiseaseGroup = 3.

• The variable MTXUsed is an indicator for whether a patient received a graft-versus-host prophylactic combining methotrexate (MTX) with cyclosporine and/or methylprednisolone.

• The variable TimetoPlateRec equals the time after transplant (in days) at which platelet recovery occurs, in the sense that the platelet count returns to a self-sustaining level of at least 40×10^9/l. This variable is subject to right censoring, which is shown by a 0 value for PlateRec. We will not regard TimetoPlateRec as the response variable. Rather, we will view platelet recovery as a time-dependent explanatory variable. The idea is that a patient's risk of death or relapse may change when platelet recovery occurs.

• The variable TimetoDeathorRelapse equals the time after transplant (in days) at which death or relapse occurs. This variable is also subject to right censoring, which is shown by a 0 value for DeathorRelapse. We will regard TimetoDeathorRelapse as the response variable.

Extending proportional hazards regression. Suppose that we wish to express the hazard function for TimetoDeathorRelapse in terms of Z_1, Z_2, and platelet recovery. A naive attempt would entail fitting the proportional hazards regression model

$$\log[h_{z_1,z_2,\mathrm{PlateRec}}(t)] = \alpha(t) + \beta_1 z_1 + \beta_2 z_2 + \beta_3 \mathrm{PlateRec}, \tag{5}$$

where PlateRec $= 1$ for a patient who eventually experienced platelet recovery and PlateRec $= 0$ for a patient who never experienced platelet recovery.

The problem with (5) is that it does not distinguish between patients who experience platelet recovery at different times.

To be concrete, suppose that Patient A experiences platelet recovery at time $t = 10$ and that Patient B experiences platelet recovery at time $t = 40$. If Patient A and Patient B both have DiseaseGroup $= 2$, then (5) implies that Patient A and Patient B have the same hazard function, namely $h_0(t)\exp[\beta_1 + \beta_3]$.

This does not seem right. At time $t = 20$, when Patient A has experienced platelet recovery but Patient B has not, their hazards ought to be different.

What we really need is a model

$$\log[h_{z_1,z_2,\text{PlateRec}(t)}(t)] = \alpha(t) + \beta_1 z_1 + \beta_2 z_2 + \beta_3 \text{PlateRec}(t), \qquad (6)$$

where $\text{PlateRec}(t) = 1$ if platelet recovery has occurred by time t and $\text{PlateRec}(t) = 0$ if platelet recovery has not occurred by time t.

Then Patients A and B would have different hazards at time $t = 20$. Specifically, Patient A would have hazard $h_0(20)\exp[\beta_1 + \beta_3]$, while Patient B would have hazard $h_0(20)\exp[\beta_1]$.

Such an extension of proportional hazards regression is available via SAS's PROC PHREG with a few extra lines of code following the MODEL statement. Note that $\text{PlateRec}(t)$ cannot be created in a DATA step.

Results for the motivating example. The SAS output for proportional hazards regression with time-dependent explanatory variables looks just like the SAS output for ordinary proportional hazards regression.

Page 1 of {BMTExamples.rtf} tells us that there were 137 patients, of which 83 died or relapsed while in the study, that $-2\log L = 723.876$ for model (6), and that $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is rejected with p-value less than 0.0001.

Page 2 of {BMTExamples.rtf} shows that, when we control for platelet recovery, neither AML Low Risk nor AML High Risk patients can be statistically distinguished from ALL patients. However, the negative $\widehat{\beta}_1$ with p-value 0.0854 and the positive $\widehat{\beta}_2$ with p-value 0.1546 suggest that AML Low Risk patients may be statistically distinguishable from AML High Risk patients. We could investigate this possibility by adding a TEST statement following the MODEL statement in our invocation of PROC PHREG.

In any case, the time-dependent explanatory variable platelet recovery is highly significant with p-value 0.0007.

Comments on interpretation with time-dependent explanatory variables. One interpretation of the estimated hazard ratio $0.326 = \exp[-1.11943]$ is that the hazard at time t is an estimated 67.4% smaller for a patient who has experienced platelet recovery by time t than for a patient in the same disease group who has not experienced platelet recovery by time t.

A second interpretation of the 0.326 is that the hazard of death or relapse for a given patient is estimated to decline by 67.4% the moment his or her platelet count returns to a self-sustaining level of at least $40 \times 10^9/\mathrm{l}$.

The second interpretation does not seem plausible biologically. Unless the patient experiences a sudden catastrophic event (such as a myocardial infarction), we anticipate that his or her hazard should change gradually over time. In particular, we do not anticipate that 67.4% of the hazard should disappear instantaneously. The hazard should become incrementally smaller as the platelet count gets closer and closer to $40 \times 10^9/\mathrm{l}$.

However, the second interpretation is a consequence of how we defined platelet recovery. Whenever we have a time-dependent explanatory variable in which a naturally continuous phenomenon has been discretized, an interpretation of the exponentiated partial slope coefficient as an instantaneous hazard reduction (or increase) is automatic.

Thus, replacing PlateRec(t) by PlateCount(t), where PlateCount(t) denotes the platelet count at time t, might be desirable. Unfortunately, the information that would be required to create PlateCount(t) is not contained in {BMTdata.xls}.

Testing the proportional hazards assumption

A useful technique. One way to see whether the proportional hazards assumption is met — i.e., whether hazard ratios truly do not depend on t — is to fit a proportional hazards regression model with artificial time-dependent explanatory variables.

Let $V_1(t) := X_1 \times g(t)$, $V_2(t) := X_2 \times g(t)$, and so forth, where $g(t)$ is an increasing function of time such as $\log t$. Note that $V_1(t), \ldots, V_k(t)$ essentially represent interactions of the explanatory variables with time.

The model

$$\log[h_{x_1,\ldots,v_k(t)}(t)] = \alpha(t) + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma_1 v_1(t) + \cdots + \gamma_k v_k(t) \quad (7)$$

can be expressed as

$$\log[h_{x_1,\ldots,v_k(t)}(t)] = \alpha(t) + x_1\{\beta_1 + \gamma_1 g(t)\} + \cdots + x_k\{\beta_k + \gamma_k g(t)\}, \quad (8)$$

revealing that the hazard ratio associated with a one-unit increase in X_1 is

$$\exp[\beta_1 + \gamma_1 g(t)]. \quad (9)$$

The hazard ratio has no t-dependence if and only if $\gamma_1 = 0$.

Hence, we can see whether the proportional hazards assumption is met by testing the null hypothesis that $\gamma_1 = \cdots = \gamma_k = 0$. Rejection of the null hypothesis implies that the proportional hazards assumption is not met.

Since the hazard ratio for one explanatory variable may be t-dependent while the hazard ratio for another explanatory variable is not, another option besides testing the null hypothesis that $\gamma_1 = \cdots = \gamma_k = 0$ is to test individual null hypotheses that $\gamma_1 = 0$, $\gamma_2 = 0$, and so forth.

The rationale for the second option is that we can more easily deal with a violation of the proportional hazards assumption if we know which explanatory variables yield t-dependent hazard ratios.

Illustration with the motivating example. Beginning on page 3 of {BMTExamples.rtf} are results for a proportional hazards regression model with explanatory variables Z_1, Z_2, MTXUsed and artificial time-dependent explanatory variables $Z_1 \times \log t$, $Z_2 \times \log t$, MTXUsed $\times \log t$. As before, TimetoDeathorRelapse is the response variable.

On page 4 we find that $\gamma_1 = 0$ is not rejected (p-value 0.7317) and that $\gamma_2 = 0$ is not rejected (p-value 0.3393). However, $\gamma_3 = 0$ is rejected (p-value 0.0329). Other than saying that the proportional hazards assumption does not seem to be met, how can we interpret the latter result?

Discussion questions.
1. What is the estimated hazard ratio at time t, where in this instance we refer to the hazard at time t for a patient on whom methotrexate was used divided by the hazard at time t for a patient in the same disease group on whom methotrexate was not used?
2. Evaluate the estimated hazard ratio at $t = 1$, $t = 10$, and $t = 100$.
3. What can we say about the manner in which methotrexate use elevates the risk of relapse or death?

Stratified proportional hazards regression

Data analysis options when hazards are not proportional. One option when the proportional hazards assumption is not satisfied is to proceed as we did in the discussion questions, namely by describing the t-dependence of the hazard ratio for any explanatory variable whose interaction term achieved a small p-value.

On the other hand, we would not want to bother describing the t-dependence of the hazard ratios for explanatory variables whose interaction terms did not achieve small p-values.

So, in practice we might pare down the model before reporting and interpreting results. In the above illustration, this would entail refitting the model without $Z_1 \times \log t$ and $Z_2 \times \log t$ (but with MTXUsed $\times \log t$).

A second data analysis option is to perform stratified proportional hazards regression. Continuing with the above illustration, this entails constructing one hazard function for patients with MTXUsed $= 1$,

$$\log[h_{z_1,z_2}^{MTXUsed=1}(t)] = \alpha(t) + \beta_1 z_1 + \beta_2 z_2, \tag{10}$$

and another hazard function for patients with MTXUsed $= 0$,

$$\log[h_{z_1,z_2}^{MTXUsed=0}(t)] = \alpha^*(t) + \beta_1^* z_1 + \beta_2^* z_2. \tag{11}$$

If we wish, we can constrain $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$. However, we do not constrain $\alpha(t) = \alpha^*(t)$. Indeed, a constraint of the form $\alpha(t) = \alpha^*(t) + C$ would imply the proportional hazards assumption that we are trying to avoid!

The preceding formulation — and all material that now follows — can be modified if the number of explanatory variables is different or if the stratification variable has more than two categories.

Results for the motivating example with partial slope constraints. Pages 5 and 6 of {BMTExamples.rtf} show results for (10) and (11) with the constraints that $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$.

The estimates of β_1 and β_2, based on all 137 observations, are -0.50018 and 0.43855 respectively. The corresponding p-values are 0.0874 and 0.1037. Again, these p-values do not imply that disease group is altogether irrelevant. Rather, they imply that AML Low Risk patients cannot be statistically distinguished from ALL patients and that AML High Risk patients cannot be statistically distinguished from ALL patients.

Especially since $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have opposite signs, we are inclined to think that AML Low Risk patients can be statistically distinguished from AML High Risk patients. We can confirm this directly by adding a TEST statement to our invocation of PROC PHREG, but some indirect evidence is already present: the null hypothesis that $\beta_1 = \beta_2 = 0$, which corresponds to complete irrelevance of disease group, is easily rejected with p-value 0.0018.

We note that $-2 \log L = 632.047$.

Results for the motivating example without partial slope constraints. Pages 7 through 10 of {BMTExamples.rtf} show results for (10) and (11) without constraints on β_1, β_2, β_1^*, and β_2^*.

The estimates of β_1^* and β_2^*, based on 97 observations, are -0.43100 and 0.64983 respectively. The corresponding p-values are 0.2441 and 0.0636.

The estimates of β_1 and β_2, based on only 40 observations, are -0.46836 and 0.05096 respectively. The corresponding p-values are 0.3522 and 0.9112.

We note that $-2 \log L$ is the sum of two pieces, one contributed from each stratum of MTXUsed, $-2 \log L = 461.699 + 168.995 = 630.694$.

To constrain or not to constrain. In statistics we impose constraints not because we believe that they are true but because, if the constraints are not grossly false, they simplify and/or optimize data analysis.

Recall from your introductory statistics course that there were actually two versions of the independent samples T test for comparing two population means. One version constrained the population variances to be equal,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s^2(1/n_1 + 1/n_2)}},$$

while the other did not,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_1 + s_y^2/n_2}}.$$

As a student you may have preferred the equal-variances version because you knew that the degrees of freedom were $n_1 + n_2 - 2$, whereas with the unequal-variances version a gruesome computation was required,

$$df = \frac{(s_x^2/n_1 + s_y^2/n_2)^2}{(s_x^2/n_1)^2/(n_1 - 1) + (s_y^2/n_2)^2/(n_2 - 1)}.$$

However, a better reason for preferring the equal-variances version is that df as defined above is always less than $n_1 + n_2 - 2$, sometimes much less. This implies that the critical value for rejecting the null hypothesis with the unequal-variances version, $t_{df,1-\alpha/2}$, is larger than the critical value for rejecting the null hypothesis with the equal-variances version, $t_{n_1+n_2-2,1-\alpha/2}$. Hence, rejecting the null hypothesis is potentially more difficult with the unequal-variances version. So, if the constraint of equal variances is not grossly false, as judged by an auxiliary F test for comparing two population variances, we impose it even though we do not really believe it to be true.

Likewise, we impose the constraints $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$ in stratified proportional hazards regression unless they seem unreasonable. Reporting one set of parameter estimates and p-values is more palatable than reporting two sets of comparatively imprecise parameter estimates and p-values.

A likelihood ratio test can be used to decide whether the constraints $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$ are reasonable. In our example the likelihood ratio test statistic is $632.047 - 630.694 = 1.353$. Since there are two constraints, the reference distribution is chi-square on two degrees of freedom. The test statistic 1.353 being (far) less than $\chi^2_{2,0.95} = 5.99$ suggests that there is no real harm in imposing the constraints.