The HAZPRED Procedure

The Cleveland Clinic Foundation.¹

¹Inquiries concerning this procedure should be directed to Eugene H. Blackstone, MD at (216) 444-6712. Our general e-mail address for inquiries about program availability, installation or functioning is hazard@bio.ri.ccf.org. Mail to that address is simultaneously received by Dr. Blackstone and Mr. John Ehrlinger, the programmer currently supporting the procedures. The procedures, along with this documentation, examples, and useful macros, are available on the Internet at www.clevelandclinic.org/heartcenter/hazard. Specific data application questions are welcomed, directed either to Dr. Blackstone at blackse@ccf.org or to hazard support at hazard@bio.ri.ccf.org.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>2</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>4</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>4</td>
</tr>
<tr>
<td>SYNTAX</td>
<td>4</td>
</tr>
<tr>
<td>SPECIFICATIONS</td>
<td>4</td>
</tr>
<tr>
<td><strong>THE PROC HAZPRED STATEMENT</strong></td>
<td>5</td>
</tr>
<tr>
<td>PROC HAZPRED OPTIONS;</td>
<td>5</td>
</tr>
<tr>
<td><strong>THE TIME STATEMENT</strong></td>
<td>6</td>
</tr>
<tr>
<td>TIME VARIABLE NAME;</td>
<td>6</td>
</tr>
<tr>
<td>CAUTIONS</td>
<td>7</td>
</tr>
<tr>
<td><strong>TIME ZERO</strong></td>
<td>7</td>
</tr>
<tr>
<td><strong>CONFLICTING SAS VARIABLE NAMES</strong></td>
<td>7</td>
</tr>
<tr>
<td>DETAILS</td>
<td>7</td>
</tr>
<tr>
<td><strong>MISSING VALUES</strong></td>
<td>7</td>
</tr>
<tr>
<td><strong>STRUCTURE OF INPUT DATA SET OF ESTIMATES</strong></td>
<td>7</td>
</tr>
<tr>
<td>COMPUTATIONAL METHOD</td>
<td>8</td>
</tr>
<tr>
<td><strong>POINT ESTIMATES</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>CONFIDENCE LIMITS AROUND POINT ESTIMATES</strong></td>
<td>9</td>
</tr>
<tr>
<td><strong>SURVIVORSHIP FUNCTION</strong></td>
<td>9</td>
</tr>
<tr>
<td><strong>HAZARD FUNCTION</strong></td>
<td>10</td>
</tr>
<tr>
<td>EXAMPLES</td>
<td>10</td>
</tr>
</tbody>
</table>
ABSTRACT

PROC HAZPRED is a procedure for calculation of time-related estimates (predictions) from a parametric equation previously established by PROC HAZARD.

INTRODUCTION

The HAZPRED procedure uses an estimates data set produced by the OUTHAZ option in PROC HAZARD and an input data set of values for variables to generate predictions from that parametric equation specification in the form of an output data set. The estimates data set contains equation specifications, parameter estimates, status flags, and the variance-covariance matrix. By input of specific values for the time variable and for each of the covariables that may be defined in the equation, PROC HAZPRED calculates maximum-likelihood survivorship and hazard function estimates and their confidence limits. In addition, survivorship and hazard function estimates are provided in the output data set for each hazard phase.

With both survivorship and hazard function estimates available, other survival functions may be calculated subsequently, such as the death density function, cumulative hazard function, lifetime function, and cumulative incidence functions (for competing risks of events).

There is no printed output from PROC HAZPRED.

SYNTAX

Since PROC HAZPRED is now an external program, rather than integrated into SAS® by means of its Toolkit, the procedure must be enclosed within a macro calling sequence as follows:

```sas
%HAZPRED(
   PROC HAZPRED etc;
);
```

SPECIFICATIONS

The following statements can be used with the HAZPRED procedure:

```sas
PROC HAZPRED options;
   TIME variable_name;
```

The TIME statement is required.
The PROC HAZPRED Statement

PROC HAZPRED options;
This statement invokes the procedure. The following options are available:

CLIMITS = value
     CL=value  specifies a value between 0 and 1 for the asymptotic confidence
     coefficient. The default value is 0.683..., representing the
equivalent of one standard deviation (roughly 70% confidence
limits). For 95% confidence limits, specify the value 0.95. If
NOCL is specified CL is ignored.

DATA = SASdataset  names the SAS data set containing variables to be used for
prediction. Each observation in the data set must define TIME
(see below) and a value for each concomitant variable specified in
the model. If DATA= is omitted, the most recently created SAS
data set is used.

ID=variable_name  informs procedure that all consecutive observations with the same
ID value have identical concomitant information, and only the
calculations for time will be performed. There is no verification
that this assumption is valid. It is recommended that the data set
be sorted according to variable_name (which may be numeric or
character) since complete calculations always occur when the
procedure detects a different ID value. This option is used
primarily to accelerate computations. By default, if ID is not
specified, complete calculations are done on each observation.

INHAZ = SASdataset  names the SAS data set containing the model specification,
parameter estimates, and variance-covariance matrix. See
Structure of Input Data Set of Estimates later in this chapter for
details of the structure of this estimates data set variables to be
used for prediction. The INHAZ data set must be named or the
procedure will terminate without performing any predictions.

NOCL  suppresses default calculation of confidence limits for the
estimates. _CLLSURV, _CLUSURV, _CLLHAZ, and _CLUHAZ
are set to missing.

NOHAZ  NOH  suppresses default calculation of the hazard function, its
confidence limits, and its phase components. _HAZARD,
 _CLLHAZ, _CLUHAZ, _EARLYH, CONSTH, and _LATEH are
set to missing.

NOSURV  NOS  suppresses default calculation of the survivorship function, its
confidence limits, and its phase components. _SURVIV,
 _CLLSURV, _CLUSURV, _EARLYS, CONSTS, and _LATES
are set to missing.
OUT = SASdataset names the output SAS data set that contains the prediction variables and all the input variables. If an output data set is not specified, the predictions are added to the data set specified in the DATA= statement. The variable names added to the data set are as follows:

- _SURVIV survivorship function
- _CLLSURV lower confidence limit of survivorship function
- _CLUSURV upper confidence limit of survivorship function
- _EARLYS early hazard phase conditional survivorship function
- _CONSTS constant hazard phase conditional survivorship function
- _LATES late hazard phase conditional survivorship function
- _HAZARD hazard function
- _CLLHAZ lower confidence limit of hazard function
- _CLUHAZ upper confidence limit of hazard function
- _EARLYH hazard function for early hazard phase
- _CONSTH hazard function for constant hazard phase
- _LATEH hazard function for late hazard phase

Any of these variables that are not applicable to the model or that are not calculated based upon user selection options are set to missing values.

The TIME Statement

TIME variable_name;

The TIME statement names a variable in the input data set that contains values for the time of the event. It must be expressed in the same units of time that were used in PROC HAZARD for the data analysis. (The time of event variable can be any other positive-valued variable whose distribution was modeled parametrically). The values must be positive numbers, greater than zero, otherwise calculations are not performed and missing values are generated for all variables output from PROC HAZPRED.
CAUTIONS

Time Zero
The model is defined within the interval zero to infinity, but at exactly zero or infinity the model is defined only by limits. The survivorship function is 1.0 at $t=0$. However, the hazard function will be exactly zero or infinity in many cases, although certain combinations of parameters will generate finite values, as might also occur with the inclusion of the constant hazard phase. Consequently, PROC HAZPRED outputs missing values for the hazard function when $t=0$.

Conflicting SAS Variable Names
The variable names output by PROC HAZARD have been selected to avoid conflict with usual SAS variable naming conventions. You should consider these HAZPRED names as restricted ones. In particular, if predictions are generated for a data set that has already been processed by HAZPRED, these names will be flagged as being duplicate.

DETAILS

Missing Values
The HAZPRED procedure does not perform calculations if an observation has missing or inadmissible values for variables contained in the model or in the TIME statement. For such observations, missing values for variables generated by PROC HAZPRED are output.

Structure of Input Data Set of Estimates
The input data set of parameter estimates and fixed values, model specifiers, and variance-covariance elements is output by PROC HAZARD and is structure as follows:

Observations 1 through 6: Model Specification Flags
G1FLAG early phase equation flag
FIXDEL0 DELTA is to be fixed at zero (default)
FIXMNU1 |NU-Μ|=1 flag
G3FLAG late phase equation flag
FIXGE2 GAMMA·ETA=2 flag
FIXGAE2 (GAMMA·ETA)/ALPHA=2 flag

Observations 7 through 14: Shaping Model Parameter Estimates
This includes parameter estimates (_EST_) and a status flag (_STATUS_) set to 1 if estimated and 0 if the parameter was fixed or not in the model, followed by the variance-covariance matrix components.
DELTA
THALF
Observations 15 through \((17 + 3p)\): Intercepts and Covariables \((p \text{ unique variables})\)

Each variable is followed by its estimate \(_\text{EST}_\), status flag \(_\text{STATUS}_\) for each of the phases, and variance-covariance components. The number of unique variables is found by adding all variables with a different \textit{variable name} in all hazard phases.

E0
Concomitant information variable 1 (\textit{variable name} used)
.
.
.
Concomitant information variable \(p\) (\textit{variable name} used)
C0
Concomitant information variable 1 (uses C01 as \textit{variable name})
C02
.
.
.
C0\((p-1)\)
Concomitant information variable \(p\) (used C0\(p\) as \textit{variable name})
L0
Concomitant information variable 1 (uses L01 as \textit{variable name})
L02
.
.
.
L0\((p-1)\)
Concomitant information variable \(p\) (used L0\(p\) as \textit{variable name})

COMPUTATIONAL METHOD

Point Estimates

Point estimates (predictions) are obtained by direct algebraic solution of the cumulative hazard and hazard function equations cited in the introductory section of this document, \textit{Parametric Analysis of Time-Related Events}. The survivorship function is obtained by exponential transformation of the cumulative hazard function. When concomitant information is taken into account, so-called risk-adjusted and patient-specific estimates are formed.
Other useful functions can be calculated from the point estimates provided, including death density function, lifetime function, and cumulative incidence function for competing risk problems. In addition, various goodness-of-fit statistics can be generated using the estimates for individual patients in the study group. We have provided a SAS® macro for accomplishing some of these (%MACRO HAZPLOT) as part of the available software for downloading. In addition, we have programs or macros available for looking at differences between survival and hazard functions, at internal verification of time-related shape for recognizable subsets, for bootstrapping, and so forth. Make inquiries, and if we have done it, we will provide examples.

**Confidence Limits Around Point Estimates**

Confidence limits for these point estimates are estimated using the method of statistical differentials (Ku, 1966). The formulation is made separately for the confidence limits of survivorship and hazard function estimates, based on transformations to an unbounded scale.

**Survivorship Function**

For confidence limits of the survivorship function \( S(t, \Theta) \), a logistic transformation is used (NOTE: we will abbreviate \( Z(t, \Theta) \) as \( Z \) as originally suggested by Hazelrig and colleagues (1982) to simplify notation):

\[
S(t, \Theta) = \frac{1}{1 + e^Z}
\]

or

\[
P(t, \Theta) = \frac{1}{1 + e^{-Z}}
\]

where \( P(t, \Theta) = 1 - S(t, \Theta) \). Then, in terms of cumulative hazard \( \Lambda(t, \Theta) \):

\[
\Lambda(t, \Theta) = -\ln(S(t, \Theta))
\]

and

\[
Z = \ln(e^{\Lambda(t, \Theta)} - 1).
\]

Partial derivatives of \( Z \) with respect to the parameter vector theta required for the statistical differential are:

\[
\frac{\partial Z}{\partial \Theta} = \frac{\partial \Lambda(t, \Theta)}{\partial \Theta} \cdot \frac{1}{1 - e^{-\Lambda(t, \Theta)}}.
\]

The first order approximation to the variance of \( Z \), \( Var(Z) \), is

\[
Var(Z) = \left[ \frac{\partial Z}{\partial \Theta} \right]^{\text{transpose}} \cdot Cov(\Theta) \cdot \left[ \frac{\partial Z}{\partial \Theta} \right]^{\text{transpose}}
\]

Then the confidence limits of \( Z \) are:

\[
Z^+ = Z + t_{1-\alpha/2} \sqrt{Var(Z)}
\]

and

\[
Z^- = Z - t_{1-\alpha/2} \sqrt{Var(Z)}
\]

where \( t_{1-\alpha/2} \) is the inverse normal of the confidence coefficient \( \alpha \). Finally, \( Z \) and \( Z^+ \) are transformed to \( S(t, \Theta)^+ \) and \( S(t, \Theta)^- \), respectively.
These survivorship confidence limits are generally asymmetrical except at \( S(t)=0.5 \). They converge to 1.0 as time \( t \) approaches zero, and to 0.0 as \( t \) approaches infinity. Also, they are consistent in width above and below \( S(t)=0.5 \) in that the confidence limits for the same variance of \( Z \) are the same as for \( S(t) \) and \( 1-S(t) \).

**Hazard Function**

For confidence limits of the hazard function \( \lambda(t, \Theta) \), a logarithmic transformation is used:

\[
\lambda(t, \Theta) = e^Z
\]

and

\[
Z = \ln(\lambda(t, \Theta))
\]

The partial derivatives of \( Z \) with respect to the parameter vector \( \Theta \) are

\[
\frac{\partial Z}{\partial \Theta} = \frac{\partial \lambda(t, \Theta)}{\partial \Theta} \frac{1}{\lambda(t, \Theta)}
\]

The general scheme used for the survivorship estimates is then followed. The transformation guarantees that the upper and lower confidence limits of the hazard function are always positive and asymmetric. The confidence limits are not consistent between the hazard and the survivorship functions, since the calculations use different transformations and asymptotic approximations. However, the results are reasonably consistent for data set with approximately one hundred or more observations.

**EXAMPLES**

Several examples are available in the accompanying data file.

**REFERENCES**
