# Cox Proportional Hazards Survival Regression

(Contributed by John Pezzullo)

This program analyzes survival-time data by the method of Proportional Hazards regression (Cox). Given survival times, final status (alive or dead) , and one or more covariates, it produces a baseline survival curve, covariate coefficient estimates with their standard errors, risk ratios, 95% confidence intervals, and significance levels.

> *A patient asked his surgeon what the odds were of him surviving an impending operation. The doctor replied they were 50/50 but he'd be all right because the first fifty had already died!!*

## Background Information (just what is Proportional Hazards Survival Regression, anyway?)

Survival analysis takes the survival times of a group of subjects (usually with some kind of medical condition) and generates a survival curve, which shows how many of the members remain alive over time. Survival time is usually defined as the length of the interval between diagnosis and death, although other "start" events (such as surgery instead of diagnosis), and other "end" events (such as recurrence instead of death) are sometimes used.

The major mathematical complication with survival analysis is that you usually do not have the luxury of waiting until the very last subject has died of old age; you normally have to analyze the data while some subjects are still alive. Also, some subjects may have moved away, and may be lost to follow-up. In both cases, the subjects were known to have survived for some amount of time (up until the time you last saw them), but you don't know how much longer they might ultimately have survived. Several methods have been developed for using this "at least this long" information to preparing unbiased survival curve estimates, the most common being the Life Table method and the method of Kaplan and Meier.

We often need to know whether survival is influenced by one or more factors, called "predictors" or "covariates", which may be categorical (such as the kind of treatment a patient received) or continuous (such as the patient's age, weight, or the dosage of a drug). For simple situations involving a single factor with just two values (such as drug vs placebo), there are methods for comparing the survival curves for the two groups of subjects. But for more complicated situations we need a special kind of regression that lets us assess the effect of each predictor on the shape of the survival curve.

To understand the method of proportional hazards, first consider a "baseline" survival curve. This can be thought of as the survival curve of a hypothetical "completely average" subject -- someone for whom each predictor variable is equal to the average value of that variable for the entire set of subjects in the study. This baseline survival curve doesn't have to have any particular formula representation; it can have any shape whatever, as long as it starts at 1.0 at time 0 and descends steadily with increasing survival time.

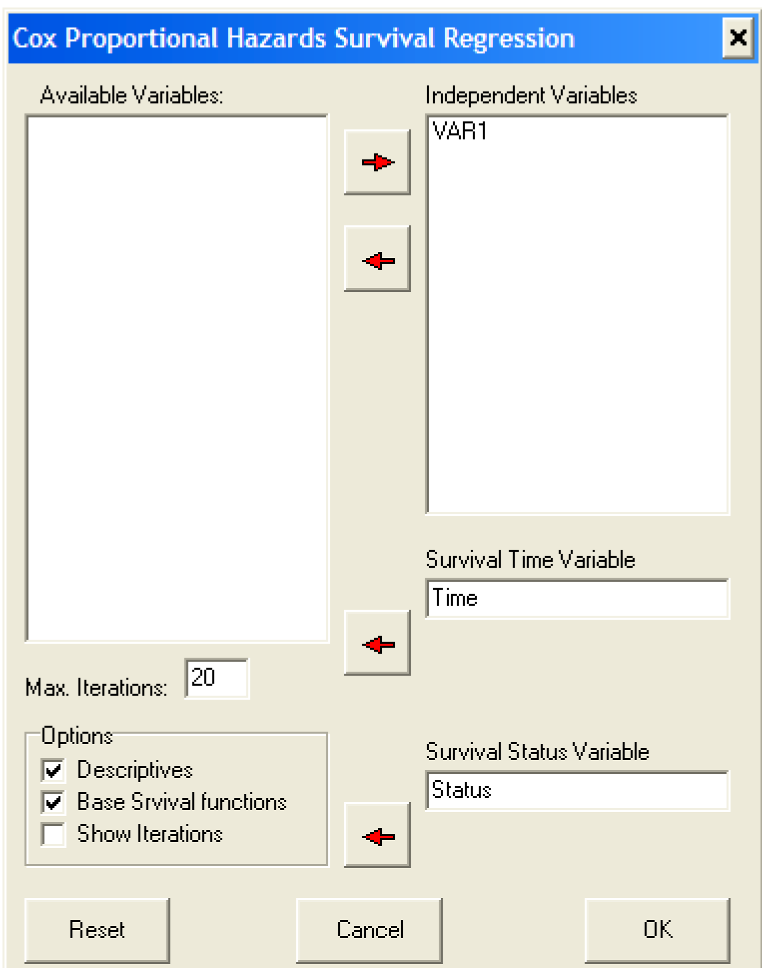
The baseline survival curve is then systematically "flexed" up or down by each of the predictor variables, while still keeping its general shape. The proportional hazards method computes a coefficient for each predictor variable that indicates the direction and degree of flexing that the predictor has on the survival curve. Zero means that a variable has no effect on the curve -- it is not a predictor at all; a positive variable indicates that larger values of the variable are associated with greater mortality. Knowing these coefficients, we could construct a "customized" survival curve for any particular combination of predictor values. More importantly, the method provides a measure of the sampling error associated with each predictor's coefficient. This lets us assess which variables' coefficients are significantly different from zero; that is: which variables are significantly related to survival.

The log-likelihood function is minimized by Newton's method, with a very simple elimination algorithm to invert and solve the simultaneous equations. Central-limit estimates of parameter standard errors are obtained from the diagonal terms of the inverse matrix. 95% confidence intervals around the parameter estimates are obtained by a normal approximation. Risk ratios (and their confidence limits) are computed as exponential functions of the parameters (and their confidence limits). The baseline survival function is generated for each time point at which an event (death) occurred.

No special convergence-acceleration techniques are used. For improved precision, the independent variables are temporarily converted to "standard scores" ( value - Mean ) / StdDev. The *Null Model* (all parameters = 0 )is used as the starting guess for the iterations. Convergence is not guaranteed, but this page should work properly with most real-world data.

There are no predefined limits to the number of variables or cases this page can handle. The actual limits are probably dependent on your computer's available memory.

The specification form for this analysis is shown below with variables entered for a sample file labeled COXREG.LAZ:




Results for the above sample are as follows:

Cox Proportional Hazards Survival Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/prophaz.html

```
Descriptive Statistics
Variable  Label            Average     Std.Dev.
     1            VAR1    51.1818    10.9778

Converged

Overall Model Fit...
Chi Square =   7.3570 with d.f. 1 and probability =   0.0067

Coefficients, Std Errs, Signif, and Confidence Intervals

Var             Coeff.    StdErr       p      Lo95%    Hi95%
      VAR1     0.3770     0.2542   0.1379  -0.1211   0.8752

Risk Ratios and Confidence Intervals

Variable      Risk Ratio   Lo95%     Hi95%
      VAR1     1.4580     0.8859     2.3993

Baseline Survivor Function (at predictor means)...
    2.0000     0.9979
    7.0000     0.9820
    9.0000     0.9525
   10.0000     0.8310
```