Sample Statistical Workflow:
Survival Analysis

Adam Davey

March 22, 2015

Cox proportional hazards is the workhorse model for time to event data. Interpretation shares similarities to logistic regression where the outcome is modeled across instantaneous times.

Adam

1 Preamble

Survival analysis models the association between one or more predictors and *time to event*, where not everyone may have experienced the event by the end of the study. These observations are said to be *"censored."* Survival analysis assumes a constant ratio in the *log odds* of an event occurring across all time periods. Other assumptions such as correct model specification, independence of observations, and that predictors are measured without error also apply. The objective of our analysis is to predict relative risk of an event over time as a function of the predictor(s). Because of their importance for medical and public health problems, these models have undergone a tremendous amount of development over the past 20 years or so.

2 Steps in a Survival Analysis Workflow

Below, we present a set of sample steps that should be performed whenever you are performing a survival analysis. It is not necessary for you to literally perform every one of these steps with every single analyses. However, this workflow should essentially cover the range of requirements and assumptions that you are expected to be responsible for under most circumstances. Know, however, that it is in no way exhaustive and that research conventions in your area of research may differ in important ways from the expectations for

biostatistics. Likewise, many areas of biostatistics are developing very quickly and so current conventions are subject to change.

2.1 Frequencies/Summary Statistics

Censoring is of critical importance in survival analysis, and survival data require special handling because of this form of informative missing data.

Prior to any analyses, you need to declare your data as survival-time. This means linking the time variable and the censoring variable (and its value if an observation is censored). In Stata, this requires a single line: `stset timevar, fail(censorvar==censorval)`. Note that many software packages code the *censored* category, but Stata codes the *observed value* category. Something else to note is that Stata excludes observations that experience an event in the very first interval (e.g., time=0). If you do not want these cases to be excluded, recode the 0s to a small value (e.g., 0.01).

You can obtain summary statistics for survival time data in Stata with the `stsum` and `stdescribe` commands, which can also be run by levels of categorical predictor variables, e.g., `stsum, by(x)`. These commands will provide incidence rates, quantiles, mean (if appropriate, i.e., if the longest observed case is not censored), and the proportion of (un)censored cases.

Life tables are very useful for examining the distribution of survival times. The syntax is quite simple in Stata: `ltable timevar censorvar, survival` and can also be used with the `by` option.

Examine the univariate distribution of your predictor(s) using frequencies and/or summary statistics (M, SD, Min, Max, etc.), as with other analyses we have considered.

2.2 Univariate plots (Y and X)

Survival analysis has some of the best plots for statistical analyses, and they are very useful to understand the risk of an event and how this risk changes over time. Plots for predictors are similar to the other analyses we have considered. Here we focus on plots for the time-to-event variable.

The simplest plot is the Kaplan-Meier survival plot (`sts graph, survival ci censored(number)`). Failure and hazard plots can be obtained using the `failure` and `hazard` options, respectively. The `ci` option may be omitted if you do not wish for confidence intervals. It is important to indicate censored observations. Instead of `number`, you can also enter `single` for a single hash mark indicating censoring or `multiple` for one hash mark per censored observation. Plots can also be estimated with the `by` option. A table of the number of observations at risk of the even can also be requested using the `risktable` option. Stata picks default values of time at which to display these values, but

you can also specify specific values in parentheses, e.g., `risktable(2 4 8 16)` or `risktable(0(5)15)`.

In terms of what you are looking for, consider the following. Is the last observation censored? If so, you cannot estimate mean survival time. This is equivalent to asking whether the survival function goes to zero. It is not necessary that the graph be smooth or a straight line. How the survival function changes over time will tell you something about how quickly or slowly individuals experience the event, as well as the proportion of individuals who experience the event during the course of the study.

There are a number of methods for bivariate comparisons of survival times across predictors. The basic Stata syntax is of the form `sts test x, logrank`. By default a log-rank test is performed (an extension of the Maentel-Haenszel test), but other options include the Wilcoxon (`wilcoxon`), Tarone-Ware (`tware`), Peto-Peto-Prentice (`peto`), and Fleming-Harrington (`fh(p q)`) which allows the user to choose weights for earlier and later failures (when $p = q$, it is the same as the log-rank test).

2.3 Testing Normality of Distributions

Normality of the response variable is not a concern for survival analysis.

2.4 Scatterplots

Scatterplots are also tricky for survival analysis, and should only be plotted when you can indicate censoring. Even so, they can be misleading because you will not know how far beyond a censoring point an event occurred. Proceed with caution. Sample syntax to accomplish this for survival data is `twoway (scatter timevar byvar, sort mlabel(censorvar))`.

2.5 Lowess

Just as we first transformed out response variable prior to estimating a lowess curve, we do something similar for survival analysis. First, we estimate Martingale residuals in an empty model (one with no covariates). Next, we plot a lowess curve of the Martingale residuals by each predictor to evaluate the functional form. In Stata, one line estimates the residuals (`stcox, mgale(mg0) efron estimate`) and a second plots the values (`lowess mg0 age`).

2.6 Estimate Model

Because you have already declared the survival time structure of the data, it is only necessary to specify the predictor(s) to estimate a Cox proportional hazards model. As with

simple linear regression, Stata allows model estimation using robust standard errors. In Stata, type `stcox x, vce(robust) efron`. In contrast to logistic regression, the hazard ratio (analogous to the odds ratio in logistic regression) is displayed by default. To obtain the unexponentiated coefficient, add the `nohr` option to the above syntax. Notice that in these models, no intercept (the baseline hazard rate) is displayed by default, since it is the *hazard ratio* that is of primary interest and the hazard itself is free to vary over time (and usually does in most applications).

Recall that tied observations require special treatment since the survival model assumes that time is measured with perfect precision and therefore no two observations should share exactly the same values but differ on covariate values. The Breslow method is the default; a better option is Efron's method (`efron`). There are also two (approximately) exact methods based on marginal-likelihood (`exactm`) and partial-likelihood(`exactp`) methods. These latter methods cannot be specified with robust standard errors. How much the approach for handling tied observations matters depends on the nature and number of tied observations, which is partly determined by the precision with which time-to-event is measured (e.g., months, weeks, days).

With one predictor, that's almost always all there is to model estimation. In the multivariable framework, we will considerably elaborate the steps and tools available to us.

2.7 Evaluate Model

Model evaluation for survival analysis has much in common with the procedures for logistic regression. The formal test that all regression coefficients are 0 is provided by a likelihood ratio χ^2 with regular standard errors and by a Wald χ^2 with robust standard errors. If the overall model χ^2 is not significant at our α, we stop. There is no association to interpret. Something similar to an R^2 statistic, called Herrell's C can be obtained with the `estat concordance` command. Next, the model will provide you with estimated hazard ratios, standard errors, and the corresponding z-statistics and their p-values. Keep the following in mind when evaluating your model.

1. Did your model converge? Survival analysis is an iterative procedure. You can examine model convergence by looking at the iteration history along with any warning messages. If you see messages like "Not concave" or "Backing up" this is an indication that there are difficulties with estimation of your model and you should proceed carefully. It is usually a good idea to change the tolerance (a function of accuracy, or the change in parameter estimates from one iteration to the next) of estimation to a smaller value to make it more likely that estimation has converged at a global maximum. Stata has many estimation options that you can use to "tweak" estimation including the ability to change the number of iterations, the estimation procedure, and the tolerance. More important is to try and identify the source of estimation difficulties in the first place, such as empty cells.
2. Were any combinations/characteristics excluded/dropped from the analysis? This is

an indication of empty cells that can cause problems for analyses and interpretation. When identified, they need to be diagnosed before proceeding.

3. Are the effects in the correct (expected) direction and of a reasonable magnitude? If your predictor is participation in the treatment group, does the direction of the effect go in the right direction? Is participation in the treatment associated with better outcomes? If the effect is not in the expected direction, there may be something wrong, such as with the coding of your variables. This is the time to check. Likewise, if there were problems with estimation, Stata will usually flag these by providing coefficients (often very large), but no standard error. These are indications of problems with your analyses.

4. Do the standard errors look reasonable? If you have a large sample size, you should expect very precise estimates on the sampling distribution of the regression coefficients. If the standard errors are very large, there may be something wrong with the model and estimation. This is something we will revisit again when considering multivariable models. On the other hand, very small standard errors can also be a sign of problems with variable coding and estimation. Beware of small hazard ratios associated with large z-values. Although most statistics packages pay very careful attention to numerical accuracy, computers use something called floating point arithmetic which can sometimes introduce a great deal of imprecision.

2.8 Survival Analysis Diagnostics

As with logistic regression, we can begin with a hat test (`linktest`), looking for a significant coefficient for `_hatsq` as evidence of model mis-specification.

Much of the Cox proportional hazards model rests upon – you guessed it! – proportionality of hazards. Stata includes several tests that can be used to evaluate this assumption. First, there is a formal statistical test for each predictor (in multivariable models) and an overall test. Request it with `estat phtest, detail`. Any significant predictors can be plotted using `estat phtest, plot(x)`. Two other visual displays include `stphplot, by(old)` which plots -log(-log(Pr(Survival) by log(time) and `stcoxkm, by(old)` which plots observed and predicted survival by time.

Evaluating goodness of fit (a measure of the discrepancy between observed and predicted values), is typically evaluated using a plot of Cox-Snell residuals. First, we estimate the model of interest, requesting Martingale residuals (`stcox x, efron mgale(mg1)`). (Note that our Martingale residuals from the empty model were labeled as mg0.) Next, we obtain the predicted Cox-Snell residuals (`predict cs, csnell`) and cumulative hazard function (`sts gen chazfun = na`). We then plot the Cox-Snell residuals and cumulative hazard function by the Cox-Snell residuals (which produces a 45 degree line).

Outliers and influential points also need to be considered in survival models. To begin, estimate the model of interest saving "efficient score residuals" which we will transform into dfbetas (`stcox x, efron esr(e*)` – this syntax will be set up to with with one or many

predictors). We convert these variables into matrix form (`mkmat q*, matrix(q)`), save the estimated variance covariance matrix of the corresponding regression coefficients (`mat V = e(V)`), and then multiply them together (`mat Inf = esr*V`) and save the results as variables in our data set (`svmat Inf, names(s)`) which can be plotted (`scatter s1 _t, yline(0) mlabel(id) msymbol(i)`).

2.9 Tabling Results

Different disciplines have different expectations for how results from survival analysis models should be presented. In Public Health, critical information typically includes unexponentiated estimated regression coefficients, upper and lower confidence limits, and a p-value to 4 decimals. Almost everyone relies exclusively on hazard ratios for interpretations (along with great plots to see exactly what is happening in terms of time-to-even) along with their upper and lower confidence limits, z-values, and associated p-values. Tabling results is essentially the same as for logistic regression output. See the Stata do-file for syntax to export results into an Excel file.

Table 1: Sample Cox Proportional Hazards Model Output in Stata

Variable	**Coefficient**	**(Std. Err.)**
age	0.047	(0.004)

2.10 Writing Up Results

Prior to analysis, distribution of survival times was examined using Kaplan-Meier estimates and life-tables. Lowess plots of Martingale residuals were used to evaluate the functional form of the association between each predictor and survival time. Overall, 37.6% of cases were censored at the last time of observation. Median survival time was 4 months (mean 3.49 months), and the overall incidence rate was 0.18. A Cox proportional hazards model was used to evaluate the association between age and mortality risk using the Efron method for tied observations. The overall model was significant (LR $\chi^2(1) = 120.74, p < .0001$ and the model did an adequate job of predicting observed mortality order (Harrell's C = 0.63). Older age was associated with greater hazard of mortality
($HR = 1.05, 95\%CI[1.04 - 1.06], z = 10.94, p < .0001$) suggesting each additional 5 years of age [age is rounded in 5-year bins] is associated with a 5% higher mortality risk. No significant evidence of model mis-specifications or violations of assumptions were observed according to a hat test, graphical and statistical evaluation of the proportional hazards assumption, and outliers and influential points.

Checklist: Survival Analysis

☐ Survival time descriptive statistics and life tables for time-to-event variable. Frequencies and summary statistics for predictors. What proportion of cases is censored? Longest-observed value censored?

☐ Kaplan-Meier plots of survival, failure, and hazard. Log-rank or related tests for bivariate associations with survival time.

☐ Lowess plot of Martingale residuals by predictor. Scatterplots only if informative and indicate censored observations.

☐ Estimate model using Efron method for tied data and robust standard errors as appropriate.

☐ Evaluate model. Proceed if χ^2 test significant; otherwise stop. Verify that model converged, coefficient is in the expected direction and that hazard ratios and standard errors appear to have been estimated appropriately. What is Harrell's C statistic?

☐ Evaluate regression diagnostics. Perform hat test; evaluate proportional hazards assumption graphically and statistically; examine outliers and influential points.

☐ Table results. Results should include hazard ratios, standard errors, z-values, and p-values, plus number of observations used in analysis and model χ^2. Format your results in accordance with the style and conventions for your area of research.

☐ Write up results; note any violations of assumptions and their effects, if any, on results. Refer to model table.

Stata Syntax: Survival Analysis

```
/*****************************************************************
* Sample Cox Proportional Hazards Regression Syntax
* PH 8012, Spring 2015
* Adam Davey
* Requires outreg2
* Type: findit outreg2 to install
*****************************************************************/

#delimit;
clear all;
capture log close;
log using "mysurvival.log", replace;

* Below simulates some data for the workflow;
set seed 83633;
local lambdat = 0.2;
local gammat = 2;
local lambdac = 0.4;
local gammac = -1;
set obs 1000;
gen id = _n;
gen age = int(rnormal(50,10)) + 1;
gen time = int(10*((log(1-uniform()))/-‘lambdat’*exp(-.05*age))^(1/‘gammat’));
gen censor = int(10*((log(1-uniform()))/-‘lambdac’)^(1/‘gammac’));
gen died = censor >= time;
replace time = min(time,censor);
drop censor;
replace age = round(age,5);
gen old = age>=65;

* Step 2.1 Frequencies / Summary Statistics;
* Recode 0 survival times to small value;
replace time = 0.01 if time==0;

* Declaring survival time data;
stset time, fail(died==1);
stsum;
stdescribe;

tab age, missing;
summarize age;

* Step 2.2 Univariate Plots;
```

```
histogram age, by(died);
graph export "hist_age_died.pdf", replace;
graph box age, by(died);
graph export "box_age_died.pdf", replace;
qnorm age if died==0;
graph export "qnorm_age_died0.pdf", replace;
qnorm age if died==1;
graph export "qnorm_age_died1.pdf", replace;

sts graph, survival ci censored(number);
graph export "km_survival.pdf", replace;
sts graph, failure ci censored(number);
graph export "km_failure.pdf", replace;
sts graph, hazard ci;
graph export "km_hazard.pdf", replace;

* Step 2.3 Testing Normality of Distributions;
* NA;

* Step 2.4 Scatterplot;
* Totally useless, right?;
twoway (scatter time age, sort mlabel(died));
graph export "scatter_time_age.pdf", replace;

* Step 2.5 Lowess;
lowess died time, logit;
graph export "lowess_died_time.pdf", replace;
* Martingale Residuals;
* Estimate model without predictors;
stcox, mgale(mg0) efron estimate;
* Use them in place of Lowess plot above;
lowess mg0 age;
graph export "lowess_martingales.pdf", replace;


* Step 2.6 Estimate Model;
* First, estimate and store model with normal theory SEs;
stcox age;
stcox age, nohr;
stcox age, vce(robust);
stcox age, breslow;
stcox age, efron;
* Save the preferred model;
estimates store coxph;
stcox age, exactm;
```

```
stcox age, exactp;
* Restore estimates from default model;
estimates restore coxph;
* Try the model below;
*stcox i.age, efron;
estat concordance;

* Step 2.7 Evaluate Model;
* Please see workflow text for what to look for;
* Here's a margins plot;
margins, at(age=(20(5)80));
marginsplot, xlabel(20(5)80) recast(line) recastci(rarea);
graph export "margins_age.pdf", replace;

* Step 2.8 Regression Diagnostics;
* Specification;
linktest;

* Proportional Hazard tests and plots;
estat phtest, detail;
estat phtest, plot(age);
graph export "schoenfeld_age.pdf", replace;
stphplot, by(old);
graph export "phtest_old.pdf", replace;
stcoxkm, by(old);
graph export "coxkm_old.pdf", replace;

* Cox and Snell GOF;
stcox age, efron;
predict cs, csnell;
stset cs, fail(died==1);
sts gen cumhaz = na;
line cumhaz cs cs, sort ytitle("") legend(cols(1));
graph export "cox-snell_gof.pdf", replace;

* Outliers and Influential Points;
stcox age, efron esr(e*);
mkmat e*, matrix(esr);
mat V = e(V);
mat Inf = esr*V;
svmat Inf, names(s);
scatter s1 _t, yline(0) mlabel(id) msymbol(i);
graph export "inf_age.pdf", replace;

* Step 2.9 Tabling Results;
```

```
* Example 1 -- direct to Excel;
* This is different from other workflows;
* We have to reset and reestimate everything here before tabling;
stset time, fail(died==1);
stsum;
quietly: stcox age, efron;
putexcel set "survival.xls", sheet("Cox Proportional Hazards") replace;
putexcel F1=("Number of obs")          G1=(e(N));
local chi2type = e(chi2type);
putexcel F2=("`chi2type' Chi-square")  G2=(e(chi2));
putexcel F3=("df")                     G3=(e(df_m));
putexcel F4=("Prob > Chi-square")      G4=(max(0.0001,chi2tail(e(df_m),e(chi2))));
matrix a = r(table)';
matrix a = a[.,1..6];
putexcel A6=matrix(a, names);
quietly: estat concordance;
putexcel F5=("Harrell's C")       G5=(r(C));

log close;

* Convert text output to PDF;
translate mysurvival.log mysurvival.pdf, replace;
```