

## Cox Proportional Hazard Model

### Purpose:

1. Introduction to the **Cox proportional hazard model**
2. How to fit PHM using SAS procedure PHREG

### 1. Introduction to the Cox proportional hazard model

Unlike parametric methods, Cox's method does not require that you choose some particular probability distribution to represent survival times. That's why we call it semiparametric model, which makes Cox's method more robust. Another advantage of using Cox's method is that it's relatively easy to incorporate time-dependent covariates. As we know, we use the method of maximum likelihood to estimate the regression parameter in parametric model, while we use the method of *maximum partial likelihood* to estimate the parameters in Cox's model. What's remarkable about partial likelihood is that you can estimate the coefficients without having to specify the baseline hazard function  $\lambda_0(t)$ .

The Cox's model is represented as:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_1 x_{i1} + \dots + \beta_p x_{ip}\} \quad (1)$$

where  $\lambda_i(t)$  is the hazard function for the  $i$ th subject,  $\lambda_0(t)$  is the baseline hazard function, and  $\beta_1, \dots, \beta_p$  are parameters to be estimated.

### 2. How to fit PHM using SAS procedure LIFEREG

#### 2.1 Introduction

**Data description:** The **recid** data set consists of 432 males inmates who were released from Maryland state prisons in the early 1970s (Rossi et al. 1980). These men were followed for one year after their release, and the dates of any arrests were recorded. We'll only look at the first arrest here. The **WEEK** variable contains the week of the first arrest after release. The variable **ARREST** has a value of 1 for those arrested during the one-year follow-up, and it has a value of 0 for those who were not. Only 26% of the men were arrested. The data are right-censored (type I) so that all the censored cases have a value of 52 for **WEEK**. The other covariates are

**FIN:** 1 if the inmate received financial aid after release; otherwise, 0.

**AGE:** age in years at the time of release.

**RACE:** 1 for black; 0 for otherwise.

**WEXP:** 1 if the inmate had full-time work experience before incarceration; 0 otherwise.

**MAR:** 1 if the inmate was married at the time of release; 0 otherwise.

**PARO:** 1 if the inmate was released on parole; 0 otherwise.

**PRIO:** number of convictions an inmate had prior to incarceration.

One of interests of survival analysis is to understand the relationship between time to failure and other covariates measured at the studied subjects.

Let's apply PHM on recidivism data. The syntax of PROC PHREG is almost identical to that for PROC LIFEREG, except that you do not need to specify a distribution:

```
PROC PHREG DATA=recid;
  MODEL week*arrest(0)=fin age race wexp mar paro prio;
RUN;
```

The output is:

The PHREG Procedure			
Model Information			
Data Set	WORK.RECID		
Dependent Variable	week		
Censoring Variable	arrest		
Censoring Value(s)	0		
Ties Handling	BRESLOW		
Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
432	114	318	73.61
Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			
Model Fit Statistics			
Criterion	Without Covariates	With Covariates	
-2 LOG L	1351.367	1318.241	
AIC	1351.367	1332.241	
SBC	1351.367	1351.395	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.1256	7	<.0001
Score	33.3828	7	<.0001
Wald	31.9875	7	<.0001

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.37902	0.19136	3.9228	0.0476	0.685
age	1	-0.05724	0.02198	6.7798	0.0092	0.944
race	1	0.31415	0.30802	1.0402	0.3078	1.369
wexp	1	-0.15113	0.21212	0.5076	0.4762	0.860
mar	1	-0.43280	0.38180	1.2850	0.2570	0.649
paro	1	-0.08497	0.19575	0.1884	0.6642	0.919
prio	1	0.09114	0.02863	10.1331	0.0015	1.095

- The model information is almost the same as that of PROC LIFEREG, while the line TIES HANDLING: BRESLOW says the default method for handling ties is Breslow's method. Other methods include the exact method, the discrete method and the Efron's method.
- The information on Testing Global Null Hypothesis ( $H_0: \beta_1 = \dots = \beta_p = 0$ ) gives the result of the 3 tests, (partial) likelihood ratio test, score test and Wald test. For the recidivism data, all 3 tests lead to very small p-values. We can conclude that at least one of the coefficients is not 0.
- The main part of this output gives the coefficient estimates and associated estimated standard errors, chi-square statistics and p-values of Wald test ( $H_0: \beta_j = 0$ ). For example, the estimated hazard ratio for FIN is 0.685, which means that the hazard of arrest for those who received financial aid is about 69% of the hazard for those who did not receive aid. Comparing the coefficients with those in exponential model, we find that the numbers are very close, but the signs are reverse.

## 2.2 Time-dependent variable

In the recidivism data, for each of the 52 weeks of follow-up, there was a dummy variable coded 1 if the person was employed full-time during that week; otherwise it was coded 0. The 52 variables, EMP1, EMP2, ..., EMP52, are stored in separate variables. We want to combine them into one time-dependent variable, EMPLOYED, such that we can pick out the employment indicator corresponding to the particular week in which an event occurred and assign that value to the variable EMPLOYED. The SAS code is as follows:

```
PROC PHREG DATA=recid;
  MODEL week*arrest(0)=fin age race wexp mar paro prio employed
    / TIES=efron;
  ARRAY emp{*} emp1-emp52;
  employed=emp{week};
RUN;
```

The ARRAY statement makes it possible to treat the 52 variables as a single subscripted array and the line `employed=emp{week}` pick out the employment indicator corresponding to the particular week in which an event occurred and assign that value to

the variable EMPLOYED. Note that the lines defining time-dependent variables always include the time variable (in this case the time variable is WEEK).

```

The PHREG Procedure

Model Information

Data Set          WORK.RECID
Dependent Variable  week
Censoring Variable  arrest
Censoring Value(s)  0
Ties Handling       EFRON

Summary of the Number of Event and Censored Values

Total      Event      Censored      Percent
432        114         318           73.61

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion          Without          With
                   Covariates       Covariates
-2 LOG L           1350.761         1282.110
AIC                 1350.761         1298.110
SBC                 1350.761         1319.999

Testing Global Null Hypothesis: BETA=0

Test              Chi-Square      DF      Pr > ChiSq
Likelihood Ratio   68.6514         8       <.0001
Score              64.4845         8       <.0001
Wald               56.1527         8       <.0001

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable  DF      Parameter Estimate      Standard Error      Chi-Square      Pr > ChiSq      Hazard Ratio
fin       1      -0.35672                0.19113              3.4835           0.0620           0.700
age       1      -0.04633                0.02174              4.5442           0.0330           0.955
race      1       0.33867                0.30960              1.1966           0.2740           1.403
wexp      1      -0.02557                0.21142              0.0146           0.9037           0.975
mar       1      -0.29374                0.38303              0.5881           0.4432           0.745
paro      1      -0.06420                0.19468              0.1088           0.7416           0.938
prio      1       0.08515                0.02896              8.6455           0.0033           1.089
employed  1      -1.32823                0.25071             28.0679          <.0001           0.265
    
```

For the time-independent variables, the coefficients and test statistics are pretty much the same as the previous model, however the time-dependent variable EMPLOYED has the strongest effect. But we notice that sometime arrests affect employment status rather than vice versa. If someone is arrested in a week, the probability of working full time during

that week is very small. This potential reverse causation is a common problem in time-dependent variables. Let's modify the previous model a little bit:

```
PROC PHREG DATA=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;
```

We add a statement `WHERE week>1` to eliminate the observations with an arrest in the first week after release because there were no values of employment status prior to the first week. We also change the subscript of EMP to `week-1`.

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.35129	0.19181	3.3541	0.0670	0.704
age	1	-0.04977	0.02189	5.1697	0.0230	0.951
race	1	0.32149	0.30912	1.0816	0.2983	1.379
wexp	1	-0.04765	0.21323	0.0499	0.8232	0.953
mar	1	-0.34477	0.38322	0.8094	0.3683	0.708
paro	1	-0.04709	0.19630	0.0576	0.8104	0.954
prio	1	0.09201	0.02880	10.2085	0.0014	1.096
employed	1	-0.78689	0.21808	13.0195	0.0003	0.455

The magnitude of the coefficient for EMPLOYED changed from -1.33 to -0.79, but it's still statistically significant.

Another type of time-dependent variables is the interaction of a variable with time. In the recidivism data, suppose we are interested in if the effect of financial aid changes steadily with time. We can add the interaction term of  $FIN \times WEEK$  into the model:

```
PROC PHREG DATA=recid;
  MODEL week*arrest(0)=fin age race wexp mar paro prio fintime
    / TIES=efron;
  fintime=fin*week;
RUN;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.34350	0.42503	0.6532	0.4190	0.709
age	1	-0.05742	0.02200	6.8119	0.0091	0.944
race	1	0.31394	0.30799	1.0390	0.3081	1.369
wexp	1	-0.14967	0.21224	0.4973	0.4807	0.861
mar	1	-0.43381	0.38189	1.2904	0.2560	0.648
paro	1	-0.08496	0.19576	0.1883	0.6643	0.919
prio	1	0.09163	0.02868	10.2082	0.0014	1.096
fintime	1	-0.00125	0.01321	0.0089	0.9247	0.999

The interaction term is not significant, then we can say there is no evidence here that the effect of financial aid varies steadily with time.

### 2.3 Non-proportional hazards

Actually, all the models in section 2.2 are non-proportional hazard models. Because the time-dependent variables vary with time and the ratio of hazard is function of all the variables, hence is function of time. That means the hazards are not proportional to each other for two subjects. When there are time-dependent variables in the model, it is not appropriate to call it Proportional Hazard Model anymore. We can just call it Cox's model generally. Think this problem inversely, we actually gave a way to test the violation of proportional hazard assumption in section 2.2: For the suspected covariate, simply add a time-interaction term into the model. If the interaction variable is not significant, we say the assumption is not violated, while if it's significant we say the PH assumption is violated. Of course the new model incorporates the nonproportionality, i.e., the diagnosis is also the cure.

Another approach to nonproportionality is stratification, a method that is useful when the variable interacting with time is both categorical and not of direct interest. For example, for recidivism data, we can stratify the data by RACE:

```
PROC PHREG DATA=recid;
  WHERE week>1;
  MODEL week*arrest(0)=fin age wexp mar paro prio employed
    / TIES=efron;
  ARRAY emp{*} emp1-emp52;
  employed=emp{week-1};
  STRATA race;
RUN;
```

Note that there is no coefficient estimated for the strata variable. The partial likelihood for the stratified data is the product of the partial likelihood for each stratum.

Summary of the Number of Event and Censored Values						
Stratum	race	Total	Event	Censored	Percent Censored	
1	0	53	12	41	77.36	
2	1	378	101	277	73.28	
-----						
Total		431	113	318	73.78	
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
fin	1	-0.34589	0.19188	3.2497	0.0714	0.708
age	1	-0.05025	0.02192	5.2552	0.0219	0.951
wexp	1	-0.04087	0.21379	0.0366	0.8484	0.960
mar	1	-0.36337	0.38359	0.8974	0.3435	0.695
paro	1	-0.04703	0.19639	0.0573	0.8107	0.954
prio	1	0.09317	0.02887	10.4149	0.0013	1.098
employed	1	-0.78660	0.21784	13.0384	0.0003	0.455

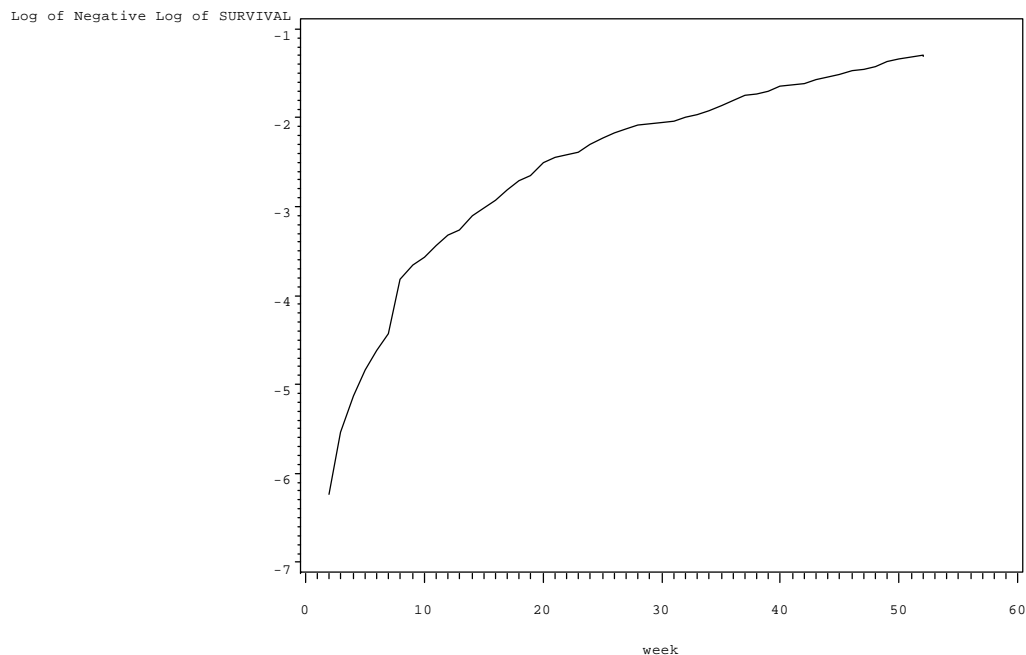
## 2.4 Estimating survivor functions

As we know, the form of the dependence of the hazard on time is left unspecified in the Cox's model. Hence we can not fully specify the survival function. But we can still get nonparametric estimates of the survival function based on a fitted proportional hazards model. We can accomplish this by adding a BASELINE statement. Let give an example using a simple model:

```
PROC PHREG DATA=recid;
  WHERE week>1;
  MODEL week*arrest(0)=fin age race wexp mar paro prio / TIES=efron;
  BASELINE OUT=a SURVIVAL=s LOGSURV=ls LOGLOGS=lls;
RUN;

PROC GPLOT DATA=a;
  SYMBOL1 VALUE=none interpol=join;
  PLOT llS*week;
RUN;
```

The `OUT=a` option in the Baseline statement asks SAS to put the survival estimates (for the subjects whose covariates are all equal to the mean of each variable) in a new data set named `a`. `SURVIVAL=s` asks that the survival probabilities be stored in a variable `s` in the data set `a`, `LOGSURV=ls` asks to save the log survival in variable `ls`, `LOGLOGS=lls` for  $\log[-\log S(t)]$ . I give the log-log survivor plot for the fitted model. You can also plot the survival function and the cumulative hazard function (negative log survival function) by yourself.



#### REFERENCE:

P. D. Allison, Survival analysis using the SAS system: a practice guide, SAS company, 1995.