

Accelerated Failure Time Model (II)

Purpose:

1. Introduction to the **accelerated failure time model**
2. Estimating parametric regression models with PROC LIFEREG in SAS

Data description: The data 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.

Let's open the data set in SAS.

```
DATA recid;  
    SET "C:\140.641\recid";  
RUN;
```

One of interests of survival analysis is to understand the relationship between time to failure and other covariates measured at the studied subjects. This can be done by using regression models. The regression model we will talk about in this lab is a parametric model. The LIFEREG procedure produces estimates of parametric regression models with censored survival data using the method of maximum likelihood. In recent years parametric model has been eclipsed by semiparametric regression model (Cox's model), which uses a method known as partial likelihood. The reasons for semi-parametric model's popularity will become apparent in the next several labs.

The class of regression models estimated by PROC LIFEREG is known as the accelerated failure time (AFT) model. We have introduced the most general form of AFT model in the last lab. What PROC LIFEREG actually estimates is a special case. Let T_i be a random variable denoting the failure time for the i th subject, and let $x_{i1}, x_{i2}, \dots, x_{ip}$ be the values of p covariates for that same subject. The model is then

$$\log T_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \sigma \varepsilon_i \quad (1)$$

where ε_i is a random disturbance term, and β_0, \dots, β_p , and σ are parameters to be estimated.

Note that the only differences between the model in (1) and the usual linear regression models are that there is a σ before ε_i the and that the dependent variable is logged. The σ can be omitted, which requires that the variance of ε_i be allowed to be different from 1. But it is simpler to fix the variance of ε_i at 1 and let σ change. This notational strategy could be used for linear regression models. As for the log transformation of T , its main purpose is to ensure that predicted values of T are positive.

If there are no censored data, we can readily estimate this model by ordinary least squares. Simply generate a new variable, $Y = \log T$, and use the linear regression model with Y as the dependent variable. This process yields the best linear unbiased estimates of coefficients, without distribution assumption on ε . If ε is normal, the OLS estimates will also be maximum likelihood estimates and will have minimum variance among all estimators, both linear and nonlinear.

But survival data typically have at least some censored observations, and these are difficult to handle with OLS. Alternatively, we can use MLE with different distribution assumption on ε . For each of the distribution of ε , there is a corresponding distribution for T .

| Distribution of ε | Distribution of T |
|-------------------------------|---------------------|
| Extreme value (2 parameters) | Weibull |
| Extreme value (1 parameter) | Exponential |
| Log-gamma | Gamma |
| Logistic | Log-logistic |
| Normal | Log-normal |

Note that all AFT models are named for the distribution of T rather than the distribution of ε or $\log T$. The reason for allowing different distribution assumptions is that they have different implications for the shapes of hazard function. We will briefly introduce 3 models, log-normal, exponential, and Weibull models in this lab.

1. The log-normal model

To estimate the log-normal model, we specify

```
PROC LIFEREG data=recid;
  MODEL week*arrest(0)=fin age race wexp mar paro prio
    / dist=lnormal;
RUN;
```

As we mentioned, if there is no censored data this model will give exactly the same estimates as in linear regression model. For this data, the results are as follows:

| The LIFEREG Procedure | | | | | | |
|---------------------------------|----|----------|----------------|------------|--------|--------------|
| Model Information | | | | | | |
| Data Set | | | | | | WORK.RECID |
| Dependent Variable | | | | | | Log(week) |
| Censoring Variable | | | | | | arrest |
| Censoring Value(s) | | | | | | 0 |
| Number of Observations | | | | | | 432 |
| Noncensored Values | | | | | | 114 |
| Right Censored Values | | | | | | 318 |
| Left Censored Values | | | | | | 0 |
| Interval Censored Values | | | | | | 0 |
| Name of Distribution | | | | | | LNORMAL |
| Log Likelihood | | | | | | -322.6945851 |
| Algorithm converged. | | | | | | |
| Analysis of Parameter Estimates | | | | | | |
| Variable | DF | Estimate | Standard Error | Chi-Square | Pr > | ChiSq Label |
| Intercept | 1 | 4.26767 | 0.46169 | 85.4438 | <.0001 | Intercept |
| fin | 1 | 0.34285 | 0.16409 | 4.3657 | 0.0367 | |
| age | 1 | 0.02720 | 0.01576 | 2.9806 | 0.0843 | |
| race | 1 | -0.36316 | 0.26469 | 1.8824 | 0.1701 | |
| wexp | 1 | 0.26813 | 0.17889 | 2.2466 | 0.1339 | |
| mar | 1 | 0.46035 | 0.29515 | 2.4328 | 0.1188 | |
| paro | 1 | 0.05588 | 0.16911 | 0.1092 | 0.7411 | |
| prio | 1 | -0.06552 | 0.02709 | 5.8489 | 0.0156 | |
| Scale | 1 | 1.29457 | 0.09895 | | | Normal scale |

e^β has the interpretation of the estimated ratio of the expected survival times for two groups. The “scale” is an estimate of the σ of equation (1). For some distributions, changes in the value of this parameter can produce qualitative difference in the shape of the hazard function. For the log-normal model, however, changes in σ merely compress or stretch the hazard function.

2. The exponential model

This model specifies that ε has a standard extreme-value distribution, and constrains $\sigma = 1$. Under exponential assumption, equation (1) is equivalent to

$$\log \lambda_i(t) = \beta_0^* + \beta_1^* x_{i1} + \dots + \beta_p^* x_{ip} \quad (2)$$

where $\beta_j^* = -\beta_j$ for all j .

The change in signs makes intuitive sense. If the hazard is high, then events occur quickly and survival times are short. Let's fit this model using PROC LIFEREG

```
PROC LIFEREG data=a;
  MODEL week*arrest(0)=fin age race wexp mar paro prio
    / dist=exponential;
RUN;
```

| The LIFEREG Procedure | | | | | | |
|---------------------------------|----|------------|----------------|------------|------------|---------------------|
| Model Information | | | | | | |
| Data Set | | | | | | WORK.RECID |
| Dependent Variable | | | | | | Log(week) |
| Censoring Variable | | | | | | arrest |
| Censoring Value(s) | | | | | | 0 |
| Number of Observations | | | | | | 432 |
| Noncensored Values | | | | | | 114 |
| Right Censored Values | | | | | | 318 |
| Left Censored Values | | | | | | 0 |
| Interval Censored Values | | | | | | 0 |
| Name of Distribution | | | | | | EXPONENT |
| Log Likelihood | | | | | | -325.8259007 |
| Algorithm converged. | | | | | | |
| Analysis of Parameter Estimates | | | | | | |
| Variable | DF | Estimate | Standard Error | Chi-Square | Pr > ChiSq | Label |
| Intercept | 1 | 4.05069 | 0.58604 | 47.7754 | <.0001 | Intercept |
| fin | 1 | 0.36626 | 0.19112 | 3.6728 | 0.0553 | |
| age | 1 | 0.05560 | 0.02184 | 6.4798 | 0.0109 | |
| race | 1 | -0.30494 | 0.30794 | 0.9806 | 0.3220 | |
| wexp | 1 | 0.14675 | 0.21170 | 0.4805 | 0.4882 | |
| mar | 1 | 0.42699 | 0.38138 | 1.2535 | 0.2629 | |
| paro | 1 | 0.08265 | 0.19560 | 0.1785 | 0.6726 | |
| prio | 1 | -0.08566 | 0.02831 | 9.1531 | 0.0025 | |
| Scale | 0 | 1.00000 | 0 | | | Extreme value scale |
| Lagrange Multiplier Statistics | | | | | | |
| Variable | | Chi-Square | | Pr > ChiSq | | |
| Scale | | 24.9302 | | <.0001 | | |

The coefficient for AGE is twice as large in the exponential model, and its p-value declines from .08 to .01. The coefficient for PRIO increases somewhat in magnitude, and its p-value also goes down substantially. The p-value for FIN increases to slightly above the .05 level. The SCALE parameter σ is forced equal to 1.0. The last line is a 1 df test for $H_0 : \sigma = 1$. Here the null hypothesis is rejected, indicating that the hazard function is not constant over time. It implies that we might need some more complex distribution assumption in terms of the shape of hazard.

3. The Weibull model

The Weibull model is a slight modification of the exponential model. We retain the assumption that ε has a standard extreme-value distribution, but we relax the assumption that $\sigma = 1$. When $\sigma > 1$, the hazard decreases with time. When $0.5 < \sigma < 1$, the hazard is increasing at a decreasing rate. When $0 < \sigma < 0.5$, the hazard is increasing at an increasing rate. And when $\sigma = 0.5$, the hazard function is an increasing straight line with an origin at 0. Under Weibull assumption, equation (1) is equivalent to

$$\log \lambda_i(t) = \alpha \log t + \beta_0^* + \beta_1^* x_{i1} + \dots + \beta_p^* x_{ip} \tag{3}$$

where $\beta_j^* = -\beta_j/\sigma$ for all j and $\alpha = 1/\sigma - 1$.

```
PROC LIFEREG data=recid;
  MODEL week*arrest(0)=fin age race wexp mar paro prio
    / dist=weibull;
RUN;
```

| The LIFEREG Procedure | | | | | | |
|---------------------------------|----|----------|----------------|------------|--------|--------------|
| Model Information | | | | | | |
| Data Set | | | | | | WORK.RECID |
| Dependent Variable | | | | | | Log(week) |
| Censoring Variable | | | | | | arrest |
| Censoring Value(s) | | | | | | 0 |
| Number of Observations | | | | | | 432 |
| Noncensored Values | | | | | | 114 |
| Right Censored Values | | | | | | 318 |
| Left Censored Values | | | | | | 0 |
| Interval Censored Values | | | | | | 0 |
| Name of Distribution | | | | | | WEIBULL |
| Log Likelihood | | | | | | -319.3765238 |
| Algorithm converged. | | | | | | |
| Analysis of Parameter Estimates | | | | | | |
| Variable | DF | Estimate | Standard Error | Chi-Square | Pr > | ChiSq Label |
| Intercept | 1 | 3.99013 | 0.41910 | 90.6462 | <.0001 | Intercept |
| fin | 1 | 0.27216 | 0.13796 | 3.8917 | 0.0485 | |
| age | 1 | 0.04071 | 0.01600 | 6.4722 | 0.0110 | |
| race | 1 | -0.22480 | 0.22016 | 1.0426 | 0.3072 | |
| wexp | 1 | 0.10656 | 0.15154 | 0.4944 | 0.4820 | |
| mar | 1 | 0.31127 | 0.27330 | 1.2972 | 0.2547 | |
| paro | 1 | 0.05883 | 0.13964 | 0.1775 | 0.6735 | |

| | | | | | | |
|-------|---|----------|---------|--------|--------|---------------------|
| prio | 1 | -0.06582 | 0.02094 | 9.8787 | 0.0017 | |
| Scale | 1 | 0.71241 | 0.06342 | | | Extreme value scale |

Compared with the exponential model, the coefficients are all somewhat attenuated. But the standard errors are also smaller, so the chi-square statistics and p-values are hardly affected at all. If we convert the coefficients to the format in equation (3) by changing sign and dividing by the estimate of σ , we get coefficients much closer to the coefficients in exponential model. Since the SCALE, estimate of σ , is between 0 and 1, we conclude that the hazard is increasing at a decreasing rate.

Although we have already discussed the log-normal model and applied to the recidivism data, we have not yet considered the shape of its hazard function. Unlike the Weibull model, the log-normal model has a nonmonotonic hazard function. The hazard is 0 when $t = 0$, and increase to a peak and then declines to 0 as t goes to infinity. The log-normal model can not be expressed as a proportional-hazard-like form as exponential or Weibull model. As we will see later that log-normal model is not nested in exponential or Weibull models.

4. Model selection

We have seen that the AFT model encompasses a number of submodels that differ in the assumed distribution for T . Clearly, we need some way of deciding between the models, i.e., the shapes of hazard. Here we will introduce the likelihood-ratio test for comparing nested models. A model is said to be nested within another model if the first model is a special case of the second. More precisely, model A is nested within model B if A can be obtained by imposing restrictions on the parameters in B. For example, the exponential model is nested within both the Weibull and the standard gamma models.

Let's calculate the likelihood ratio test for the recidivism data.

H_0 : exponential model

H_1 : Weibull model

The log-likelihood for exponential model is -325.83 and the log-likelihood for Weibull model is -319.38. The likelihood-ratio Chi-square statistic is $-2(-325.83 - (-319.38)) = 12.9$ with 1 degree of freedom. Clearly, we can reject the null hypotheses, that is to say Weibull model fits the data better than exponential model. As we know, the generalized gamma model is more general than Weibull, so we can further fit generalized gamma model on our data and test if Weibull model is enough.

Other techniques to discriminate between different models include the graphic methods, which can be found in reference books.

REFERENCE:

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