

**Survival Models in SAS**  
**Part 4: PROC LIFEREG -**  
**Part 2**

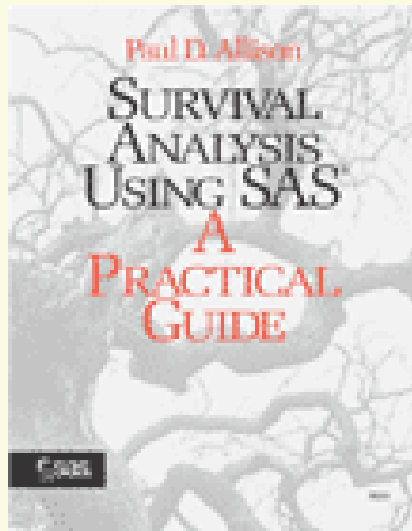
January 16, 2008

Charlie Hallahan

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG

These talks are based on the book “**Survival Analysis Using the SAS System: A Practical Guide**” (1995) by Paul Allison.

The book is part of the SAS Books-by-Users series and can be found at <http://www.sas.com/apps/pubscat/bookdetails.jsp?catid=1&pc=55233>



## **Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG**

This series of talks will cover

Chapter 1: Introduction

Chapter 2: Basic Concepts of Survival Analysis

Chapter 3: Estimating and Comparing Survival Curves with PROC LIFETEST

**Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG**

Chapter 5: Estimating Cox Regression Models with PROC PHREG

Chapter 6: Competing Risks

## **Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG**

### **Topics in Chapter 4:**

*Introduction*

*The Accelerated Failure Time Model*

*Alternative Distributions*

*Categorical Variables and the CLASS Statement*

*Maximum Likelihood Estimation*

*Hypothesis tests*

*Goodness-of-Fit Tests with the Likelihood-Ratio Statistic*

*Graphical Methods of Evaluating Model Fit*

*Left Censoring and Interval Censoring*

*The Piecewise Exponential Model*

*Generating Predictions and Hazard Functions*

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Categorical Variables and the CLASS Statement

The SAS **CLASS** statement automatically creates dummy variables associated with categorical variables. Otherwise, the user must create the dummy variables in a DATA step.

**LIFEREG** has a **CLASS** statement, while **PHREG** does not.

The recidivism dataset has a variable `educ` coded as follows:

The FREQ Procedure

| educ | Frequency | Percent | Cumulative<br>Frequency | Cumulative<br>Percent |
|------|-----------|---------|-------------------------|-----------------------|
| 2    | 24        | 5.56    | 24                      | 5.56                  |
| 3    | 239       | 55.32   | 263                     | 60.88                 |
| 4    | 119       | 27.55   | 382                     | 88.43                 |
| 5    | 39        | 9.03    | 421                     | 97.45                 |
| 6    | 11        | 2.55    | 432                     | 100.00                |

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Categorical Variables and the CLASS Statement

A **DATA** step will collapse these 5 categories to 3 by folding the two extreme cases into the neighboring categories. The new variable is called `educ3`.

```
title "Estimation of the Weibull model with the CLASS statement";  
proc lifereg data=recid;  
    class educ3;  
    model week*arrest(0)=fin age race wexp mar paro prio educ3/  
        dist=weibull covb;  
run;
```

The `covb` option requests that the covariance matrix of the parameter estimates be printed.

By default, the highest formatted value for a class variable (here, its `educ3 = 5`) is treated as the reference category.

This can be controlled with the `ORDER` option.

# Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Categorical Variables and the CLASS Statement

## Type III Analysis of Effects

| Effect       | DF       | Wald          |               |
|--------------|----------|---------------|---------------|
|              |          | Chi-Square    | Pr > ChiSq    |
| mar          | 1        | 1.2160        | 0.2701        |
| paro         | 1        | 0.2221        | 0.6374        |
| prio         | 1        | 7.5308        | 0.0061        |
| <b>educ3</b> | <b>2</b> | <b>3.2023</b> | <b>0.2017</b> |

## Analysis of Parameter Estimates

| Parameter     | DF       | Estimate       | Standard Error | 95% Confidence Limits |               | Chi-Square  | Pr > ChiSq    |
|---------------|----------|----------------|----------------|-----------------------|---------------|-------------|---------------|
|               |          |                |                |                       |               |             |               |
| Intercept     | 1        | 4.4680         | 0.5171         | 3.4544                | 5.4816        | 74.65       | <.0001        |
| fin           | 1        | 0.2690         | 0.1379         | -0.0012               | 0.5392        | 3.81        | 0.0510        |
| age           | 1        | 0.0392         | 0.0159         | 0.0079                | 0.0705        | 6.04        | 0.0140        |
| race          | 1        | -0.2524        | 0.2229         | -0.6893               | 0.1845        | 1.28        | 0.2575        |
| wexp          | 1        | 0.0773         | 0.1522         | -0.2209               | 0.3755        | 0.26        | 0.6114        |
| mar           | 1        | 0.3013         | 0.2732         | -0.2342               | 0.8368        | 1.22        | 0.2701        |
| paro          | 1        | 0.0658         | 0.1396         | -0.2078               | 0.3394        | 0.22        | 0.6374        |
| prio          | 1        | -0.0585        | 0.0213         | -0.1004               | -0.0167       | 7.53        | 0.0061        |
| <b>educ3</b>  | <b>3</b> | <b>-0.5116</b> | <b>0.3090</b>  | <b>-1.1172</b>        | <b>0.0941</b> | <b>2.74</b> | <b>0.0978</b> |
| <b>educ3</b>  | <b>4</b> | <b>-0.3536</b> | <b>0.3243</b>  | <b>-0.9892</b>        | <b>0.2819</b> | <b>1.19</b> | <b>0.2755</b> |
| <b>educ3</b>  | <b>5</b> | <b>0.0000</b>  | <b>.</b>       | <b>.</b>              | <b>.</b>      | <b>.</b>    | <b>.</b>      |
| Scale         | 1        | 0.7119         | 0.0634         | 0.5979                | 0.8476        |             |               |
| Weibull Shape | 1        | 1.4047         | 0.1251         | 1.1798                | 1.6726        |             |               |

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Maximum Likelihood Estimation

Maximum likelihood estimation is popular for two reasons:

1. It has attractive **asymptotic properties**.
2. It has **wide application**. For example, with censored data.

Applying MLE requires two steps:

1. **Constructing the likelihood function** (i.e., assume a probability distribution for the data).
2. **Maximize the likelihood function** (typically an iterative numerical method).

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Maximum Likelihood Estimation

Given  $n$  independent observations ( $i = 1, \dots, n$ ) where the data consists of three parts:

$t_i$  = time of the event

$\delta_i$  = indicator variable equal to 1 if observation not censored and 0 if censored

$\mathbf{x}_i = [1 \ x_{i1} \ \dots \ x_{ik}]$  = vector of covariate values

For an **uncensored observation**, its contribution to the likelihood function is  $f(t_i)$ , where  $f()$  is the probability density function for  $T$ .

For a **censored observation**, its contribution to the likelihood function is  $S(t_i)$ , the survivor function.

$$\text{Thus, } L = \prod_{i=1}^n [f_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i}$$

The above likelihood function is used by **LIFEREG** for all right-censored data.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Maximum Likelihood Estimation

For example, for the **exponential model**:

$$f_i(t_i) = \lambda_i e^{-\lambda_i t_i} \quad \text{and} \quad S_i(t_i) = e^{-\lambda_i t_i} \quad \text{where} \quad \lambda_i = \exp[-\boldsymbol{\beta} \mathbf{x}_i]$$

$$\text{Thus, } L = \prod_{i=1}^n [\lambda_i e^{-\lambda_i t_i}]^{\delta_i} [e^{-\lambda_i t_i}]^{1-\delta_i} = \prod_{i=1}^n \lambda_i^{\delta_i} e^{-\lambda_i t_i}$$

$$\text{The function maximized in practice is: } \log L = \sum_{i=1}^n \delta_i \log \lambda_i - \sum_{i=1}^n \lambda_i t_i = -\boldsymbol{\beta} \sum_{i=1}^n \delta_i \mathbf{x}_i - \sum_{i=1}^n t_i e^{-\boldsymbol{\beta} \mathbf{x}_i}$$

Since  $\mathbf{x}_i$  is a vector, this is actually a system of  $k + 1$  nonlinear equations, one for each element of  $\boldsymbol{\beta}$ .

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Maximum Likelihood Estimation

**LIFEREG** uses the **Newton - Raphson** iterative method.

$U(\boldsymbol{\beta}) = \frac{\partial \log L}{\partial \boldsymbol{\beta}}$  is called the *score* or *gradient* vector.

$I(\boldsymbol{\beta}) = \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$  is called the *Hessian* matrix.

The **Newton - Raphson** *iterative step* is:  $\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - I^{-1}(\boldsymbol{\beta}_j)U(\boldsymbol{\beta}_j)$

Starting values  $\boldsymbol{\beta}_0$  are obtained by **OLS** and ignoring any censoring.

Upon convergence,  $\text{Cov}(\hat{\boldsymbol{\beta}}) = -I^{-1}(\hat{\boldsymbol{\beta}})$ .

This is the matrix printed out with the COVB option.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Maximum Likelihood Estimation

Convergence problems can occur for several reasons:

- small sample
- heavy censoring
- many parameters
- if there is a categorical variable  $X$  listed in the CLASS statement such that every observation with a particular value of  $X$  is censored.

The reason for this last situation causing problems is that the coefficient of a dichotomous variable is a function of the logarithm of the ratio of the hazards for the two groups.

If all cases for a category of  $X$  are censored, then the ML estimate for the hazard in that group is 0.

If this 0 is in the denominator of the ratio, then the coefficient estimate tends toward plus infinity. In the numerator, the estimate of  $\log(0)$  tends to minus infinity.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Maximum Likelihood Estimation

The same problem can occur for a categorical variable included in the CLASS statement. Since a set of dummy variables are created, the situation on the previous page can arise.

Some possible solutions for variables X with several values are to combine categories or to treat the variable as continuous.

The default number of iterations in **LIFEREG** is 50 and there is a default convergence criterion.

These could be changed with the **MAXITER=** and **CONVERGE=** options on the **MODEL** statement, but doing this could only be masking the real problem, namely, that an optimal solution may not exist in this case.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

As Allison states in 1995: “*PROC LIFEREG is somewhat skimpy in its facilities for hypothesis tests.*”

The situation doesn't seem to have changed much since then (see next page).

Aside from the usual Wald tests for significance of individual parameter estimates and for the joint significance for a **CLASS** variable, **LIFEREG** doesn't offer much in the way of hypothesis testing.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

I sent an email to SAS Support and this is part of the reply:

*“Unfortunately, the situation Allison(1995) mentions regarding hypothesis testing in PROC LIFEREG has not changed. PROC LIFEREG still does not have a TEST statement.*

...

*In any event, I have entered a suggestion for the 2007 SASWare Ballot (upon which users can vote in favor of specific suggestions) for adding a TEST and CONTRAST statement to PROC LIFEREG.*

...

*See FAQ 4311 for annotated survival analysis references including Allison(1995) which we highly recommend for anyone doing survival analysis in SAS – <http://support.sas.com/faq/043/FAQ04311.html>”*

# Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

The three types of likelihood-based hypothesis tests are:

**Wald tests:** quadratic forms of parameter estimates and their estimated variances and covariances.

**Lagrange tests:** (also called **score tests**) are quadratic forms using the first and second derivatives of the likelihood function.

**Likelihood-ratio tests:** twice the difference in the log-likelihood values of the full and restricted models.

All three tests are **asymptotically equivalent** and the test statistics are all distributed as **chi-square** with  $df =$  the number of restrictions.

There is some evidence that the **LR-test** best approximates a chi-square distribution and is the preferred method.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

**Example:** To perform a **LR test** that all the coefficients in a model are 0, one must fit the full model and the null model and do the calculation by hand from the listing.

For example, using the Weibull model for the recidivism data we've already fit, the log-likelihood for the full model is -321.85. The null model can fit with the following model statement:

```
model week*arrest(0)= / dist=weibull;
```

This results in a log-likelihood value of -338.59 and a test statistic of 33.48. With  $df = 7$ , the p-value is less than 0.001.

So the null hypothesis is rejected and we conclude that at least one of the coefficients is significant.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

Calculating a **Wald statistic** for survival models in **SAS** is much more involved.

**First**, estimate the model and save the parameter estimates and their covariance matrix in output datasets. (Unfortunately, if there are any **CLASS** variables in the model, then **SAS** won't create an output dataset with the parameter estimates. However, the recently added **ODS** facility should get around this problem).

**Next**, read the parameter estimates and covariance matrix into **IML** and do the necessary calculation.

This is obviously a lot of work for something that could be easily done by the **PROC**.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

The book gives an example of a **Wald test** for the equality of two coefficients.

In the recidivism example with the **CLASS** variable *educ* with three levels, the default reference category was *educ* = 5 and the estimates for categories 3 and 4 are basically estimates of the differences of these coefficients from that of category 5.

Suppose we want to test whether there is a significant difference between the coefficients for levels 3 and 4. In the model, these would be the parameters  $\beta_3$  and  $\beta_4$ .

The **Wald chi - square test statistic** is: 
$$\frac{(\beta_3 - \beta_4)^2}{\text{Var}(\beta_3) + \text{Var}(\beta_4) - 2\text{Cov}(\beta_3, \beta_4)}$$

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Hypothesis Tests

This calculation could be done "*by-hand*" by using the **COVB** option and recording the various estimates from the output.

The **test statistic** becomes: 
$$\frac{(-0.5116 - (-0.3536))^2}{0.09549 + 0.1051 - 2(0.08666)} = 0.9154$$

The 0.05 critical value for a chi-square with  $df = 1$  is 3.84. Thus we do not reject the hypothesis that  $\beta_3 = \beta_4$ .

To carry out this same test as a **LR test**, we would need to estimate the restricted model along with the unrestricted model.

For the restricted model, we would need to construct a new dataset where

```
if educ = 3 then educ = 4;
```

This results in a test statistic of 0.94, very close to the **Wald statistic**.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: GOF Tests with the LR Statistic

The choice of a distribution for the parametric survival models can have a profound effect on the shape of the hazard function, ranging from a constant hazard for the exponential distribution many non-constant shapes of the generalized gamma family.

Since the exponential, log-normal, Weibull, and standard gamma distributions are all nested in the generalized gamma distribution, we can test the various restrictions using likelihood-ratio tests. (Note that the log-logistic model is not a submodel of the generalized gamma model).

Previous results with the recidivism data show that the log-normal parameter estimates are different from the other models for some of the covariates.

Recall that the **generalized gamma distribution** has two parameters:

$\sigma$  = scale parameter

$\delta$  = shape parameter

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: GOF Tests with the LR Statistic

Restrictions imposed on the generalized gamma model for its various submodels:

|                          |                |
|--------------------------|----------------|
| $\sigma = \delta$        | standard gamma |
| $\delta = 1$             | Weibull        |
| $\sigma = 1, \delta = 1$ | exponential    |
| $\delta = 0$             | log-normal     |

The log-likelihood values from the models fitted earlier:

|         |                   |
|---------|-------------------|
| -325.83 | exponential       |
| -319.38 | Weibull           |
| -322.69 | log-normal        |
| -319.46 | standard gamma    |
| -319.40 | log-logistic      |
| -319.38 | generalized gamma |

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: GOF Tests with the LR Statistic

The various LR test statistics are:

|       |                                |        |
|-------|--------------------------------|--------|
| 12.90 | exponential vs. Weibull        | df = 1 |
| 12.74 | exponential vs. standard gamma | df = 1 |
| 12.90 | exponential vs. g. gamma       | df = 2 |
| 0.0   | Weibull vs. g. gamma           | df = 1 |
| 6.62  | log-normal vs. g. gamma        | df = 1 |
| 0.16  | standard gamma vs. g. gamma    | df = 1 |

Conclusions:      Reject the exponential model (i.e., no constant hazard)  
                          Reject the log-normal ( $p = 0.01$ )  
                          Both the Weibull and standard gamma provide good fits  
                          relative to the generalized gamma.

While the Weibull and standard gamma have monotonic hazards, the standard gamma has an upper limit while the Weibull does not. Since the Weibull is simpler mathematically, in these cases the Weibull is usually preferred.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: GOF Tests with the LR Statistic

Since the **log-logistic model** is not nested in the **generalized gamma model**, we cannot use the above LR-methodology to compare them.

Since their likelihoods are very close, the **log-logistic** should still be considered as a viable alternative model.

The above tests rely on the generalized gamma model itself being a “good” model. Since we don’t have it nested in an even more general model, we don’t know if this is true or not.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

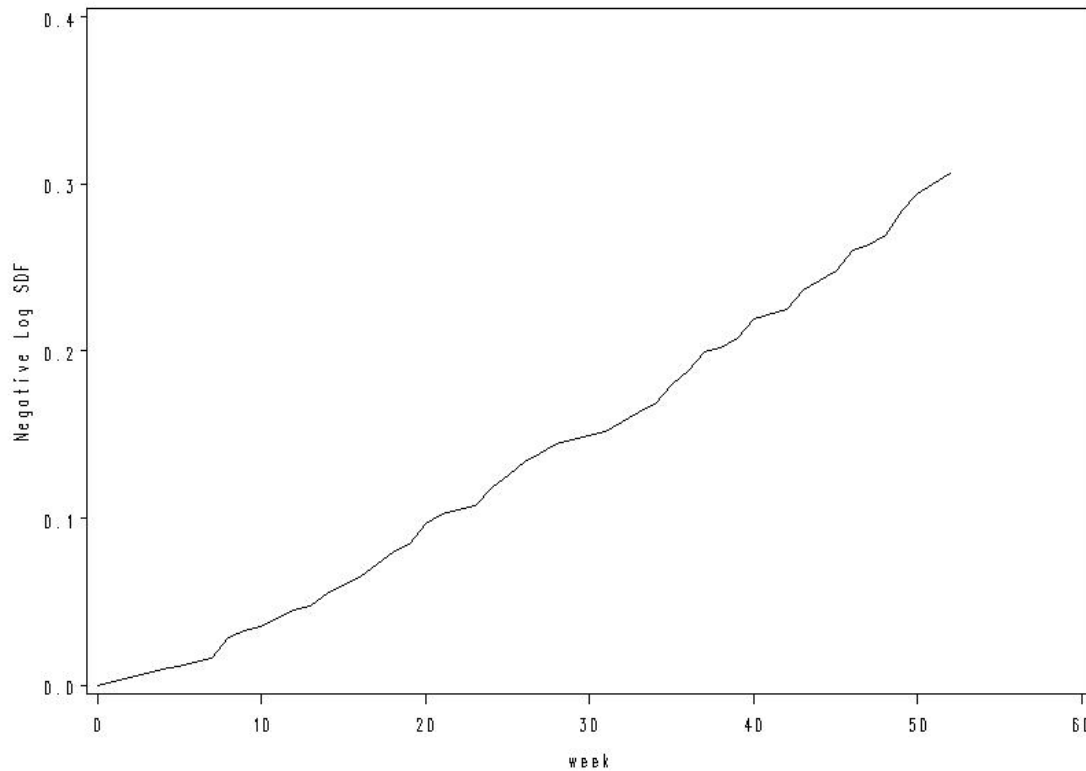
Recall that the **PLOTS = (LS, LLS)** option in **PROC LIFETEST** can be used to produce the *log - survivor plot* (**LS**, plots  $-\log \hat{S}(t)$  versus  $t$ ) and the *log - log - survivor plot* (**LLS**, plots  $\log[-\log \hat{S}(t)]$  versus  $t$ ).

```
title "Log-Survivor Plot for the Recidivism Data";  
proc lifetest data=survival.recid plots=(LS);  
    time week*arrest(0) ;  
    symbol1 v=none;  
run;
```

```
title "Log-Log-Survivor Plot for the Recidivism Data";  
proc lifetest data=survival.recid plots=(LLS);  
    time week*arrest(0) ;  
    symbol1 v=none;  
run;
```

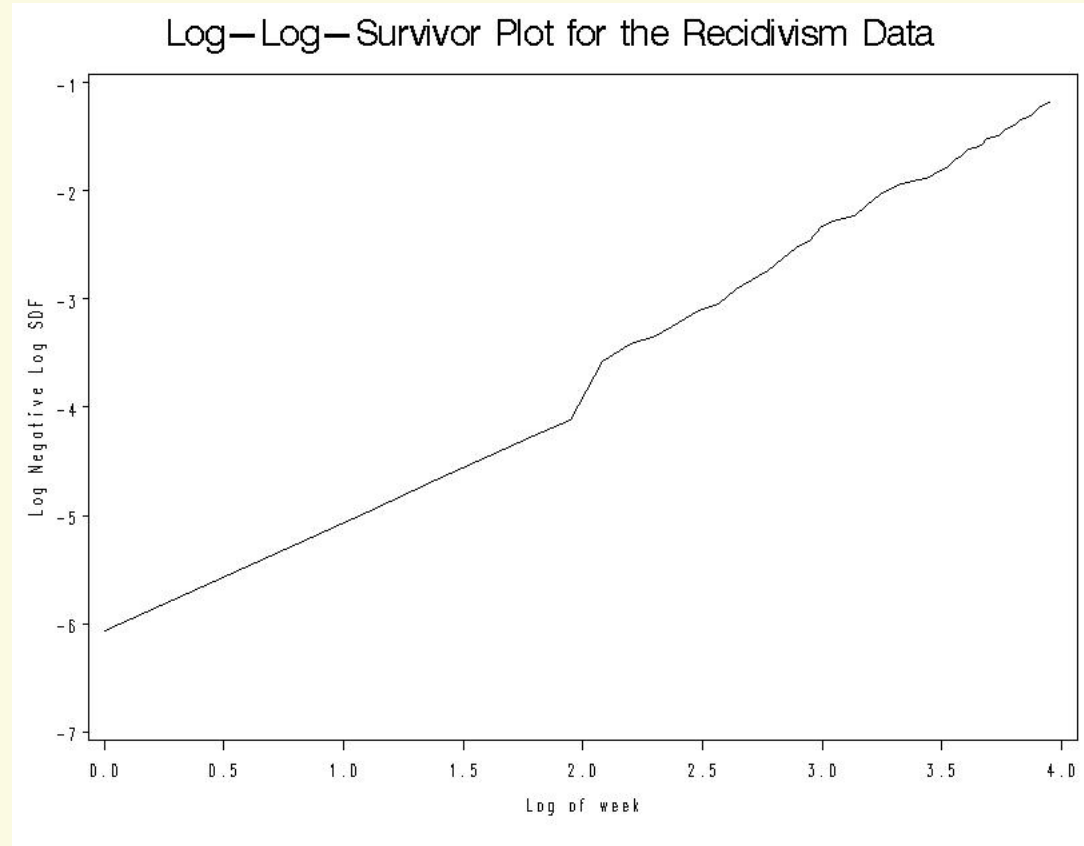
## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

Log-Survivor Plot for the Recidivism Data



An exponential distribution results in a straight line through the origin for the log-survivor plot.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit



A Weibull distribution results in a straight line through the origin for the log-log-survivor plot. Both graphs appear to be reasonably linear.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

For the **log-normal** and **log-logistic** distributions, a little more work is necessary.

1. Use **PROC LIFETEST** to get the Kaplan-Meier estimate of the survivor function and output it to a **SAS** dataset.
2. In a new **DATA** step, apply appropriate transformations to the survivor estimates.
3. Use the **GPLOT** procedure to produce the desired graphs.

For the **log-normal** distribution, a plot of  $\Phi^{-1} [1 - \hat{S}(t)]$  versus  $\log t$  should be linear.

For the **log-logistic** distribution, a plot of  $\log \left[ (1 - \hat{S}(t)) / \hat{S}(t) \right]$  versus  $\log t$  should be linear

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

```
proc lifetest data=survival.recid outsurv=a;  
  time week*arrest(0);  
run;
```

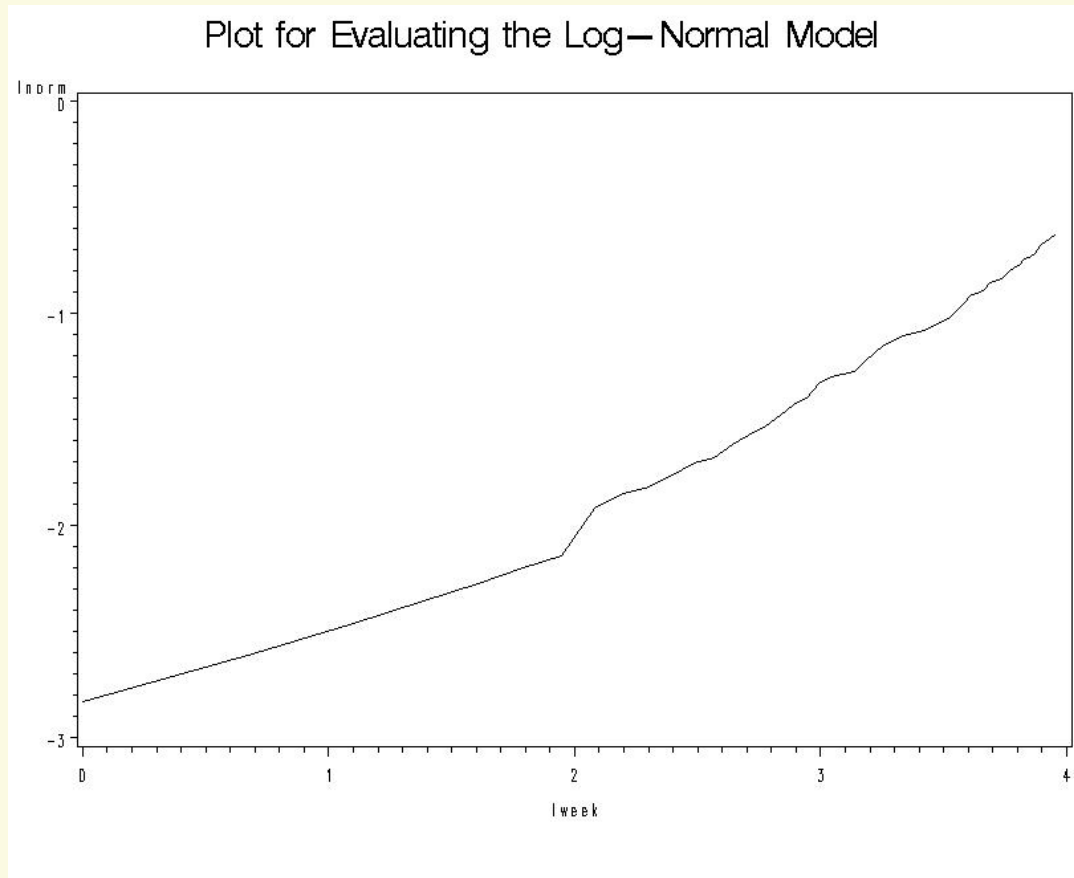
(In the dataset **a**, the variable called survival is the **Kaplan-Meier** estimate of the survivor function)

```
data;  
  set a;  
  s=survival;  
  logit=log((1-s)/s);  
  lnorm=probit(1-s);  
  lweek=log(week);  
run;
```

```
title "Plot for Evaluating the Log-Normal Model";  
proc gplot;  
  symbol1 value=none i=join;  
  plot lnorm*lweek;  
run;
```

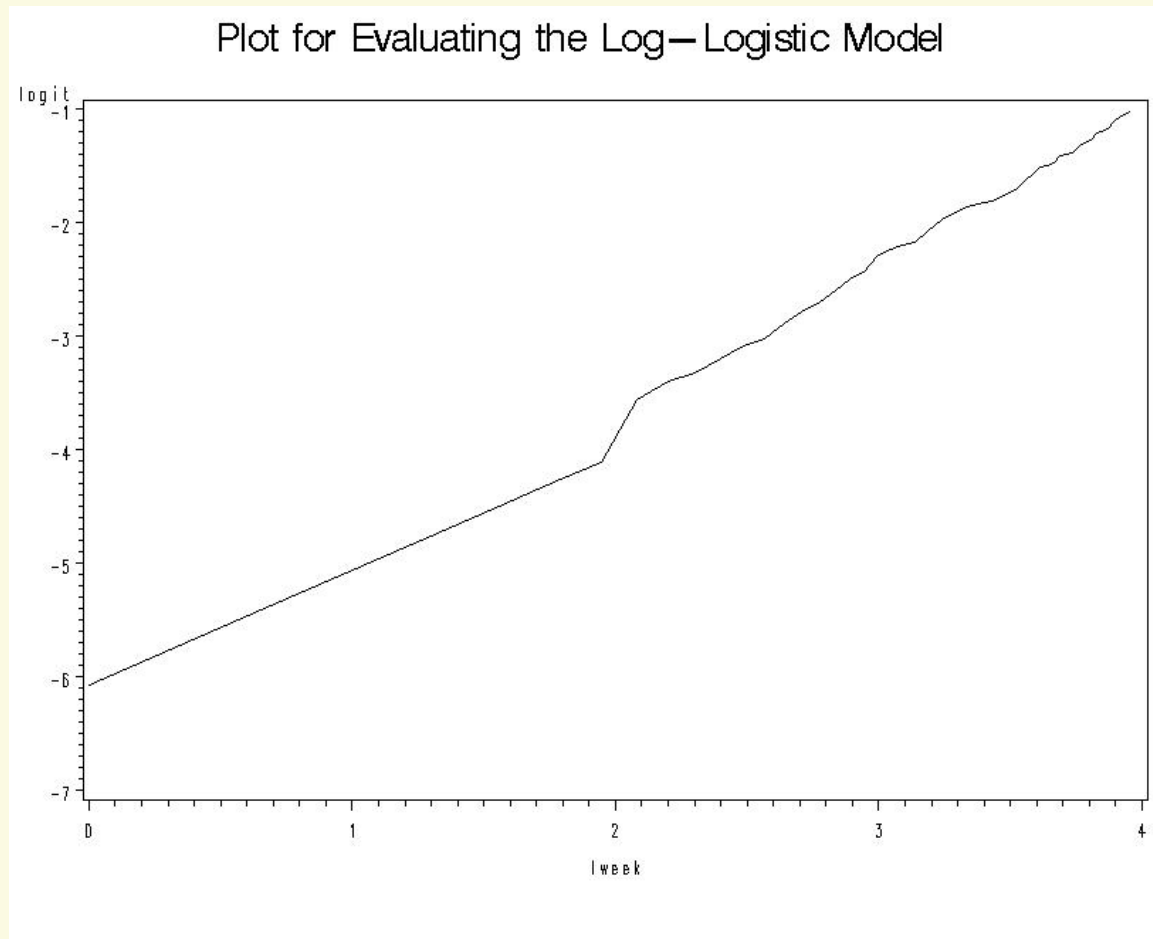
```
title "Plot for Evaluating the Log-Logistic Model";  
proc gplot;  
  symbol1 value=none i=join;  
  plot logit*lweek ;  
run;
```

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit



Plot appears somewhat bowed upward (according to the author).

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit



Plot shows some minor deviations from linearity.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

A caveat when interpreting these plots is that they assume that the data is from a homogenous population, i.e., no covariates are in the picture.

Taking into account the heterogeneity induced by conditioning on covariates can lead to different conclusions.

One solution is to look at “residual” plots.

The “most suitable” residuals (according to the author) are *Cox-Snell residuals*.

They are defined as:  $e_i = -\log \hat{S}(t_i | \mathbf{x}_i)$

Note that these “residuals” are always positive.

**Key point:** if the model is correct, then the  $e_i$ 's have (approximately) an **exponential distribution** with parameter  $\lambda = 1$ . If  $t_i$  is a censoring time, then  $e_i$  is also treated as a censored observation.

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

As we've already seen, we can graphically assess a potential exponential distribution by plotting the Kaplan-Meier estimate of its survivor function.

i.e., plot minus the log of its estimated survivor function versus  $t$  (note that the residual  $e$  plays the role of time  $t$  in this case).

The resulting graph should be a straight line through the origin with a slope of 1.

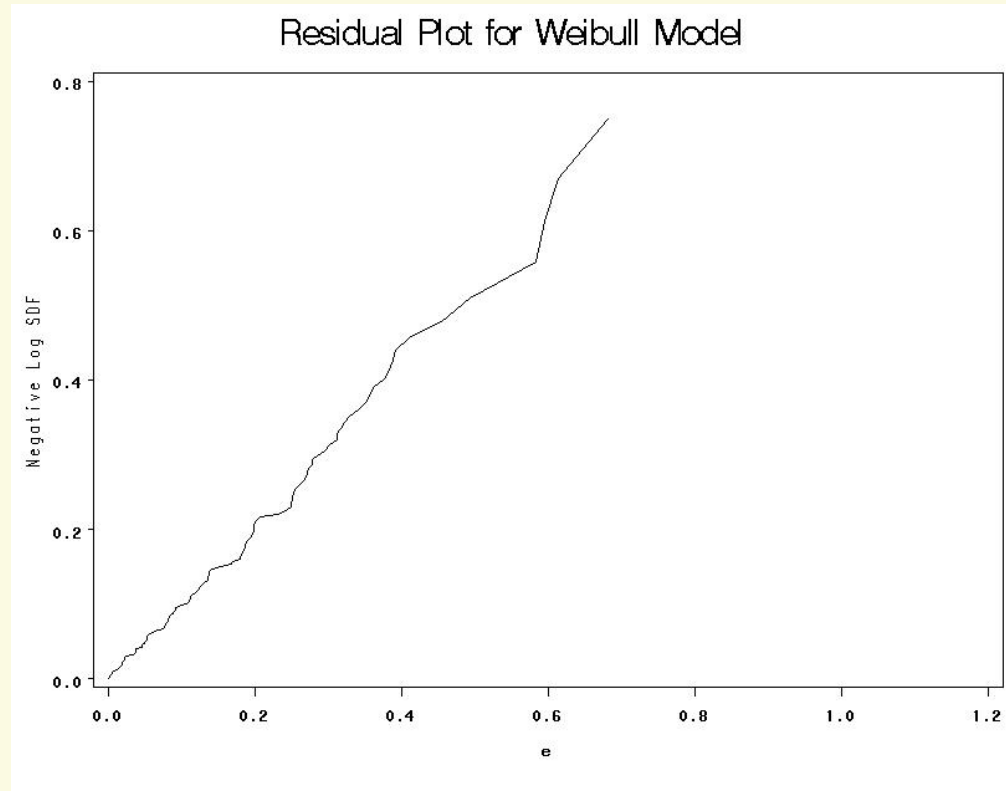
## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

```
proc lifereg data=survival.recid;  
  model week*arrest(0)=fin age race wexp mar paro prio  
    / dist=weibull;  
  output out=a cdf=f;  
run;
```

```
data b;  
  set a;  
  e=-log(1-f);  
run;
```

```
title "Residual Plot for Weibull Model";  
proc lifetest data=b plots=(ls) notable graphics;  
  time e*arrest(0);  
  symbol1 v=none;  
run;
```

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit



Is this a straight line through the origin with slope = 1?

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit

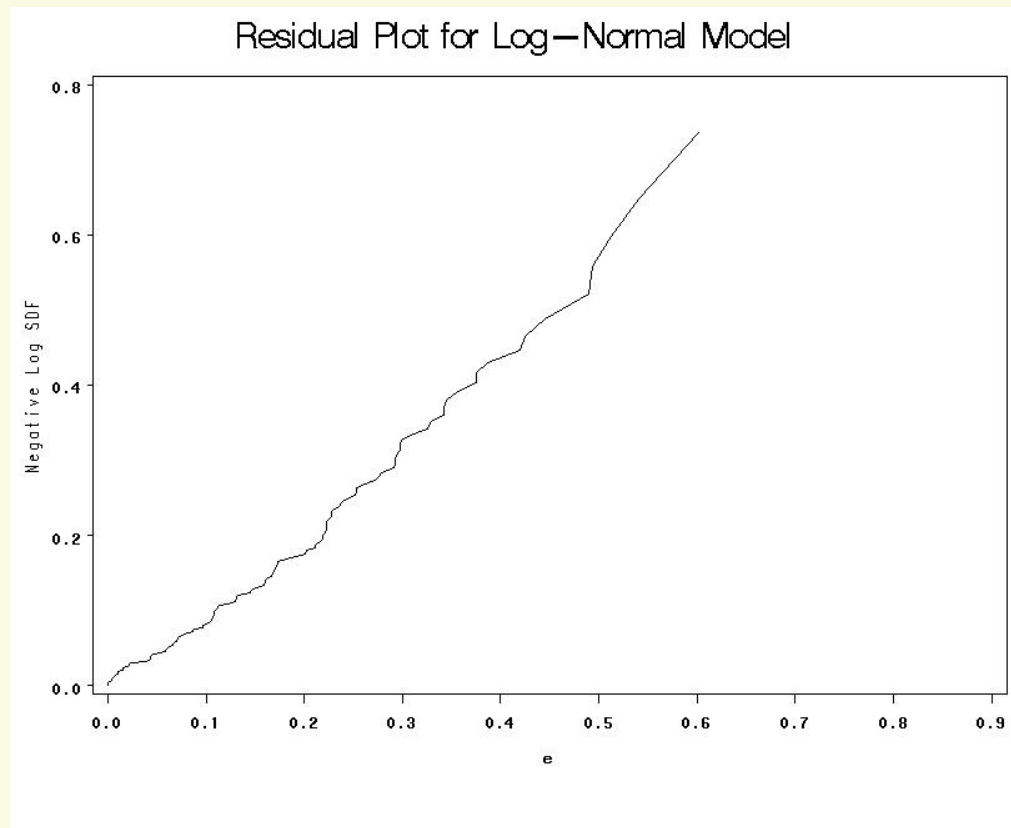
```
proc lifereg data=survival.recid;  
  model week*arrest(0)=fin age  
  race wexp mar paro prio  
    / dist=lnormal;  
  output out=a cdf=f;  
run;
```

Now do the same for the log-normal distribution.

```
data b;  
  set a;  
  e=-log(1-f);  
run;
```

```
title "Residual Plot for Log-Normal  
Model";  
proc lifetest data=b plots=(ls)  
  notable graphics;  
  time e*arrest(0);  
  symbol1 v=none;  
run;
```

## Chapter 4: Estimating Parametric Regression Models with PROC LIFEREG: Graphical Methods for Evaluating Model Fit



As the author states “... while this method is attractive in theory and is easy to implement, I have not found it to be sensitive to differences in model fit.” While the above graphs for both the Weibull and log-normal look good, the previous LR test definitely favors the Weibull over the log-normal.