Data Science & Statistical Research with SAS

Sri Sowbhagya Vidya Yaddanapudi

SAS STAT to fit parametric models

Abstract- **Data Science and Statistics are two components which go hand in hand. There is no point in having lot of data without having statistical analysis done on it. SAS is one such tool which is widely used for doing statistical analysis since it has many inbuilt procedures which help in calculating many parametric models and fit the curve. In this paper we discuss more about the SAS/STAT and SAS/QA products of the SAS software. We discuss procedures that can fit parametric models to failure time data that can be uncensored, right censored, left censored, or interval censored. The models for the response variable consist of a linear effect composed of the covariates and a random disturbance term. The distribution of the random disturbance can be taken from a class of distributions that includes the extreme value, normal, logistic, and, by using a log transformation, the exponential, Weibull, lognormal, loglogistic, and three-parameter gamma distribution.**

INTRODUCTION

The Weibull distribution is one of the most widely used lifetime distributions in reliability engineering. It is a versatile distribution that can take on the characteristics of other types of distributions, based on the value of the shape parameter.

The 3-Parameter Weibull

The 3-parameter Weibull pdf is given by:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}}$$

where:

$$f(t) \geq 0,\ t \geq \gamma$$
$$\beta > 0$$
$$\eta > 0$$
$$-\infty < \gamma < +\infty$$

and:

scale parameter, or characteristic life

shape parameter (or slope)

location parameter (or failure free life)

and:

$\eta =$ scale parameter, or characteristic life

$\beta =$ shape parameter (or slope)

$\gamma =$ location parameter (or failure free life)

The 2-Parameter Weibull

The 2-parameter Weibull *pdf* is obtained by setting $\gamma = 0$, and is given by:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}$$

Calculating Weibull 3p Using Different Procedures

PROC LIFEREG

The LIFEREG procedure fits parametric models to failure time data that can be uncensored, right censored, left censored, or interval censored. The models for the response variable consist of a linear effect composed of the covariates and a random disturbance term. The distribution of the random disturbance can be taken from a class of distributions that includes the extreme value, normal, logistic, and, by using a log transformation, the exponential, Weibull, lognormal, log-logistic, and three-parameter gamma distributions.

The following examples demonstrate how you can use the LIFEREG procedure to fit a parametric model to failure time data.

Suppose you have a response variable *y* that represents failure time; a binary variable, *censor*, with *censor*=0 indicating censored values; and two linearly independent variables, *x1* and *x2*. The following statements perform a typical accelerated failure time model analysis. Higher-order effects such as interactions and nested effects are allowed in the independent variables list, but they are not shown in this example.

```
proc lifereg;
model y*censor(0) = x1 x2;
run;
```

PROC LIFEREG can fit models to interval-censored data. The syntax for specifying interval-censored data is as follows:

```
proc lifereg;
model (begin, end) = x1 x2;
run;
```

You can also model binomial data by using ***the events/trials*** syntax for the response, as illustrated in the following statements:

```
proc lifereg;
model r/n=x1 x2;
run;
```

WEIBULL 3P USING PROC RELIABILITY

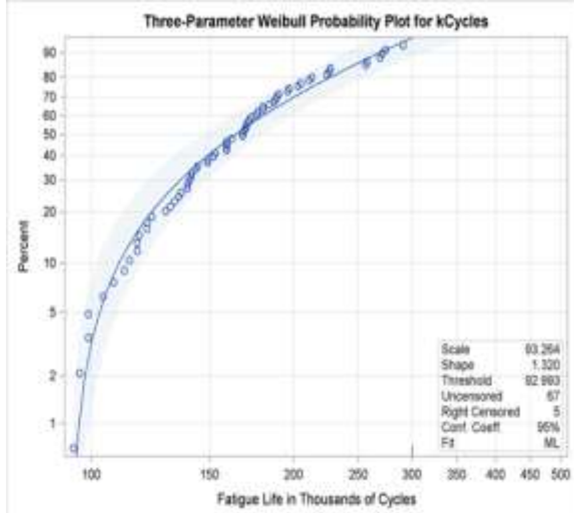
```
proc Reliability data=Alloy;
distribution Weibull3;
Pplot kCycles*Cen (1) / Profile (noconf range=(50,100)) LifeUpper =500;
run;
```

Below figure shows the maximum likelihood estimates of the Weibull threshold, shape and scale parameters, and the corresponding extreme value location and scale parameter estimates.

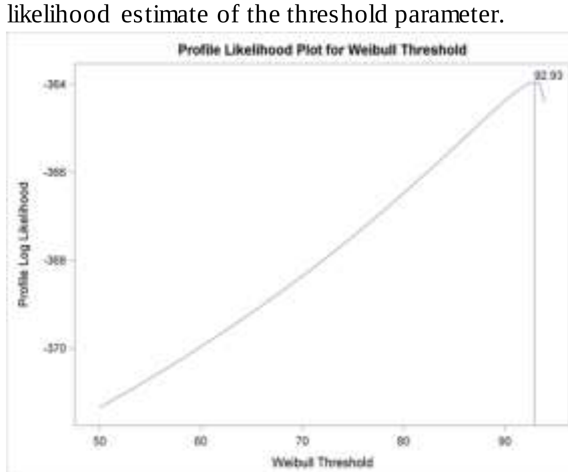
THE RELIABILITY PROCEDURE

Three-Parameter Weibull Parameter Estimates				
			Asymptotic Normal 95% Confidence Limits	
Parameter	Estimate	Standard Error	Lower	Upper
EV Location	4.5354	0.1009	4.3377	4.7332
EV Scale	0.7575	0.0898	0.6005	0.9556
Weibull Scale	93.2642	9.4082	76.5329	113.6531
Weibull Shape	1.3202	0.1565	1.0465	1.6654
Weibull Threshold	92.9928	1.9516	89.1676	96.8179

A probability plot of the failure lifetimes and the fitted three-parameter Weibull distribution is shown below.



A profile likelihood plot for the threshold parameter is shown below. The threshold value at the maximum log likelihood corresponds to the maximum likelihood estimate of the threshold parameter.



Weibull, Log Normal and G-Gamma using Proc Univariate:

The UNIVARIATE Procedure

To determine an appropriate model for a data distribution, you should consider curves from several distribution families. As shown in this example, you can use the HISTOGRAM statement to fit more than one distribution and display the density curves on a histogram.

The gap between two plates is measured (in cm) for each of 50 welded assemblies selected at random from the output of a welding process. The following statements save the measurements (Gap) in a data set named Plates:

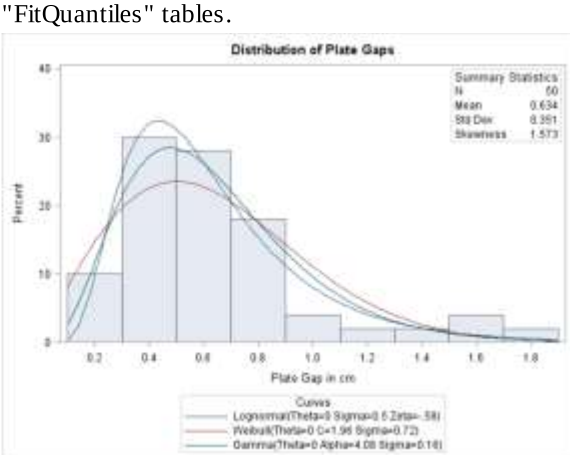
```
data Plates;
 label Gap = 'Plate Gap in cm';
 input Gap @@;
```

```
datalines;
0.746 0.357 0.376 0.327 0.485 1.741 0.241 0.777
0.768 0.409
0.252 0.512 0.534 1.656 0.742 0.378 0.714 1.121
0.597 0.231
0.541 0.805 0.682 0.418 0.506 0.501 0.247 0.922
0.880 0.344
0.519 1.302 0.275 0.601 0.388 0.450 0.845 0.319
0.486 0.529
1.547 0.690 0.676 0.314 0.736 0.643 0.483 0.352
0.636 1.080;
```

The following statements fit three distributions (lognormal, Weibull, and gamma) and display their density curves on a single histogram:

```
title 'Distribution of Plate Gaps';
ods graphics on;
ods select Histogram ParameterEstimates
GoodnessOfFit FitQuantiles;
proc univariate data=Plates;
var Gap;
histogram / midpoints=0.2 to 1.8 by 0.2
lognormal
weibull
gamma
odstitle = title;
inset n mean(5.3) std='Std Dev'(5.3) skewness(5.3)/
pos = ne header = 'Summary Statistics';
run;
```

The ODS SELECT statement restricts the output to the "ParameterEstimates," "GoodnessOfFit," and "FitQuantiles" tables.



Distribution of Plate Gaps

The UNIVARIATE Procedure

Fitted Lognormal Distribution for Gap (Plate Gap in cm)

Parameters for Lognormal Distribution		
Parameter	Symbol	Estimate
Threshold	Theta	0
Scale	Zeta	-0.58375
Shape	Sigma	0.499546
Mean		0.631932
Std Dev		0.336436

Goodness-of-Fit Tests for Lognormal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.06441431	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.02823022	Pr > W-Sq	>0.500
Anderson-Darling	A-Sq	0.24308402	Pr > A-Sq	>0.500

Quantiles for Lognormal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.23100	0.17449
5.0	0.24700	0.24526
10.0	0.29450	0.29407
25.0	0.37800	0.39825
50.0	0.53150	0.55780
75.0	0.74600	0.78129
90.0	1.10050	1.05807
95.0	1.54700	1.26862
99.0	1.74100	1.78313

Distribution of Plate Gaps

The UNIVARIATE Procedure

Fitted Weibull Distribution for Gap (Plate Gap in cm)

Parameters for Weibull Distribution

Parameter	Symbol	Estimate
Threshold	Theta	0
Scale	Sigma	0.719208
Shape	C	1.961159
Mean		0.637641
Std Dev		0.339248

Goodness-of-Fit Tests for Weibull Distribution				
Test	Statistic		p Value	
Cramer-von Mises	W-Sq	0.15937281	Pr > W-Sq	0.016
Anderson-Darling	A-Sq	1.15693542	Pr > A-Sq	<0.010

Quantiles for Weibull Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.23100	0.06889
5.0	0.24700	0.15817
10.0	0.29450	0.22831
25.0	0.37800	0.38102
50.0	0.53150	0.59661
75.0	0.74600	0.84955
90.0	1.10050	1.10040
95.0	1.54700	1.25842
99.0	1.74100	1.56691

Below figure provides two EDF goodness-of-fit tests for the Weibull distribution: the Anderson-Darling and the Cramér–von Mises tests. The p-values for the EDF tests are all less than 0.10, indicating that the data do not support a Weibull model.

Distribution of Plate Gaps
The UNIVARIATE Procedure
Fitted Gamma Distribution for Gap (Plate Gap in cm)

Parameters for Gamma Distribution		
Parameter	Symbol	Estimate
Threshold	Theta	0
Scale	Sigma	0.155198
Shape	Alpha	4.082646
Mean		0.63362
Std Dev		0.313587

Goodness-of-Fit Tests for Gamma Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.09695325	Pr > D	>0.250
Cramer-von Mises	W-Sq	0.07398467	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.58106613	Pr > A-Sq	0.137

Quantiles for Gamma Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.23100	0.13326
5.0	0.24700	0.21951
10.0	0.29450	0.27938
25.0	0.37800	0.40404
50.0	0.53150	0.58271
75.0	0.74600	0.80804
90.0	1.10050	1.05392
95.0	1.54700	1.22160
99.0	1.74100	1.57939

At the $\alpha = 0.10$ significance level, all tests support the conclusion that the gamma distribution with scale parameter $\sigma = 0.16$ and shape parameter $\alpha = 4.08$ provides a good model for the distribution of plate gaps.
Based on this analysis, the fitted lognormal distribution and the fitted gamma distribution are both good models for the distribution of plate gaps.

CONCLUSION

All the above mentioned are tested on SAS Base, SAS QC , SAS /STAT software machine which was running on the Linux box.
Data need to be censored before using any of the procedures above.

REFERENCES

[1] https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_lifereg_sect003.htm
[2] Weibull Wiki : https://en.wikipedia.org/wiki/Weibull_distribution