Chapter 566

Parametric Survival (Weibull) Regression

Introduction

This module fits the regression relationship between a positive-valued dependent variable (often time to failure) and one or more independent variables. The distribution of the residuals (errors) is assumed to follow the exponential, extreme value, logistic, log-logistic, lognormal, lognormal10, normal, or Weibull distribution. The data may include failed, left censored, right censored, and interval observations. This type of data often arises in the area of *accelerated life testing*.

When testing highly reliable components at normal stress levels, it may be difficult to obtain a reasonable amount of failure data in a short period of time. For this reason, tests are conducted at higher than expected stress levels. The models that predict failure rates at normal stress levels from test data on items that fail at high stress levels are called *acceleration models.*

The basic assumption of acceleration models is that failures happen faster at higher stress levels. That is, the failure mechanism is the same, but the time scale has been changed (shortened).

Technical Details

The linear regression equation is

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + Se$$

Here, *S* represents the value of a constant standard deviation, *Y* is a transformation of time (either *ln(t), log(t),* or just *t*), the *X's* are one or more independent variables, the *B*'s are the regression coefficients, and *e* is the residual (error) that is assumed to follow a particular probability distribution. The problem reduces to estimating the *B*'s and *S*. The density functions of the eight distributions that are fit by this module were given in the Distribution Fitting section and will not be repeated here.

So that you can get the general idea, we will give detailed results for the lognormal distribution. The results for other distributions follow a similar pattern.

The lognormal probability density function may be written as

$$f(t|M,S) = \frac{1}{tS\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(t)-M}{S}\right)^2}$$

If we replace the location parameter, *M*, with the regression model, the density now becomes

$$f\left(t\middle|B_0 \cdots B_p, S\right) = \frac{1}{tS\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(t) - \sum_{i=0}^{p} B_iX_i}{S}\right)^2\right\}$$

Maximum likelihood estimation consists of finding the values of the distribution parameters that maximize the log-likelihood of the data values. Loosely speaking, these are the values of the parameters which maximize the probability that the current set of data values occur.

The general form of the log-likelihood function is given by

$$L(\underline{P}) = \sum_{F} \ln\left(f(\underline{P}, t_k)\right) + \sum_{R} \ln\left(S(\underline{P}, t_k)\right) + \sum_{L} \ln\left(F(\underline{P}, t_k)\right) + \sum_{I} \ln\left(f(\underline{P}, t_{uk}) - f(\underline{P}, t_{lk})\right)$$

where *F* represents the set of failed items, *R* represents the set of right censored items, *L* represents the set of left censored items, and *I* represents the set of interval censored items. In the case of interval censored observations, t_{lk} represents the first time of the interval and t_{uk} represents the last time of the interval. Also, $\underline{P}$ represents the parameters, including *S* and the *B*'s.

We employ the Newton-Raphson algorithm with numerical differentiation to obtain the maximum likelihood estimates. These estimates have been shown to have optimality characteristics in large samples (number of failures greater than 20). They have been shown to be competitive estimates even for sample sizes less than 20.

Data Structure

Survival data are somewhat more difficult to enter because of the presence of various types of censoring.

Time Variable(s)

One (or two in the case of interval data) variable is needed to contain the time values.

Censor Variable

Another variable is needed to indicate the type of censoring.

Failed or Complete

A failed observation is one in which the time until the terminal event was measured exactly; for example, the machine stopped working or the mouse died of the disease being studied.

Right Censored

A right censored observation provides a lower bound for the actual failure time. All that is known is that the failure occurred (or will occur) at some point after the given time value. Right censored observations occur when a study is terminated before all items have failed. They also occur when an item fails due to an event other than the one of interest.

Left Censored

A left censored observation provides an upper bound for the actual failure time. All we know is that the failure occurred at some point before the time value. Left censoring occurs when the items are not checked for failure until sometime after the study has begun. When a failed item is found, we do not know exactly when it failed, only that it was at some point before the left censor time.

Interval Censored or Readout

An interval censored observation is one in which we know that the failure occurred between two time values, but we do not know exactly when. This type of data is often called *readout* data. It occurs in situations where items are checked periodically for failures.

Independent Variable(s)

One or more independent variables must be supplied also.

Data Structure

The following data, found in Nelson (1990), are quoted in many books and articles on accelerated testing. These data come from a temperature-accelerated life test of a Class-B insulation for electric motors. Ten motorettes were tested at each of four temperatures. When the testing was stopped, the following failure times were recorded. These data are stored in the Motors dataset.

Motors Dataset

Hours	Censor	Count	Temperature
	1	1	130
8064	0	10	150
1764	1	1	170
2772	1	1	170
3444	1	1	170
3542	1	1	170
3780	1	1	170
4860	1	1	170
5196	1	1	170
5448	0	3	170
408	1	2	190
1344	1	2	190
1440	1	1	190
1680	0	5	190
408	1	2	220
504	1	3	220
528	0	5	220

Example 1 – Lognormal Regression

This section presents an example of how to fit a lognormal regression. The data used were shown above and are found in the Motors dataset.

Setup

To run this example, complete the following steps:

1 Open the Motors example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Motors** and click **OK**.

2 Specify the Parametric Survival (Weibull) Regression procedure options

- Find and open the **Parametric Survival (Weibull) Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Y: Time Variable ..... **Hours**
Censor Variable ..... **Censor**
Stress Variable ..... **Temp** (Note that the default values of Stress A and Stress B are appropriate for this problem.)
Frequency Variable..... **Count**
Distribution..... **Log10normal**

Reports Tab

Percent Failing Report Calculation Times..... **10000:100000(10000)**

Plots Tab

Stress – Time Plot Format (*Click the Button*)

Y Axis Tab
Scale..... **Log**
Format ..... **Powers**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Data Summary

Data Summary

Type of Observation	Rows	Count	Hours Minimum	Hours Maximum
Missing or Prediction	1			
Failed	12	17	408	5196
Right Censored	4	23	528	8064
Left Censored	0	0		
Interval Censored	0	0		
Total (Nonmissing)	16	40	408	8064

Means (Not adjusted for Censoring)

Variable	Mean
Hours	1919.412
Temp	190.5882

This report displays a summary of the data that were analyzed. Scan this report to determine if there are any obvious data-entry errors by double-checking the counts and the minimum and maximum.

The means given for each variable are for the noncensored rows.

Maximum Likelihood Parameter Estimation

Maximum Likelihood Parameter Estimation

Parameter Name	Estimate	Standard Error	Z-Value	Two-Sided P-Value	95% Confidence Interval Limits Lower	95% Confidence Interval Limits Upper
Intercept	-6.018404	0.9474726	-6.3521	0.000000	-7.875416	-4.161391
Temp	4.31048	0.4369486	9.8650	0.000000	3.454076	5.166883
Sigma	0.2591772	0.04733526	5.4754	0.000000	0.1811908	0.3707297

Model Fit

R-Squared	0.531405
Log-Likelihood	-148.5373
Iterations	55

This report displays parameter estimates along with standard errors, significance tests, and confidence limits. Note that the significance levels and confidence limits all use large sample formulas. How large is a large sample? We suggest that you only use these results when the number of failed items is greater than twenty.

Parameter Estimates

These are the maximum likelihood estimates (MLE) of the parameters. They are the estimates that maximize the likelihood function. Details are found in Nelson (1990) pages 287 - 295.

Standard Error

The standard errors are the square roots of the diagonal elements of the estimated Variance Covariance matrix.

Z-Value

The z-value is equal to the parameter estimate divided by the estimated standard error. This ratio, for large samples, follows the normal distribution. It is used to test the hypothesis that the parameter value is zero. This value corresponds to the t value that is used in multiple regression.

Two-Sided P-Value

This is the two-tailed p-value for testing the significance of the corresponding parameter. You would deem independent variables with small p-values (less than 0.05) important in the regression equation.

Upper and Lower Confidence Interval Limits

These are the lower and upper confidence limits for the corresponding parameters. They are large sample limits. They should be ignored when the number of failed items is less than thirty. For the regression coefficients *B*, the formulas are

$$CL_i = \hat{B}_i \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{B}_i} \quad i = 0, \cdots, p$$

where $\hat{B}_i$ is the estimated regression coefficient, $\hat{\sigma}_{\hat{B}_i}$ is its standard error, and *z* is found from tables of the standard normal distribution.

For the estimate of sigma, the formula is

$$CL = \hat{S}\exp\left\{\frac{\pm z_{1-\alpha/2}\hat{\sigma}_{\hat{S}}}{\hat{S}}\right\}$$

R-Squared

R-Squared reflects the percent of variation in log(time) explained by the independent variables in the model. A value near zero indicates a complete lack of fit, while a value near one indicates a perfect fit.

Note that this R-Squared value is computed for the failed observations only. Censored observations are ignored.

Log-Likelihood

This is the value of the log likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Iterations

This is the number of iterations that were required to solve the likelihood equations. If this is greater than the maximum you specified, you will receive a warning message. You should then increase the Maximum Iterations and rerun the analysis.

Variance-Covariance Matrix

Variance-Covariance Matrix

	Intercept	Temp	Sigma
Intercept	0.8977042	-0.4132871	-0.008281598
Temp	-0.4132871	0.1909241	0.004406308
Sigma	-0.008281598	0.004406308	0.002240627

This table gives an estimate of the asymptotic variance covariance matrix which is the inverse of the Fisher information matrix. The elements of the Fisher information matrix are calculated using numerical differentiation.

Percent Failing

Percent Failing

Row	Temp	Hours	Percent Failing	95% Confidence Interval Limits	
				Lower	Upper
1	130	10000	0.4689%	0.0159	12.2623
1	130	20000	7.5434%	0.9584	40.7554
1	130	30000	22.4510%	4.5175	63.9185
1	130	40000	39.1662%	9.9782	78.9012
1	130	50000	53.9401%	15.9649	87.8329
1	130	60000	65.7053%	21.6618	92.9946
1	130	70000	74.6251%	26.7472	95.9493
1	130	80000	81.2324%	31.1609	97.6408
1	130	90000	86.0786%	34.9545	98.6139
1	130	100000	89.6239%	38.2164	99.1777

This report displays the estimated percent failing at the time values that were specified in the Report Times box of the Reports Tab for each observation with a missing time value. In our example, the first row of the Motors database is missing. The value of Temp (the stress variable) equal to 130 degrees. Reliability is one minus probability of failure. Thus, the reliability at 80,000 hours at a temperature of 130 degrees is 100-81.2324 which is 18.7676%. The confidence limits for reliability may also be converted from the percent failing confidence limits by subtracting from 100.

Percent Failing

The percent failing at a particular temperature is calculated as

$$100 \times \hat{F}(t|X_i) = 100 \times \hat{F} = 100 \times F\left(\frac{log(t) - \sum_{i=0}^{p} x_{ki}\hat{B}_i}{\hat{S}}\right)$$

where *F(z)* is the cumulative distribution of *f(z)*, the probability density function. That is,

$$F(z) = \int_0^z f(t|B_0, B_1, S, X)dt$$

Confidence Interval Limits for Percent Failing

The confidence limits for this estimate are computed using the following formulas from Nelson (1990) page 296. Note that these estimates are large sample estimates based on the assumption that the distribution of *F* is asymptotically normal. We recommend that the number of failures be at least thirty when using these estimates.

$$\hat{F}_{lower}(t|X_i) = \frac{\hat{F}}{\hat{F} + (1-\hat{F})exp\left\{\frac{z_{1-\alpha/2}\sigma_{\hat{F}}}{\hat{F}(1-\hat{F})}\right\}}$$

$$\hat{F}_{upper}(t|X_i) = \frac{\hat{F}}{\hat{F} + (1-\hat{F})exp\left\{\frac{-z_{1-\alpha/2}\sigma_{\hat{F}}}{\hat{F}(1-\hat{F})}\right\}}$$

where

$$\sigma_{\hat{F}} = \sqrt{\sum_{i=0}^{p+1}\sum_{j=0}^{p+1} h_i h_j vc_{ij}}$$

$$h_i = \frac{-x_i g\left(\frac{y(t) - \sum x_i \hat{B}_i}{\hat{S}}\right)}{\hat{S}} \quad i = 0, \cdots, p$$

$$h_{p+1} = -\frac{y(t) - \sum x_i \hat{B}_i}{\hat{S}^2}$$

and vc_{ij} is the corresponding element from the variance covariance matrix. The function *y(t)* is *ln(t)* for the Weibull, log-logistic, exponential, and lognormal distributions, *log(t)* for the lognormal10 distribution, and simply *t* for the normal, extreme value, and logistic distributions. The value of *g(x)* depends on the distribution. For the Weibull, exponential, and extreme value distributions

$$g(z) = e^{z-e^z}$$

For the normal, lognormal, and lognormal10 distributions

$$g(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$

For the logistic and log-logistic distributions

$$g(z) = \frac{e^z}{(1+e^z)^2}$$

Failure Time Percentiles

Failure Time Percentiles

Row	Temp	Percentile	Estimated Hours	95% Confidence Interval Limits	
				Lower	Upper
1	130	10	21937.2	11776.2	40865.6
1	130	50	47133.6	24093.8	92205.6
1	130	90	101269.8	44847.0	228678.9

This report displays failure time percentiles and confidence intervals for those percentiles specified in the Report Percentiles box of the Report tab. For example, the median failure time is 47,135.1 hours. The 95% confidence limits for the median time are 24,106.6 to 92,162.2 hours.

The confidence limits rely on the asymptotic normality of the distribution of the percentiles. The sample size should be greater than thirty failed items before you use these confidence limits. The formulas for these limits are given in Nelson (1990) page 295.

Percentile

This is the percentile being found. For example, the value of 50 here refers to the median failure time.

Estimated Hours

The estimated time value (dependent variable) at which *100P* of the items are expected to fail. The percentile is found by solving the equation

$$P = F\left(\frac{y(t) - \sum_{i=0}^{p} x_{ki}\hat{B}_i}{\hat{S}}\right)$$

for *y(t)*. The function *y(t)* is *ln(t)* for the Weibull, log-logistic, exponential, and lognormal distributions, *log(t)* for the lognormal10 distribution, and simply *t* for the normal, extreme value, and logistic distributions. *F(t)* is the cumulative distribution function.

Confidence Interval Limits for a Percentile

The confidence limits are computed as follows. First compute

$$u_p = F^{-1}(P)$$

Next compute

$$\sigma_{\hat{t}_p} = \sqrt{\sum_{i=0}^{p+1}\sum_{j=0}^{p+1} x_i x_j vc_{ij}}$$

where vc_{ij} is the corresponding element of the variance covariance matrix and

$$x_0 = 1$$
$$x_1 = X_1$$
$$.$$
$$.$$
$$.$$
$$x_p = X_p$$
$$x_{p+1} = u_P$$

Finally, for the lognormal, Weibull, exponential, and log-logistic distributions, compute

$$\hat{t}_{lower,p} = e^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} - z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

$$\hat{t}_{upper,p} = e^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} + z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

For the lognormal10 distribution, compute

$$\hat{t}_{lower,p} = 10^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} - z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

$$\hat{t}_{upper,p} = 10^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} + z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

For the normal, extreme value, and logistic distributions, compute

$$\hat{t}_{lower,p} = \sum_{i=0}^{p} x_i B_i + u_p \hat{S} - z_{1-\alpha/2}\sigma_{\hat{t}_p}$$

$$\hat{t}_{upper,p} = \sum_{i=0}^{p} x_i B_i + u_p \hat{S} + z_{1-\alpha/2}\sigma_{\hat{t}_p}$$

Residuals

Residuals

Row*	Hours (T)	Log(T)		Residual		
		Actual	Predicted	Raw	Standardized	Cox-Snell
1			4.673331			
2R	8064	3.906550	4.168003	-0.2614523	-1.008778	0.1702434
3	1764	3.246499	3.708286	-0.4617874	-1.781744	0.03811264
4	2772	3.442793	3.708286	-0.2654927	-1.024368	0.1658548
5	3444	3.537063	3.708286	-0.1712228	-0.6606401	0.293595
6	3542	3.549248	3.708286	-0.1590374	-0.6136243	0.3143434
7	3780	3.577492	3.708286	-0.1307942	-0.5046515	0.3665836
8	4860	3.686636	3.708286	-0.0216497	-0.08353243	0.6286976
9	5196	3.715669	3.708286	0.007383174	0.02848698	0.7161356
10R	5448	3.736237	3.708286	0.02795113	0.1078456	0.7829427
11	408	2.610660	3.288272	-0.6776116	-2.614472	0.004478281
12	1344	3.128399	3.288272	-0.1598725	-0.6168463	0.3128878
13	1440	3.158362	3.288272	-0.1299092	-0.5012372	0.3683169
14R	1680	3.225309	3.288272	-0.06296246	-0.2429321	0.5175633
15	408	2.610660	2.722126	-0.1114662	-0.4300772	0.4058198
16	504	2.702430	2.722126	-0.01969582	-0.07599366	0.6343352
17R	528	2.722634	2.722126	0.0005075629	0.001958363	0.694711

* R = Right Censored, L = Left Censored, I = Interval Censored

This report displays the predicted value and residual for each row. If the analysis is being run on logarithms of time, all values are in logarithms. The report provides predicted values for all rows with values for the independent variables. Hence, you can add rows of data with missing time values to the bottom of your database and obtain the predicted values for them from this report. The report also allows you to obtain predicted values for censored observations.

You should ignore the residuals for censored observations, since the residual is calculated as if the time value was a failure.

Row

This is the number of the observation being reported on. Censored observations have a letter appended to the row number.

Hours (T)

This is the original value of the dependent variable.

Actual Log(T)

This is the transformed value of the dependent variable.

Predicted Log(T)

This is the predicted transformed value of the dependent variable (usually time). Note that *y* depends on the distribution being fit. For the Weibull, exponential, lognormal, and log-logistic distributions, the *y* is *ln(t)*. For the lognormal10 distribution, *y* is *log(t)*. For the extreme value, normal, and logistic distributions, *y* is *t*. The formula for *y* is

$$\hat{y} = \sum_{i=0}^{p} x_i B_i$$

Raw Residual

This is the residual in the *y* scale. The formula is

$$r_k = y_k - \sum_{i=0}^{p} x_i B_i$$

Note that the residuals of censored observations are not directly interpretable, since there is no obvious value of *y*. The row is displayed so that you can see the predicted value for this censored observation.

Standardized Residual

This is the residual standardized by dividing by the standard deviation. The formula is

$$r_k' = \frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}$$

Cox-Snell Residual

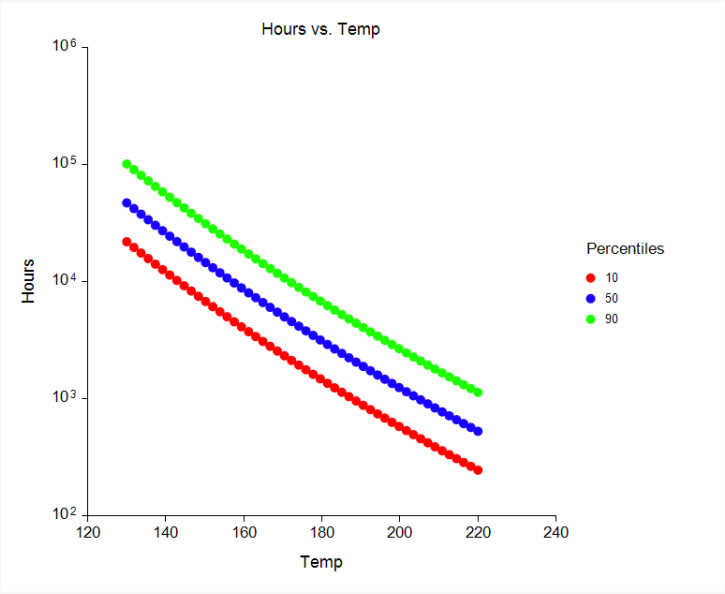
The Cox-Snell residual is defined as

$$r_k'' = -\log\left\{1 - F\left(\frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}\right)\right\}$$

Here again, the residual does not have a direct interpretation for censored values.

Stress Plot

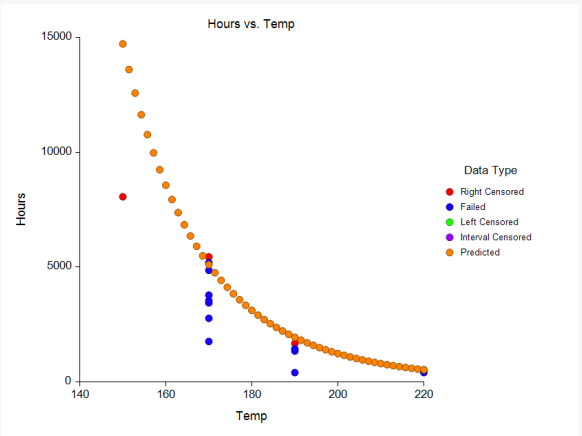
Stress Plot

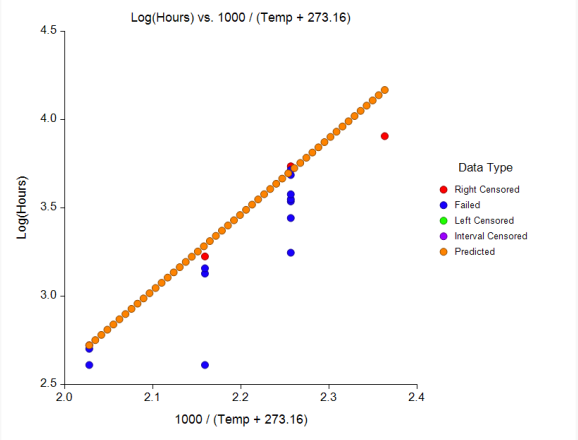


This plot displays the time on the vertical axis and the stress variable on the horizontal axis. The plotted lines represent the percentiles specified on the Reports tab window. This allows you to quickly view the percentiles for a wide range of stress values.

X-Y Plots

X's vs Y Plots

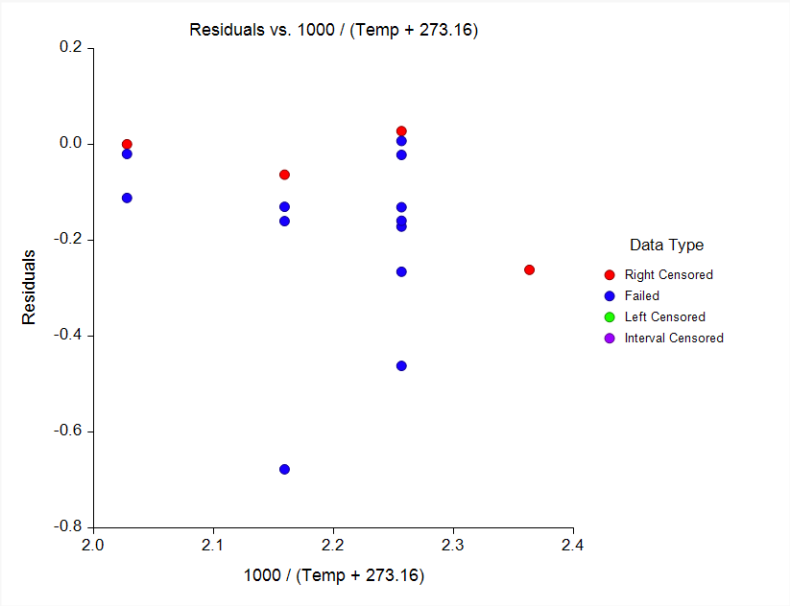




These plots show the data values from which the analysis was run. The plot on the left shows time versus stress in the original scale. The plot on the right shows time versus stress in the transformed metric. The prediction equation is also show on the chart. This lets you decide whether predictions are accurate. It also lets you study the goodness of fit.

X's versus Residuals Plots

X's vs Residuals Plots



This plot shows the residuals in the transformed scale. You would study this chart just as you would any other residual versus independent variable plot from a multiple regression analysis. You are especially interested in finding outliers, as they can distort your results.

Discussion of Example

This example will look at an analysis of the electric motor data that was presented above and for which all of the above sample reports were generated. As mentioned earlier, a temperature-accelerated life test of a Class-B insulation was conducted using ten motors tested at each of four temperatures. When the testing was stopped, the failure times were recorded. These data are stored in the MOTORS database.

The purpose of this study was to determine the reliability of these motors at the normal operating temperature of 130°C by testing the reliability at higher temperatures. Note that at 150°C, none of the motors failed during the duration of the test.

The first step in the analysis is to determine if the fit is adequate. We look at the plots, the value of R-Squared, and the estimated value of sigma to determine this. The plots do not show any alarming points, although the residual plots show what might be a mild outlier in the 190°C batch.

Once the adequacy of the fit has been substantiated, we look at the Failure Time Percentile Section. This report provides the 10^{th}, 50^{th}, and 90^{th} percentiles. The estimated failure times for these percentiles are 21,937 hours, 47,134 hours, and 101,270 hours. That is, we would expect about 10% of the motors to fail by 22,000 hours and 90% of the machines to have failed by 100,000 hours. No further calculations are necessary.