Life Data Regression

Summary

The **Life Data Regression** procedure is designed to fit a parametric statistical model relating failure times to one or more predictor variables. The predictors may be either quantitative or categorical. First or second order models can be fit, with or without interactions. The distribution of failure times may take any of seven different forms, including a Weibull, exponential, normal, lognormal, logistic, loglogistic, or smallest extreme value distribution. Failure times may be censored or uncensored.

The output of the procedure includes an estimate of the hazard function and failure time percentiles. Predictions may be made from the fitted model and unusual residuals detected.

Sample StatFolio: *lifedata reg.sgp*

Sample Data:

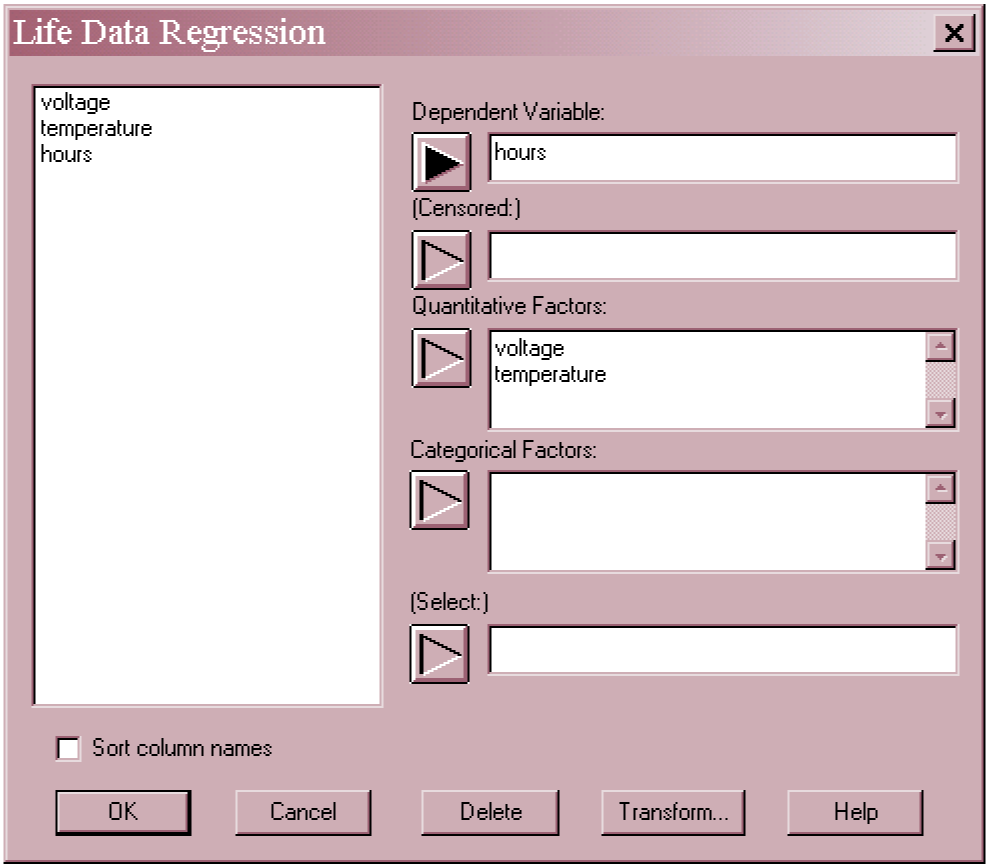
The file *capacitors.sgd* contains data from an experiment performed to determine the effect of *Voltage* and *Temperature* on the failure times of glass capacitors, reported by Meeker and Escobar (1998). A total of n = 32 capacitors were tested, four at each combination of 4 voltages and 2 temperatures. A portion of the file is shown below.

Voltage	*Temperature*	*Hours*
200	170	439
200	170	904
200	170	1092
200	170	1105
250	170	572
250	170	690
250	170	904
250	170	1090
300	170	315
300	170	315
300	170	439
300	170	628
350	170	258
350	170	258
350	170	347
350	170	588
200	180	959
200	180	1065
…	…	…

All of the observed failure times are uncensored.

Data Input

The data input dialog box requests information about the failure times and the predictor variables:



- **Dependent Variable**: a numeric variable containing Y, the failure times (for uncensored data) or censoring times (for censored data).

- **(Censored)**: an optional column indicating whether or not each data value has been censored. Enter a 0 if the value of the dependent variable represents an uncensored failure time. Enter a 1 if the value has been right-censored (the true failure time is greater than the value entered).

- **Quantitative Factors**: numeric columns containing the values of any quantitative factors to be included in the model.

- **Categorical Factors**: numeric or non-numeric columns containing the levels of any categorical factors to be included in the model.

- **Select**: subset selection.

Statistical Model

STATGRAPHICS fits two types of parametric life data regression models, location-scale regression models and log-location-scale regression models.

Location-Scale Models

For this type of model, the percentiles of the lifetime distribution are related to the predictor variables through a linear function of the form

$$Y_P = \mu + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \Phi^{-1}(p)\sigma \tag{1}$$

where μ is a location parameter that depends on the predictor variables, σ is a scale parameter, and $\Phi^{-1}(p)$ is the standardized inverse cdf of the lifetime distribution, i.e.,

$$F(Y) = \Phi\left(\frac{Y-\mu}{\sigma}\right) \tag{2}$$

For such a model, lifetimes may be assumed to follow either a normal, logistic, or smallest extreme value distribution.

Log-Location-Scale Models

For this type of model, the percentiles of the lifetime distribution are related to the predictor variables through a log-linear function of the form

$$\log(Y_P) = \mu + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \Phi^{-1}(p)\sigma \tag{3}$$

where

$$F(Y) = \Phi\left(\frac{\log(Y)-\mu}{\sigma}\right) \tag{4}$$

For such a model, lifetimes may be assumed to follow either a lognormal, loglogistic, Weibull, or exponential distribution.

Analysis Summary

The *Analysis Summary* displays a table showing the estimated model and likelihood ratio tests for the significance of the model coefficients.

Life Data Regression - hours

Dependent variable: hours
Factors:
  voltage
  temperature

Number of uncensored values: 32
Number of right-censored values: 0

Estimated Regression Model - Weibull

		Standard	Lower 95.0%	Upper 95.0%
Parameter	*Estimate*	*Error*	*Conf. Limit*	*Conf. Limit*
CONSTANT	11.6981	1.96481	7.84716	15.5491
voltage	-0.00660564	0.000883368	-0.00833701	-0.00487426
temperature	-0.0200546	0.0110668	-0.0417451	0.00163591
SIGMA	0.312591	0.0432654	0.238321	0.410007

Log likelihood = -211.019

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
voltage	29.3505	1	0.0000
temperature	3.06457	1	0.0800

The table includes:

- **Data Summary:** a summary of the input data, including the number of observations *n* used to fit the model.

- **Estimated Regression Model:** estimates of the coefficients in the regression model, with standard errors and approximate confidence intervals.

- **Likelihood Ratio Tests:** tests run to determine whether or not the coefficients are significantly different from 0. Two-sided P-values are displayed. Small P-values (less than 0.05 if operating at the 5% significance level) correspond to statistically significant variables.

The above table shows the result of fitting a first-order model to the capacitor data, assuming a Weibull distribution for the failure times at fixed values of the predictor variables. The estimated model has parameters:

$$\mu = 11.6981 - 0.00660564\ \textit{voltage} - 0.0200546\ \textit{temperature} \qquad (5)$$

$$\sigma = 0.312591 \qquad (6)$$

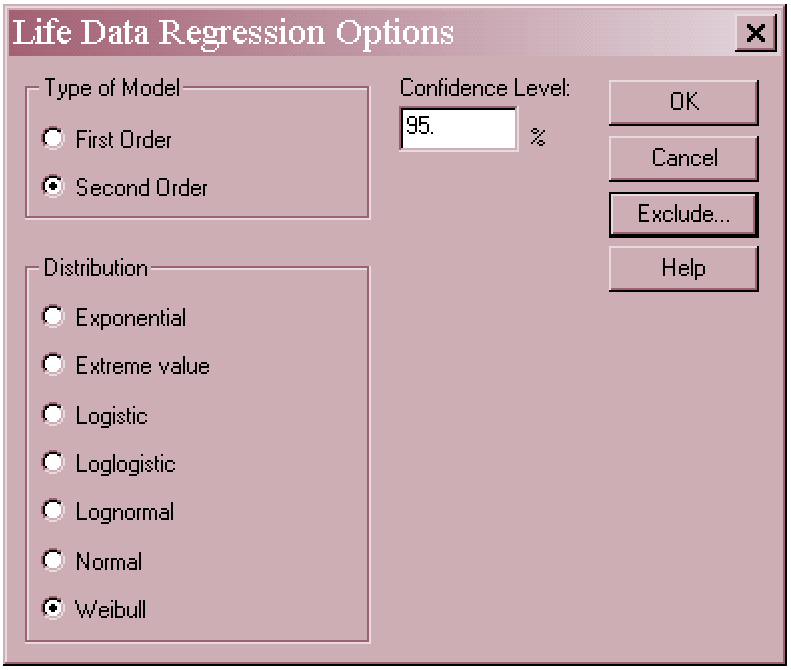
based on a log-linear model. The equation for the *p*-th percentile is

$$\text{time}_p = \exp(11.6981 - 0.00660564\ \textit{voltage} - 0.0200546\ \textit{temperature} + 0.312591 \log(-\log(1-p))) \qquad (7)$$

Both *voltage* and *temperature* have a negative effect on capacitor lifetimes. Voltage is highly significant, while temperature is significant at the 10% level but not at the 5% level.

Analysis Options

The statistical model to be fit is specified using *Analysis Options*:

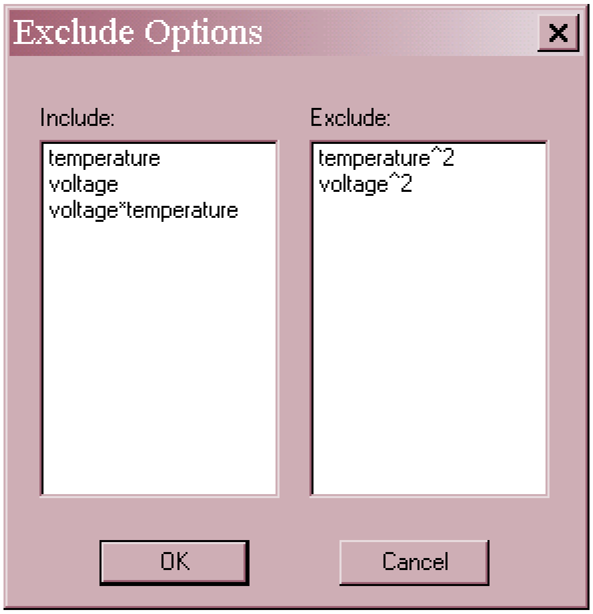


- **Type of Model:** Select *First Order* to fit a model involving only main effects of each factor. Select *Second Order* to include quadratic effects for the quantitative factors and 2-factor interactions between all of the variables.

- **Distribution:** the assumed distribution for the failure times at fixed values of the predictor variables.

- **Confidence Level:** percentage confidence for the interval estimates of the model coefficients.

- **Exclude:** Press this button to exclude specific terms from the model. A dialog box of the form shown below will be displayed:



Double click on an effect to move it from the *Include* field to the *Exclude* field or back again.

Example: Fitting a Model with an Interaction

To add an interaction to the model, select *Second Order* on the *Analysis Options* dialog box. Then press the *Exclude* button to remove *temperature*^{2} and *voltage*^{2} from the model, leaving the main effects and the cross-product *voltage*temperature*. The results of the fit are shown below.

Life Data Regression - hours

Dependent variable: hours
Factors:
  voltage
  temperature

Number of uncensored values: 32
Number of right-censored values: 0

Estimated Regression Model - Weibull

		Standard	*Lower 95.0%*	*Upper 95.0%*
Parameter	*Estimate*	*Error*	*Conf. Limit*	*Conf. Limit*
CONSTANT	9.06005	8.98877	-8.55765	26.6778
voltage	0.00297857	0.0319249	-0.0595933	0.0655504
temperature	-0.00508477	0.0510078	-0.105058	0.0948888
voltage*temperature	-0.0000543878	0.000181097	-0.000409332	0.000300556
SIGMA	0.311771	0.0432319	0.237577	0.409137

Log likelihood = -210.974

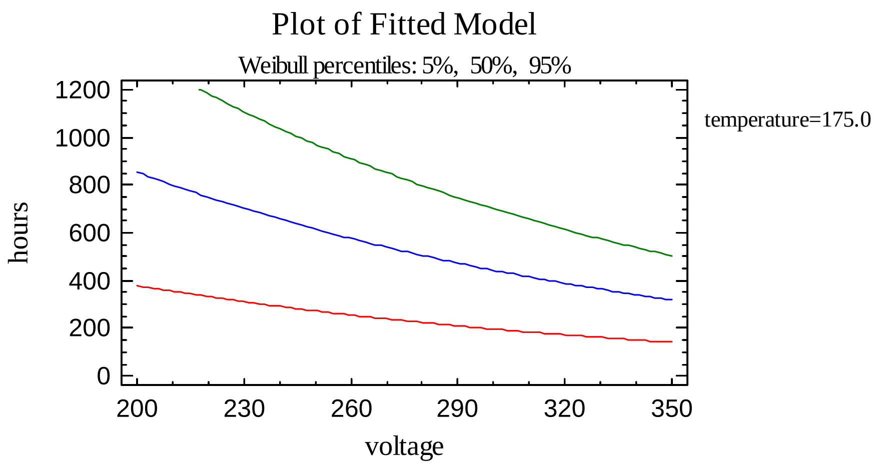
Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
voltage	0.00869322	1	0.9257
temperature	0.00995633	1	0.9205
voltage*temperature	0.0900258	1	0.7641

The likelihood ratio test for the cross-product term has a large P-Value, indicating that there is no significant interaction between *voltage* and *temperature*.

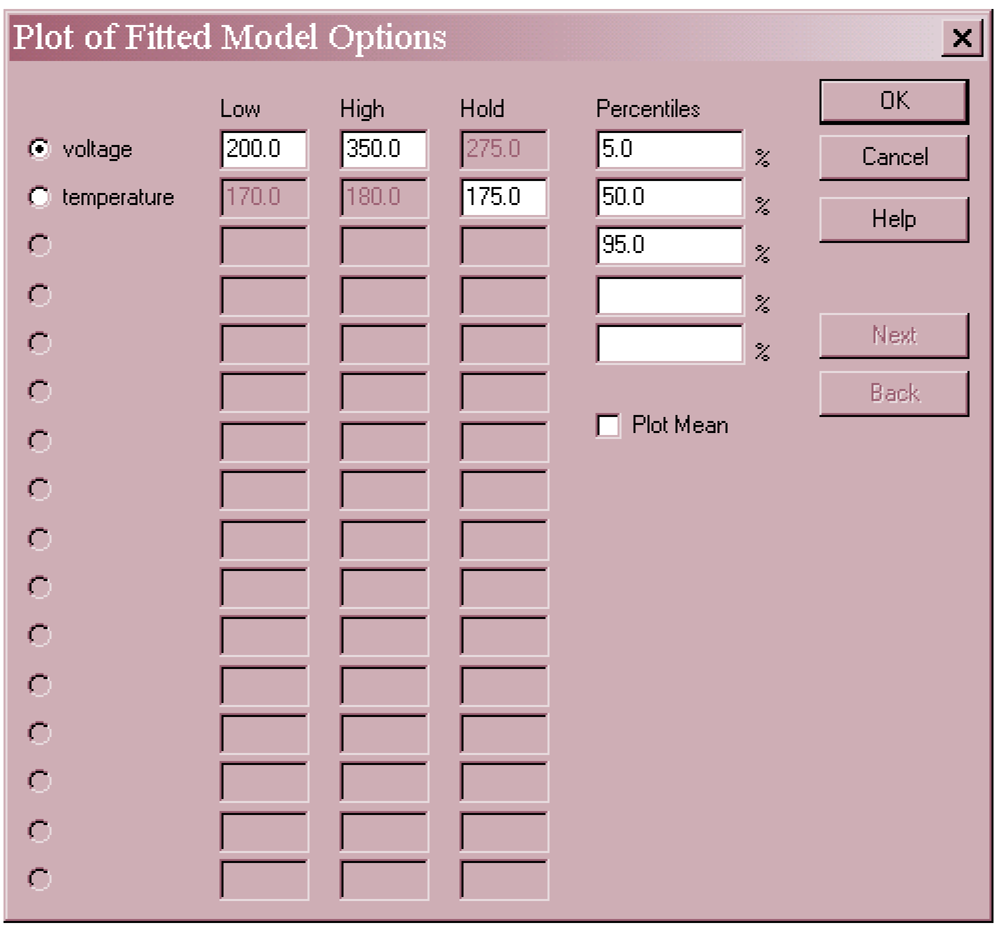
Plot of Fitted Model

The *Plot of Fitted Model* pane displays the percentiles as a function of any single variable X with all other variables set fixed at specified values.



For example, the above plot shows how the 5-th, 50-th, and 95-th percentiles vary as a function of *voltage*, with *temperature* set equal to 175. The mean failure time decreases as the voltage increases, with the variability decreasing as well.

Pane Options



- **Factors:** select one factor to plot on the horizontal axis, with lower and upper limits for the plot. For all other factors, specify values at which they should be held fixed.

- **Percentiles**: percentages of the desired percentiles.

- **Plot Mean**: include a line at the estimated mean failure time.

- **Next** and **Back:** used to display other factors when more than 16 are present.

Percentiles

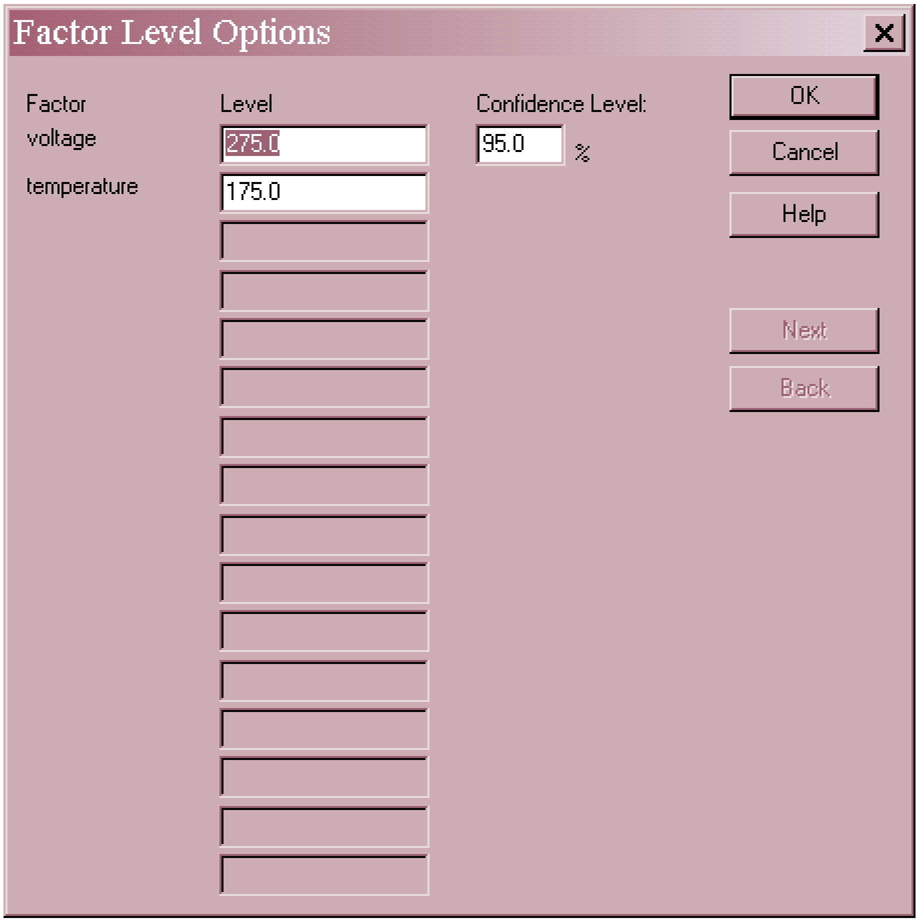
The *Percentiles* pane displays a table of estimated percentiles at a selected combination of the predictor variables.

Table of Percentiles for hours
voltage=275.0
temperature=175.0

		Standard	*Lower 95.0%*	*Upper 95.0%*
Percent	*Percentile*	*Error*	*Conf. Limit*	*Conf. Limit*
0.1	67.5506	21.7934	35.8931	127.13
0.5	111.788	28.418	67.9212	183.985
1.0	138.942	31.2525	89.4071	215.922
2.0	172.831	33.8385	117.751	253.675
3.0	196.498	35.1376	138.404	278.978
4.0	215.334	35.921	155.281	298.612
5.0	231.266	36.4308	169.834	314.92
6.0	245.231	36.7737	182.781	329.016
7.0	257.762	37.0057	194.543	341.525
8.0	269.198	37.1599	205.387	352.835
9.0	279.764	37.2571	215.494	363.203
10.0	289.623	37.3114	224.996	372.813
15.0	331.643	37.2028	266.186	413.197
20.0	366.192	36.7907	300.739	445.89
25.0	396.457	36.2715	331.376	474.321
30.0	424.014	35.7291	359.463	500.156
35.0	449.788	35.2101	385.811	524.374
40.0	474.4	34.7469	410.959	547.633
45.0	498.307	34.3677	435.302	570.432
50.0	521.89	34.1008	459.156	593.195
55.0	545.493	33.9785	482.801	616.325
60.0	569.466	34.0404	506.508	640.25
65.0	594.205	34.3382	530.575	665.466
70.0	620.206	34.9425	555.366	692.616
75.0	648.154	35.9563	581.377	722.602
80.0	679.11	37.5429	609.374	756.828
85.0	714.934	39.9931	640.693	797.777
90.0	759.558	43.931	678.156	850.732
91.0	770.256	45.0125	686.898	863.73
92.0	781.841	46.2402	696.267	877.932
93.0	794.534	47.6507	706.42	893.638
94.0	808.653	49.2971	717.581	911.283
95.0	824.682	51.2605	730.092	931.527
96.0	843.412	53.6757	744.506	955.457
97.0	866.285	56.7912	761.83	985.061
98.0	896.428	61.1553	784.233	1024.67
99.0	943.323	68.4714	818.231	1087.54
99.5	985.587	75.5529	848.093	1145.37
99.9	1070.79	91.0336	906.442	1264.94

Confidence intervals are included based on a large-sample normal approximation. For example, at a voltage = 275 and a temperature = 175, it is estimated that 50% of the capacitors will have failed after approximately 522 hours. The 95% confidence interval for that the 50-th percentile ranges from 459 hours to 593 hours.

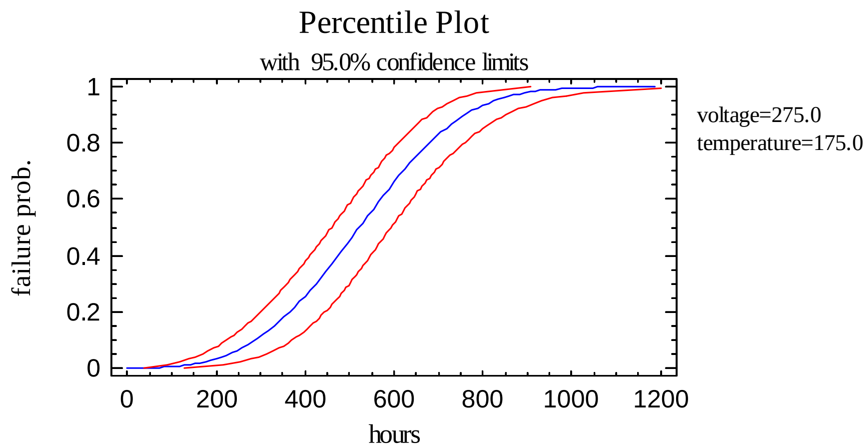
Pane Options



- **Level**: values of the predictor variables at which the percentiles are to be estimated.
- **Confidence Level**: percentage confidence for the interval estimates.
- **Next** and **Back:** used to display other factors when more than 16 are present.

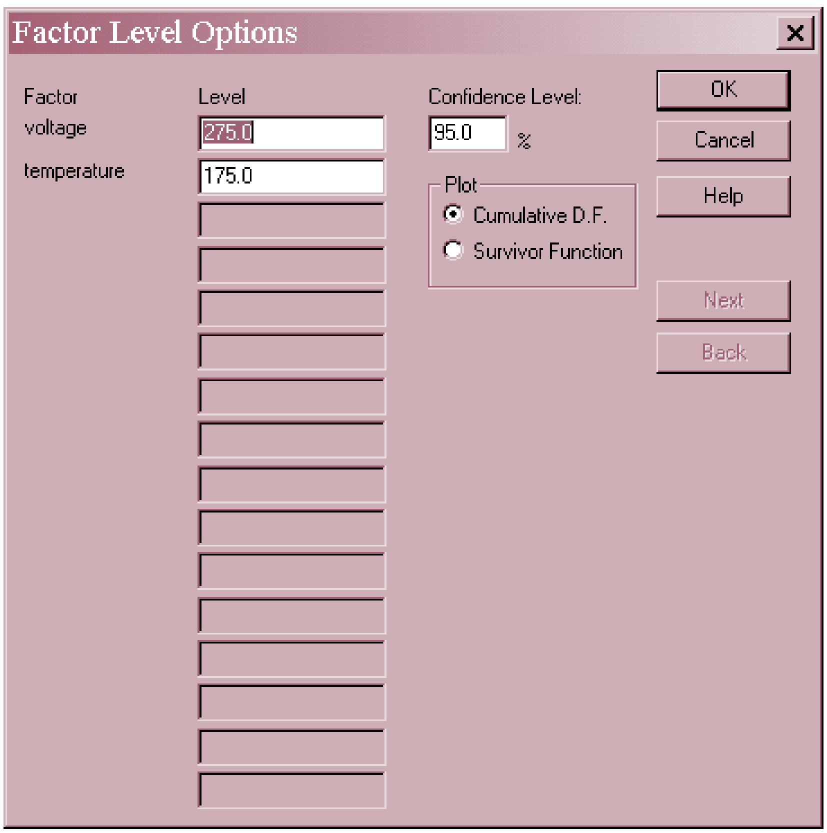
Percentile Plot

The *Percentile Plot* graphs the estimated percentiles at a selected combination of the predictor variables.



Confidence intervals are included based on a large-sample normal approximation.

Pane Options



- **Level**: value of the predictor variable at which the percentiles are to be estimated.

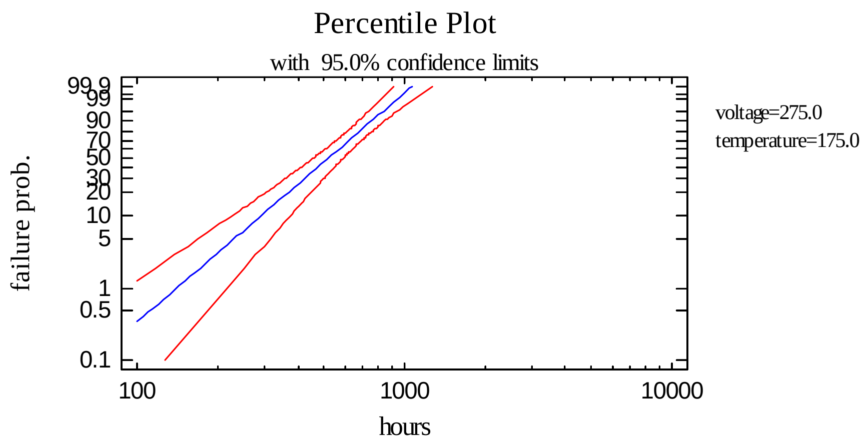
- **Confidence Level**: percentage confidence for the interval estimates.

- **Plot:** select *Cumulative D.F.* to plot the percentiles or *Survivor Function* to plot the estimated survival probabilities.

- **Next** and **Back:** used to display other factors when more than 16 are present.

Percentile Probability Plot

This plots graphs the estimated percentiles on a chart scaled so that the cumulative distribution function is a straight line.



Pane Options
The options are the same as for the *Percentile Plot*.

Predictions

The *Predictions* pane creates predictions using the fitted model. By default, the table includes a line for each row in the datasheet that has complete information on the X variables and a missing value for the Y variable. This allows you to add columns to the bottom of the datasheet corresponding to levels at which you want predictions without affecting the fitted model.

For example, suppose a prediction is desired for a capacitor subjected to a voltage of 275 and a temperature of 175. In row #33 of the datasheet, these values would be added but the *Hours* column would be left blank. The resulting table is shown below:

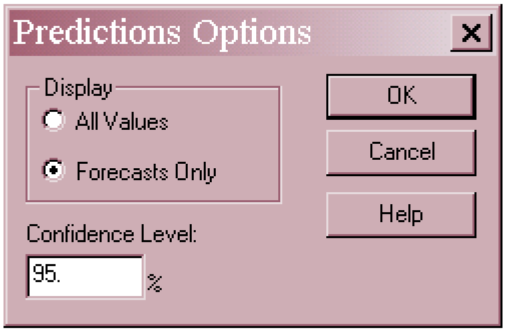
Predictions for hours

Row	*Observed Value*	*Fitted Value*	*Standard Error*	*Lower 95.0% CL for Mean*	*Upper 95.0% CL for Mean*
33		585.242	0.0584386	521.906	656.264

Included in the table are:

- **Row** - the row number in the datasheet.
- **Observed Value** - the observed values, Y_i**.**
- **Fitted Value** - the fitted values, given by $\hat{\mu}_i$ for location-scale models and $\exp(\hat{\mu}_i)$ for log-location-scale models.
- **Standard Error** – the standard errors corresponding to $\hat{\mu}_i$.
- **Confidence Limits** – approximate confidence limits for the *Fitted Values*.

Pane Options

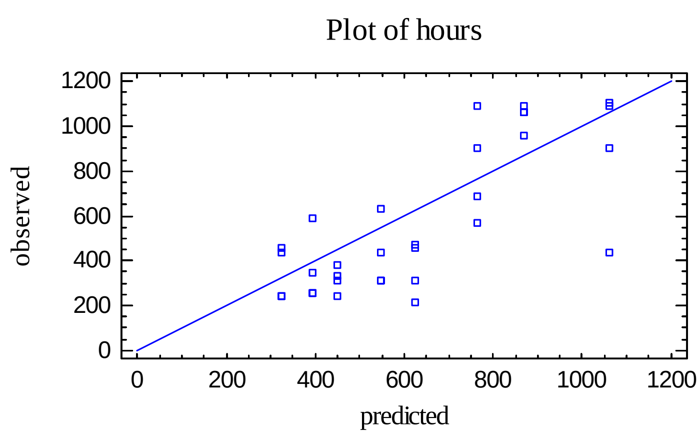


- **Display**: All rows may be displayed, or *Forecasts Only* (only those rows with missing values for the dependent variable).

- **Confidence Level**: percentage confidence for the interval estimates.

Observed versus Predicted

The *Observed versus Predicted* pane plots the observed failure times Y_i versus $\hat{\mu}_i$ for location-scale models and $\exp(\hat{\mu}_i)$ for log-location-scale models.



If the model fits well, the points should be randomly scattered around the diagonal line.

Residual Probability Plot

In all regression applications, it is important to calculate and plot the residuals. The *Life Data Regression* procedure creates three different types of residuals:

1. *Ordinary residuals*:

for location-scale models: $r_i = y_i - \hat{\mu}_i$ (8)

for log-location-scale models: $r_i = y_i - \exp(\hat{\mu}_i)$ (9)

2. *Standardized residuals*:

for location-scale models: $$e_i = \frac{y_i - \hat{\mu}}{\hat{\sigma}} \quad (10)$$

for log-location-scale models: $$e_i = \exp\left(\frac{\ln(y_i) - \hat{\mu}_i}{\hat{\sigma}}\right) \quad (11)$$

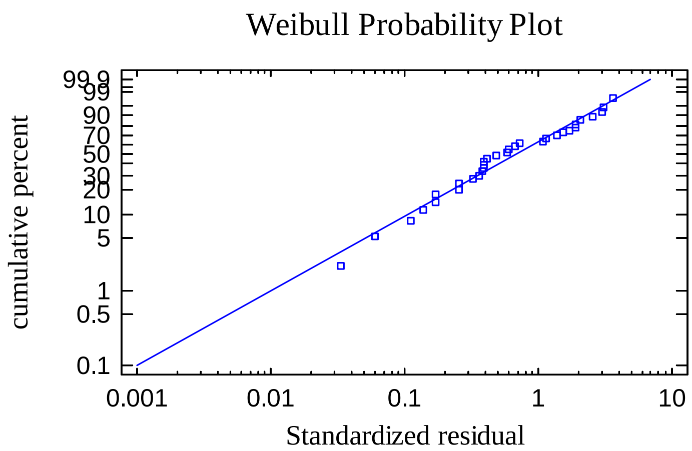
3. *Cox-Snell Residuals* – a type of Cox-Snell residuals constrained to lie between 0 and 1, defined by:

$$\hat{u}_i = \hat{F}(Y_i) \quad (12)$$

which is the estimated cumulative lifetime distribution evaluated at the observed failure time.

The *ordinary residuals* quantify the difference between the observed data values and the fitted values. The *standardized residuals* are scaled so that they should follow a standardized form of the assumed failure time distribution. The Cox-Snell residuals can be useful in identifying outliers.

The *Residual Probability Plot* displays the standardized residuals on a plot designed to help determine whether the assumed distribution of lifetimes is reasonable for the data:



If the selected distribution is adequate for the data, the points should lie along the diagonal reference line.

Unusual Residuals

The *Unusual Residuals* pane lists all observations that have unusually large residuals.

Unusual Residuals for hours

		Predicted		*Standardized*	*Cox-Snell*
Row	*Y*	*Y*	*Residual*	*Residual*	*Residual*

The table displays:

- *Row* – the row number in the data sheet.
- *Y* – the observed failure time (possibly censored).
- *Predicted Y* - the fitted values, given by $\hat{\mu}_i$ for location-scale models and $\exp(\hat{\mu}_i)$ for log-location-scale models.
- *Residual* – the ordinary residuals.
- *Standardized Residuals* – the standardized e_i.
- *Cox-Snell Residuals* – the Cox-Snell residuals constrained $\hat{u}_i$.

A row is added to the list corresponding to all Cox-Snell residuals that are less than 0.025 or greater than 0.975, i.e., any residuals outside of the central 95% of the estimated lifetime distribution. Particular attention should be given to any residuals outside of the interval

$$0.00135 \leq \hat{u}_i \leq .99865$$

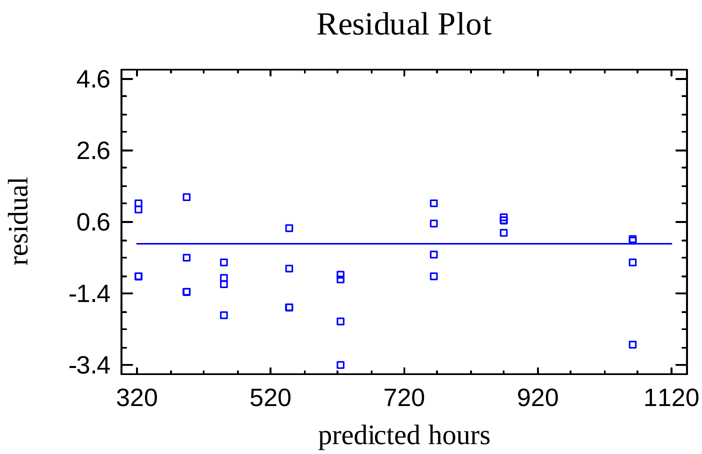
since that would be equivalent to being beyond 3 standard deviations if the distribution was Gaussian.

Residual Plots

Several other types of residual plots can be created:

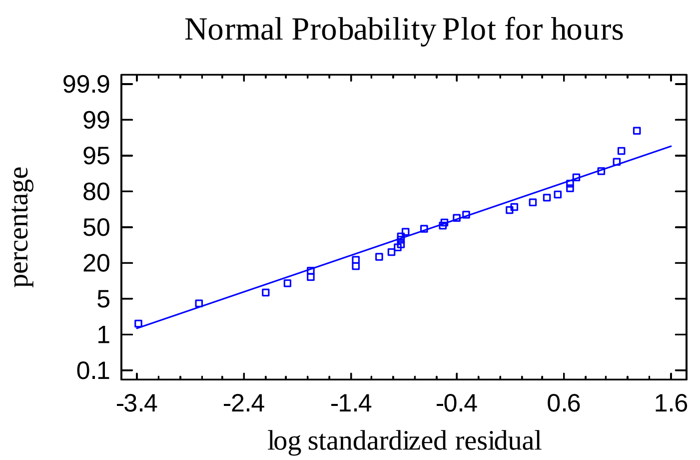
Scatterplot versus Predicted Value

This plot is helpful in visualizing whether the variability is constant or varies according to the magnitude of Y.



Normal Probability Plot

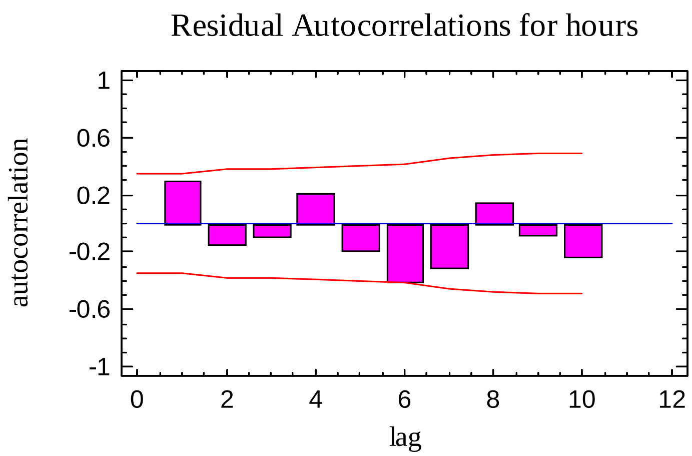
This plot can be used to determine whether or not the deviations around the line follow a normal distribution.



Although this plot is created in all regression procedures, the special *Residual Probability Plot* described earlier is more useful for life data residuals.

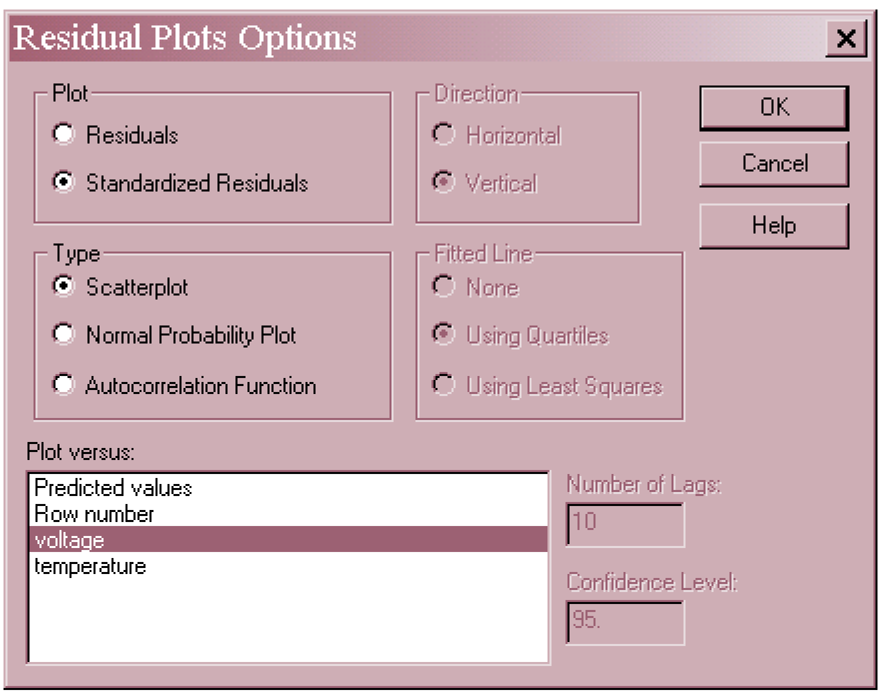
Residual Autocorrelations

This plot calculates the autocorrelation between residuals as a function of the number of rows between them in the datasheet.



It is only relevant if the data have been collected sequentially. Any bars extending beyond the probability limits would indicate significant dependence between residuals separated by the indicated "lag".

Pane Options



- **Plot:** the type of residuals to plot.

- **Type:** the type of plot to be created. A *Scatterplot* is used to test for curvature. A *Normal Probability Plot* is used to determine whether the model residuals come from a normal distribution. An *Autocorrelation Function* is used to test for dependence between consecutive residuals.

- **Plot Versus**: for a *Scatterplot*, the quantity to plot on the horizontal axis.

- **Number of Lags**: for an *Autocorrelation Function*, the maximum number of lags. For small data sets, the number of lags plotted may be less than this value.

- **Confidence Level:** for an *Autocorrelation Function*, the level used to create the probability limits.

Correlation Matrix

The *Correlation Matrix* displays estimates of the correlation between the estimated coefficients.

Correlation matrix for coefficient estimates

	CONSTANT	voltage	temperature
CONSTANT	1.0000	-0.1737	-0.9920
voltage	-0.1737	1.0000	0.0516
temperature	-0.9920	0.0516	1.0000

This table can be helpful in determining how well the effects of different independent variables have been separated from each other.

Save Results

The following results may be saved to the datasheet:

1. *Predicted Values* – the fitted values corresponding to each of the *n* observations.
2. *Standard Errors of Means* – the standard errors for the *n* fitted values.
3. *Lower Limits for Forecast Means* – the lower confidence limits for the fitted values.
4. *Upper Limits for Forecast Means* – the upper confidence limits for the fitted values.
5. *Residuals* – the *n* residuals r_i.
6. *Standardized Residuals* – the *n* standardized residuals e_i.
7. *Cox-Snell Residuals* - the *n* Cox-Snell residuals $\hat{u}_i$.
8. *Coefficients* – the estimated model coefficients.
9. *Percentages* – the percentages at which percentiles were calculated.
10. *Percentiles* – the estimated percentiles.
11. *Stnd. Error of Percentiles* – the standard errors of the estimated percentiles.
12. *Lower Percentile Conf. Limits* – lower confidence limits for the percentiles.
13. *Upper Percentile Conf. Limits* – upper confidence limits for the percentiles.

Calculations

Standardized Distributions

Logistic, loglogistic: $\Phi(z) = \exp(z)/[1+\exp(z)]$ (13)

Normal, lognormal: $\Phi(z) = \int_{-\infty}^{z} \left(1/\sqrt{2\pi}\right)\exp\left(-z^2/2\right)$ (14)

Smallest extreme value, Weibull, exponential: $\Phi(z) = 1-\exp[-\exp(z)]$ (15)

Likelihood Functions

Let δ_i = 1 for an exact failure time and 0 for a right-censored observation.

Location-Scale Models: $L(\beta,\sigma) = \prod_{i=1}^{n}\left[\frac{1}{\sigma}\phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]^{\delta_i}\left[1-\Phi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]^{1-\delta_i}$ (16)

Log-Location-Scale Models: $L(\beta,\sigma) = \prod_{i=1}^{n}\left[\frac{1}{\sigma}\phi\left(\frac{\log(y_i)-\mu_i}{\sigma}\right)\right]^{\delta_i}\left[1-\Phi\left(\frac{\log(y_i)-\mu_i}{\sigma}\right)\right]^{1-\delta_i}$ (17)

Standard Errors for Coefficients

Determined from the partial derivatives evaluated at the maximum likelihood estimates. Confidence intervals are based on a large-sample normal approximation.

Mean Failure Times

Distribution	*E(Y)*
Normal	μ
Lognormal	$\exp(\mu + \sigma^2/2)$
Logistic	μ
Loglogistic	$\exp(\mu)\Gamma(1+\sigma)\Gamma(1-\sigma)$
Smallest extreme value	$\mu-0.5772\sigma$
Weibull	$\exp(\mu)\Gamma(1+\sigma)$
Exponential	$\exp(\mu)\Gamma(2)$

