

Cox Proportional Hazards

Summary

The **Cox Proportional Hazards** procedure is designed to fit a semi-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. Unlike the *Life Data Regression* procedure, no assumption is made about the distribution of failure times. The only assumption is that the predictors act in a multiplicative manner on the hazard function. Failure times may be censored or uncensored.

The output of the procedure includes an estimate of the hazard function and tests of significance on each of the predictor variables. Plots of the estimated survival and hazard functions may be created, as well as several residual plots.

Sample StatFolio: *coxph.sgp*

Sample Data:

The file *nephrectomy.sgd* contains data from a study of $n = 36$ patients with malignant tumors in their kidney, reported by Collett (1994). The patients are divided into three age groups. Some patients in each group had undergone a nephrectomy (surgical removal of the kidney), while others had not. A portion of the file is shown below.

<i>Survival time</i>	<i>Censored</i>	<i>Nephrectomy</i>	<i>Age</i>
9	0	0	1
6	0	0	1
21	0	0	1
15	0	0	2
8	0	0	2
17	0	0	2
12	0	0	3
104	1	1	1
9	0	1	1
56	0	1	1
35	0	1	1
35	0	1	1
52	0	1	1
68	0	1	1
77	1	1	1
84	0	1	1
8	0	1	1
38	0	1	1
...	

Survival time is in months. *Censored* equals 1 if *Survival time* is a censored observation and 0 otherwise. *Nephrectomy* is coded as 1 for those patients whose kidneys were removed and 0 for those who did not have the surgery. *Age* is coded as a 1 for patients under 60 years of age, 2 for those between 60 and 70, and 3 for patients more than 70 years old.

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.

In the example, there are two categorical factors.

Statistical Model

The Cox proportional hazards model assumes that the effect of the predictor variables on the hazard function can be expressed as a product of a term involving the predictor variables X and a baseline hazard function. STATGRAPHICS fits a model of the form

$$h_x(t) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k) h_0(t) \tag{1}$$

where $h_0(t)$ is a *baseline hazard function* appropriate when all predictor variables equal 0. The function $\exp(X\beta)$ is thus the ratio of the hazard function for an observation with observed predictor variables X and the baseline hazard function.

Analysis Summary

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

<u>Cox Proportional Hazards - Survival time</u>				
Dependent variable: Survival time				
Censoring: Censored				
Factors:				
Nephrectomy				
Age				
Number of uncensored values: 32				
Number of right-censored values: 4				
Estimated Regression Model				
		Standard	Lower 95.0%	Upper 95.0%
Parameter	Estimate	Error	Conf. Limit	Conf. Limit
Nephrectomy=1	-1.41108	0.377288	-2.15056	-0.67161
Age=2	0.0124456	0.330352	-0.635033	0.659924
Age=3	1.34132	0.448089	0.463082	2.21956
Log likelihood = -82.7542				
Likelihood Ratio Tests				
Factor	Chi-Squared	Df	P-Value	
Nephrectomy	6.66386	1	0.0098	
Age	4.73827	2	0.0936	

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:** maximum likelihood estimates of the coefficients β 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.

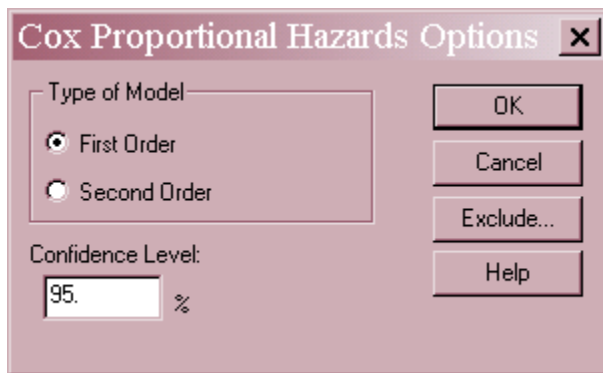
The above table shows the result of fitting a first-order model to the example data. The estimated model is:

$$h_x(t) = \exp[-1.41108 (Nephrectomy=1) + 0.0124456 (Age=2) + 1.34132 (Age=3)] h_0(t)$$

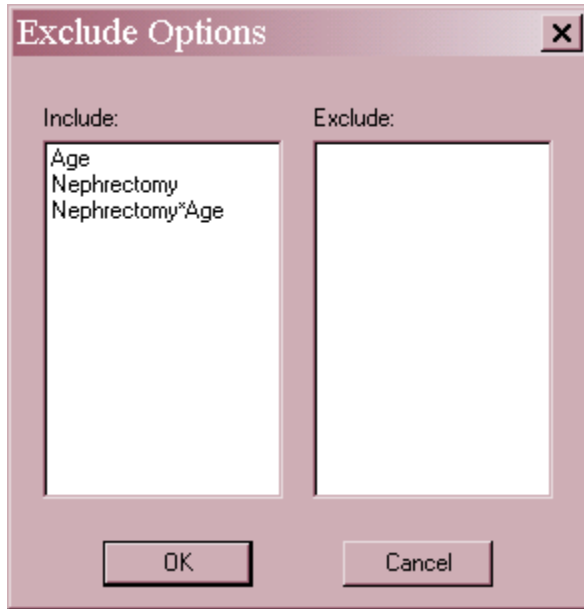
The terms *Nephrectomy=1*, *Age=2*, and *Age=3* are Boolean indicator variables that take the value 0 if false and 1 if true. The likelihood ratio tests indicate a significant effect from having a nephrectomy, but the effect of age is not significant at the 5% significance level. Note that the effect of *Nephrectomy* is negative, indicating that it successfully *reduces* the hazard function and thus improves survival.

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.
- **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. The results of the fit are shown below.

Cox Proportional Hazards - Survival time
 Dependent variable: Survival time
 Censoring: Censored
 Factors:
 Nephrectomy
 Age

Number of uncensored values: 32
 Number of right-censored values: 4

Estimated Regression Model

Parameter	Estimate	Standard Error	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Nephrectomy=1	-1.94365	0.562645	-3.04642	-0.840886
Age=2	0.00491685	0.730185	-1.42622	1.43606
Age=3	0.0650905	1.03264	-1.95885	2.08903
Nephrectomy=1*Age=2	-0.0504223	0.819766	-1.65714	1.55629
Nephrectomy=1*Age=3	2.00278	1.14729	-0.245857	4.25143

Log likelihood = -81.2395

Likelihood Ratio Tests

Factor	Chi-Squared	Df	P-Value
Nephrectomy	5.42807	1	0.0198
Age	0.00322947	2	0.9984
Nephrectomy*Age	3.0294	2	0.2199

Two additional terms have been added to the model, equal to cross-products between the indicator variables for the main effects. The likelihood ratio test indicates that the estimated interactions are not significant.

Baseline Functions

The baseline hazard function is the hazard function corresponding to the situation when all of the predictor variables equal 0. For a categorical factor, this equates to the first level of the factor, when all of the associated indicator variables are set to 0. The *Baseline Functions* table displays the baseline hazard and related functions.

Baseline Functions				
Survival time	Alpha	Hazard Function	Survivor Function	Cumulative Hazard
0.0-		0.0	1.0	0.0
5.0-	0.950245	0.0497546	0.950245	0.051035
6.0-	0.896152	0.103848	0.851564	0.160681
8.0-	0.886837	0.113163	0.755198	0.280775
9.0-	0.762896	0.237104	0.576137	0.551409
10.0-	0.926728	0.0732716	0.533923	0.627504
12.0-	0.910413	0.0895868	0.48609	0.721361
14.0-	0.8918	0.1082	0.433495	0.835875
15.0-	0.883764	0.116236	0.383107	0.95944
17.0-	0.868258	0.131742	0.332636	1.10071
18.0-	0.715287	0.284713	0.23793	1.43578
21.0-	0.814989	0.185011	0.193911	1.64036
26.0-	0.617936	0.382064	0.119824	2.12173
35.0-	0.768142	0.231858	0.0920421	2.38551
36.0-	0.557136	0.442864	0.05128	2.97045
38.0-	0.720924	0.279076	0.036969	3.29768
48.0-	0.70074	0.29926	0.0259056	3.65329
52.0-	0.440307	0.559693	0.0114064	4.47358
56.0-	0.617835	0.382165	0.00704729	4.95511
68.0-	0.579434	0.420566	0.00408344	5.50082
72.0-	0.532782	0.467218	0.00217558	6.13046
84.0-	0.402576	0.597424	0.000416131	7.78451
108.0-	0.191974	0.808026	0.0000247665	10.606
115.0-	0.0		0.0	

The hazard function is a step function that changes immediately after each failure. Displayed in the table are:

- **Survival time** – the time t corresponding to each unique failure.
- **Alpha** – $\hat{\alpha}_j$, where $(1 - \hat{\alpha}_j)$ is the contribution to the hazard function corresponding to the indicated failure time (see the *Calculations* section).
- **Hazard function** – the baseline hazard function, computed from

$$\hat{h}_0(t_j) = 1 - \hat{\alpha}_j \tag{2}$$

- **Survivor function** – $\hat{S}_0(t)$, the estimated probability of surviving until time t , computed from

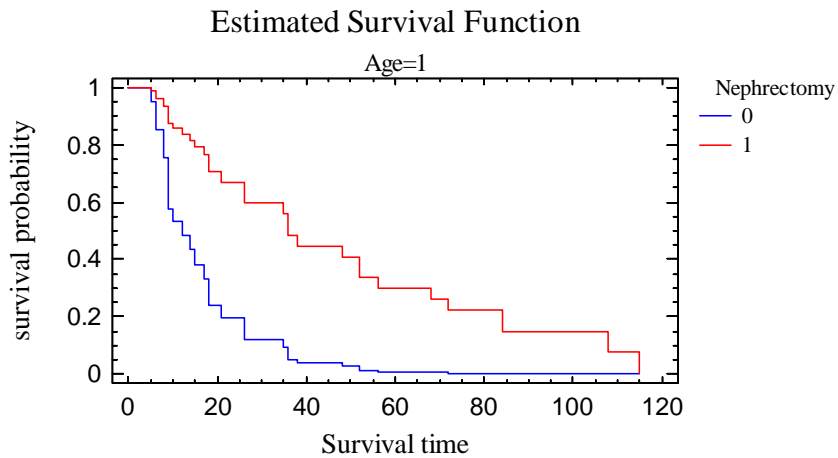
$$\hat{S}_0(t) = \prod_{j|t_{(j)} < t} \hat{\alpha}_j \tag{3}$$

- **Cumulative hazard** – the cumulative hazard function, computed from

$$\hat{H}_0(t) = -\ln[\hat{S}_0(t)] \quad (4)$$

Survival Function

The *Survival Function* pane displays the estimated survival function for selected levels of one predictor variable, at fixed values of the other predictors.



If the selected factor is categorical, a separate curve will be displayed for each level of that factor. If the selected factor is quantitative, curves will be displayed at the low and high values displayed on the *Pane Options* dialog box.

For example, the plot above shows the estimated survival function for patients in the group who had a nephrectomy and for those who did not, for $Age = 1$. Note the considerable improvement in survival for patients that had the surgery.

Pane Options

	Low	High	Hold
<input checked="" type="radio"/> Nephrectomy			0
<input type="radio"/> Age			1
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			
<input type="radio"/>			

Confidence Level: 95.0 %

Buttons: OK, Cancel, Help, Next, Back

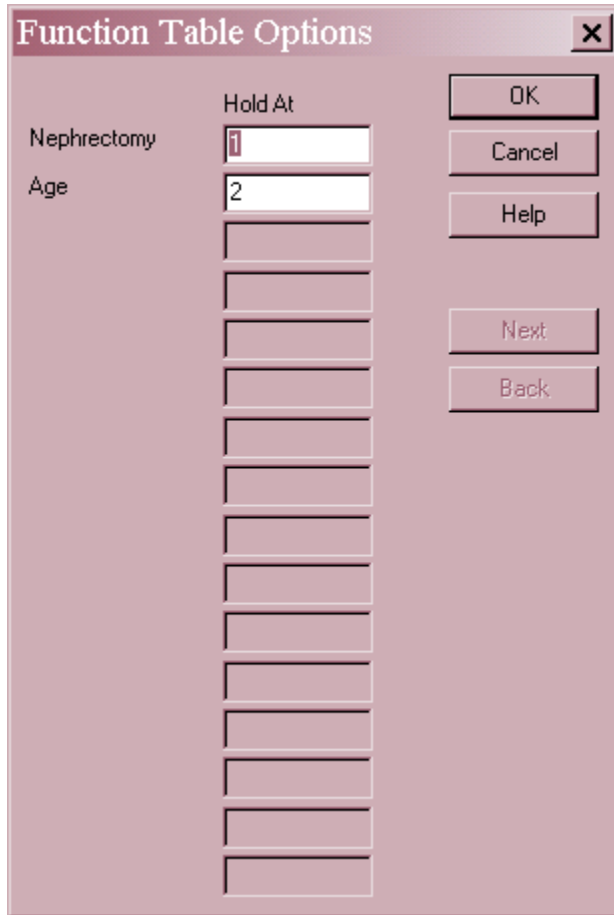
- **Factor:** Select the factor to be displayed at multiple levels.
- **Low and High:** If the selected factor is quantitative, the two levels at which to display the functions.
- **Hold:** for unselected factors, the value to fix the variable at when estimating the functions.
- **Next and Back:** used to display other factors when more than 16 are present.

Function Table

The *Function Table* displays the estimated hazard function, survival function, and cumulative hazard function at specified values of the predictor variables.

Function Table			
Nephrectomy=1			
Age=2			
	<i>Hazard</i>	<i>Survivor</i>	<i>Cumulative</i>
<i>Survival time</i>	<i>Function</i>	<i>Function</i>	<i>Hazard</i>
0.0-	0.0	1.0	0.0
5.0-	0.0122861	0.987477	0.0126022
6.0-	0.0256436	0.961099	0.0396774
8.0-	0.0279438	0.933016	0.0693327
9.0-	0.0585489	0.872702	0.136161
10.0-	0.0180932	0.856457	0.154952
12.0-	0.022122	0.836835	0.178128
14.0-	0.0267183	0.813503	0.206405
15.0-	0.0287026	0.789056	0.236918
17.0-	0.0325314	0.762006	0.271801
18.0-	0.0703052	0.701495	0.354541
21.0-	0.0456853	0.666937	0.405059
26.0-	0.0943443	0.592191	0.523925
35.0-	0.0572534	0.554848	0.589061
36.0-	0.109358	0.480223	0.733504
38.0-	0.0689132	0.442947	0.814306
48.0-	0.0738971	0.405709	0.90212
52.0-	0.138207	0.331319	1.10467
56.0-	0.0943693	0.294175	1.22358
68.0-	0.103852	0.257089	1.35833
72.0-	0.115372	0.220069	1.51381
84.0-	0.147524	0.146277	1.92225
108.0-	0.199528	0.0728773	2.61898
115.0-		0.0	

The values of the predictor variables are set using *Pane Options*.

Pane Options

The image shows a dialog box titled "Function Table Options" with a close button (X) in the top right corner. The dialog is divided into two main sections. On the left, there is a list of predictor variables: "Nephrectomy" and "Age". To the right of this list is a column labeled "Hold At" containing a series of input boxes. The first box for "Nephrectomy" contains the number "1", and the second box for "Age" contains the number "2". Below these are 14 empty input boxes. On the right side of the dialog, there are several buttons: "OK", "Cancel", "Help", "Next", and "Back".

- **Hold At:** the values of the predictor variables at which the functions are to be estimated.
- **Next and Back:** used to display other factors when more than 16 are present.

Residuals

The *Residuals* table displays several types of residuals.

Residuals					
		<i>Cox-Snell</i>	<i>Modified C.S.</i>	<i>Martingale</i>	<i>Deviance</i>
<i>Row</i>	<i>Survival time</i>	<i>Residual</i>	<i>Residual</i>	<i>Residual</i>	<i>Residual</i>
1	9.0	0.551409	0.551409	0.448591	0.541641
2	6.0	0.160681	0.160681	0.839319	1.40643
3	21.0	1.64036	1.64036	-0.640358	-0.53934
4	15.0	0.971456	0.971456	0.0285445	0.0288207
5	8.0	0.284291	0.284291	0.715709	1.0412
6	17.0	1.11449	1.11449	-0.114491	-0.110392
7	12.0	2.75855	2.75855	-1.75855	-1.21971
8	104.0	1.71699	2.71699	-1.71699	-1.8531
9	9.0	0.134477	0.134477	0.865523	1.51052
10	56.0	1.20845	1.20845	-0.208448	-0.195505
11	35.0	0.581776	0.581776	0.418224	0.496882
12	52.0	1.09101	1.09101	-0.0910117	-0.0883886
13	68.0	1.34153	1.34153	-0.341534	-0.308934
14	77.0	1.49509	2.49509	-1.49509	-1.72921
15	84.0	1.71699	1.71699	-0.716988	-0.593998
16	8.0	0.0684752	0.0684752	0.931525	1.8707
17	38.0	0.804234	0.804234	0.195766	0.210233
18	72.0	1.49509	1.49509	-0.49509	-0.431054
19	36.0	0.724431	0.724431	0.275569	0.30594
20	48.0	0.890962	0.890962	0.109038	0.113274
21	26.0	0.517445	0.517445	0.482555	0.593796
22	108.0	1.71699	1.71699	-0.716988	-0.593998
23	5.0	0.0124464	0.0124464	0.987554	2.60721
24	108.0	1.73849	2.73849	-1.73849	-1.86467
25	26.0	0.523925	0.523925	0.476075	0.583663
26	14.0	0.206405	0.206405	0.793595	1.25245
27	115.0				
28	52.0	1.10467	1.10467	-0.104675	-0.101231
29	5.0	0.0126022	1.0126	-0.0126022	-0.158759
30	18.0	0.354541	0.354541	0.645459	0.884841
31	36.0	0.733504	0.733504	0.266496	0.294708
32	9.0	0.136161	0.136161	0.863839	1.50338
33	10.0	0.58522	0.58522	0.41478	0.491909
34	9.0	0.514253	0.514253	0.485747	0.598821
35	18.0	1.33903	1.33903	-0.339029	-0.306869
36	6.0	0.149853	0.149853	0.850147	1.44772

Modified Cox-Snell residuals based on delta = 1.0

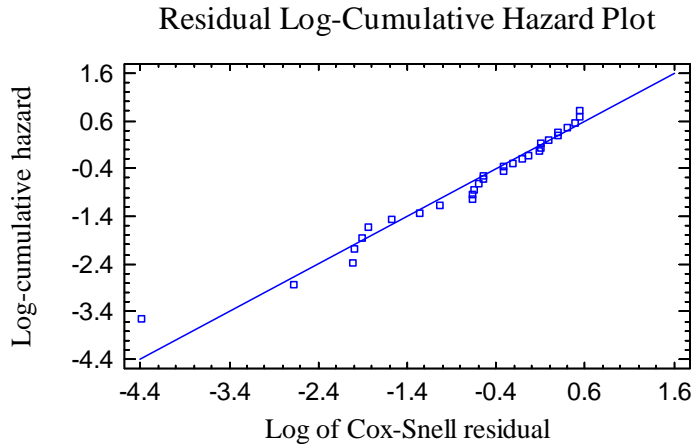
Included are:

- *Cox-Snell Residuals* – These residuals should behave like a sample from an exponential distribution with a mean equal to 1. Censored observations will yield censored residuals.
- *Modified Cox-Snell Residuals* – These residuals are created by adding a positive quantity Δ to all censored residuals. By default, $\Delta = 1$ but may be changed using *Pane Options* on the *Residual Plots* pane. Another popular value is $\Delta = 0.693$. Note: these residuals are not displayed if there are no censored observations.
- *Martingale residuals* – These residuals are derived from the *Cox-Snell Residuals* by multiplying them by -1 and adding 1.0 for each uncensored observation to shift the mean to 0. These residuals behave similarly to residuals encountered when fitting other linear models, although they are not symmetrically distributed around zero.

- *Deviance Residuals* – They are related to the Martingale residuals but are transformed so as to have a more symmetric distribution.

Residual Log-Cumulative Hazard Plot

If the model fits the data well, the Cox-Snell residuals should behave like a sample from an exponential distribution with a mean of 1. A useful plot for testing this hypothesis is the *Residual Log-Cumulative Hazard Plot*.



The log of the Cox-Snell residuals for each of the uncensored observations is plotted on the horizontal axis. The vertical axis shows the log of the estimated cumulative hazard function of the Cox-Snell residuals, estimated using the Kaplan-Meier procedure. If the residuals act like a sample from a unit exponential distribution, they should lie along the 45° diagonal line.

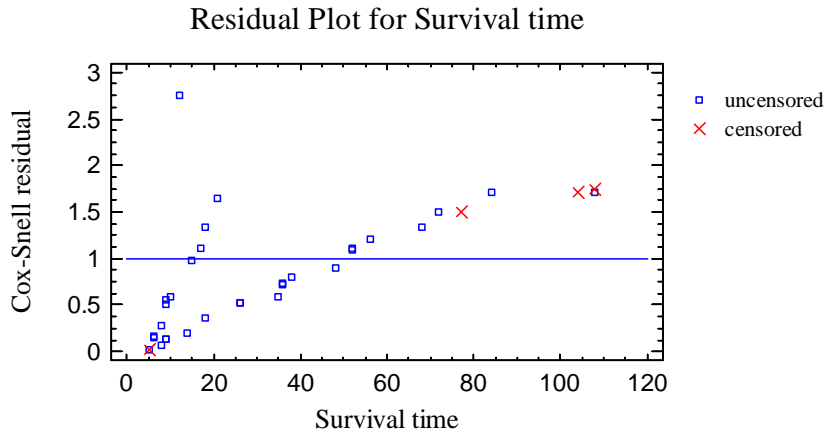
Except for possibly one residual at the lower end, the points in the plot above lie reasonably close to the line.

Residual Plots

The residuals may also be plotted versus other quantities.

Scatterplot versus Survival Time

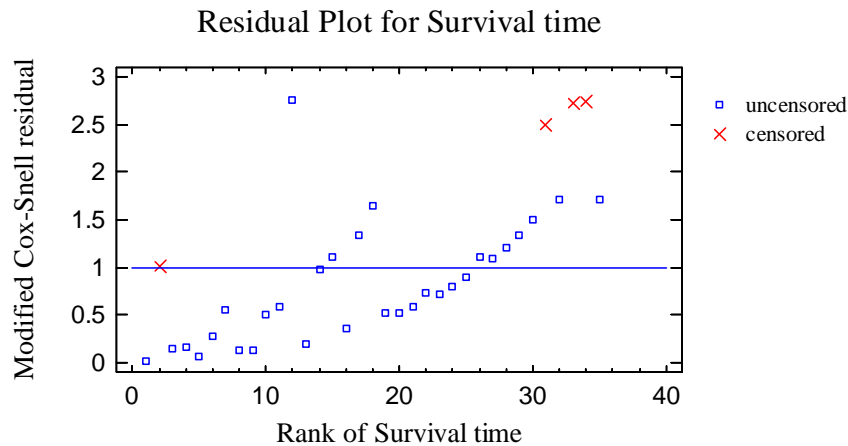
This plot is helpful in visualizing whether there are unusual outliers.



For Cox-Snell and modified Cox-Snell residuals, a reference line is added to the plot at 1, since the residuals should follow a unit exponential distribution.

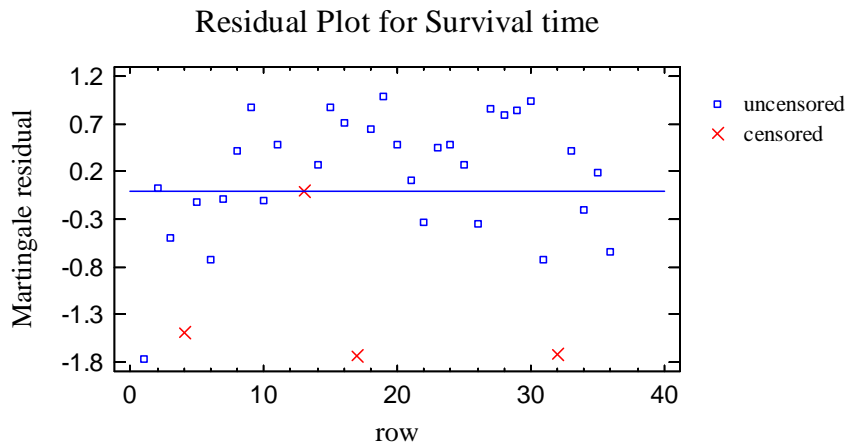
Scatterplot versus Rank Survival Time

Since the rank order of the survival times affects the estimates of the model coefficients, a plot of the residuals versus the ranks of the survival times may help in judging whether the current model is adequate for the data.



Scatterplot versus Row Number

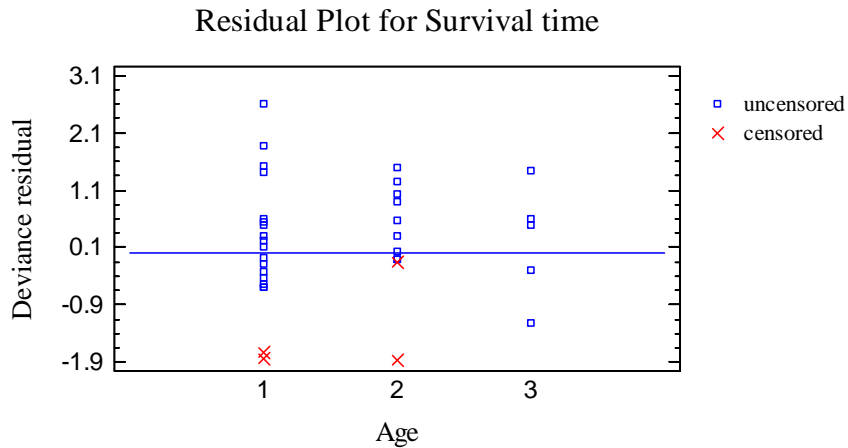
This plot is helpful in visualizing any dependence on the order within the datasheet.

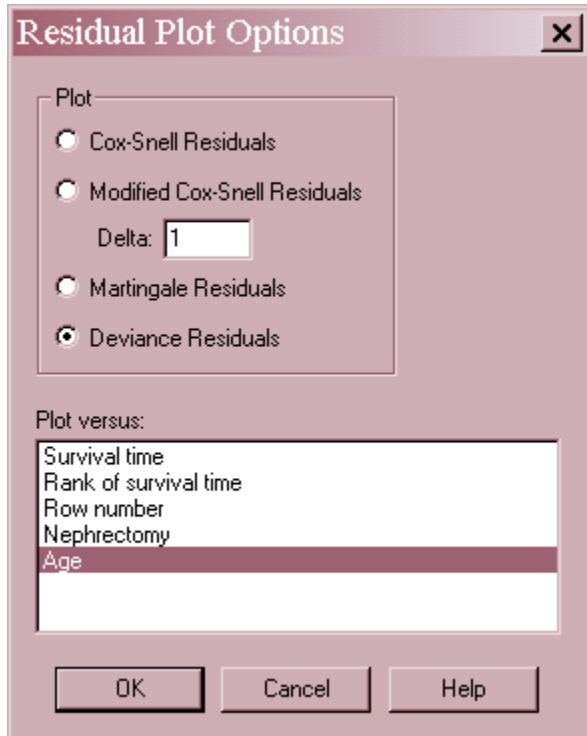


For Martingale and Deviance residuals, a reference line is added to the plot at 0.

Scatterplot versus Factor

This plot is helpful in visualizing any lack of fit with respect to a given variable.



Pane Options

- **Plot:** type of residuals to plot. For Modified Cox-Snell Residuals, Delta is added to each residual that corresponds to a censored observation.
- **Plot versus:** the quantity plotted on the horizontal axis.

Save Results

The following results may be saved to the datasheet:

1. *Survival Function* – the estimated survival function corresponding to each of the n observations.
2. *Hazard Function* – the estimated hazard function corresponding to each of the n observations.
3. *Cumulative Hazard Function* – the estimated cumulative hazard function corresponding to each of the n observations.
4. *Cox-Snell Residuals* - the n Cox-Snell residuals.
5. *Modified Cox-Snell Residuals* - the n modified Cox-Snell residuals.
6. *Martingale Residuals* - the n Martingale residuals.
7. *Deviance Residuals* - the n deviance residuals.
8. *Coefficients* – the estimated model coefficients.

Calculations

Let d_j be the number of failures at each of r unique ordered failure times $y_{(j)}$. Let $R(y_{(j)})$ be the set of items at risk at time $y_{(j)}$ and $D(y_{(j)})$ be the set of items that failed at time $y_{(j)}$.

Likelihood Function

$$L = \prod_{j=1}^r \left\{ \frac{\exp(s_j \beta)}{\left[\sum_{l \in R(y_{(j)})} \exp(x_l \beta) \right]^{d_j}} \right\} \quad (5)$$

where s_j is a vector containing the sum of the predictor variables for all failures at time $y_{(j)}$.

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.

Baseline Hazard Function

$$h_0(y_{(j)}) = 1 - \hat{\alpha}_j \quad (6)$$

where the α 's are found by solving

$$\sum_{l \in D(y_{(j)})} \frac{\exp(x_l \hat{\beta})}{1 - \hat{\alpha}_j^{\exp(x_l \hat{\beta})}} = \sum_{l \in R(y_{(j)})} \exp(x_l \hat{\beta}) \quad (7)$$

Residuals

Let $\delta_i = 1$ for a failure time and 0 for a censored observation.

$$\text{Cox-Snell: } r_i = \exp(x_i \beta) \hat{H}_o(y_i) \quad (8)$$

$$\text{Modified Cox-Snell: } r'_i = \begin{cases} r_i & \text{if } \textit{uncensored} \\ r_i + \Delta & \text{if } \textit{censored} \end{cases} \quad (9)$$

$$\text{Martingale: } m_i = \delta_i - r_i \quad (10)$$

$$\text{Deviance: } d_i = \text{sgn}(m_i) \sqrt{-2 \{m_i + \delta_i \log(\delta_i - m_i)\}} \quad (11)$$