

## Chapter 565

# Cox Regression

---

## Introduction

This procedure performs Cox (proportional hazards) regression analysis, which models the relationship between a set of one or more covariates and the hazard rate. Covariates may be discrete or continuous. Cox's proportional hazards regression model is solved using the method of marginal likelihood outlined in Kalbfleisch (1980).

This routine can be used to study the impact of various factors on survival. You may be interested in the impact of diet, age, amount of exercise, and amount of sleep on the survival time after an individual has been diagnosed with a certain disease such as cancer. Under normal conditions, the obvious statistical tool to study the relationship between a response variable (survival time) and several explanatory variables would be multiple regression. Unfortunately, because of the special nature of survival data, multiple regression is not appropriate. Survival data usually contain censored data, and the distribution of survival times is often highly skewed. These two problems invalidate the use of multiple regression. Many alternative regression methods have been suggested. The most popular method is the proportional hazard regression method developed by Cox (1972). Another method, Weibull regression, is available in **NCSS** in the Distribution Regression procedure.

---

## Further Reading

Several books provide in depth coverage of Cox regression. These books assume a familiarity with basic statistical theory, especially with regression analysis. Collett (1994) provides a comprehensive introduction to the subject. Hosmer and Lemeshow (1999) is almost completely devoted to this subject. Therneau and Grambsch (2000) provide a complete and up-to-date discussion of this subject. We found their discussion of residual analysis very useful. Klein and Moeschberger (1997) provides a very readable account of survival analysis in general and includes a lucid account of Cox regression.

---

## The Cox Regression Model

*Survival analysis* refers to the analysis of elapsed time. The response variable is the time between a *time origin* and an *end point*. The end point is either the occurrence of the event of interest, referred to as a *death* or *failure*, or the end of the subject's participation in the study. These elapsed times have two properties that invalidate standard statistical techniques, such as t-tests, analysis of variance, and multiple regression. First of all, the time values are often positively skewed. Standard statistical techniques require that the data be normally distributed. Although this skewness could be corrected with a transformation, it is easier to adopt a more realistic data distribution.

The second problem with survival data is that part of the data are *censored*. An observation is censored when the end point has not been reached when the subject is removed from study. This may be because the study ended before the subject's response occurred, or because the subject withdrew from active participation. This may be because the subject died for another reason, because the subject moved, or

## Cox Regression

because the subject quit following the study protocol. All that is known is that the response of interest did not occur while the subject was being studied.

When analyzing survival data, two functions are of fundamental interest—the *survivor function* and the *hazard function*. Let  $T$  be the survival time. That is,  $T$  is the elapsed time from the beginning point, such as diagnosis of cancer, and death due to that disease. The values of  $T$  can be thought of as having a *probability distribution*. Suppose the *probability density function* of the random variable  $T$  is given by  $f(T)$ . The *probability distribution function* of  $T$  is then given by

$$\begin{aligned} F(T) &= \Pr(t < T) \\ &= \int_0^T f(t) dt \end{aligned}$$

The *survivor function*,  $S(T)$ , is the probability that an individual survives past  $T$ . This leads to

$$\begin{aligned} S(T) &= \Pr(T \geq t) \\ &= 1 - F(T) \end{aligned}$$

The *hazard function* is the probability that a subject experiences the event of interest (death, relapse, etc.) during a short time interval given that the individual has survived up to the beginning of that interval. The mathematical expression for the hazard function is

$$\begin{aligned} h(T) &= \lim_{\Delta T \rightarrow 0} \frac{\Pr(T \leq t < (T + \Delta T) | T \leq t)}{\Delta T} \\ &= \lim_{\Delta T \rightarrow 0} \frac{F(T + \Delta T) - F(T)}{\Delta T} \\ &= \frac{f(T)}{S(T)} \end{aligned}$$

The cumulative hazard function  $H(T)$  is the sum of the individual hazard rates from time zero to time  $T$ . The formula for the cumulative hazard function is

$$H(T) = \int_0^T h(u) du$$

Thus, the hazard function is the derivative, or slope, of the cumulative hazard function. The cumulative hazard function is related to the cumulative survival function by the expression

$$S(T) = e^{-H(T)}$$

or

$$H(T) = -\ln(S(T))$$

## Cox Regression

We see that the distribution function, the hazard function, and the survival function are mathematically related. As a matter of convenience and practicality, the hazard function is used in the basic regression model.

Cox (1972) expressed the relationship between the hazard rate and a set of covariates using the model

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^p x_i \beta_i$$

or

$$h(T) = h_0(T) e^{\sum_{i=1}^p x_i \beta_i}$$

where  $x_1, x_2, \dots, x_p$  are covariates,  $\beta_1, \beta_2, \dots, \beta_p$  are regression coefficients to be estimated,  $T$  is the elapsed time, and  $h_0(T)$  is the baseline hazard rate when all covariates are equal to zero. Thus, the linear form of the regression model is

$$\ln \left[ \frac{h(T)}{h_0(T)} \right] = \sum_{i=1}^p x_i \beta_i$$

Taking the exponential of both sides of the above equation, we see that this is the ratio between the actual hazard rate and the baseline hazard rate, sometimes called the *relative risk*. This can be rearranged to give the model

$$\begin{aligned} \frac{h(T)}{h_0(T)} &= \exp \left( \sum_{i=1}^p x_i \beta_i \right) \\ &= e^{x_1 \beta_1} e^{x_2 \beta_2} \dots e^{x_p \beta_p} \end{aligned}$$

The regression coefficients can thus be interpreted as the relative risk when the value of the covariate is increased by one unit.

Note that unlike most regression models, this model does not include an intercept term. This is because if an intercept term were included, it would become part of  $h_0(T)$ .

Also note that the above model does not include  $T$  on the right-hand side. That is, the relative risk is constant for all time values. This is why the method is called *proportional hazards*.

An interesting attribute of this model is that you only need to use the ranks of the failure times to estimate the regression coefficients. The actual failure times are not used except to generate the ranks. Thus, you will achieve the same regression coefficient estimates regardless of whether you enter the time values in days, months, or years.

## Cumulative Hazard

Under the proportional hazards regression model, the cumulative hazard is

$$\begin{aligned}
 H(T, X) &= \int_0^T h(u, X) du \\
 &= \int_0^T h_0(u) e^{\sum_{i=1}^p x_i \beta_i} du \\
 &= e^{\sum_{i=1}^p x_i \beta_i} \int_0^T h_0(u) du \\
 &= H_0(T) e^{\sum_{i=1}^p x_i \beta_i}
 \end{aligned}$$

Note that the survival time  $T$  is present in  $H_0(T)$ , but not in  $e^{\sum_{i=1}^p x_i \beta_i}$ . Hence, the cumulative hazard up to time  $T$  is represented in this model by a baseline cumulative hazard  $H_0(T)$  which is adjusted by the covariates by multiplying by the factor  $e^{\sum_{i=1}^p x_i \beta_i}$ .

## Cumulative Survival

Under the proportional hazards regression model, the cumulative survival is

$$\begin{aligned}
 S(T, X) &= \exp(-H(T, X)) \\
 &= \exp\left(-H_0(T) e^{\sum_{i=1}^p x_i \beta_i}\right) \\
 &= \left[e^{-H_0(T)}\right] e^{\sum_{i=1}^p x_i \beta_i} \\
 &= S_0(T) e^{\sum_{i=1}^p x_i \beta_i}
 \end{aligned}$$

Note that the survival time  $T$  is present in  $S_0(T)$ , but not in  $e^{\sum_{i=1}^p x_i \beta_i}$ .

---

## A Note On Using e

The discussion that follows uses the terms  $\exp(x)$  and  $e^x$ . These terms are identical. That is

$$\begin{aligned}\exp(x) &= e^x \\ &= (2.71828182846)^x\end{aligned}$$

The decision as to which form to use depends on the context. The preferred form is  $e^x$ . But often, the expression used for  $x$  becomes so small that it cannot be printed. In these situations, the  $\exp(x)$  form will be used.

One other point needs to be made while we are on this subject. People often wonder why we use the number  $e$ . After all,  $e$  is an unfamiliar number that cannot be expressed exactly. Why not use a more common number like 2, 3, or 10? The answer is that it does matter because the choice of the base is arbitrary in that you can easily switch from one base to another. That is, it is easy to find constants  $a$ ,  $b$ , and  $c$  so that

$$e = 2^a = 3^b = 10^c$$

In fact,  $a$  is  $1/\ln(2) = 1.4427$ ,  $b$  is  $1/\ln(3) = 0.9102$ , and  $c$  is  $1/\ln(10) = 0.4343$ . Using these constants, it is easy to switch from one base to another. For example, suppose a calculate only computes  $10^x$  and we need the value of  $e^3$ . This can be computed as follows

$$\begin{aligned}e^3 &= (10^{0.4343})^3 \\ &= 10^{3(0.4343)} \\ &= 10^{1.3029} \\ &= 20.0855\end{aligned}$$

The point is, it is simple to change from base  $e$  to base 3 to base 10. The number  $e$  is used for mathematical convenience.

---

## Maximum Likelihood Estimation

Let  $t = 1, \dots, M$  index the  $M$  unique failure times  $T_1, T_2, \dots, T_M$ . Note that  $M$  does not include duplicate times or censored observations. The set of all failures (deaths) that occur at time  $T_t$  is referred to as  $D_t$ . Let  $c$  and  $d = 1, \dots, m_t$  index the members of  $D_t$ . The set of all individuals that are at risk immediately before time  $T_t$  is referred to as  $R_t$ . This set, often called the *risk set*, includes all individuals that fail at time  $T_t$  as well as those that are censored or fail at a time later than  $T_t$ . Let  $r = 1, \dots, n_t$  index the members of  $R_t$ . Let  $X$  refer to a set of  $p$  covariates. These covariates are indexed by the subscripts  $i, j$ , or  $k$ . The values of the covariates at a particular failure time  $T_d$  are written  $x_{1d}, x_{2d}, \dots, x_{pd}$  or  $x_{id}$  in general. The regression coefficients to be estimated are  $\beta_1, \beta_2, \dots, \beta_p$ .

## The Log-Likelihood

When there are no ties among the failure times, the log-likelihood is given by Kalbfleisch and Prentice (1980) as

$$\begin{aligned} LL(\beta) &= \sum_{t=1}^M \left\{ \left( \sum_{i=1}^p x_{it}\beta_i \right) - \ln \left( \sum_{r \in R_t} \exp \left( \sum_{i=1}^p x_{ir}\beta_i \right) \right) \right\} \\ &= \sum_{t=1}^M \left\{ \sum_{i=1}^p x_{it}\beta_i - \ln(G_{R_t}) \right\} \end{aligned}$$

where

$$G_R = \sum_{r \in R} \exp \left( \sum_{i=1}^p x_{ir}\beta_i \right)$$

The following notation for the first-order and second-order partial derivatives will be useful in the derivations in this section.

$$\begin{aligned} H_{jR} &= \frac{\partial G_R}{\partial \beta_j} \\ &= \sum_{r \in R} x_{jr} \exp \left( \sum_{i=1}^p x_{ir}\beta_i \right) \end{aligned}$$

$$\begin{aligned} A_{jkR} &= \frac{\partial^2 G_R}{\partial \beta_j \partial \beta_k} \\ &= \frac{\partial H_{jR}}{\partial \beta_k} \\ &= \sum_{r \in R} x_{jr} x_{kr} \exp \left( \sum_{i=1}^p x_{ir}\beta_i \right) \end{aligned}$$

## Cox Regression

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first and second order partial derivatives. The first order partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j}$$

$$= \sum_{t=1}^M \left\{ x_{jt} - \frac{H_{jR_t}}{G_{R_t}} \right\}$$

The second order partial derivatives, which are the information matrix, are

$$I_{jk} = \sum_{t=1}^M \frac{1}{G_{R_t}} \left( A_{jkR_t} - \frac{H_{jR_t} H_{kR_t}}{G_{R_t}} \right)$$

When there are failure time ties (note that censor ties are not a problem), the exact likelihood is very cumbersome. **NCSS** allows you to select either the approximation proposed by Breslow (1974) or the approximation given by Efron (1977). Breslow's approximation was used by the first Cox regression programs, but Efron's approximation provides results that are usually closer to the results given by the exact algorithm, and it is now the preferred approximation (see for example Homer and Lemeshow (1999)). We have included Breslow's method because of its popularity. For example, Breslow's method is the default method used in SAS.

### Breslow's Approximation to the Log-Likelihood

The log-likelihood of Breslow's approximation is given by Kalbfleisch and Prentice (1980) as

$$LL(\beta) = \sum_{t=1}^M \left\{ \left( \sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i \right) - m_t \ln \left[ \sum_{r \in R_t} \exp \left( \sum_{i=1}^p x_{ir} \beta_i \right) \right] \right\}$$

$$= \sum_{t=1}^M \left\{ \sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - m_t \ln(G_{R_t}) \right\}$$

where

$$G_R = \sum_{r \in R} \exp \left( \sum_{i=1}^p x_{ir} \beta_i \right)$$

## Cox Regression

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first-order and second-order partial derivatives. The first order partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j}$$

$$= \sum_{t=1}^M \left\{ \left( \sum_{d \in D_t} x_{jd} \right) - \left( m_t \frac{H_{jR_t}}{G_{R_t}} \right) \right\}$$

The negative of the second-order partial derivatives, which form the information matrix, are

$$I_{jk} = \sum_{t=1}^M \frac{m_t}{G_{R_t}} \left( A_{jkR_t} - \frac{H_{jR_t} H_{kR_t}}{G_{R_t}} \right)$$

**Efron's Approximation to the Log Likelihood**

The log-likelihood of Efron's approximation is given by Kalbfleisch and Prentice (1980) as

$$LL(\beta) = \sum_{t=1}^M \left\{ \sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - \sum_{d \in D_t} \ln \left[ \sum_{r \in R_t} \exp \left( \sum_{i=1}^p x_{ir} \beta_i \right) - \frac{d-1}{m_t} \sum_{c \in D_t} \exp \left( \sum_{i=1}^p x_{ic} \beta_i \right) \right] \right\}$$

$$= \sum_{t=1}^M \left\{ \sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - \sum_{d \in D_t} \ln \left[ G_{R_t} - \frac{d-1}{m_t} G_{D_t} \right] \right\}$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first and second order partial derivatives. The first partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j}$$

$$= \sum_{t=1}^M \sum_{d \in D_t} \left( x_{jd} - \frac{H_{jR_t} - \left( \frac{d-1}{m_t} \right) H_{jD_t}}{G_{R_t} - \left( \frac{d-1}{m_t} \right) G_{D_t}} \right)$$

$$= \sum_{t=1}^M \sum_{d \in D_t} x_{jd} - \sum_{t=1}^M \sum_{d=1}^{m_t} \left( \frac{H_{jR_t} - \left( \frac{d-1}{m_t} \right) H_{jD_t}}{G_{R_t} - \left( \frac{d-1}{m_t} \right) G_{D_t}} \right)$$



The second partial derivatives provide the information matrix which estimates the covariance matrix of the estimated regression coefficients. The negative of the second partial derivatives are

$$I_{jk} = -\frac{\partial^2 LL(\beta)}{\partial \beta_j \partial \beta_k}$$

$$= \frac{\sum_{t=1}^M \sum_{d=1}^{m_t} \left( G_{R_t} - \left( \frac{d-1}{m_t} \right) G_{D_t} \right) \left( A_{jkR_t} - \left( \frac{d-1}{m_t} \right) A_{jkD_t} \right) - \left( H_{jR_t} - \left( \frac{d-1}{m_t} \right) H_{jD_t} \right) \left( H_{kR_t} - \left( \frac{d-1}{m_t} \right) H_{kD_t} \right)}{\left( G_{R_t} - \left( \frac{d-1}{m_t} \right) G_{D_t} \right)^2}$$

## Estimation of the Survival Function

Once the maximum likelihood estimates have been obtained, it may be of interest to estimate the survival probability of a new or existing individual with specific covariate settings at a particular point in time. The methods proposed by Kalbfleisch and Prentice (1980) are used to estimate the survival probabilities.

### Cumulative Survival

This estimates the cumulative survival of an individual with a set of covariates all equal to zero. The survival for an individual with covariate values of  $X_0$  is

$$S(T|X_0) = \exp(H(T|X_0))$$

$$= \exp\left( H_0(T|X_0) \exp \sum_{i=1}^p x_{i0} \beta \right)$$

$$= [S_0(T)]^{\exp \sum_{i=1}^p x_{i0} \beta}$$

The estimate of the baseline survival function  $S_0(T)$  is calculated from the cumulated hazard function using

$$S_0(T_0) = \prod_{T_t \leq T_0} \alpha_t$$

where

$$\alpha_t = \frac{S(T_t)}{S(T_{t-1})}$$

$$= \left[ \frac{S_0(T_t)}{S_0(T_{t-1})} \right]^{\exp(\sum_{i=1}^p x_{it} \beta_i)}$$

$$= \left[ \frac{S_0(T_t)}{S_0(T_{t-1})} \right]^{\theta_t}$$

## Cox Regression

and

$$\theta_r = \exp\left(\sum_{i=1}^p x_{ir}\beta_i\right)$$

The value of  $\alpha_t$ , the conditional baseline survival probability at time  $T$ , is the solution to the conditional likelihood equation

$$\sum_{d \in D_t} \frac{\theta_d}{1 - \alpha_t^{\theta_d}} = \sum_{r \in R_t} \theta_r$$

When there are no ties at a particular time point,  $D_t$  contains one individual and the above equation can be solved directly, resulting in the solution

$$\hat{\alpha}_t = \left[1 - \frac{\hat{\theta}_t}{\sum_{r \in R_t} \hat{\theta}_r}\right]^{\hat{\theta}_t^{-1}}$$

When there are ties, the equation must be solved iteratively. The starting value of this iterative process is

$$\hat{\alpha}_t = \exp\left(\frac{-m_t}{\sum_{r \in R_t} \hat{\theta}_r}\right)$$

## Baseline Hazard Rate

Hosmer and Lemeshow (1999) estimate the baseline hazard rate  $h_0(T_t)$  as follows

$$h_0(T_t) = 1 - \alpha_t$$

They mention that this estimator will typically be too unstable to be of much use. To overcome this, you might smooth these quantities using lowess function of the Scatter Plot program.

## Cumulative Hazard

An estimate of the cumulative hazard function  $H_0(T)$  derived from relationship between the cumulative hazard and the cumulative survival. The estimated baseline survival is

$$\hat{H}_0(T) = -\ln(\hat{S}_0(T))$$

This leads to the estimated cumulative hazard function is

$$\hat{H}(T) = -\exp\left(\sum_{i=1}^p x_i \hat{\beta}_i\right) \ln(\hat{S}_0(T))$$

## Cumulative Survival

The estimate of the cumulative survival of an individual with a set of covariates values of  $X_0$  is

$$\hat{S}(T|X_0) = \hat{S}_0(T) \exp \sum_{i=1}^p x_{i0} \hat{\beta}_i$$

---

## Statistical Tests and Confidence Intervals

Inferences about one or more regression coefficients are all of interest. These inference procedures can be treated by considering hypothesis tests and/or confidence intervals. The inference procedures in Cox regression rely on large sample sizes for accuracy.

Two tests are available for testing the significance of one or more independent variables in a regression: the likelihood ratio test and the Wald test. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation.

These two testing procedures will be described next.

### Likelihood Ratio and Deviance

The *Likelihood Ratio* test statistic is -2 times the difference between the log-likelihoods of two models, one of which is a subset of the other. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. The test is named as a ratio rather than a difference since the difference between two log-likelihoods is equal to the log of the ratio of the two likelihoods. That is, if  $L_{\text{full}}$  is the log-likelihood of the full model and  $L_{\text{subset}}$  is the log-likelihood of a subset of the full model, the likelihood ratio is defined as

$$\begin{aligned} LR &= -2[L_{\text{subset}} - L_{\text{full}}] \\ &= -2 \left[ \ln \left( \frac{l_{\text{subset}}}{l_{\text{full}}} \right) \right] \end{aligned}$$

Note that the -2 adjusts  $LR$  so the chi-square distribution can be used to approximate its distribution.

The likelihood ratio test is the test of choice in Cox regression. Various simulation studies have shown that it is more accurate than the Wald test in situations with small to moderate sample sizes. In large samples, it performs about the same. Unfortunately, the likelihood ratio test requires more calculations than the Wald test, since it requires the fitting of two maximum-likelihood models.

### Deviance

When the full model in the likelihood ratio test statistic is the saturated model,  $LR$  is referred to as the *deviance*. A saturated model is one which includes all possible terms (including interactions) so that the predicted values from the model equal the original data. The formula for the deviance is

$$D = -2[L_{\text{Reduced}} - L_{\text{Saturated}}]$$

The deviance in Cox regression is analogous to the residual sum of squares in multiple regression. In fact, when the deviance is calculated in multiple regression, it is equal to the sum of the squared residuals.

## Cox Regression

The change in deviance,  $\Delta D$ , due to excluding (or including) one or more variables is used in Cox regression just as the partial  $F$  test is used in multiple regression. Many texts use the letter  $G$  to represent  $\Delta D$ . Instead of using the  $F$  distribution, the distribution of the change in deviance is approximated by the chi-square distribution. Note that since the log-likelihood for the saturated model is common to both deviance values,  $\Delta D$  can be calculated without actually fitting the saturated model. This fact becomes very important during subset selection. The formula for  $\Delta D$  for testing the significance of the regression coefficient(s) associated with the independent variable  $X_1$  is

$$\begin{aligned}\Delta D_{X_1} &= D_{\text{without } X_1} - D_{\text{with } X_1} \\ &= -2[L_{\text{without } X_1} - L_{\text{Saturated}}] + 2[L_{\text{with } X_1} - L_{\text{Saturated}}] \\ &= -2[L_{\text{without } X_1} - L_{\text{with } X_1}]\end{aligned}$$

Note that this formula looks identical to the likelihood ratio statistic. Because of the similarity between the change in deviance test and the likelihood ratio test, their names are often used interchangeably.

### Wald Test

The Wald test will be familiar to those who use multiple regression. In multiple regression, the common  $t$ -test for testing the significance of a particular regression coefficient is a Wald test. In Cox regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$z_j = \frac{b_j}{s_{b_j}}$$

where  $s_{b_j}$  is an estimate of the standard error of  $b_j$  provided by the square root of the corresponding diagonal element of the covariance matrix,  $V(\hat{\beta}) = I^{-1}$ .

With large sample sizes, the distribution of  $z_j$  is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as "adequate."

The Wald test is used in **NCSS** to test the statistical significance of individual regression coefficients.

### Confidence Intervals

Confidence intervals for the regression coefficients are based on the Wald statistics. The formula for the limits of a  $100(1 - \alpha)\%$  two-sided confidence interval is

$$b_j \pm |z_{\alpha/2}|s_{b_j}$$

## $R^2$

Hosmer and Lemeshow (1999) indicate that at the time of the writing of their book, there is no single, easy to interpret measure in Cox regression that is analogous to  $R^2$  in multiple regression. They indicate that if such a measure “must be calculated” they would use

$$R_p^2 = 1 - \exp\left[\frac{2}{n}(L_0 - L_p)\right]$$

where  $L_0$  is the log-likelihood of the model with no covariates,  $n$  is the number of observations (censored or not), and  $L_p$  is the log-likelihood of the model that includes the covariates.

---

## Subset Selection

Subset selection refers to the task of finding a small subset of the available regressor variables that does a good job of predicting the dependent variable. Because Cox regression must be solved iteratively, the task of finding the best subset can be time consuming. Hence, techniques which look at all possible combinations of the regressor variables are not feasible. Instead, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Before discussing the details of these two algorithms, it is important to comment on a couple of issues that can come up. The first issue is what to do about the binary variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms used here search on model terms rather than on the individual variables. Thus, the whole set of binary variables associated with a given term are considered together for inclusion in, or deletion from, the model. It's all or none. Because of the time-consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and designate them as Numeric Variables.

## Hierarchical Models

A second issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term  $A*B*C$  is not included unless the terms  $A$ ,  $B$ ,  $C$ ,  $A*B$ ,  $A*C$ , and  $B*C$  are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to only consider hierarchical models during its search. Thus, if  $C$  is not in the model, interactions involving  $C$  are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

## Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of  $R$ -squared. Enter this term into the model.
3. Continue adding terms until a preset limit on the maximum number of terms in the model is reached.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations so that other, more time-consuming methods, are not feasible, or when you have far too many possible regressor variables and you want to reduce the number of terms in the selection pool.

## Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of  $R$ -squared. If a switch can be found, it is made, and the candidate terms are again searched to determine if another switch can be made.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached, or all terms are included in the model.

## Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some F-to-enter and F-to-remove tests whose properties are not well understood to begin with.

---

## Residuals

The following presentation summarizes the discussion on residuals found in Klein and Moeschberger (1997) and Hosmer and Lemeshow (1999). For a more thorough treatment of this topic, we refer you to either of these books.

In most settings in which residuals are studied, the dependent variable is predicted using a model based on the independent variables. In these cases, the residual is simply the difference between the actual value and the predicted value of the dependent variable. Unfortunately, in Cox regression there is no obvious analog this actual minus predicted. Realizing this, statisticians have looked at how residuals are used and then, based on those uses, developed quantities that meet those needs. They call these quantities *residuals* because they are used in place of residuals. However, you must remember that they are not equivalent to usual the residuals that you see in multiple regression, for example.

## Cox Regression

In the discussion that follows, the formulas will be simplified if we use the substitution

$$\theta_r = \exp\left(\sum_{i=1}^p x_{ir}\beta_i\right)$$

## Cox-Snell Residuals

The Cox-Snell residuals were used to assess the goodness-of-fit of the Cox regression. The Cox-Snell residuals are defined as

$$r_t = H_{B0}(T_t)\theta_t$$

where the  $b$ 's are the estimated regression coefficients and  $H_0(T_t)$  is Breslow's estimate of the cumulative baseline hazard function. This value is defined as follows

$$H_{B0}(T_t) = \sum_{T_i \leq T_t} \left[ \frac{m_i}{\sum_{j \in R_{T_i}} \theta_j} \right]$$

The Cox-Snell residuals were the first to be proposed in the literature. They have since been replaced by other types of residuals and are now only of historical interest. See, for example, the discussion of Marubini and Valsecchi (1996) who state that the use of these residuals on distributional grounds should be avoided.

## Martingale Residuals

Martingale residuals cannot be used to assess goodness-of-fit as are the usual residuals in multiple regression. The best model need not have the smallest sum of squared martingale residuals. Martingale residuals follow the unit exponential distribution. Some authors suggested analyzing these residuals to determine how close they are to the exponential distribution, hoping that a lack of exponentiality indicated a lack of fit. Unfortunately, just the opposite is the case since in a model with no useful covariates, these residuals are exactly exponential in distribution. Another diagnostic tool used in regular multiple regression is a plot of the residuals versus the fitted values. Here again, the martingale residuals cannot be used for this purpose since they are negatively correlated with the fitted values.

So, of what use are martingale residuals? They have two main uses. First, they can be used to find outliers—individuals who are poorly fit by the model. Second, martingale residuals can be used to determine the functional form of each of the covariates in the model.

### Finding Outliers

The martingale residuals are defined as

$$M_t = c_t - r_t$$

where  $c_t$  is one if there is a failure at time  $T_t$  and zero otherwise. The martingale residual measures the difference between whether an individual experiences the event of interest and the expected number of events based on the model. The maximum value of the residual is one and the minimum possible value is negative infinity. Thus, the residual is highly skewed. A large negative martingale residual indicates a high-risk individual who still had a long survival time.

## Finding the Function Form of Covariates

Martingale residuals can be used to determine the functional form of a covariate. To do this, you generate the Martingale residuals from a model without the covariates. Next, you plot these residuals against the value of the covariate. For large datasets, this may be a time-consuming process. Therneau and Grambsch (2000) suggest that the martingale residuals from a model with no covariates be plotted against each of the covariates. These plots will reveal the appropriate functional form of the covariates in the model so long as the covariates are not highly correlated among themselves.

## Deviance Residuals

Deviance residuals are used to search for outliers. The deviance residuals are defined as

$$DEV_t = \text{sign}(M_t) \sqrt{-2[M_t + c_t \ln(c_t - M_t)]}$$

or zero when  $M_t$  is zero. These residuals are plotted against the risk scores given by

$$\exp\left(\sum_{i=1}^p x_{it} b_i\right)$$

When there is slight to moderate censoring, large absolute values in these residuals point to potential outliers. When there is heavy censoring, there will be a large number of residuals near zero. However, large absolute values will still indicate outliers.

## Schoenfeld's Residuals

A set of  $p$  Schoenfeld residuals is defined for each noncensored individual. The residual is missing when the individual is censored. The Schoenfeld residuals are defined as follows

$$\begin{aligned} r_{it} &= c_t \left[ x_{it} - \frac{\sum_{r \in R_t} x_{ir} \theta_r}{\sum_{r \in R_t} \theta_r} \right] \\ &= c_t \left[ x_{it} - \sum_{r \in R_t} x_{ir} w_r \right] \end{aligned}$$

where

$$w_r = \frac{\sum_{r \in R_t} x_{ir} \theta_r}{\sum_{r \in R_t} \theta_r}$$

Thus, this residual is the difference between the actual value of the covariate and a weighted average where the weights are determined from the risk scores.

These residuals are used to estimate the influence of an observation on each of the regression coefficients. Plots of these quantities against the row number or against the corresponding covariate values are used to study these residuals.



## Scaled Schoenfeld's Residuals

Hosmer and Lemeshow (1999) and Therneau and Grambsch (2000) suggest that scaling the Schoenfeld residuals by an estimate of their variance gives quantities with greater diagnostic ability. Hosmer and Lemeshow (1999) use the covariance matrix of the regression coefficients to perform the scaling. The scaled Schoenfeld residuals are defined as follows

$$r_{kt}^* = m \sum_{i=1}^p V_{ik} r_{it}$$

where  $m$  is the total number of deaths in the dataset and  $V$  is the estimated covariance matrix of the regression coefficients.

These residuals are plotted against time to validate the proportional hazards assumption. If the proportional hazards assumption holds, the residuals will fall randomly around a horizontal line centered at zero. If the proportional hazards assumption does not hold, a trend will be apparent in the plot.

## Data Structure

Survival data sets require up to three components for the survival time: the ending survival time, the beginning survival time during which the subject was not observed, and an indicator of whether the observation was censored or failed.

Based on these three components, various types of data may be analyzed. *Right censored* data are specified using only the ending time variable and the censor variable. *Left truncated* and *Interval* data are entered using all three variables.

The table below shows survival data ready for analysis. These data are from a lung cancer study reported in Kalbfleisch (1980), page 223. These data are in the LungCancer dataset. The variables are

<b>TIME</b>	Days of survival
<b>CENSOR</b>	Censor indicator
<b>STATUS</b>	Performance status
<b>MONTHS</b>	Months from diagnosis
<b>AGE</b>	Age in years
<b>THERAPY</b>	Prior therapy

**LungCancer Dataset (Subset)**

<b>TIME</b>	<b>CENSOR</b>	<b>STATUS</b>	<b>MONTHS</b>	<b>AGE</b>	<b>THERAPY</b>
72	1	60	7	69	0
411	1	70	5	64	10
228	1	60	3	38	0
126	1	60	9	63	10
118	1	70	11	65	10
10	1	20	5	49	0
82	1	40	10	69	10
110	1	80	29	68	0
314	1	50	18	43	0
100	0	70	6	70	0
42	1	60	4	81	0
8	1	40	58	63	10
144	1	30	4	63	0
25	0	80	9	52	10
11	1	70	11	48	10

## Example 1 – Cox Regression Analysis

This section presents an example of how to run a Cox regression analysis. The data used are found in the LungCancer dataset. These data are an excerpt from a lung cancer study reported in Kalbfleisch (1980). The variables used in the analysis are

<b>TIME</b>	Days of survival
<b>CENSOR</b>	Censor indicator
<b>STATUS</b>	Karnofsky rating performance status
<b>MONTHS</b>	Months from diagnosis
<b>AGE</b>	Age in years
<b>THERAPY</b>	Prior therapy: 0 no, 10 yes

The purpose of this analysis is to study the relationship between length of patient survival and the covariates.

### Setup

To run this example, complete the following steps:

#### 1 Open the LungCancer example dataset

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

#### 2 Specify the Cox Regression procedure options

- Find and open the **Cox 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, Model Tab

Time.....**Time**  
 Ties Method.....**Efron**  
 Censor.....**Censor**  
 Numeric X's.....**Status-Therapy**

Reports Tab

All Available Reports.....**Checked** (click the *Check All* button)

Plots Tab

All Available Plots.....**Checked** (click the *Check All* button)

#### 3 Run the procedure

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

## Run Summary

Run Summary			
Item	Value	Rows	Value
Time Variable	Time	Rows Processed	15
Censor Variable	Censor	Rows Used in Estimation	15
Frequency Variable	None	Rows with X's Missing	0
Independent Variables Available	4	Rows with Y Missing	0
Number of X's in the Model	4	Rows Failed	13
Final Log-Likelihood	-20.1143	Rows Censored	2
Number of Likelihood Iterations	7 of 20		
Convergence Criterion	1E-09		
Achieved Convergence	1.206491E-15		
Completion Status	Normal completion		
Starting B's	0		
Subset Selection Method	None		

This report summarizes the characteristics of the dataset and provides useful information about the reports to follow. It should be studied to make sure that the data were read in properly and that the estimation algorithm terminated normally. We will only discuss those parameters that need special explanation.

### Final Log-Likelihood

This is the log-likelihood of the model.

### Number of Likelihood Iterations

This is the number of iterations used by the maximum likelihood procedure. This value should be compared against the value of the Maximum Iterations option to see if the iterative procedure terminated early.

### Achieved Convergence

This is the maximum of the relative changes in the regression coefficients on the last iteration. If this value is less than the Convergence Criterion, the procedure converged normally. Otherwise, the specified convergence precision was not achieved.

### Rows Processed

This is the number of rows processed during the run. Check this count to make certain it agrees with what you anticipated.

## Regression Coefficients

### Regression Coefficients

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	P-Value	Pseudo R <sup>2</sup>
B1: Status	-0.032415	0.020324	0.9681	57.33333	-1.5949	0.1107	0.2203
B2: Months	0.064557	0.033056	1.0667	12.6	1.9530	0.0508	0.2975
B3: Age	0.039805	0.035232	1.0406	60.33333	1.1298	0.2586	0.1241
B4: Therapy	0.013967	0.068384	1.0141	4.666667	0.2042	0.8382	0.0046

### Estimated Cox Regression Model

Exp( -0.0324152392634531\*Status + 0.0645571159984993\*Months + 0.0398048128120681\*Age + 0.0139668973406698\*Therapy )

#### Transformation Note:

Regular transformations must be less the 255 characters. If this expression is longer the 255 characters, copy this expression and paste it into a text file, then use the transformation FILE(filename.txt) to access the text file.

This report displays the results of the proportional hazards estimation. Following are the detailed definitions:

### Independent Variable

This is the variable from the model that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the binary variable is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the *Stagger label and output* option of the Report Options tab.

### Regression Coefficient (B)

This is the estimate of the regression coefficient,  $\beta_i$ . Remember that the basic regression equation is

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^p x_i \beta_i$$

Thus, the quantity  $\beta_i$  is the amount that the log of the hazard rate changes when  $x_i$  is increased by one unit. Note that a positive coefficient implies that as the value of the covariate is increased, the hazard increases, and the prognosis gets worse. A negative coefficient indicates that as the variable is increased, the hazard decreases, and the prognosis gets better.

### Standard Error of B

This is  $s_{b_j}$ , the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is provided by the square root of the corresponding diagonal element of the covariance matrix,  $V(\hat{\beta}) = I^{-1}$ . It is also used as the denominator of the Wald test.

## Cox Regression

**Risk Ratio Exp(B)**

This is the value of  $e^{\beta_i}$ . This value is often called the *risk ratio* since it is the ratio of two hazards whose only difference is that  $x_i$  is increased by one unit. That is,

$$\frac{h(T|x_i = a + 1)}{h(T|x_i = a)} = e^{\beta_i}$$

In this example, if Months is increased by one, the hazard rate is increased by 6.67%. If you want to calculate the effect of increasing Months by three, the hazard rate is increased by  $1.0667^3 = 1.2137$ , or 21.37%. Note that is not equal to 3.0 times 6.67.

**Mean**

This is the average of this independent variable. The means are especially important in interpreting the baseline hazard rates. Unless you have opted otherwise, the independent variables have been centered by subtracting these mean values. Hence, the baseline hazard rate occurs when each independent variable is equal to its mean.

**Wald Z-Value**

This is the z value of the Wald test used for testing the hypothesis that  $\beta_i = 0$  against the alternative  $\beta_i \neq 0$ . The Wald test is calculated using the formula

$$z_i = \frac{b_{ij}}{s_{b_i}}$$

The distribution of the Wald statistic is closely approximated by the normal distribution in large samples. However, in small samples, the normal approximation may be poor. For small samples, likelihood ratio tests perform better and are preferred.

**P-Value**

This is the two-sided probability level. This is the probability of obtaining a z-value larger in absolute value than the one obtained. If this probability is less than the specified significance level (say 0.05), the regression coefficient is significantly different from zero.

**Pseudo R<sup>2</sup>**

An index value, similar to R<sup>2</sup> in regression, representing the relative influence of this variable. If  $C = z^2$ ,  $n$  = sample size, and  $p$  = number of variables, then  $R^2 = C/(n-p+C)$ .

**Estimated Cox Regression Model**

This section gives the Cox regression model in a regular text format that can be used as a transformation formula. The regression coefficients are displayed in double precision because a single-precision formula does not include the accuracy necessary to calculate the hazard rates.

Note that transformation must be less than 255 characters. Since these formulas are often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file,

## Cox Regression

including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

## Confidence Interval Limits for Regression Coefficients

Confidence Interval Limits for Regression Coefficients

Independent Variable	Regression Coefficient (B)	95% Confidence Interval Limits for B		Risk Ratio Exp(B)	95% Confidence Interval Limits for Exp(B)	
		Lower	Upper		Lower	Upper
B1: Status	-0.032415	-0.072250	0.007420	0.9681	0.9303	1.0074
B2: Months	0.064557	-0.000230	0.129345	1.0667	0.9998	1.1381
B3: Age	0.039805	-0.029249	0.108859	1.0406	0.9712	1.1150
B4: Therapy	0.013967	-0.120062	0.147996	1.0141	0.8869	1.1595

This report provides the confidence intervals for the regression coefficients and the risk ratios. The confidence coefficient, in this example 95%, was specified on the Format tab.

### Independent Variable

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by *Stagger label and output* option of the Report Options tab.

### Regression Coefficient (B or Beta)

This is the estimate of the regression coefficient,  $\beta_i$ . Remember that the basic regression equation is

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^p x_i \beta_i$$

Thus, the quantity  $\beta_i$  is the amount that the log of the hazard rate changes when  $x_i$  is increased by one unit. Note that a positive coefficient implies that as the value of the covariate is increased, the hazard increases and the prognosis gets worse. A negative coefficient indicates that as the variable is increased, the hazard decreases and the prognosis gets better.

### 95% Confidence Interval Limits for B (Lower and Upper)

A 95% confidence interval for  $\beta_i$  is given by an upper and lower limit. These limits are based on the Wald statistic using the formula

$$b_i \pm z_{1-\alpha/2} s_{b_i}$$

Since they are based on the Wald test, they are only valid for large samples.

## Cox Regression

**Risk Ratio Exp(B)**

This is the value of  $e^{\beta_i}$ . This value is often called the *risk ratio* since it is the ratio of two hazards whose only difference is that  $x_i$  is increased by one unit. That is,

$$\frac{h(T|x_i = a + 1)}{h(T|x_i = a)} = e^{\beta_i}$$

In this example, if Months is increased by one, the hazard rate is increased by 6.67%. If you want to calculate the effect of increasing Months by three, the hazard rate is increased by  $1.0667^3 = 1.2137$ , or 21.37%. Note that is not equal to 3.0 times 6.67.

**95% Confidence Interval Limits for Exp(B) (Lower and Upper)**

A 95% confidence interval for  $e^{\beta_i}$  is given by an upper and lower limit. These limits are based on the Wald statistic using the formula

$$\exp(b_i \pm z_{1-\alpha/2} s_{b_i})$$

Since they are based on the Wald test, they are only valid for large samples.

**Analysis of Deviance****Analysis of Deviance**

Term(s) Omitted	DF	-2 Log-Likelihood	Increase from Model Deviance (Chi <sup>2</sup> )	P-Value
All Terms	4	46.6698	6.4413	0.1685
Status	1	42.7787	2.5501	0.1103
Months	1	44.3928	4.1642	0.0413
Age	1	41.5943	1.3657	0.2426
Therapy	1	40.2704	0.0419	0.8379
None(Model)	4	40.2286		

The P-Value is for testing the significance of each term after adjusting for all other terms.

This report is the Cox regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms in the regression are statistically significant after adjusting for all other terms in the model.

This report is not produced during a subset selection run.

Note that this report requires that a separate regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

**Term(s) Omitted**

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.



## Cox Regression

The “All” line refers to a no-covariates model. The “None(Model)” refers to the complete model with no terms removed.

Note that it is usually not advisable to include an interaction term in a model when one of the associated main effects is missing—which is what happens here. However, in this case, we believe this to be a useful test.

Note also that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the *Stagger label and output* option of the Report Options tab. This should create a better looking report when the names are extra-long.

**DF**

This is the degrees of freedom of the chi-square test displayed on this line. DF is equal to the number of individual independent variables in the term.

**-2 Log-Likelihood**

This is the log-likelihood achieved by the model being described on this line of the report.

**Increase from Model Deviance (Chi<sup>2</sup>)**

This is the difference between the deviance for the model described on this line and the deviance of the complete model. This value follows the chi<sup>2</sup> distribution in medium to large samples. This value can be thought of as the analog of the residual sum of squares in multiple regression. Thus, you can think of this value as the increase in the residual sum of squares that occurs when this term is removed from the model.

Another way to interpret this test is as a *redundancy test* because it tests whether this term is redundant after considering all of the other terms in the model.

**P-Value**

This is the significance level of the chi<sup>2</sup> test. This is the probability that a chi<sup>2</sup> value with degrees of freedom DF is equal to this value or greater. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

---

## Log-Likelihood and R<sup>2</sup> Section

**Log-Likelihood and R<sup>2</sup>**

Term(s) Omitted	DF	Log-Likelihood	R <sup>2</sup> of Remaining Term(s)	Reduction from Model R <sup>2</sup>
All Terms	4	-23.3349	0.0000	0.3491
Status	1	-21.3893	0.2285	0.1206
Months	1	-22.1964	0.1408	0.2083
Age	1	-20.7971	0.2871	0.0620
Therapy	1	-20.1352	0.3473	0.0018
None(Model)	4	-20.1143	0.3491	0.0000

This report provides the log-likelihoods and R<sup>2</sup> values of various models. This report is not produced during a subset selection run.

## Cox Regression

Note that this report requires that a separate Cox regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

**Term(s) Omitted**

This is the term that is omitted from the model. The “All” line refers to no-covariates model. The “None(Model)” refers to the complete model with no terms removed. The “None(Model)” refers to the complete model with no terms removed.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the *Stagger label and output* option of the Report Options tab. This should create a better looking report when the names are extra-long.

**DF**

This is the degrees of freedom of the term displayed on this line.

**Log-Likelihood**

This is the log-likelihood of the model displayed on this line. Note that this is the log-likelihood of the logistic regression without the term listed.

**R<sup>2</sup> of Remaining Term(s)**

This is the  $R^2$  of the model displayed on this line. Note that the model does not include the term listed at the beginning of the line. This  $R^2$  is analogous to the  $R^2$  in multiple regression, but it is not the same.

Hosmer and Lemeshow (1999) indicate that at the time of the writing of their book, there is no single, easy to interpret measure in Cox regression that is analogous to  $R^2$  in multiple regression. They indicate that if such a measure “must be calculated” they would use

$$R_p^2 = 1 - \exp\left[\frac{2}{n}(L_0 - L_p)\right]$$

where  $L_0$  is the log-likelihood of the model with no covariates,  $n$  is the number of observations (censored or not), and  $L_p$  is the log-likelihood of the model that includes the covariates.

**Reduction from Model R<sup>2</sup>**

This is amount that  $R^2$  is reduced when the term is omitted from the regression model. This reduction is calculated from the  $R^2$  achieved by the full model.

This quantity is used to determine if removing a term causes a large reduction in  $R^2$ . If it does not, then the term can be safely removed from the model.

## Baseline Hazard and Survival

Baseline Hazard and Survival				
Centered Baseline				
Time	Cumulative Survival	Cumulative Hazard	Hazard Rate	Alpha
8	0.9654	0.0352	0.0346	0.9654
10	0.8912	0.1152	0.0768	0.9232
11	0.8183	0.2006	0.0819	0.9181
42	0.7449	0.2945	0.0897	0.9103
72	0.6717	0.3980	0.0983	0.9017
82	0.5934	0.5220	0.1166	0.8834
110	0.4942	0.7048	0.1671	0.8329
118	0.3904	0.9407	0.2102	0.7898
126	0.2911	1.2341	0.2543	0.7457
144	0.1843	1.6915	0.3670	0.6330
228	0.0922	2.3841	0.4997	0.5003
314	0.0288	3.5461	0.6872	0.3128
411	0.0288	3.5461	1.0000	0.0000

This report displays various estimated survival and hazard values. These are centered if the Centered X's option is selected.

### Baseline Cumulative Survival

This estimates the cumulative survival probability of an individual with all covariates equal to their means or to zero depending on whether the data are centered or not. It is the value of  $S_0(T)$  which is estimated using the formula

$$S_0(T) = \prod_{T_t \leq T} \alpha_t$$

### Baseline Cumulative Hazard

This estimates the cumulative baseline hazard of an individual with a set of covariates all equal to zero. It is the value of  $H_0(T)$  which is calculated using the formula

$$H_0(T) = -\ln(S_0(T))$$

### Baseline Hazard Rate

This is the estimate of the baseline hazard rates  $h_0(T_t)$  which are calculated as follows

$$h_0(T_t) = 1 - \alpha_t$$

### Alpha

This is the value of the conditional baseline survival probabilities at the times listed. These values are used to calculate  $S_0(T)$ .

## Residuals

Residuals							
Row	Time	Cox-Snell Residual		Martingale Residual		Deviance Residual	
12	8	1.3862	.....	-0.3862	.....	-0.3454	.....
6	10	0.1411	.....	0.8589	.....	1.4828	.....
15	11	0.0791	.....	0.9209	.....	1.7978	.....
14+	25	0.0590	.....	-0.0590	.....	-0.3434	.....
11	42	0.3307	.....	0.6693	.....	0.9352	.....
1	72	0.3364	.....	0.6636	.....	0.9229	.....
7	82	1.1774	.....	-0.1774	.....	-0.1679	.....
10+	100	0.3112	.....	-0.3112	.....	-0.7890	.....
8	110	1.2387	.....	-0.2387	.....	-0.2220	.....
5	118	0.7300	.....	0.2700	.....	0.2991	.....
4	126	1.0748	.....	-0.0748	.....	-0.0730	.....
13	144	2.4532	.....	-1.4532	.....	-1.0543	.....
3	228	0.4531	.....	0.5469	.....	0.6996	.....
9	314	2.9953	.....	-1.9953	.....	-1.3403	.....
2	411	1.7951	.....	-0.7951	.....	-0.6481	.....

The various residuals were discussed in detail earlier in this chapter. Only a brief definition will be given here.

### Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

### Time

This is the value of the elapsed time.

### Cox-Snell Residual

Cox-Snell residuals were created to assess the goodness-of-fit of the Cox regression. They have since been replaced by other types of residuals and are now only of historical interest. See, for example, the discussion of Marubini and Valsecchi (1996) who state that the use of these residuals on distributional grounds should be avoided.

### Martingale Residual

The martingale residuals are defined as

$$M_t = c_t - r_t$$

where  $c_t$  is one if there is a failure at time  $T_t$  and zero otherwise. The martingale residual measures the difference between whether an individual experiences the event of interest and the expected number of events based on the model. The maximum value of the residual is one and the minimum possible value is negative infinity. Thus, the residual is highly skewed. A large negative martingale residual indicates a high-risk individual who still had a long survival time.

Martingale residuals cannot be used to assess goodness-of-fit as are the usual residuals in multiple regression. They have two main uses. First, they can be used to find outliers—individuals who are poorly fit

by the model. Second, martingale residuals can be used to determine the functional form of each of the covariates in the model.

Martingale residuals can be used to determine the functional form of a covariate. To do this, you generate the Martingale residuals from a model without the covariate. Next, you plot these residuals against the value of the covariate.

### Deviance Residual

Deviance residuals are used to search for outliers. The deviance residuals are defined as

$$DEV_t = \text{sign}(M_t) \sqrt{-2[M_t + c_t \ln(c_t - M_t)]}$$

or zero when  $M_t$  is zero. These residuals are plotted against the risk scores given by

$$\exp\left(\sum_{i=1}^p x_{it} b_i\right)$$

When there is slight to moderate censoring, large absolute values in these residuals point to potential outliers. When there is heavy censoring, there will be a large number of residuals near zero. However, large absolute values will still indicate outliers.

## Martingale Residuals

Martingale Residuals				
Row	Time	Null Martingale Residual		Martingale Residual
12	8	0.9310	.....	-0.3862   .....
6	10	0.8569	.....	0.8589      .....
15	11	0.7769	.....	0.9209      .....
14+	25	-0.2231	.....	-0.0590  .....
11	42	0.6815	.....	0.6693      .....
1	72	0.5762	.....	0.6636      .....
7	82	0.4584	.....	-0.1774  .....
10+	100	-0.5416	.....	-0.3112  .....
8	110	0.3043	.....	-0.2387  .....
5	118	0.1219	.....	0.2700  .....
4	126	-0.1012	.....	-0.0748  .....
13	144	-0.3889	.....	-1.4532      .....
3	228	-0.7944	.....	0.5469    .....
9	314	-1.4875	.....	-1.9953      .....
2	411	-1.4875	.....	-0.7951      .....

The various residuals were discussed in detail earlier in this chapter. Only a brief definition will be given here.

### Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

### Time

This is the value of the elapsed time.

### Null Martingale Residual

These are the null martingale residuals. They are computed from a null (no covariate) model. Therneau and Grambsch (2000) suggest that the null-model martingale residuals can show the ideal functional form of the covariates so long as the covariates are not highly correlated among themselves. To find the appropriate functional form, each covariate is plotted against these residuals.

### Martingale Residual

The martingale residuals are repeated here. They were defined in the Residuals Section.

## Schoenfeld Residuals

### Schoenfeld Residuals

Row	Time	Residual Status	Residual Months	Residual Age
12	8	-3.4327	11.8140	-0.1121
6	10	-33.7298	-5.7472	-14.4483
15	11	12.7982	-0.3388	-16.9356
11	42	3.8458	-7.4120	15.1299
1	72	4.2736	-5.2365	4.8130
7	82	-15.3358	-2.7151	5.2528
8	110	20.6225	14.7012	6.6884
5	118	18.4375	2.2723	6.2230
4	126	12.1419	0.7288	5.4733
13	144	-14.3230	-4.0590	7.0669
3	228	2.1964	-8.8792	-11.2819
9	314	-7.4946	4.8715	-7.8693
2	411	0.0000	0.0000	0.0000

### Schoenfeld Residuals (Continued)

Row	Time	Residual Therapy
12	8	1.5295
6	10	-3.8822
15	11	5.7181
11	42	-3.9309
1	72	-4.3682
7	82	5.2326
8	110	-3.3664
5	118	5.3579
4	126	6.4343
13	144	-1.6923
3	228	-3.2851
9	314	-3.7473
2	411	0.0000

Cox Regression

This report displays the Schoenfeld residuals for each non-censored individual. Note that most authors suggest using the scaled Schoenfeld residuals rather than these residuals. Since these residuals were discussed earlier in this chapter, only a brief definition will be given here.

**Row**

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

**Time**

This is the value of the elapsed time.

**Schoenfeld Residuals**

The Schoenfeld residuals are defined as follows

$$r_{it} = c_t \left[ x_{it} - \sum_{r \in R_t} x_{ir} w_r \right]$$

where

$$w_r = \frac{\sum_{r \in R_t} x_{ir} \theta_r}{\sum_{r \in R_t} \theta_r}$$

Thus, this residual is the difference between the actual value of the covariate and a weighted average where the weights are determined from the risk scores. These residuals are used to estimate the influence of an observation on each of the regression coefficients. Plots of these quantities against the row number or against the corresponding covariate values are used to study these residuals.

**Scaled Schoenfeld Residuals**

**Scaled Schoenfeld Residuals**

Row	Time	Residual Status	Residual Months	Residual Age
12	8	-0.0569	0.1828	0.0276
6	10	-0.1280	-0.0528	-0.1660
15	11	0.0701	-0.0731	-0.3476
11	42	0.0355	-0.0812	0.2516
1	72	0.0487	-0.0873	0.0959
7	82	-0.1048	0.0410	0.0482
8	110	0.0753	0.1616	0.1586
5	118	0.0611	0.0237	0.0300
4	126	0.0280	0.0206	0.0116
13	144	-0.0690	-0.0020	0.1376
3	228	0.0663	-0.1822	-0.1832
9	314	-0.0262	0.0489	-0.0642
2	411	0.0000	0.0000	0.0000

Cox Regression

**Scaled Schoenfeld Residuals (Continued)**

Row	Time	Residual Therapy	
12	8	0.1550	.....
6	10	0.0223	.....
15	11	0.4537	
11	42	-0.4288	]
1	72	-0.3501	....
7	82	0.3223	....
8	110	-0.2982	.....
5	118	0.1971	.....
4	126	0.2903	.....
13	144	-0.1256	.....
3	228	-0.1361	.....
9	314	-0.1018	.....
2	411	0.0000	.....

This report displays the scaled Schoenfeld residuals for each non-censored individual. These residuals are often used to find influential observations. Since these residuals were discussed earlier in this chapter, only a brief definition will be given here.

**Row**

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

**Time**

This is the value of the elapsed time.

**Scaled Schoenfeld Residuals**

The scaled Schoenfeld residuals are defined as follows

$$r_{kt} * = m \sum_{i=1}^p V_{ik} r_{it}$$

where  $m$  is the total number of deaths in the dataset and  $V$  is the estimated covariance matrix of the regression coefficients. Hosmer and Lemeshow (1999) and Therneau and Grambsch (2000) suggest that scaling the Schoenfeld residuals by an estimate of their variance gives quantities with greater diagnostic ability. Hosmer and Lemeshow (1999) use the covariance matrix of the regression coefficients to perform the scaling.

These residuals are plotted against time to validate the proportional hazards assumption. If the proportional hazards assumption holds, the residuals will fall randomly around a horizontal line centered at zero. If the proportional hazards assumption does not hold, a trend will be apparent in the plot.



## Predicted Values

Predicted Values						
Row	Time	Cumulative Baseline Hazard	Linear Predictor XB	Relative Risk Exp(XB)	Cumulative Hazard H(T X)	Cumulative Survival S(T X)
12	8	0.0352	3.6734	39.3853	1.3862	0.2500
6	10	0.1152	0.2032	1.2254	0.1411	0.8684
15	11	0.2006	-0.9303	0.3944	0.0791	0.9239
14+	25	0.2006	-1.2244	0.2939	0.0590	0.9427
11	42	0.2945	0.1158	1.1228	0.3307	0.7184
1	72	0.3980	-0.1682	0.8452	0.3364	0.7143
7	82	0.5220	0.8135	2.2557	1.1774	0.3081
10+	100	0.5220	-0.5171	0.5963	0.3112	0.7326
8	110	0.7048	0.5640	1.7577	1.2387	0.2897
5	118	0.9407	-0.2536	0.7760	0.7300	0.4819
4	126	1.2341	-0.1382	0.8709	1.0748	0.3414
13	144	1.6915	0.3718	1.4503	2.4532	0.0860
3	228	2.3841	-1.6603	0.1901	0.4531	0.6356
9	314	3.5461	-0.1688	0.8447	2.9953	0.0500
2	411	3.5461	-0.6808	0.5062	1.7951	0.1661

This report displays various values estimated by the model. These are centered if the Centered X's option is selected.

### Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

### Time

This is the value of the elapsed time.

### Baseline Cumulative Hazard

This estimates the cumulative baseline hazard of this individual. The baseline hazard occurs when all covariates are equal to zero (or to their means if centering is used). It is the value of  $H_0(T)$  which is calculated using the formula

$$H_0(T) = -\ln(S_0(T))$$

### Linear Predictor (XB)

This is the value of the linear portion of the Cox regression model. It is the logarithm of the ratio of the hazard rate to the baseline hazard rate. That is, it is the logarithm of the hazard ratio (or relative risk). The formula for the linear predictor is

$$\ln \left[ \frac{h(T)}{h_0(T)} \right] = \sum_{i=1}^p x_i \beta_i$$

This value is occasionally suggested for use in plotting.

## Cox Regression

**Relative Risk Exp(XB)**

This is the ratio between the actual hazard rate and the baseline hazard rate, sometimes called the *risk ratio* or the *relative risk*. The formula for this quantity is

$$\begin{aligned}\frac{h(T)}{h_0(T)} &= \exp\left(\sum_{i=1}^p x_i \beta_i\right) \\ &= e^{x_1 \beta_1} e^{x_2 \beta_2} \dots e^{x_p \beta_p}\end{aligned}$$

**Cumulative Hazard H(T|X)**

Under the proportional hazards regression model, the cumulative hazard is the sum of the individual hazard rates from time zero to time  $T$ .

$$\begin{aligned}H(T, X) &= \int_0^T h(u, X) du \\ &= \int_0^T h_0(u) e^{\sum_{i=1}^p x_i \beta_i} du \\ &= e^{\sum_{i=1}^p x_i \beta_i} \int_0^T h_0(u) du \\ &= H_0(T) e^{\sum_{i=1}^p x_i \beta_i}\end{aligned}$$

Note that the time survival time  $T$  is present in  $H_0(T)$ , but not in  $e^{\sum_{i=1}^p x_i \beta_i}$ . Hence, the cumulative hazard up to time  $T$  is represented in this model by a baseline cumulative hazard  $H_0(T)$  which is adjusted for the covariates by multiplying by the factor  $e^{\sum_{i=1}^p x_i \beta_i}$ .

**Cumulative Survival S(T|X)**

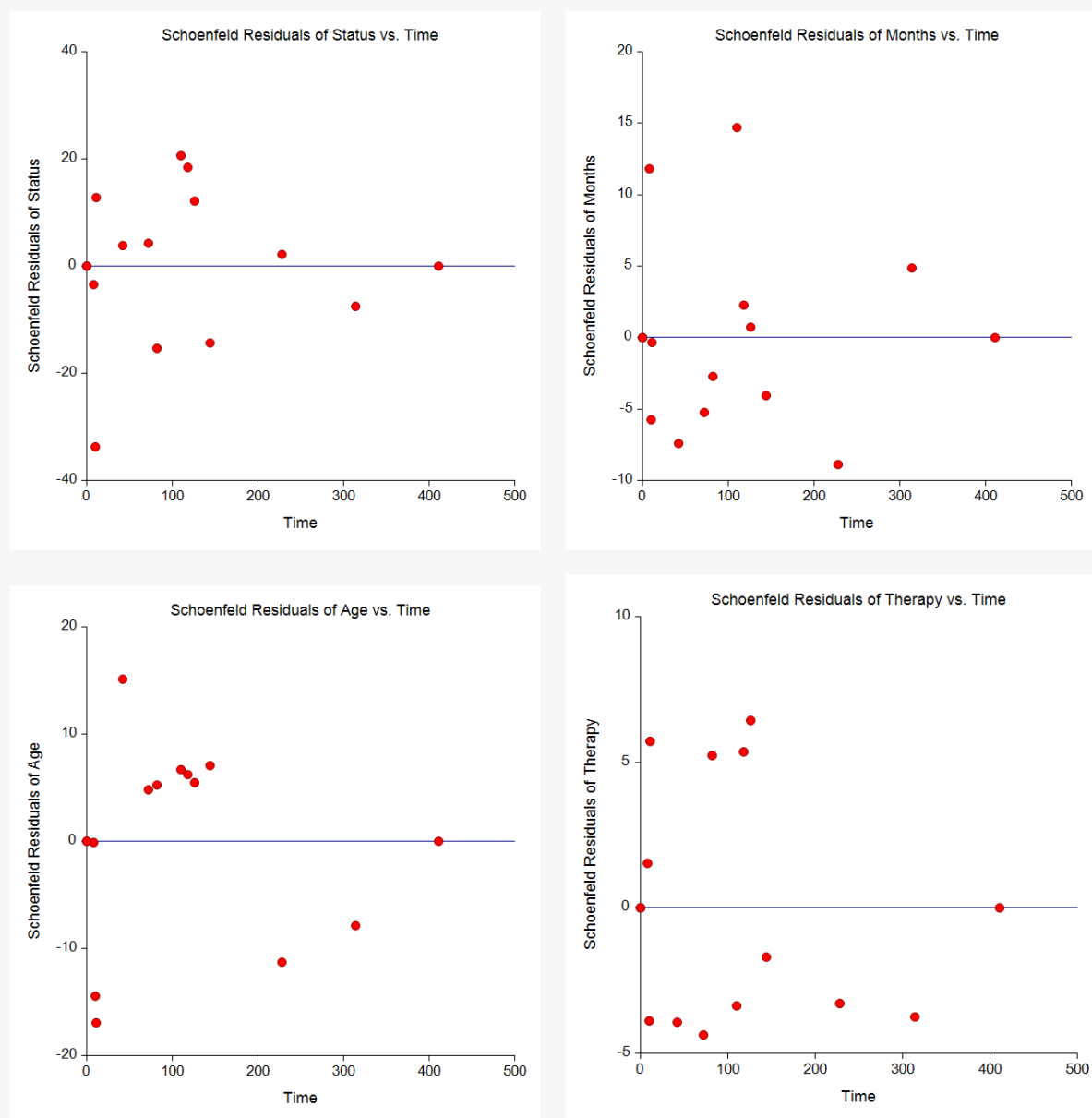
Under the proportional hazards regression model, the cumulative survival is the probability that an individual survives past  $T$ . The formula for the cumulative survival is

$$\begin{aligned}S(T, X) &= \exp(-H(T, X)) \\ &= \exp\left(-H_0(T) e^{\sum_{i=1}^p x_i \beta_i}\right) \\ &= \left[e^{-H_0(T)}\right] e^{-\sum_{i=1}^p x_i \beta_i} \\ &= S_0(T) e^{-\sum_{i=1}^p x_i \beta_i}\end{aligned}$$

Note that the time survival time  $T$  is present in  $S_0(T)$ , but not in  $e^{-\sum_{i=1}^p x_i \beta_i}$ .

## Schoenfeld Residuals vs Time Plot(s)

Schoenfeld Residuals vs Time Plot(s)



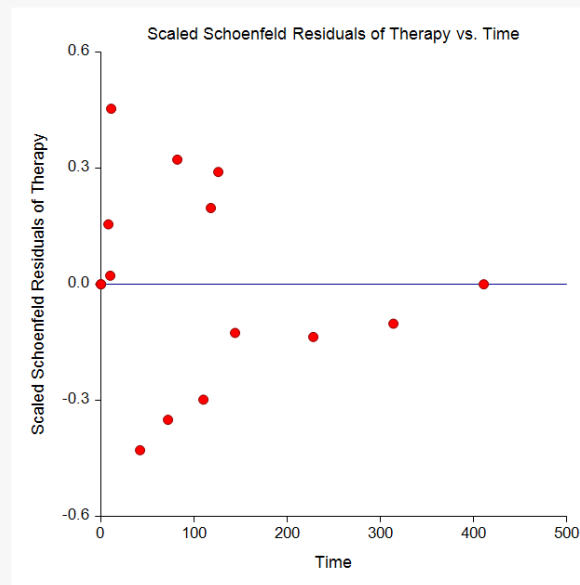
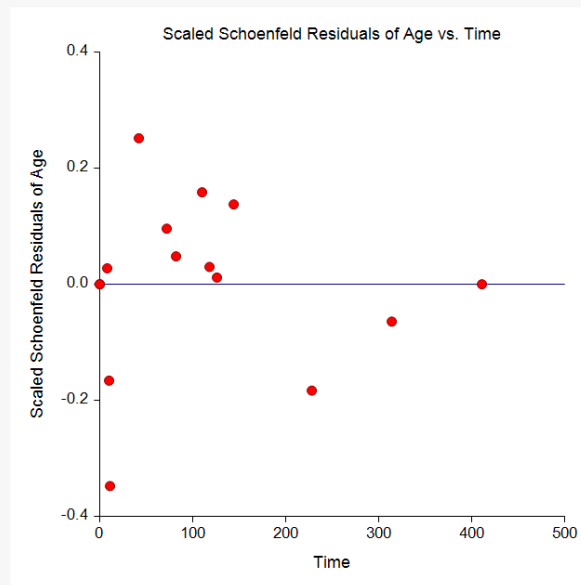
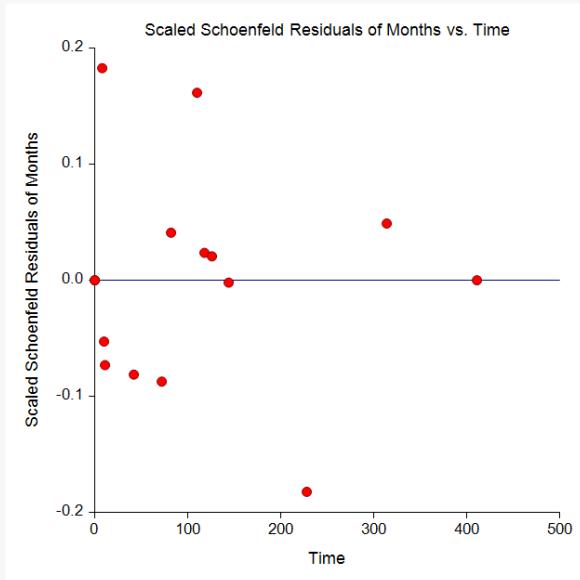
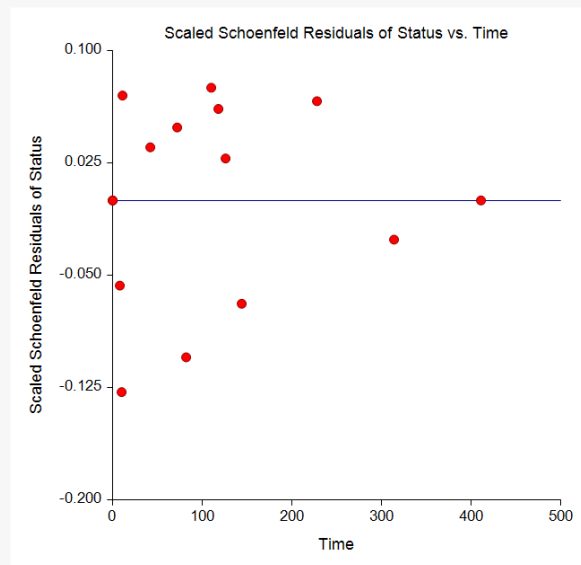
The Schoenfeld residuals are plotted for two reasons. First of all, these plots are useful in assessing whether the proportional hazards assumption is met. If the least squares line is horizontal and the lowest curve seems to track the least squares line fairly well, the proportional hazard assumption is reasonable.

Second, points that are very influential in determining the estimated regression coefficient for a covariate show up as outliers on these plots. When influential points are found, it is important to make sure that the data associated with these points are accurate. It is not advisable to remove these influential points unless a specific reason can be found for doing so.

Many authors suggest that the scaled Schoenfeld residuals are more useful than these, unscaled, residuals.

## Scaled Schoenfeld Residuals vs Time Plot(s)

Scaled Schoenfeld Residuals vs Time Plot(s)

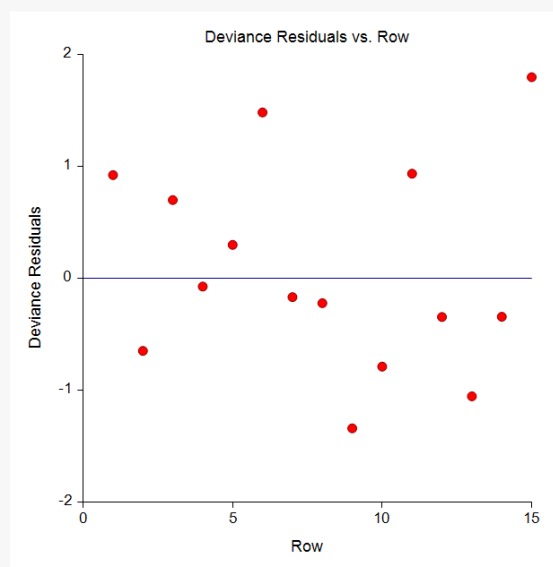


The scaled Schoenfeld residuals are plotted for two reasons. First of all, these plots are useful in assessing whether the proportional hazards assumption is met. If the least squares line is horizontal and the lowest curve seems to track the least squares line fairly well, the proportional hazard assumption is reasonable.

Second, points that are very influential in determining the estimated regression coefficient for a covariate show up as outliers on these plots. When influential points are found, it is important to make sure that the data associated with these points are accurate. It is not advisable to remove these influential points unless a specific reason can be found for doing so.

## Deviance Residuals vs Row Plot

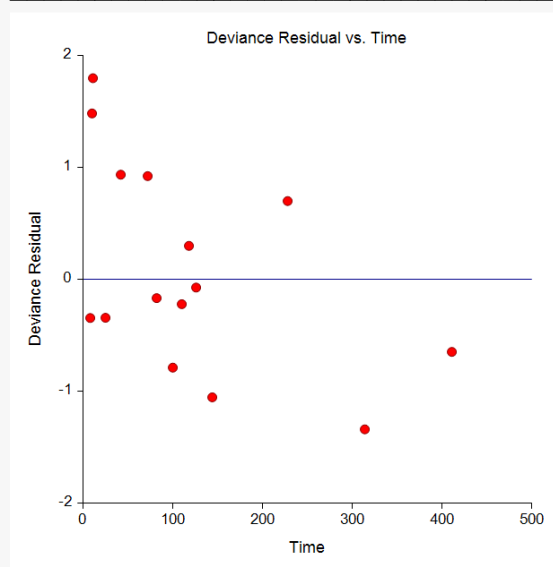
Deviance Residuals vs Row Plot



This plot allows you to find outliers. These outliers should be double-checked to be certain that the data are not in error. You should not routinely remove outliers unless you can find a good reason for doing so. Often, the greatest insight during an investigation comes while considering why these outliers are different.

## Deviance Residuals vs Time Plot

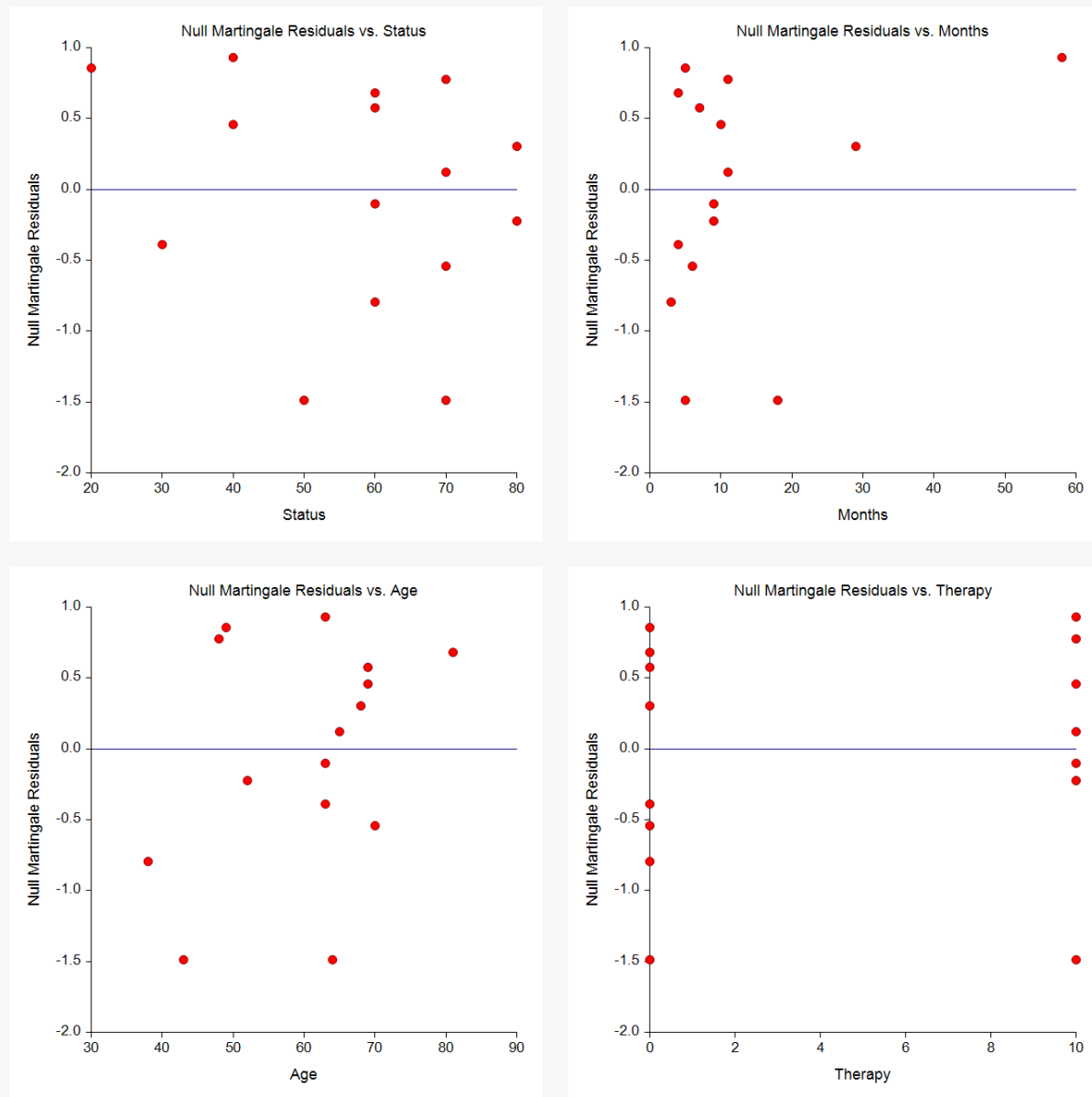
Deviance Residuals vs Time Plot



This plot allows you to find outliers. These outliers should be double-checked to be certain that the data are not in error. You should not routinely remove outliers unless you can find a good reason for doing so. Often, the greatest insight during an investigation comes while considering why these outliers are different.

## Null Martingale Residuals vs X Plots

Null Martingale Residuals vs X Plot(s)

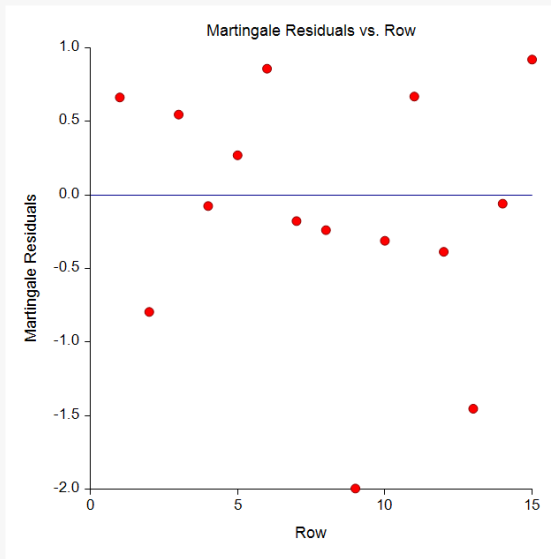


Each of the covariates are plotted against the null martingale residuals. If the covariates are not highly correlated, these plots will show the appropriate functional form of each covariate. A loess curve and a regular least squares line are added to the plot to aid the eye. Ideally, the loess curve will track along the least squares line. Be careful not to over interpret the ends of the loess curves which are based on only a few individuals.

When curvature is present, you have to decide how the model should be modified to deal with it. You might need to add the square or the logarithm of the covariate to the model.

## Martingale Residuals vs Row Plot

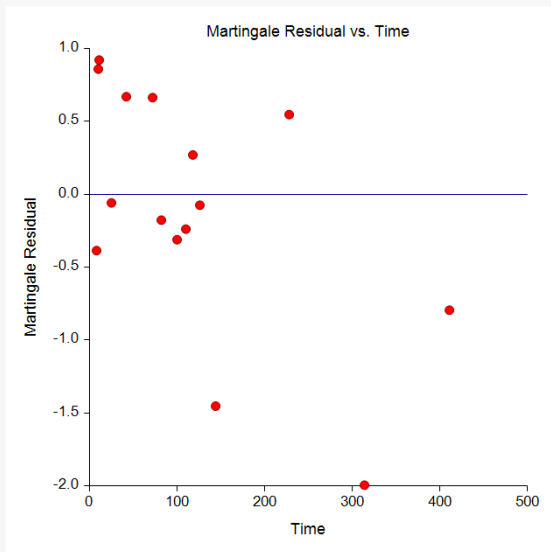
Martingale Residuals vs Row Plot



This plot allows you to find outliers. These outliers should be double-checked to be certain that the data are not in error. You should not routinely remove outliers unless you can find a good reason for doing so. Often, the greatest insight during an investigation comes while considering why these outliers are different.

## Martingale Residuals vs Time Plot

Martingale Residuals vs Time Plot



This plot allows you to find outliers. These outliers should be double-checked to be certain that the data are not in error. You should not routinely remove outliers unless you can find a good reason for doing so. Often, the greatest insight during an investigation comes while considering why these outliers are different.

## Example 2 – Subset Selection

This section presents an example of how to conduct a subset selection. We will again use the LungCancer dataset that was used in Example 1. In this run, we will be trying to find a subset of the covariates that should be kept in the regression model.

### Setup

To run this example, complete the following steps:

**1 Open the LungCancer example dataset**

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

**2 Specify the Cox Regression procedure options**

- Find and open the **Cox Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab	
Time.....	<b>Time</b>
Ties Method.....	<b>Efron</b>
Censor .....	<b>Censor</b>
Numeric X's .....	<b>Status-Therapy</b>
Search Method .....	<b>Hierarchical Forward with Switching</b>
Reports Tab	
Subset Summary .....	<b>Checked</b>
Subset Detail .....	<b>Checked</b>
All Other Reports .....	<b>Unchecked</b>

**3 Run the procedure**

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



## Subset Selection Summary

### Subset Selection Summary

Subset Selection Method = Hierarchical Forward with Switching

Number of Terms	X's	Log- Likelihood	R <sup>2</sup>	
			Value	Change
0	0	-23.3349	0.0000	0.0000
1	1	-21.8803	0.1763	0.1763
2	2	-21.0354	0.2641	0.0878
3	3	-20.1352	0.3473	0.0832
4	4	-20.1143	0.3491	0.0018

This report shows the best log-likelihood value for each subset size. In this example, it appears that a model with three terms provides the best model. Note that adding the fourth variable does not increase the R-squared value very much.

### Number of Terms

The number of terms.

### Number of X's

The number of X's that were included in the model. Note that in this case, the number of terms matches the number of X's. This would not be the case if some of the terms were categorical variables.

### Log-Likelihood

This is the value of the log-likelihood function evaluated at the maximum likelihood estimates. Our goal is to find a subset size above which little is gained by adding more variables.

### R<sup>2</sup> Value

This is the value of  $R^2$  calculated using the formula

$$R_k^2 = 1 - \exp\left[\frac{2}{n}(L_0 - L_k)\right]$$

as discussed in the introduction. We are looking for the subset size after which this value does not increase by a meaningful amount.

### R<sup>2</sup> Change

This is the increase in  $R^2$  that occurs when each new subset size is reached. Search for the subset size below which the  $R^2$  value does not increase by more than 0.02 for small samples or 0.01 for large samples.

In this example, the optimum subset size appears to be three terms.

## Subset Selection Detail

### Subset Selection Detail

Subset Selection Method = Hierarchical Forward with Switching

Step	Action	Number of		Log-Likelihood	R <sup>2</sup>	Term	
		Terms	X's			Entered	Removed
1	Begin	0	0	-23.3349	0.0000		
2	Add	1	1	-21.8803	0.1763	Months	
3	Add	2	2	-21.0354	0.2641	Status	
4	Add	3	3	-20.1352	0.3473	Age	
5	Add	4	4	-20.1143	0.3491	Therapy	

This report shows the highest log-likelihood for each subset size. In this example, it appears that three terms provide the best model. Note that adding THERAPY does not increase the R-squared value very much.

### Action

This item identifies the action that was taken at this step. A term was added, removed, or two were switched.

### Number of Terms

The number of terms.

### Number of X's

The number of X's that were included in the model. Note that in this case, the number of terms matches the number of X's. This would not be the case if some of the terms were categorical variables.

### Log-Likelihood

This is the value of the log-likelihood function after the completion of this step. Our goal is to find a subset size above which little is gained by adding more variables.

### R<sup>2</sup>

This is the value of R<sup>2</sup> calculated using the formula

$$R_k^2 = 1 - \exp\left[\frac{2}{n}(L_0 - L_k)\right]$$

as discussed in the introduction. We are looking for the subset size after which this value does not increase by a meaningful amount.

### Terms Entered and Removed

These columns identify the terms added, removed, or switched.

## Discussion of Example 2

After considering these reports, it was decided to include AGE, MONTHS, and STATUS in the final regression model. Another run is performed using only these independent variables. A complete residual analysis would be necessary before the equation is finally adopted.

### Regression Coefficients

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	P-Value	Pseudo R <sup>2</sup>
B1: Status	-0.031482	0.019680	0.9690	57.33333	-1.5997	0.1097	0.2037
B2: Months	0.063724	0.032004	1.0658	12.6	1.9912	0.0465	0.2838
B3: Age	0.041940	0.033413	1.0428	60.33333	1.2552	0.2094	0.1361

This report displays the results of the proportional hazards estimation. Note that the Wald tests indicate that only Months is statistically significant. Because of the small sample size of this example and because they add a great deal to the R-squared value, we have added Age and Status to the final model.

## Example 3 – Cox Regression with Categorical Variables

This example will demonstrate the analysis of categorical independent variables. A study was conducted to evaluate the influence on survival time of three variables: Age, Gender, and Treatment. The ages of the study participants were grouped into three age categories: 20, 40, and 60. The first age group (20) was selected as the reference group. The female group was selected as the reference group for Gender. The Treatment variable represented three groups: a control and two treatment groups. The control group was selected as the reference group for Treatment. The data for this study are contained in the **CoxReg** dataset.

### Setup

To run this example, complete the following steps:

#### 1 Open the CoxReg example dataset

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

#### 2 Specify the Cox Regression procedure options

- Find and open the **Cox Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab	
Time.....	<b>Time</b>
Ties Method.....	<b>Efron</b>
Censor.....	<b>Status</b>
Categorical X's.....	<b>Treatment(B;C) Age(B;20) Gender(B;F)</b>
Frequencies.....	<b>Count</b>
Terms.....	<b>Up to 2-Way</b>
Reports Tab	
Run Summary.....	<b>Checked</b>
Regression Coefficients.....	<b>Checked</b>
C.L. of Regression Coefficients.....	<b>Checked</b>
Analysis of Deviance.....	<b>Checked</b>
Log-Likelihood and R <sup>2</sup> .....	<b>Checked</b>

#### 3 Run the procedure

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

## Run Summary

Run Summary			
Item	Value	Rows	Value
Time Variable	Time	Rows Processed	73
Censor Variable	Status	Rows Used in Estimation	73
Frequency Variable	Count	Rows with X's Missing	0
Independent Variables Available	3	Rows with Y Missing	0
Number of X's in the Model	13	Rows Failed	54
Final Log-Likelihood	-222.9573	Rows Censored	19
Number of Likelihood Iterations	6 of 20	Sum of Frequencies	137
Convergence Criterion	1E-09	Sum of Censored Frequencies	83
Achieved Convergence	5.363288E-16	Sum of Failed Frequencies	54
Completion Status	Normal completion		
Starting B's	0		
Subset Selection Method	None		

This report summarizes the characteristics of the dataset. Note that 137 individuals were included in this study of which 83 were censored.

## Regression Coefficients

Regression Coefficients							
Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	P-Value	Pseudo R <sup>2</sup>
B1: (Treatment="T1")	0.315769	0.707474	1.3713	0.3357664	0.4463	0.6554	0.0044
B2: (Treatment="T2")	0.606087	0.674007	1.8332	0.3211679	0.8992	0.3685	0.0177
B3: (Age=40)	0.178527	0.720145	1.1955	0.3065693	0.2479	0.8042	0.0014
B4: (Age=60)	0.747377	0.652733	2.1115	0.3430657	1.1450	0.2522	0.0283
B5: (Gender="M")	0.199655	0.629734	1.2210	0.5182482	0.3170	0.7512	0.0022
B6: (Treatment="T1")*(Age=40)	-0.228646	0.872282	0.7956	0.1094891	-0.2621	0.7932	0.0015
B7: (Treatment="T1")*(Age=60)	-0.124997	0.851308	0.8825	0.1021898	-0.1468	0.8833	0.0005
B8: (Treatment="T2")*(Age=40)	-0.442119	0.843234	0.6427	0.08759124	-0.5243	0.6001	0.0061
B9: (Treatment="T2")*(Age=60)	-1.726851	0.885161	0.1778	0.1167883	-1.9509	0.0511	0.0780
B10: (Treatment="T1")*(Gender="M")	-0.976831	0.714997	0.3765	0.1751825	-1.3662	0.1719	0.0398
B11: (Treatment="T2")*(Gender="M")	-0.553592	0.721736	0.5749	0.1605839	-0.7670	0.4431	0.0129
B12: (Age=40)*(Gender="M")	0.420448	0.701899	1.5226	0.1678832	0.5990	0.5492	0.0079
B13: (Age=60)*(Gender="M")	0.366048	0.709829	1.4420	0.1970803	0.5157	0.6061	0.0059

This report displays the results of the proportional hazards estimation. Note that the names of the interaction terms are too long to fit in the space allotted, so the rest of the information appears on the next line.

### Independent Variable

It is important to understand the variable names of the interaction terms. For example, consider the term: (Treatment="T2")\*(Gender="M"). This variable was created by multiplying two indicator variables. The first indicator is "1" when the treatment is "T2" and "0" otherwise. The second indicator is "1" when the gender is "M" and "0" otherwise. This portion of the gender-by-treatment interaction is represented by the product of these two variables.

## Confidence Interval Limits for Regression Coefficients

Confidence Interval Limits for Regression Coefficients

Independent Variable	Regression Coefficient (B)	95% Confidence Interval Limits for B		Risk Ratio Exp(B)	95% Confidence Interval Limits for Exp(B)	
		Lower	Upper		Lower	Upper
B1: (Treatment="T1")	0.315769	-1.070855	1.702392	1.3713	0.3427	5.4871
B2: (Treatment="T2")	0.606087	-0.714943	1.927116	1.8332	0.4892	6.8697
B3: (Age=40)	0.178527	-1.232933	1.589986	1.1955	0.2914	4.9037
B4: (Age=60)	0.747377	-0.531956	2.026709	2.1115	0.5875	7.5891
B5: (Gender="M")	0.199655	-1.034602	1.433911	1.2210	0.3554	4.1951
B6: (Treatment="T1")*(Age=40)	-0.228646	-1.938287	1.480996	0.7956	0.1440	4.3973
B7: (Treatment="T1")*(Age=60)	-0.124997	-1.793530	1.543536	0.8825	0.1664	4.6811
B8: (Treatment="T2")*(Age=40)	-0.442119	-2.094827	1.210589	0.6427	0.1231	3.3555
B9: (Treatment="T2")*(Age=60)	-1.726851	-3.461735	0.008033	0.1778	0.0314	1.0081
B10: (Treatment="T1")*(Gender="M")	-0.976831	-2.378199	0.424538	0.3765	0.0927	1.5289
B11: (Treatment="T2")*(Gender="M")	-0.553592	-1.968169	0.860985	0.5749	0.1397	2.3655
B12: (Age=40)*(Gender="M")	0.420448	-0.955248	1.796145	1.5226	0.3847	6.0264
B13: (Age=60)*(Gender="M")	0.366048	-1.025192	1.757288	1.4420	0.3587	5.7967

This report provides the confidence intervals for the regression coefficients and the risk ratios. The confidence coefficient, in this example 95%, was specified on the Format tab. Note that the names of the interaction terms are too long to fit in the space allotted, so the rest of the information appears on the next line.

### Independent Variable

It is important to understand the variable names of the interaction terms. For example, consider the term: (Treatment="T2")\*(Gender="M"). This variable was created by multiplying two indicator variables. The first indicator is "1" when the treatment is "T2" and "0" otherwise. The second indicator is "1" when the gender is "M" and "0" otherwise. This portion of the gender-by-treatment interaction is represented by the product of these two variables.

## Analysis of Deviance

### Analysis of Deviance

Term(s) Omitted	DF	-2 Log- Likelihood	Increase from Model Deviance (Chi <sup>2</sup> )	P-Value
All Terms	13	454.5022	8.5876	0.8033
Treatment	2	446.7191	0.8044	0.6688
Age	2	447.2661	1.3515	0.5088
Gender	1	446.0147	0.1001	0.7517
Treatment*Age	4	451.1965	5.2819	0.2596
Treatment*Gender	2	447.7827	1.8681	0.3930
Age*Gender	2	446.3421	0.4275	0.8076
None(Model)	13	445.9146		

The P-Value is for testing the significance of each term after adjusting for all other terms.

This report is the Cox regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms are statistically significant after adjusting for all other terms in the model.

The DF (degrees of freedom) column indicates the number of binary variables needed to represent each term. The chi<sup>2</sup> test is used to test the significance of all binary variables associated with a particular term.

## Log-Likelihood and R<sup>2</sup>

### Log-Likelihood and R<sup>2</sup>

Term(s) Omitted	DF	Log- Likelihood	R <sup>2</sup> of Remaining Term(s)	Reduction from Model R <sup>2</sup>
All Terms	13	-227.2511	0.0000	0.0608
Treatment	2	-223.3595	0.0552	0.0055
Age	2	-223.6331	0.0514	0.0093
Gender	1	-223.0074	0.0601	0.0007
Treatment*Age	4	-225.5982	0.0238	0.0369
Treatment*Gender	2	-223.8914	0.0479	0.0129
Age*Gender	2	-223.1711	0.0578	0.0029
None(Model)	13	-222.9573	0.0608	0.0000

This report displays the Log Likelihood and R<sup>2</sup> that is achieved when each term is omitted from the regression model. The DF (degrees of freedom) column indicates the number of binary variables needed to represent each term. The chi<sup>2</sup> test is used to test the significance of all binary variables associated with a particular term.

## Example 4 – Validation of Cox Regression using Collett (1994)

Collett (1994), pages 156 and 157, present a dataset giving the results of a small study about kidney dialysis. This dataset contains two independent variables: Age and Sex. These data are contained in the **NCSS** dataset called Collett157.

Collett (1994) gives the estimated regression coefficients as 0.030 for Age and -2.711 for Sex. The chi-square test for Sex is 6.445 and the chi-square test for Age is 1.320. The Cox-Snell residual for the first patient is 0.3286. The martingale residual for this patient is 0.6714. The deviance residual for this patient is 0.9398. The Schoenfeld residuals for this patient are -1.0850 and -0.2416.

### Setup

To run this example, complete the following steps:

#### 1 Open the Collett157 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Collett157** and click **OK**.

#### 2 Specify the Cox Regression procedure options

- Find and open the **Cox Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

##### Variables, Model Tab

Time.....	<b>Time</b>
Ties Method.....	<b>Efron</b>
Censor.....	<b>Status</b>
Numeric X's.....	<b>Age-Sex</b>

##### Reports Tab

Regression Coefficients.....	<b>Checked</b>
Analysis of Deviance.....	<b>Checked</b>
Residuals.....	<b>Checked</b>
Schoenfeld Residuals.....	<b>Checked</b>

#### 3 Run the procedure

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



## Validation Report

### Regression Coefficients

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	P-Value	Pseudo R <sup>2</sup>
B1: Age	<b>0.030371</b>	0.026237	1.0308	31.46154	1.1576	0.2470	0.1181
B2: Sex	<b>-2.710762</b>	1.095898	0.0665	1.769231	-2.4736	0.0134	0.3795

### Analysis of Deviance

Term(s) Omitted	DF	-2 Log-Likelihood	Increase from Model Deviance (Chi <sup>2</sup> )	P-Value
All Terms	2	40.9454	6.4779	0.0392
Age	1	35.7880	<b>1.3204</b>	0.2505
Sex	1	40.9132	<b>6.4456</b>	0.0111
None(Model)	2	34.4676		

The P-Value is for testing the significance of each term after adjusting for all other terms.

### Residuals

Row	Time	Cox-Snell Residual	Martingale Residual	Deviance Residual
1	8	<b>0.3286</b>  .....	<b>0.6714</b>    .....	<b>0.9398</b>     .....
2	15	0.0785  .....	0.9215      .....	1.8020      .....
3	22	1.4331      .....	-0.4331  .....	-0.3828  .....
4	24	0.0939  .....	0.9061      .....	1.7087      .....
5	30	1.7736      .....	-0.7736      .....	-0.6334   .....
6+	54	0.3117  .....	-0.3117  .....	-0.7895    .....
7	119	0.2655  .....	0.7345     .....	1.0877      .....
8	141	0.5386   .....	0.4614   .....	0.5611  .....
9	185	1.6523      .....	-0.6523    .....	-0.5480  .....
10	292	1.4234      .....	-0.4234  .....	-0.3751  .....
11	402	1.4207      .....	-0.4207  .....	-0.3730  .....
12	447	2.3927      .....	-1.3927      .....	-1.0201      .....
13	536	1.5640      .....	-0.5640    .....	-0.4832  .....

Cox Regression

**Schoenfeld Residuals**

Row	Time	Residual Age	Residual Sex
1	8	<b>-1.0850</b>  .....	<b>-0.2416</b>     .....
2	15	14.4930          ..	0.6644
3	22	3.1291   .....	-0.3065     .....
4	24	-10.2215       .....	0.4341       .....
5	30	-16.5882          ..	-0.5504          ..
7	119	-17.8286	0.0000  .....
8	141	-7.6201      .....	0.0000  .....
9	185	17.0910          .	0.0000  .....
10	292	10.2390       .....	0.0000  .....
11	402	2.8575   .....	0.0000  .....
12	447	5.5338    .....	0.0000  .....
13	536	0.0000  .....	0.0000  .....

You can verify that the results matched those of Collett (1994) within rounding.