**INTRODUCTION TO SURVIVAL ANALYSES**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

Survival analyses entail a suite of statistical tests, designed to predict the duration before some event is likely to unfold. Historically, these analyses were designed to predict how long someone is likely to survive, hence the name. But, survival analyses can be valuable in a range of other circumstances, such as predicting when

* employees will leave their job
* students will complete their course or
* people will overcome some habit

In all instances, survival analyses are designed to predict when some event—such as death, departure, completion, or withdrawal—is likely to transpire. Therefore, survival analyses are sometimes called time-to-event analyses or event-history analyses.

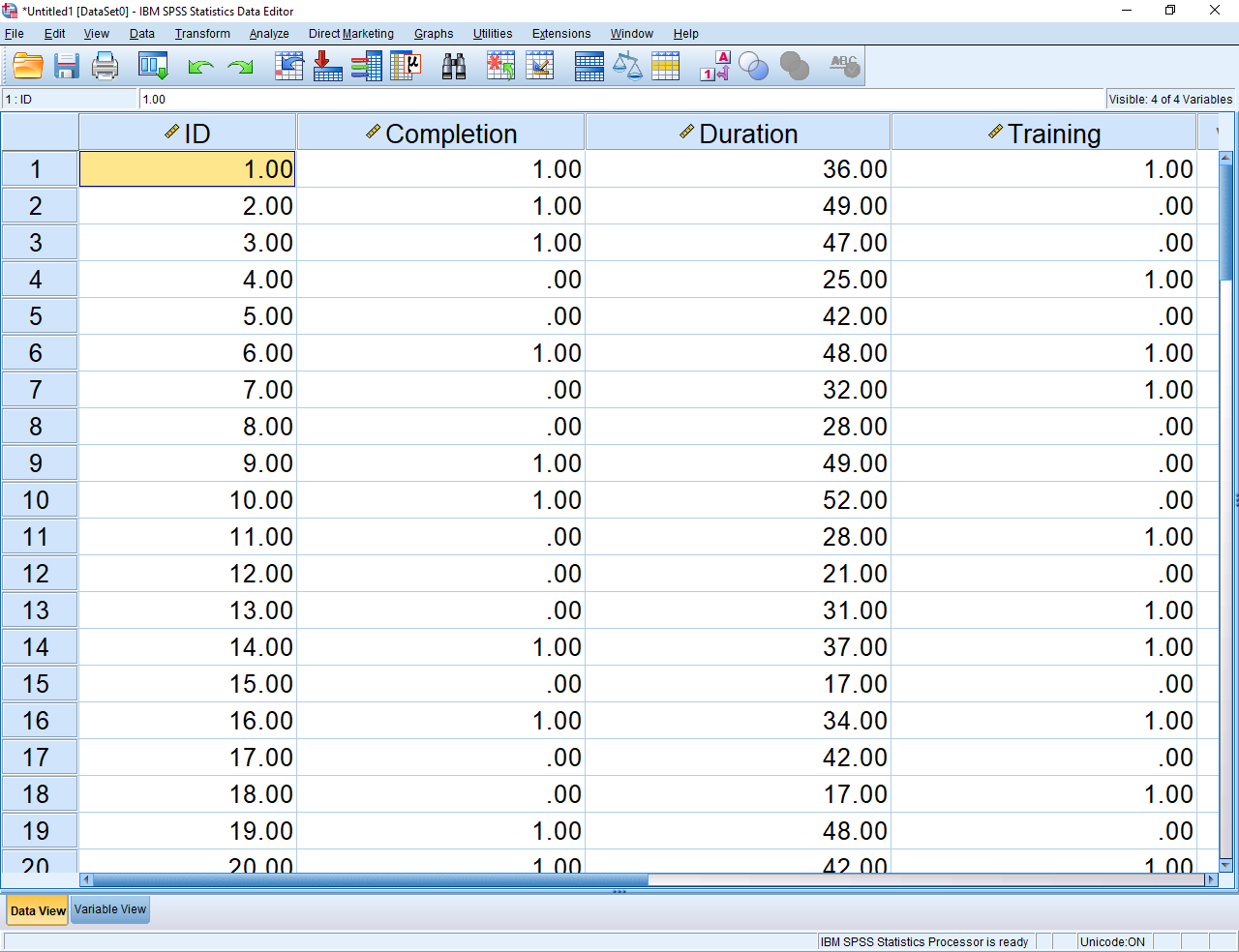
|  |
| --- |
| **Brief illustration of the Kaplan-Meier procedure** |

To learn about survival analyses, this document will first present two distinct examples: the Kaplan Meier procedure and the Cox proportional hazards regression analysis. After you understand these examples, many other concepts that are relevant to survival analyses are quite straightforward.

**Censored data**

To illustrate the Kaplan-Meier procedure, suppose you wanted to assess whether students who complete training on how to write more precisely tend to finish their thesis earlier. You might, for example, collect the data that resembles that following SPSS file. In this file

* each row corresponds to a separate person, and each person had, at least previously, been enrolled in a PhD at CDU
* completion refers to whether these individuals had completed their thesis or not. Specifically, 1 indicates completion, and 0 indicates yet to complete
* duration refers to the number of months since the individual enrolled in the PhD, full time equivalent,
* training refers to whether or not these individuals have received training on writing—specifically, training on how to write more precisely; 1 indicates they had received training.



To ascertain whether training expedites completions, some researchers merely compare people who received the training with people who did not receive training on duration, perhaps using a t-test. Unfortunately

* some of these individuals, such as the fourth and fifth participant, have not completed their thesis yet
* hence, the durations in this spreadsheet underestimates how long they might need to complete their thesis
* these individuals might not complete their theses in 25 and 42 months respectively

Alternatively, researchers could merely exclude all participants who had not completed their thesis—and then compare people who received the training with people who did not receive training on duration. But

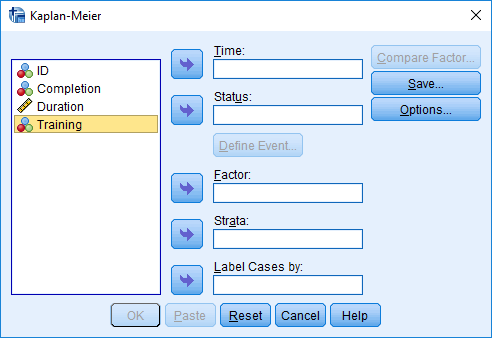
* this analysis overlooks important data
* to illustrate, consider the fifth participant. This participant had not submitted the thesis even within 42 months and did not receive the training in writing.
* this information shows that, without training in writing, participants might need 42 months or longer to complete a thesis.
* these data are, therefore, somewhat informative.

Survival analyses will utilize all data but recognize that

* if participants had yet to submit their thesis, the estimated duration in this data file is a minimum.
* that is, in practice, the participants might need a significantly longer duration to complete the thesis.
* in statistical lingo, we conclude that, if participants had not experienced the event—such as the completion of a thesis—the estimates of time are a minimum only.
* we refer to time as a censored variable; that is, the precise time is censored or unavailable if the person had yet to experience the event: thesis completion

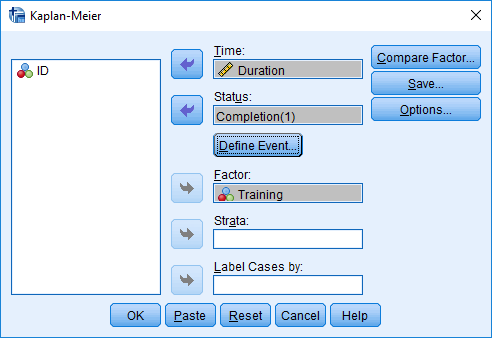
**Conducting the Kaplan Meier Procedure**

If you can access and use SPSS, you could enter the previous data to practice the Kaplan Meier Procedure. If you do not use SPSS, do not be too concerned: Follow the example anyway. The document will later offer some insights on how to conduct this procedure in R. After entering the data, select the “Analyse” menu”, “Survival”, and then “Kaplan-Meier”, to generate the following screen.



You now need to specify your variables. In particular

* use the top arrow to transfer “Duration” into the box called “Time”
* similarly, transfer “Completion” into the box called “Status”.
* click “Define Event”. Enter “1” into the box alongside “Single value”. This procedure merely informs the computer the event in which you are most interested—completion of a thesis—is coded as 1 in the data file. Then press Continue.
* transfer “Training” to the box labelled “Factor” to generate the following screen



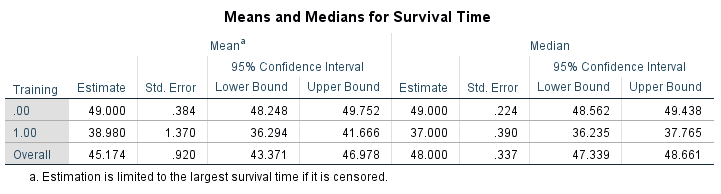
Before you execute the analysis, you should choose a couple of options. Specifically

* click “Compare Factor”. Tick the three options at the top: Log rank, Breslow, and Tarone-Ware
* then click “Options” and choose “Survival” and “Hazard” before pressing continue and OK.

**Tables generated by the Kaplan Meier Procedure**

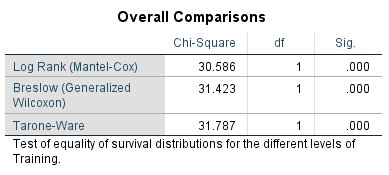
After you press OK, SPSS should generate several tables and plots. One table is called the Mean and Medians for Survival time. As the following table shows

* if participants had not received the training in writing, the procedure estimates they tend to complete the thesis within 49 months on average
* if participants had received the training in writing, the procedure estimates they tend to complete the thesis within 39.98 months on average



The next table offers some insight into whether this difference between the two estimates is significant. As this table indicates

* all the tests generate a significant result; that is, the p value is less than .05 and indeed less than .001
* consequently, training in writing does significantly diminish the duration that candidates need to complete their thesis

****

In some instances, however, these three tests do not generate the same conclusion—and so you need to decide which test to utilize. To appreciate the difference between these tests

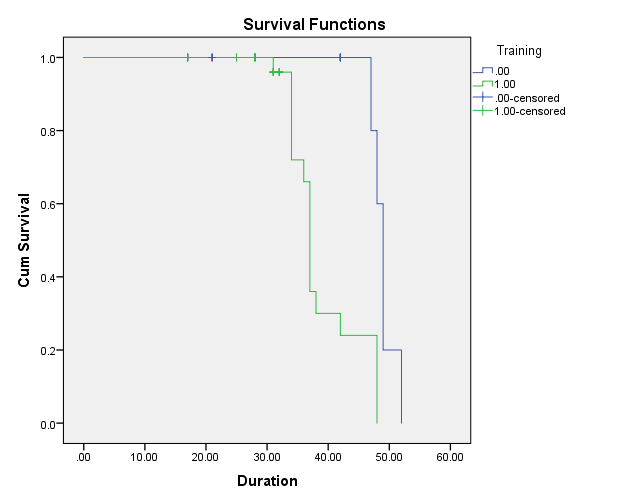
* the log rank test weights all the times or durations equally
* the Breslow test assigns more weights to times and durations that are more common in the data set. The weight assigned to a specific duration, such as 40 months, equals the number of participants who have been enrolled for 40 months
* the Tarone-Ware test also assigns more weights to times and durations that are more common in the data set. The weight assigned to a specific duration, such as 40 months, equals the square root of number of participants who have been enrolled for 40 months

So, which test should you utilize? In practice, most researchers utilize the log-rank test. However, perhaps the Breslow test or Tarone-Ware test is more applicable if some times or durations are more common than other times or durations for a reason.

**The survival function**

The Kaplan Meier Procedure can also generate some important plots. The following figure, for example, shows the survival function. To interpret this figure, consider the green line first—corresponding to the participants who had received the training on writing. To interpret the green line

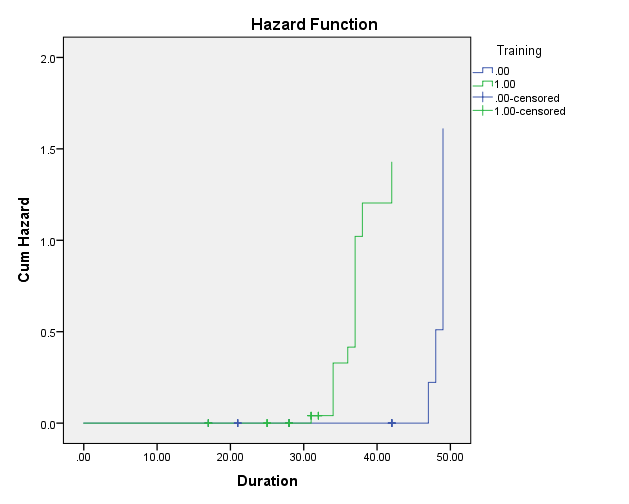
* orient your attention to the green line when duration is about 20. At this point, cumulate survival is 1.0 or 100%.
* in other words, if participants had been studying 20 months, 100% of participants had not experienced the event—that is, none had completed their thesis
* now orient your attention to the green line when duration is about 40. At this point, cumulate survival is about 0.3 or 30%
* in other words, if participants had been studying 40 months, about 30% had not experienced the event—that is, 70% had completed their thesis.
* thus, “survival” is equivalent to “not reaching the event labelled 1”
* for the blue line, individuals are less likely to have experienced the event; in other words, without training, they were not as likely to have submitted their thesis after a specific duration.



**The hazard function**

The next figure shows a similar plot called the hazard function. To interpret this figure, again consider the green line first—corresponding to the participants who had received the training on writing. To interpret the green line

* orient your attention to the green line when duration is about 40. At this point, the cumulative hazard is about 1.3.
* to understand this number, if participants had yet to submit within 40 months, they are, roughly speaking 1.3 times as likely to submit now than not submit now—to experience the event than not to experience the event
* now orient your attention to the green line when duration is about 10. At this point, the cumulative hazard is 0
* thus, if participants had yet to submit within 10 months, they are 0 times as likely to submit now than not submit now. They will thus not submit at this time.
* the green line tends to be above the blue line. So, at any time, participants who had received training are even more likely to submit now than not submit compared to participants who had not received training

****

|  |
| --- |
| **Brief illustration of Cox Regression** |

Whenever researchers explore whether some categorical predictor affects the duration before some event, the Kaplan-Meier procedure is suitable. In the previous example, the predictor—whether or not participants had received training in writing—was categorical. Therefore, the Kaplan-Meier procedure was suitable.

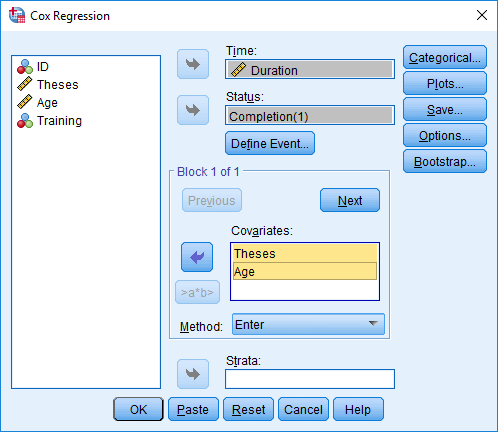
**Conducing a Cox regression**

Now suppose the researcher wants to assess whether the number of these candidates have read—a numerical variable—affects the duration they need to complete their own dissertation. In this instance, the Kaplan-Meier procedure is not as suitable. Instead, when the predictor is numerical, researchers tend to utilize Cox regression, also called the Cox proportional hazards model. If possible, you should learn about linear or multiple regression before you read about this technique; otherwise, some of the concepts might seem unclear. To illustrate Cox regression, suppose you collect the data that resembles that following SPSS file.

* This data file is similar to the previous data file
* However, “Theses” represents the number of theses each participant had read
* “Age” represents the age of participants

After entering the data, Select the “Analyse” menu”, “Survival”, and then “Cox regression”. Then

* transfer “Duration” into the box called “Time”
* transfer “Completion” into the box called “Status”.
* click “Define Event”. Enter “1” into the box alongside “Single value”.
* transfer “Theses” and “Age” into the box called “Covariates” to generate the following screen

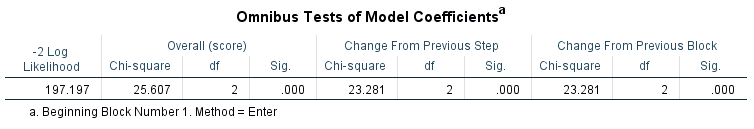


Researchers could click OK now or might

* Click “Categorical” if they also wanted to assess the impact of categorical variables—especially variables with more than two categories
* Click “Plots” if they wanted to generate survival or hazard plots

**Output from Cox regression**

Cox regression generates a series of tables and, occasionally, plots. The following example shows one of these tables: the Omnibus Tests of Model Coefficients.

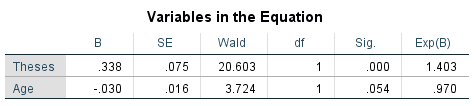


To interpret this table

* the “Change from Previous Step” significance or p value is less than 0.05—and indeed less than .001. Consequently, the overall model is significant
* thus, the number of theses participants read and their age, when combined, do affect the duration participants need to complete their own dissertation

The next table is called Variables in the Equation. To interpret this table

* note that “Theses” are significant—indicating the number of theses participants read affects the duration they need to complete their own dissertation
* this positive B value implies that thesis is positively associated with the hazard function
* remember, the hazard function is, roughly, the likelihood of experiencing the event at this time over the likelihood of not experiencing the event at this time
* therefore, this finding indicates that reading many theses is positively associated with experiencing the event: thesis completion. So, if participants read more theses, they will tend to submit their dissertation sooner.
* age is not significantly associated with thesis completion, however

****

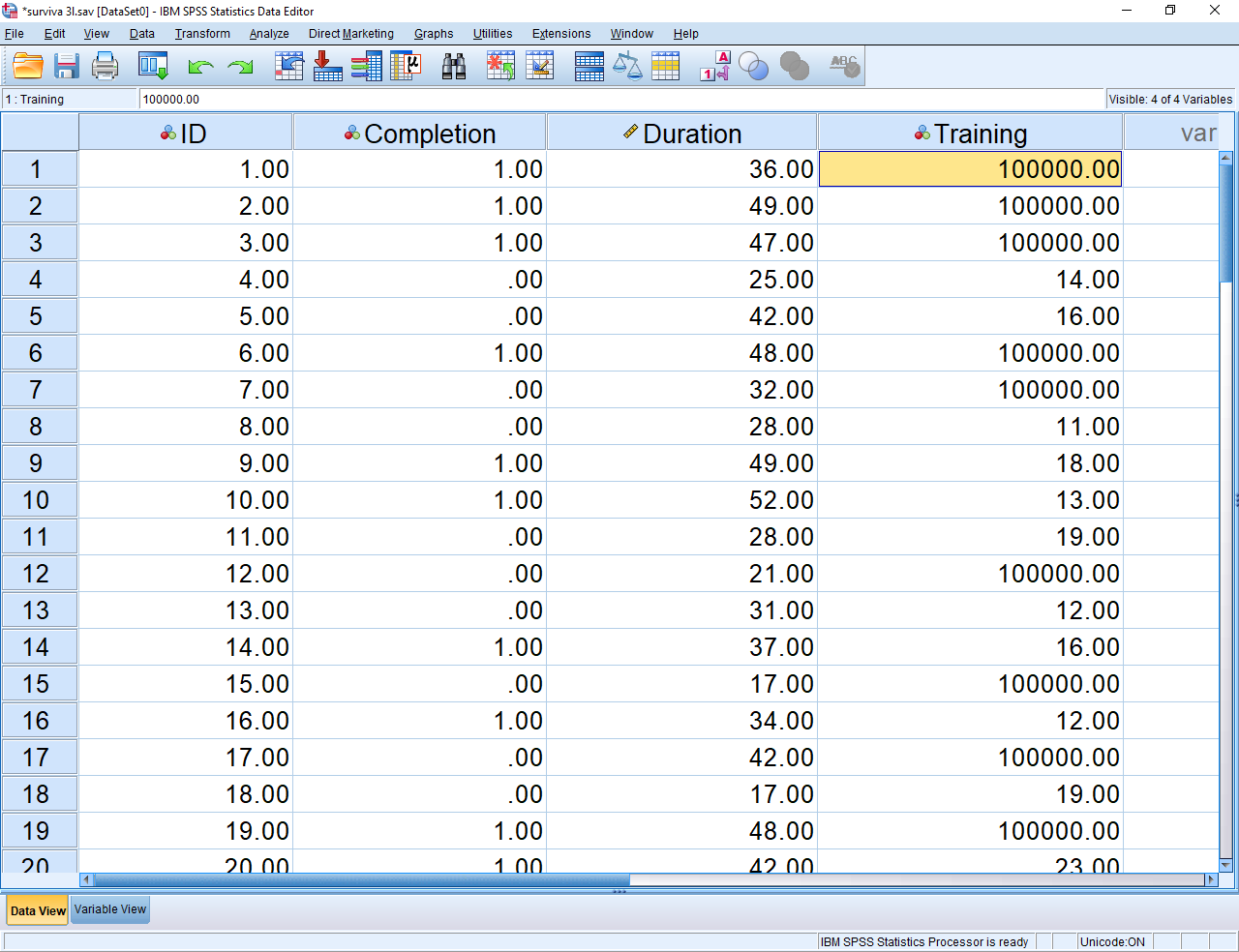
|  |
| --- |
| **Illustration of Cox Regression with time dependent co-variates** |

In the previous illustrations, the predictors—such as whether participants had received the training or the number of theses they had read—were assumed not to change over time. But, these predictors might change over time. For example, participants might not have received training in writing during the first 10 months of their candidature but then received training in writing afterwards. In these instances, researchers need to conduct a variant of Cox regression, called Cox Regression with time dependent co-variates.

**Enter the data**

To illustrate, consider the following data file. This data file is similar to a previous data file except

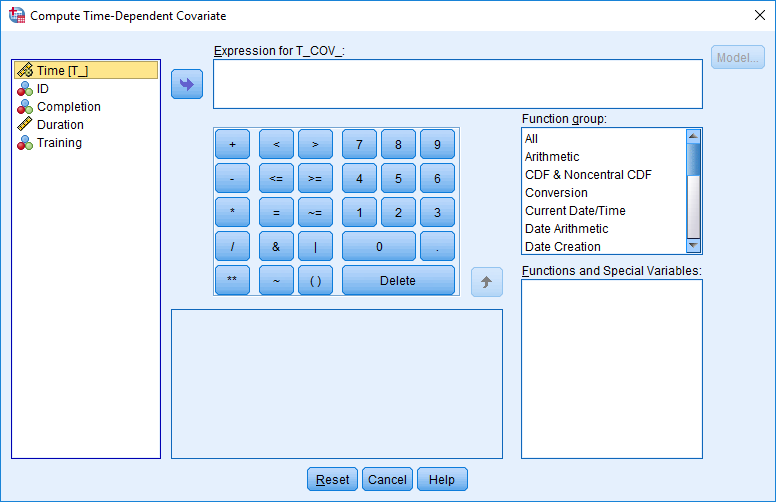
* “training” represents the month in which they received training—such as after month 14 or 16
* if participants never received training, the researcher simply enters a very large number—a number that is higher than is the maximum time or duration



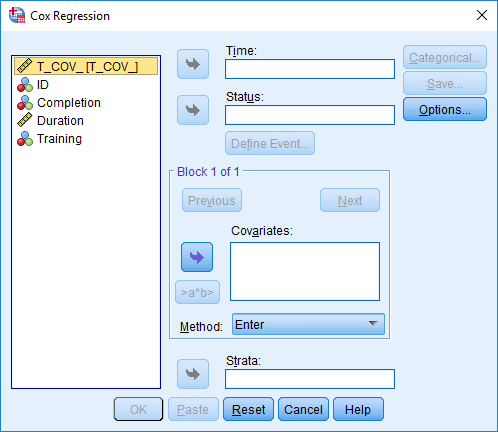
**Conduct the analysis**

To conduct the analysis, select the “Analyse” menu”, “Survival”, and then “Cox w Time-Dep Cov”, to generate the following screen. To proceed

* transfer the top label “Time [T\_]” to the box called “Expression for T\_COV\_
* then add “> Training”
* the formula in this box will thus be “T\_ > Training” but without the quotation marks
* SPSS then understands that as soon as durations exceed training, the participants received training in writing. For example, if training = 10 months, as soon as participants had been enrolled for more than 10 months, they had received the training

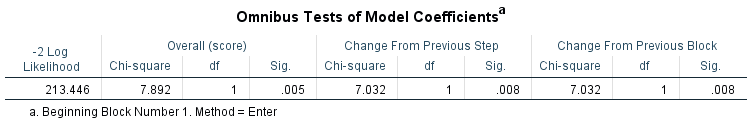


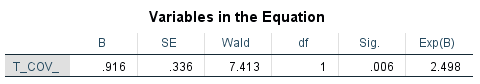
To proceed, press “Model” on the top right to generate the following screen.



In this instance

* transfer “Duration” into the box called “Time”
* transfer “Completion” into the box called “Status”.
* Click “Define Event”. Enter “1” into the box alongside “Single value”.
* transfer “T\_COV” to the box labelled “Covariates” and then Press OK to generate the following output
* you can also transfer other columns in this box, such as age





As the Table called “Variables in the Equation” shows

* T\_COV is significant and positively associated with the hazard function
* Consequently, after participants receive training in writing, the hazard function increases: they become more likely to complete rather than not complete

You can include numerical predictors that change over time as well. But, you need to learn how to code these predictors in the Expression for T\_COV box appropriately.

|  |
| --- |
| **Other considerations** |

This section outlines some of other insights and information that could extend your knowledge of survival analyses. For example, you might want to read further information on

* Survival tree analyses—a technique that can be applied if you assume the model might differ across various subsets of participants
* Life tables—a simple table that you can construct if you merely want to represent the proportion of individuals who experience the event—such as completion or death—at specific times

**Sample size**

To generate significant results, researchers need to collect a large sample size. Often, the number of events—number of completions, for example—should exceed about 100 or more. Several websites have been developed to help researchers determine an appropriate sample size. For example, if you want compare two conditions and conduct a Kaplan-Meier procedure

* visit <http://www.sample-size.net/sample-size-survival-analysis/>.
* to estimate the relative hazard, contemplate the extent to which the likelihood of an event, at some duration, is higher in one condition compared to the other condition
* alpha should .05; beta should be .02

For more complex designs, consider <http://powerandsamplesize.com/Calculators/Test-Time-To-Event-Data/Cox-PH-Equivalence>.

**Survival analyses in R**

To conduct survival analyses in R, you could utilize the “survival” package. Then, to conduct a Kaplan Meier procedure, you would enter the following commands if using the previous variables

* install.packages("survival")
* library(survival)
* km\_fit <- survfit(Surv(duration, completion) ~ training, data = database)
* autoplot(km\_fit)
* summary(km\_fit)

To conduct a Cox regression, you would enter these commands instead

* install.packages("survival")
* library(survival)
* cox <- coxph(Surv(duration, completion) ~ thesis + age, data = database)
* summary(cox)

For more information

* <https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/>
* For more advanced models, visit <https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/>