****

9B17E011

survival analysis in microsoft excel without add-ins

Matthew Lui wrote this note under the supervision of Professor Greg Zaric solely to provide material for class discussion. The authors do not intend to provide legal, tax, accounting or other professional advice. Such advice should be obtained from a qualified professional.

*This publication may not be transmitted, photocopied, digitized, or otherwise reproduced in any form or by any means without the permission of the copyright holder. Reproduction of this material is not covered under authorization by any reproduction rights organization. To order copies or request permission to reproduce materials, contact Ivey Publishing, Ivey Business School, Western University, London, Ontario, Canada, N6G 0N1; (t) 519.661.3208; (e)* [*cases@ivey.ca*](mailto:cases@ivey.ca)*;* [*www.iveycases.com*](http://www.iveycases.com)*.*

Copyright © 2017, Richard Ivey School of Business Foundation Version: 2017-10-18

Introduction to Survival Analysis

Survival analysis refers to a collection of statistical techniques used to analyze the occurrence and expected timing of events. Whereas many popular statistical analyses (e.g., linear regression, logistic regression, and artificial neural networks) provide probabilities, forecasts, or predicted values, survival analysis can provide not only the probability of an event of interest occurring but also the time at which the event is likely to occur. At its core, survival analysis is concerned with the survival of a sample (e.g., patients, machinery, customers) past certain points in time.

The ability to predict if and whenan event will occur has profound implications for organizations and has proven to be useful in many settings. Not only does this field of statistics enable analysis of the expected duration of time until an event happens, but also, certain types of survival analysis enable the study of other factors that may contribute to the likelihood of event occurrence. For example, parametric survival models can help answer the following questions:

* How do a banking customer’s credit score and account balance affect the probability of default?
* How does the amount of ambulation (i.e., walking about) and sleep affect a patient’s recovery time after a medical procedure?
* How do age and length of service affect voluntary employee attrition?

Imagine being able to tell when customers are likely to cancel their services, when an assembly line machine is likely to break down, or when a patient is likely to experience a recurrence of a disease. Given this information, one would likely take preventative measures to mitigate the risk of these events happening over both the short term and the long term. Consequently, survival analysis has found applications in banking, telecommunications, health care, hospitality, and a multitude of other industries. The topic is referred to by different names when applied to different disciplines. For example, it is known as event history analysis in sociology, duration analysis in economics, and reliability analysis in engineering.

One distinct advantage of survival analysis is its ability to handle censoring—a problem that hampers the use of traditional statistical techniques such as linear regression. Censoring refers to the situation when the value of an observation is only partially known. For example, it is not uncommon to encounter datasets that have missing values. Missing data can occur for various reasons, including nonresponse or participant attrition prior to the conclusion of a study. With classical statistical analyses, such as linear regression, missing data are commonly addressed in two ways: deletion or imputation. Listwise deletion is an approach where an entire data record is excluded from analysis when any single data value is missing. Pairwise deletion involves excluding a record when it is missing a data value that is required for a specific type of analysis but including it in analyses for which all required variables are available. Imputation is the process of replacing missing data values with substituted values such as replacing a missing value with the last observed value or the mean of the variable. These methods, while helpful, are not always ideal as they can result in incomplete analyses and introduce bias, particularly when the dataset is relatively small. By using survival analysis, however, it is possible to include censored observations and maintain the sample size.

Consider, for instance, a study in which disease recurrence is being observed (see Exhibit 1). If a patient leaves the study prior to a time of interest, *t*, it cannot be determined whether the disease recurred in that patient, and if it recurred, when it did so. The only inference that can be made is that if the disease recurs, it will happen at a time after the patient left the study. Also, at time *t*, the only conclusion that can be drawn for those patients in which the disease has not yet recurred is that if the disease recurs, it will happen after time *t*. In Exhibit 1, patients 4 and 6 are censored at the time they leave the study, and patients 1, 2, and 7 are censored at time *t*. These five patients are examples of right-censoring, when all that is known is that a data point (in this case, time of disease recurrence) is greater than a certain value but the difference between the data point and the value is unknown. Left-censoring is when a data point is known to be less than a certain value but, again, the difference between the data point and the value is unknown. An example of left-censoring is the disease recurring in a patient between regular checkpoints (see Exhibit 2). The time of recurrence is known to be greater than one value and less than another value, but the exact value of the data point is unknown.

This note explores some different applications of survival analysis and introduces two popular modelling approaches: the Kaplan-Meier estimator and the Cox proportional hazards model. Instructions are also provided on how to build these two survival models in Microsoft Excel, without add-ins.

Applications of Survival Analysis

The term “survival analysis” was coined from initial studies on the subject matter in medicine, wherein the event of interest was death. Researchers assessed the efficacy of medical treatments by monitoring the survival times of patients who had received the treatments.

Today, the scope of survival analysis has expanded and now encompasses a range of applications and analytical techniques. Outside of the medical context, survival analysis is broadly applicable, as it can be employed in any situation that involves time and an event of interest (e.g., structural or mechanical failure, onset of disease, or mortgage default). To date, many organizations have used survival analysis to increase competitiveness, improve operations, and defend against market threats.

Unicru, an American software company founded in 1987 and acquired in 2006 by Kronos, is one such company that has embraced analytical techniques to bolster its value proposition and transform its customers’ internal processes. The company specializes in a line of human resources software that evaluates applicant suitability through personality tests. Unicru’s analytic research group has also developed techniques to estimate customers’ future recruitment requirements. When employees are first hired, it is generally unknown how long they will remain in that employment. Thus, forecasting future hiring requirements to replace departing employees is difficult. However, by analyzing variables embedded in its clients’ payroll data, Unicru can analyze the survival of its clients’ hires, providing clients with the ability to estimate their employee replacement rates.[[1]](#footnote-1)

On the non-commercial side, a year-long study of four U.S. hospital emergency departments examined the effects of crowding on treatment time and length of stay. The study employed discrete-time survival analysis, in which time is not treated as a continuous variable but is divided into discrete units, to describe the relationship between the response variable (i.e., the timeliness of emergency care) and the explanatory variables (i.e., the number of patients waiting, the number being treated, the number boarding, and the inpatient occupancy rate). The study concluded that crowding has a deleterious effect on care delivery and patient outcomes, delaying the care of high-acuity (i.e., high-risk) patients and substantially delaying patients’ waiting times and boarding times.[[2]](#footnote-2)

Researchers at the University of Málaga in Spain also conducted a study that employed survival analysis techniques. The study presented findings based on survival analysis of the Spanish hotel industry. Survival of hotels in Spain is particularly relevant, given that the total contribution of tourism to Spain’s gross domestic product (GDP) is significant, relative to other countries. For example, in 2015, tourism comprised 14.2 per cent of Spanish GDP. By comparison, tourism made up 2.6 per cent of 2015 GDP in the United Kingdom, 9.1 per cent in France, 8.0 per cent in the United States, and 6.2 per cent in Canada.[[3]](#footnote-3) Parametric survival analysis was conducted to examine the influence of different variables on hotel survivability. The study concluded that hotel survival depends on the hotel’s location, size, management, and whether the hotel was launched during a time of economic prosperity. Variables that were determined to be insignificant in predicting hotel survival included hotel type and configuration of financial and economic structures.[[4]](#footnote-4)

Survival models provide data analysis practitioners with a host of tools to analyze the probability of the occurrence of an event of interest, the time to the event occurrence, and the influence of different parameters on outcomes. A particularly useful application of survival analysis is in the area of “customer churn”—the loss of clients or customers. In the face of an increasingly globalized economy and intensifying competition, customer churn rates have become a major issue in many businesses, particularly in telecommunications and financial services. Between 2015 and 2016, 49 per cent of Canadians switched to a different retailer, bank, or telecommunications provider, after a poor customer experience.[[5]](#footnote-5) The following sections will explore two different approaches to using survival analysis in Microsoft Excel to examine customer churn: the Kaplan-Meier estimator (non-parametric) and the Cox proportional hazards model (parametric).

the Kaplan-Meier Analysis: A Non-Parametric Approach

The Kaplan-Meier analysis, also known as the Kaplan-Meier estimator or product limit estimator, is a non-parametric approach to estimating the survival function of a population. The survival function, , is defined as the probability that an object of interest (e.g., a patient, customer, or product) will survive beyond a certain point in time, . A survival function is, in theory, a smooth curve. With Kaplan-Meier analysis, the survival function is approximated using a plot of the Kaplan-Meier estimator (see Exhibit 3) which, given a large enough sample size, approaches the true survival function of the population being modelled. The Kaplan-Meier plot for small samples is not smooth, but rather step-like because it has insufficient data points to create a smooth curve. Instead, the curve is approximated by straight-line segments using a step chart.

Edward Kaplan and Paul Meier jointly developed this method as a technique to estimate patient survival rates while accounting for three outcomes: some patients will die during the clinical trial, others will not complete the trial, and some will survive beyond the end of the trial.[[6]](#footnote-6) This simple and efficient method was first presented in a 1958 article in the *Journal of the American Statistical Association* and has since been credited with providing “doctors a simple statistical way of judging which treatments work best . . . [saving] millions of lives.”[[7]](#footnote-7) Today, the application of Kaplan-Meier analysis extends well beyond the field of medicine and has been used to study, for example, the length of unemployment after job loss.[[8]](#footnote-8)

A key benefit to using the Kaplan-Meier estimator is its simplicity. As the goal is to estimate a population’s survival over time, only two pieces of data are needed: the status of the last observation (i.e., whether the event of interest occurred or whether the record should be right-censored) and the time to the event or the censoring. If a comparison of survival between two or more groups is required, then a third piece of data is needed: the group assignments of the various records.

However, the Kaplan-Meier analysis is limited in its ability to adjust survival estimates in the presence of predictive variables (i.e., covariates). Also, the precision of the survival estimates is dependent on the number of observations. As the study proceeds and observations are excluded, either because they are censored or the event of interest has occurred, estimation precision declines.

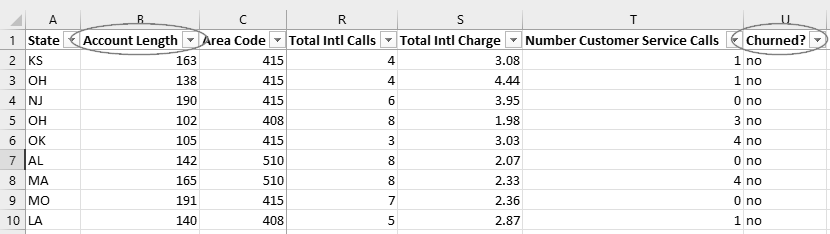
Once plotted, a Kaplan-Meier curve allows estimation of the probability of survival over time.

How to Build a Kaplan-Meier Survival Model in Microsoft Excel

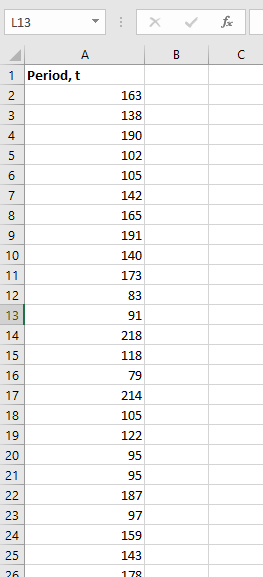
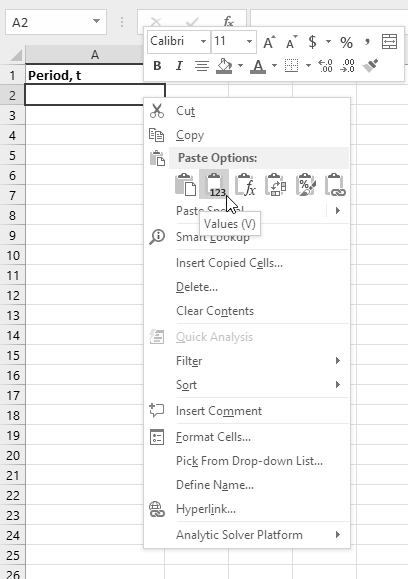
The steps enumerated below can be used, in Microsoft Excel, to produce an estimator of the survival function of any population, provided the values in the dataset indicate whether the event of interest happened (or whether the observation should be censored) and, if so, when the event of interest happened. The steps will also provide instructions for producing a Kaplan-Meier plot.

The figures below were created using a publicly available telecommunications customer churn dataset, furnished by the Machine Learning Repository at the University of California, Irvine.[[9]](#footnote-9)

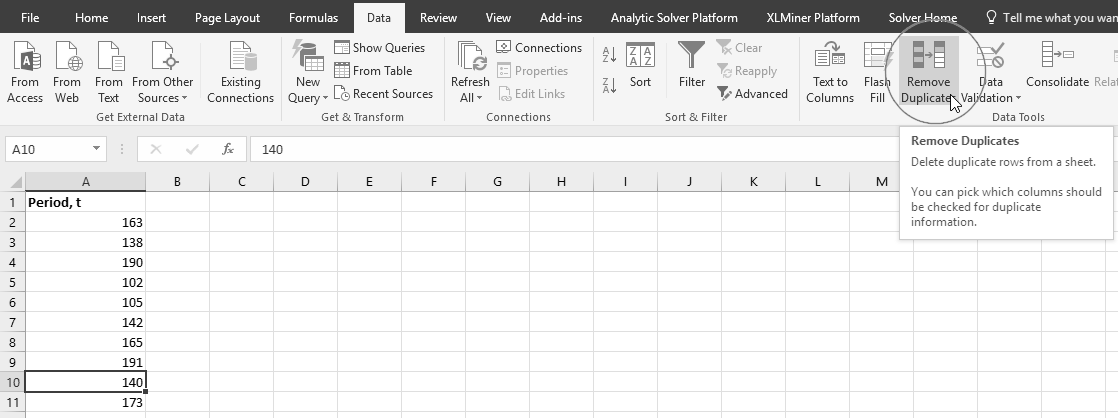
1. Open the dataset in Microsoft Excel and identify the columns that contain the time and survival data. In the dataset shown below, the columns in question are labelled “Account Length” and “Churned?”



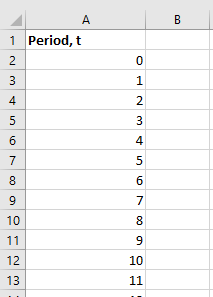
1. In a new worksheet, copy and paste the column containing time data (i.e., “Account Length”) into column A. Label this column with the heading “Period, t.”



1. To produce a data table that can be used to plot a Kaplan-Meier curve, each row needs to represent a unique time period. In other words, deduplication of the time data (i.e., the elimination of duplicated time data) is required. Click anywhere in column A and, in the ribbon, click on Data > Data Tools > Remove Duplicates.



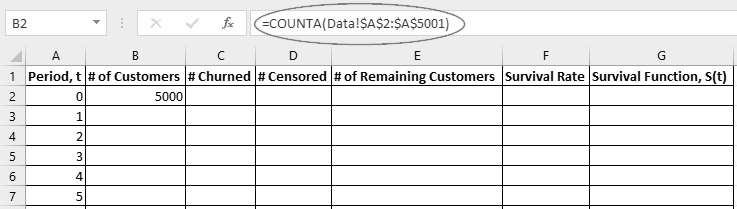
1. After the duplicates have been removed, sort the data in ascending order (i.e., in the ribbon, click on Home > Editing > Sort & Filter > Sort Smallest to Largest) and add a row for time 0 at the top of the column, below the column label.



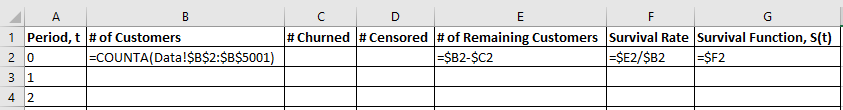
1. Create, in order, the following columns next to the “Period, t” column:

* # of Customers
* # Churned
* # Censored
* # of Remaining Customers
* Survival Rate
* Survival Function, S(t)

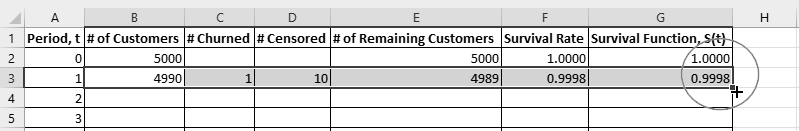
1. In cell B2, under “# of Customers,” provide the total number of subjects in the study at time 0, by counting the number of rows in the original dataset. This task can be accomplished in Excel using the COUNTA function, which counts the number of non-empty cells in the range of cells selected. In our example, the unprocessed dataset is stored in a worksheet labelled “Data.” Using the function =COUNTA(Data!$A$2:$A$5001) results in a count of all non-empty cells in the “State” column of the dataset.



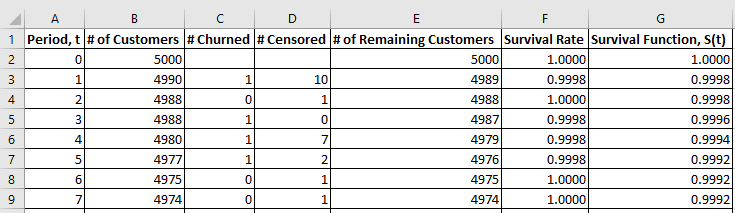
1. In cell E2, type = $B2-$C2, to indicate that the number of remaining customers is the difference between the number of customers at the beginning of the period and the number of customers that have churned.
2. In cell F2, type = $E2/$B2. The survival rate is the proportion of surviving customers at the end of the period to the number of customers at the beginning of the period.
3. In cell G2, type = $F2.



1. In cell B3, type = $E2-$D3 to represent that the number of customers at the beginning of the period is the number of customers at the end of the previous period, less those customers that have been censored.
2. In cell C3, count those customers that not only churned but churned in time period 1. Counting based on multiple criteria can be accomplished in Excel using the COUNTIFS function, which allows the use of multiple criteria, counting only those records for which all of the criteria evaluate as true. In cell C3, type = COUNTIFS(Data!$B$2:$B$5001,$A3,Data!$U$2:$U$5001,"= yes"). This command counts the number of records in the dataset that have an “Account Length” value of 1 and churned in period 1 (which is denoted by a value of “yes” in the “Churned?” column).
3. In cell D3, type = COUNTIFS(Data!$B$2:$B$5001,$A3,Data!$U$2:$U$5001,"= no"). This command counts the number of records in the dataset that have an “Account Length” value of 1 and did not churn in period 1 (which is denoted by a value of “no” in the “Churned?” column).
4. Copy and paste the formulas from cells E2 and F2 into cells E3 and F3, respectively, to calculate the number of remaining customers and survival rate at time 1.
5. In cell G3, type = $G2\*$F3 since the probability of survival past time 1 is the product of the survival rates at all times less than or equal to 1. See Exhibit 4 for a summary of all Excel formulas used thus far.
6. Click and drag to select cells B3 to G3. Copy the formulas through the rest of the table by double-clicking (when the cursor turns to a black + symbol) on the fill handle in the bottom right corner of the selected cells.



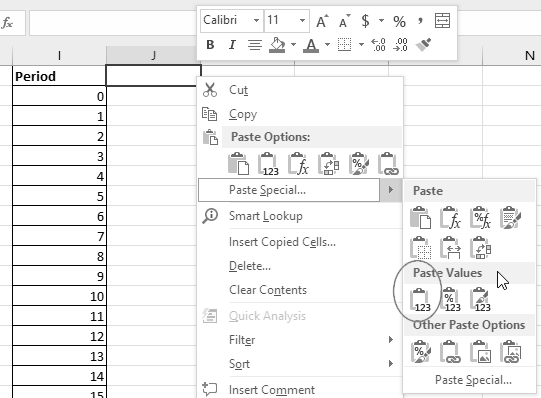
1. The result is a table of the estimated survival function that can be used to plot a Kaplan-Meier survival curve.



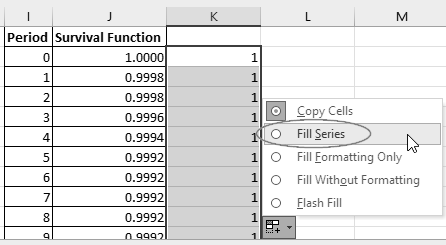
How to Plot a Kaplan-Meier Survival Curve in Microsoft Excel

To create the step chart for the survival function values obtained in the preceding section, some quick data manipulation is needed so that the data are in a format that can be easily charted by Excel.

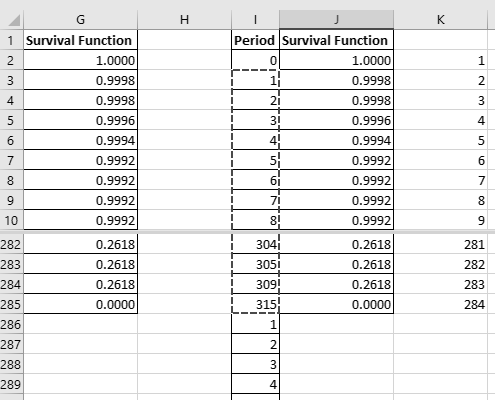
1. Copy and paste the “Period, t” and “Survival Function, S(t)” columns into columns I and J. Use Excel’s Paste Values feature to avoid copying the Excel formulas.



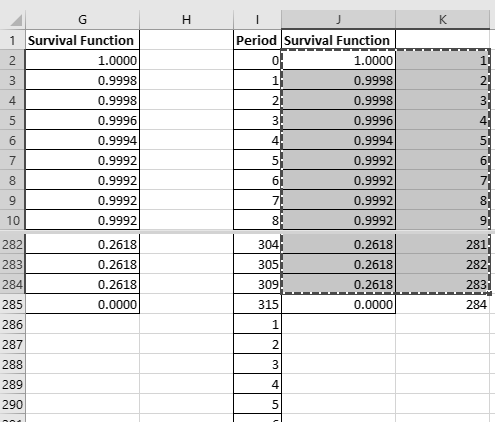
1. In column K, give each of the rows a numerical identifier starting with 1 in K2. Double-click on the fill handle to quickly populate each row. If all cells are populated with a value of 1, click on the Auto Fill Options at the bottom of the column and select Fill Series.



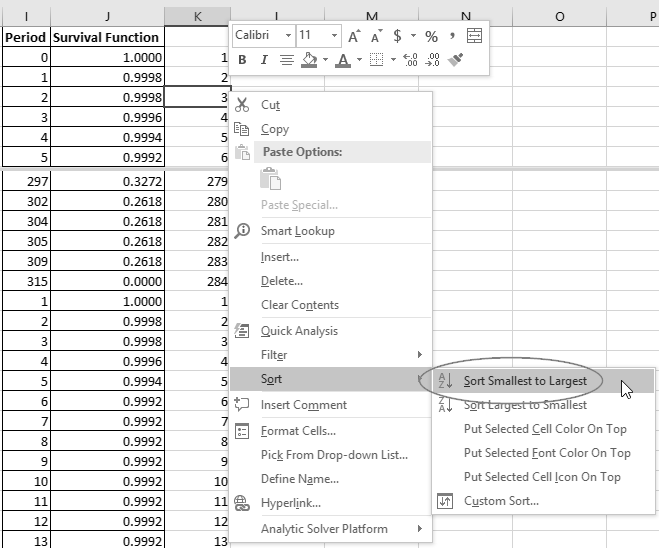
1. Excluding time 0, copy all of the time periods in column I (range I3:I285 in the screenshot below). Append these values to the end of the column I.



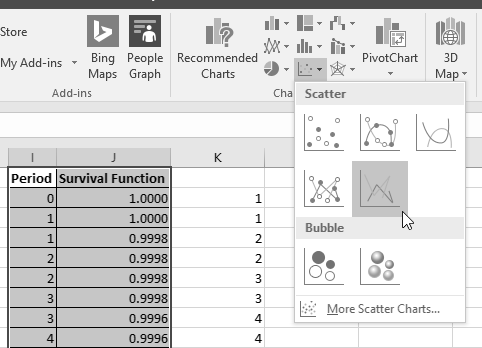
1. Excluding the last row, copy the survival function values and the numerical identifiers (range J2:K284 in the screenshot below). Append these values to the end of columns J and K.



1. Sort the resultant table in ascending order using the numerical identifiers (column K).



1. Select the data in columns I and J. Create a step chart: in the ribbon, click on Insert > Charts > Insert Scatter or Bubble Chart > Scatter with Straight Lines.



1. The result is a plot of the Kaplan-Meier estimator (see Exhibit 3) for the survival of the study population over time. The information in this plot facilitates estimation of the probability of survival.

The plot obtained can provide insights into the study population and facilitate decision-making. For example, in the chart shown in Exhibit 3, it appears that approximately 70 per cent of customers remain with the firm beyond 200 weeks and approximately 50 per cent remain beyond 250 weeks. Using this information, the firm can then formulate strategies for customer retention based on survival probabilities. Note that the chart ends abruptly at week 315 due to censoring (which was caused by the end of data collection).

the Cox Proportional Hazards Regression Analysis: A Parametric Approach

The Cox proportional hazards model (or Cox model), developed and published in the *Journal of the Royal Statistical Society* in 1972, is one of the most popular models for performing survival analysis.[[10]](#footnote-10) Its inventor, British statistician Sir David Cox, was subsequently awarded the General Motors Cancer Research Foundation’s 1990 Charles F. Kettering Prize for his application of the model in analyzing data from cancer clinical trials. The methodology, based on the Cox model, “allow[ed] medical investigators to assess and rank, with critical accuracy, the individual and combined effects of treatment and other factors on the results of clinical trials.”[[11]](#footnote-11) As with linear and logistic regression, the Cox regression model is used to analyze the effects of predictive covariates on outcomes. However, as previously mentioned, use of traditional regression techniques is not suitable in the analysis of survival data due to the presence of censored data. In addition, survival times are usually positive, a constraint that might not be accounted for by classical linear regression without first transforming the data. Lastly, the Cox model assumes a nonlinear relationship between covariates and the independent variable, which linear regression is not equipped to do.

The Cox regression is used to predict the value of the hazard function (or hazard rate) at time . The hazard function is a measure of the risk that the event of interest will occur at time , provided the object of interest has survived up to time .

The Cox model assumes a hazard function for an object of interest at time , . The overall function is the product of a baseline hazard function, , and a function of the covariates, . Thus, the hazard function is given by:

Hazard ratios are also helpful in survival analysis. The hazard ratio is a measure of relative risk and is the ratio of two hazard rates corresponding to two different observations. The hazard ratio quantifies how much more likely an object of interest is to experience the event of interest over another object at a particular time.

Of particular importance is the hazard ratio comparing object to the baseline hazard function, as it simplifies to a function of the regression coefficients and the covariates:

In other words, the natural log of the hazard ratio for object is a linear equation.

The concept of hazard ratios can be applied to churn analysis, allowing analysts to determine which customers the business is most likely to lose. To rectify the situation and retain the identified customers, proactive measures can then be taken in advance of the customers’ cancellation.

See Exhibit 5 for additional details on hazard ratios, the hazard function, and the survival function in the Cox model.

How to Build a Cox Proportional Hazards Model in Microsoft Excel

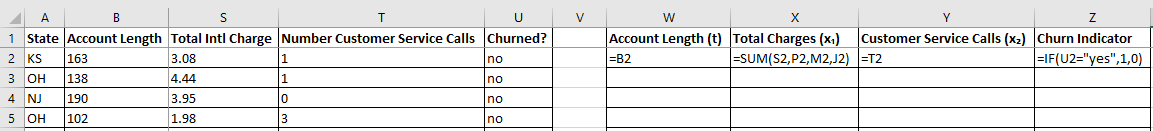
Using the customer churn dataset, the steps below describe how to calculate Cox regression coefficients in Microsoft Excel. A method for computing the hazard ratios for all of the customers in the dataset is then provided allowing analysis and comparison of the relative risk of churn across the customer portfolio.

To successfully execute the steps below, it is necessary to load the Solver Add-in in Microsoft Excel. The Solver Add-in is a default component of a standard Microsoft Excel installation.

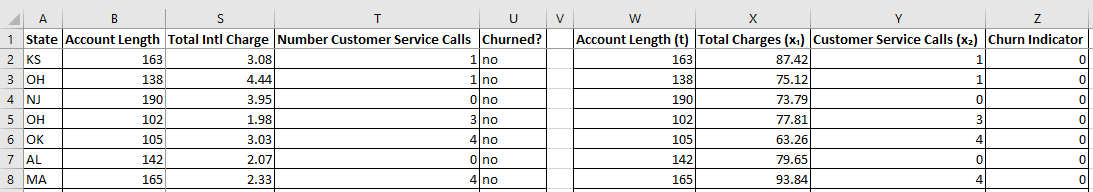
1. Open the dataset in Microsoft Excel and determine the variables whose effects on churn you want to analyze. Note that data can be manipulated to create variables that you think may be suitable for analysis. In this example, the total customer charges (i.e., total day charges, total evening charges, total night charges, and total international charges) were summed into a single variable for analysis. Also, the effect of the number of customer service calls is analyzed.
2. Starting in column W, create the following four columns beside the dataset:

* Account Length (t)
* Total Charges (x1)
* Customer Service Calls (x2)
* Churn Indicator

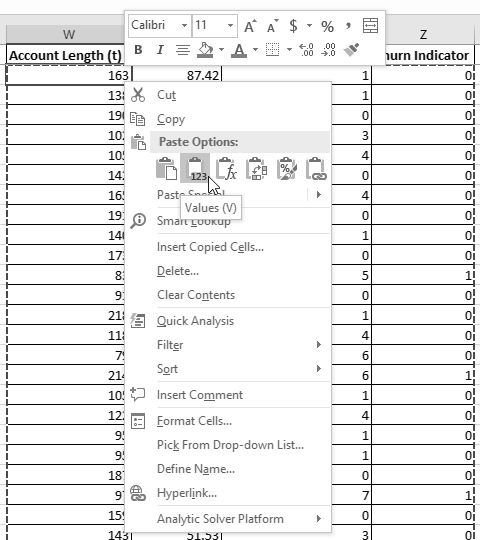
1. Under “Account Length (t),” set the cell values to be equal to the age of the account (under the “Account Length” column). Under the “Total Charges (x1)” column, sum all of the customer charges using the SUM function. Under “Customer Service Calls (x2),” set the cell values to be equal to the “Number Customer Service Calls” column. Under “Churn Indicator,” set the cell values to 1 if the customer churned, and 0 if otherwise, using the IF function.



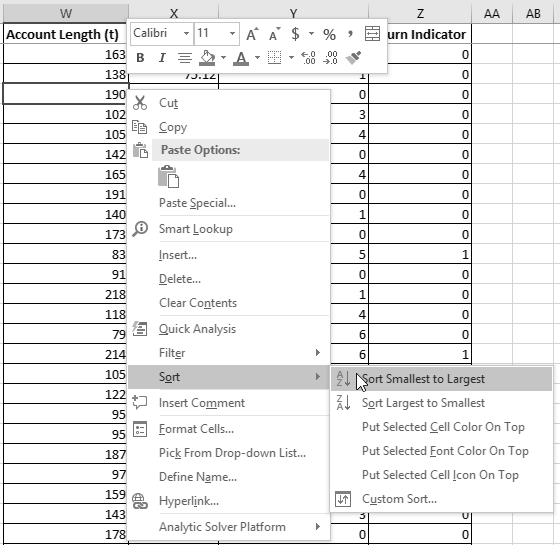
1. Copy and paste these formulas so that each customer (row) has values for the age of the account, the total charges, the total number of customer service calls, and a binary churn indicator.



1. Copy the data and replace the formulas under columns W to Z with only the data values, using Excel’s Paste Values feature.



1. Sort the data, from low to high, using the “Account Length (t)” column.



1. Beside the “Churn Indicator” column, create the following three columns:

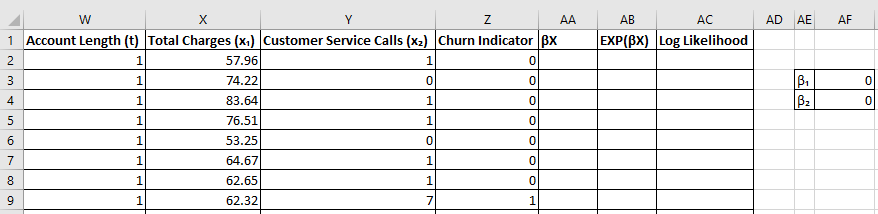
* βX
* EXP(βX)
* Log Likelihood

The “βX” column is meant to represent the sum of the products of the regression coefficients and the explanatory variates, i.e., . The “EXP(βX)” column is meant to represent raised to the power of or, equivalently, . As previously mentioned, the latter calculation is equivalent to the hazard ratio comparing observation with the baseline hazard function.

Similar to logistic regression, regression coefficients in the Cox model are calculated using maximum likelihood estimation.[[12]](#footnote-12) In the proportional hazards model, a method known as partial likelihood is used. Here, the log partial likelihood, , of the model (represented by the “Log Likelihood” column above) is

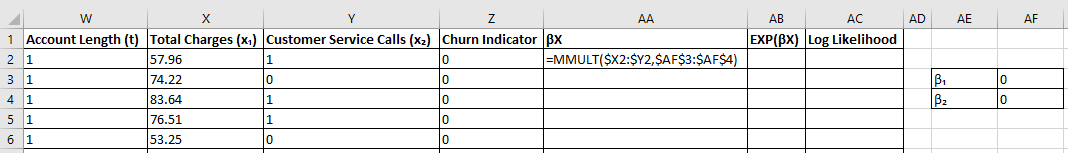
where time is divided into periods , with representing the start of the study and representing the time of the last observed event (see Exhibit 5). Note that the coefficients that maximize the partial likelihood function also maximize the log partial likelihood function.

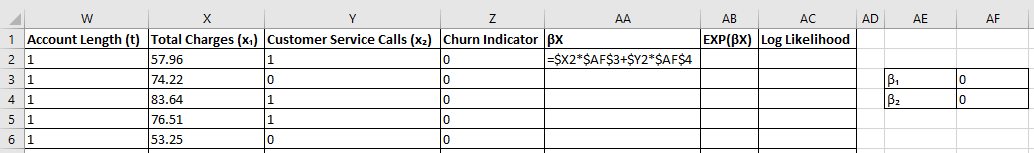
1. Create two cells, labelled β1 and β2, to the right of the table to store the regression coefficients. For now, fill these cells with 0.



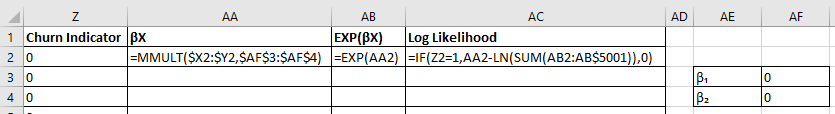
1. The exponent, , is to be calculated in the “βX” column. Using Excel’s matrix multiplication formula, MMULT, you can efficiently multiply the values (i.e., the “Total Charges (x1)” and “Customer Service Calls (x2)” columns) with the coefficients in AF3 and AF4. This function is useful when you have many values and coefficients.

In cell AA2, type =MMULT($X2:$Y2,$AF$3:$AF$4) to calculate the exponent term. Equivalently, you can also type =$X2\*$AF$3+$Y2\*$AF$4.

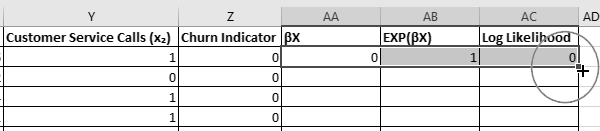




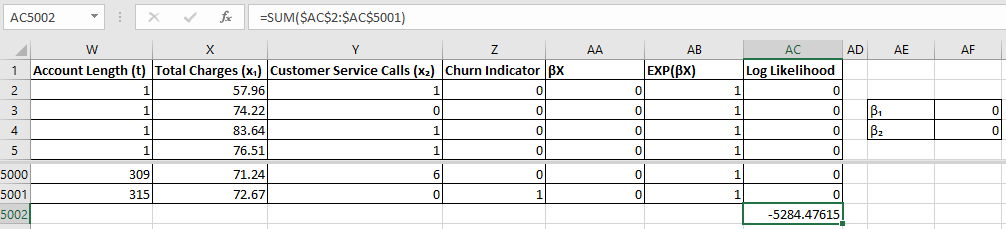
1. In the “EXP(βX)” column, it is now possible to calculate raised to the exponent determined in step 9. In cell AB2, type =EXP(AA2) to obtain the value of .
2. In cell AC2, type =IF(Z2=1,AA2-LN(SUM(AB2:AB$5001)),0) to effectively obtain the term,



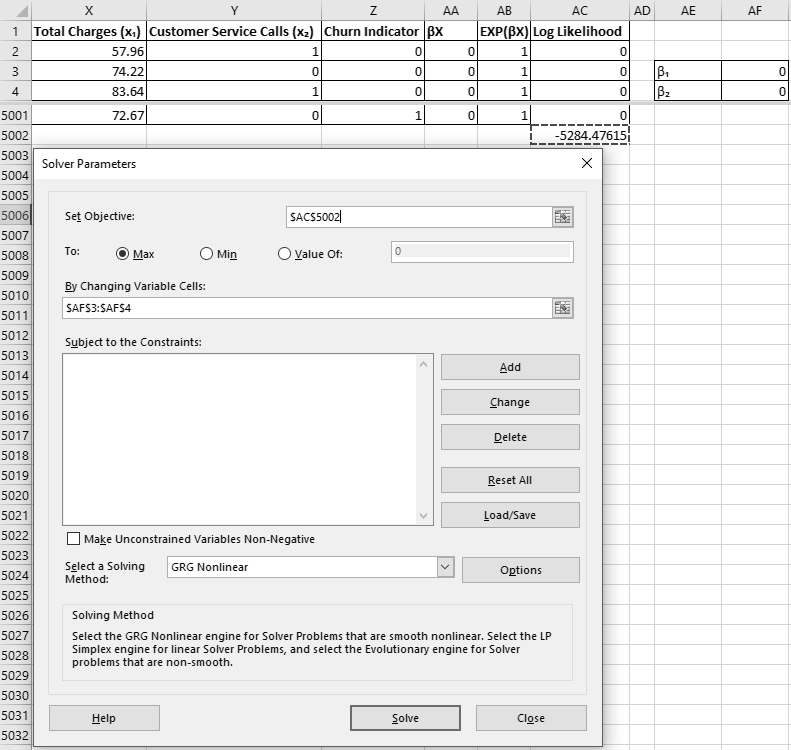
1. Propagate these formulas to the rows below, by highlighting the range AA2:AC2 and double-clicking on the fill handle.



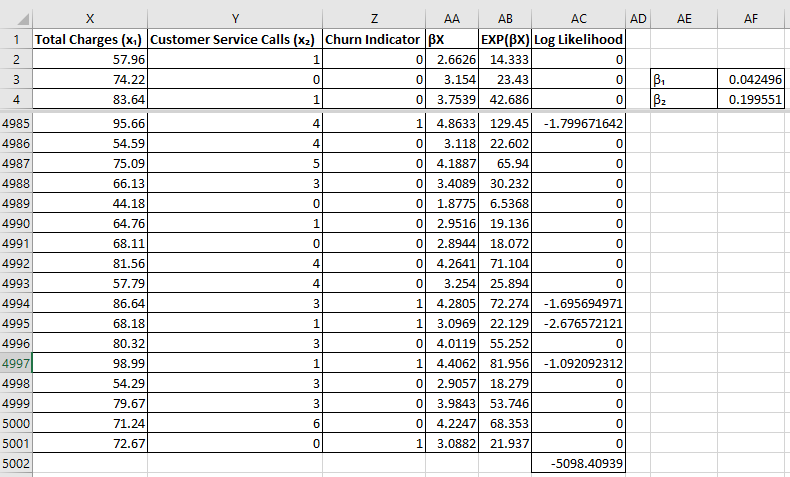
1. Type the summation formula = SUM($AC$2:$AC$5001) at the bottom of the “Log Likelihood” column to obtain the value of the full log partial likelihood term, .



1. As mentioned, the goal is to estimate the regression coefficients by maximizing the log likelihood of the model, which can be accomplished using Solver. Access Solver by navigating, in the ribbon, to Data > Analyze > Solver. Set the objective function to the cell containing the sum of the log likelihood terms in step 13. Maximize the objective function by allowing Solver to change the values of the regression coefficients in AF3:AF4. Uncheck the box labelled “Make Unconstrained Variables Non-Negative,” and select “GRG Nonlinear” as the solving method. Click “Solve.”



1. Solver will then change the values of the regression coefficients until it finds the values that maximize the log likelihood of the model. Once the solution is obtained, you can see that the value for has increased from –5284.47615 to –5098.40939. Note also that the regression coefficients are both positive, indicating that higher bill amounts and a greater number of customer service calls increase the risk that a customer will churn. In addition, the hazard ratios using the calculated coefficients are now available in the “EXP(βX)” column. Recall that the hazard ratio comparing any record to the baseline hazard function is



Recall that hazard ratios are a measure of relative risk and that those subjects with higher hazard ratios are, on average, more likely to experience the event of interest. Looking at the resultant table, it is clear that, in general, customers that churned have higher hazard ratios, which is no surprise. While only two explanatory variates were considered in the above model, additional variables can be included in the analysis to further improve the predictive power of the Cox model.

The information garnered from the Cox regression analysis can be used to improve decision-making, allowing firms to identify and attempt to retain customers who may be on the verge of cancelling their services.

Survival Analysis in Microsoft Excel with Add-Ins

Although this note is primarily concerned with the construction of survival models in Microsoft Excel without add-ins, several tools can facilitate, among other things, survival analysis in Excel. These utilities help by accelerating model development, automatically calculating pertinent test statistics, performing sensitivity analysis, and generating relevant plots and curves to visualize results. Below are two such toolkits that bear mentioning.

XLSTAT

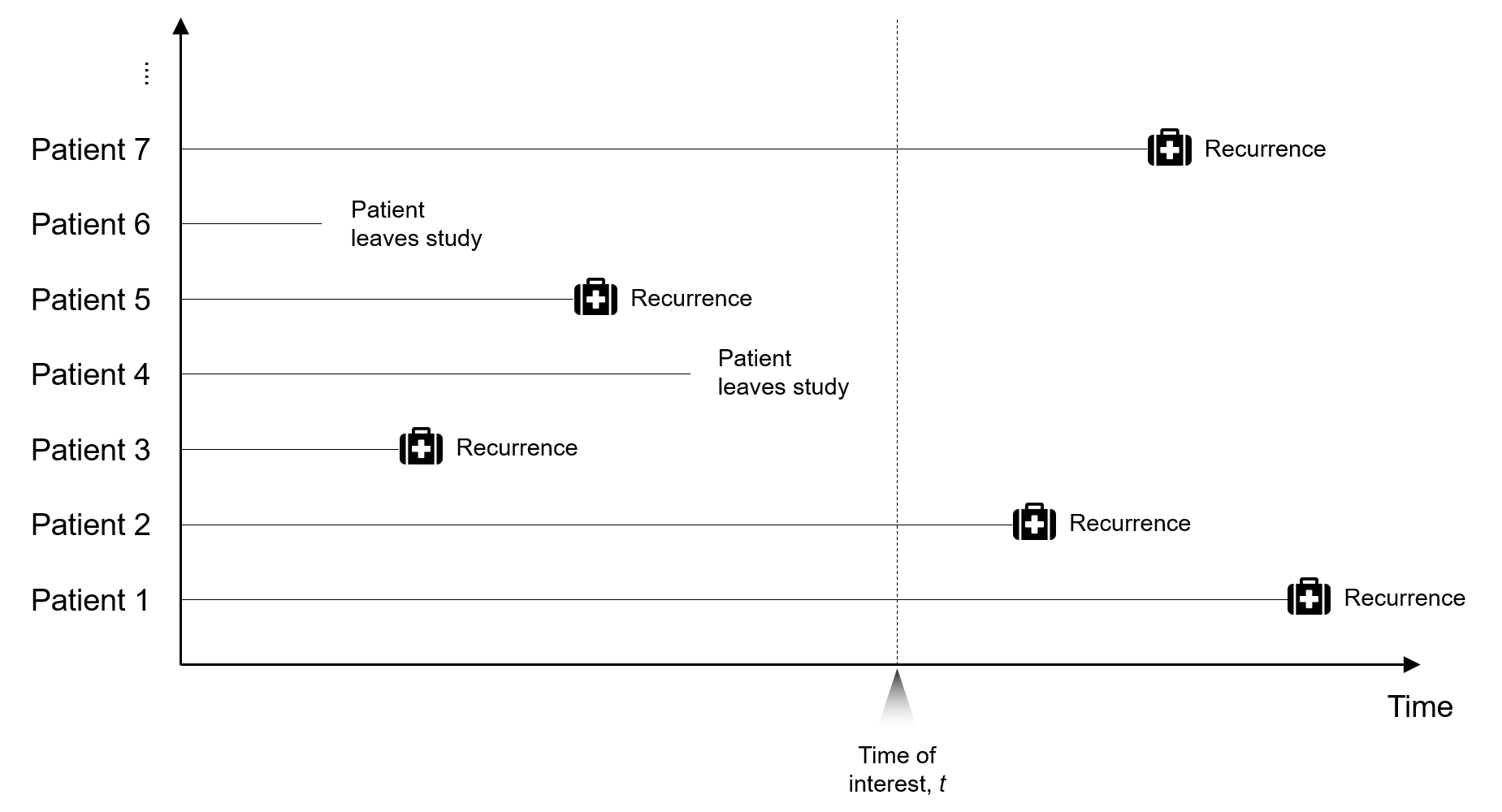
XLSTAT,[[13]](#footnote-13) available for download at www.xlstat.com, is a comprehensive statistical analysis add-in for Microsoft Excel. This paid solution is compatible with both macOS and Windows operating systems and features more than 200 statistical tools to enable both general and industry-specific analyses. For example, solutions are available for biomedical, ecological, and psychological applications. Survival analysis features include, but are not limited to, parametric survival curves, life table analysis, Cox proportional hazards models, Nelson-Aalen analysis, and Kaplan-Meier analysis. Other areas of analysis supported by XLSTAT include time series analysis, Monte Carlo simulations, and statistical process control.

Real Statistics Resource Pack

The Real Statistics Resource Pack,[[14]](#footnote-14) available for download at www.real-statistics.com, is a free Microsoft Excel add-in that extends Excel’s standard statistical capabilities to allow users to more easily perform a variety of statistical analyses. The website provides readers with a foundational understanding of statistics and includes guides on how to conduct various analyses, explanations of their significance and applications, and downloadable Excel workbooks. For example, the Real Statistics website includes material that explains probability distributions such as the normal (or Gaussian) and Student’s *t* distributions. Explanations and mathematical proofs are also provided for topics such as linear and logistic regression, time series analysis, analysis of variance (ANOVA), and cluster analysis. The Real Statistics Resource Pack allows users to quickly construct survival models such as Cox regression models (see Exhibit 6) and Kaplan-Meier plots (see Exhibit 7).

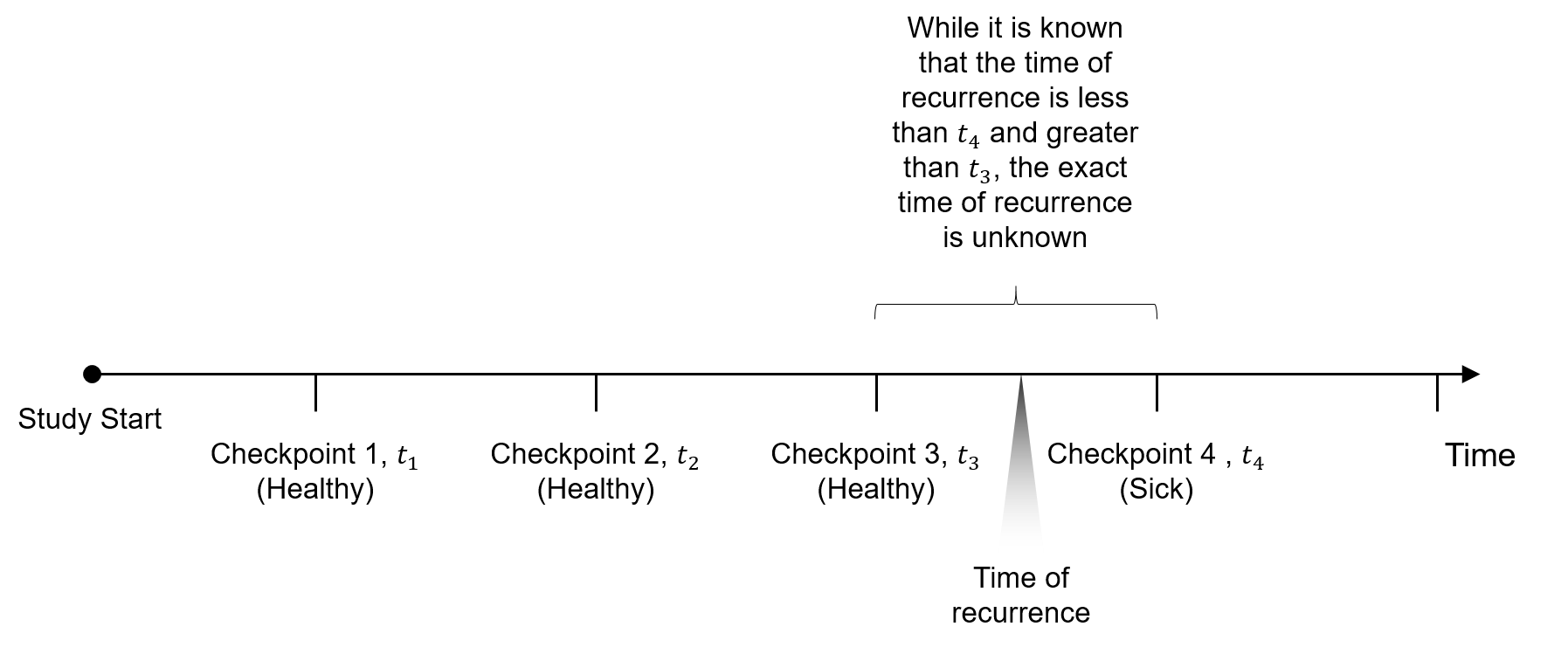
The Ivey Business School gratefully acknowledges the generous support of the Ross N. Clouston MBA Award in the development of this note.

**Exhibit 1: Right-Censoring in a prediction of Disease Recurrence**



Source: Created by the case authors.

**Exhibit 2: Left-Censoring in a prediction of Disease Recurrence**



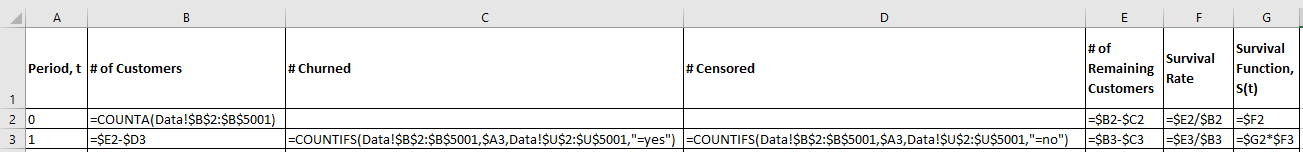
Source: Created by the case authors.

**Exhibit 3: the Kaplan-Meier Survival Plot**

A plot of the Kaplan-Meier estimator resembles a set of horizontal steps that models the survival function of a population over time.

Source: Created by the case authors.

**Exhibit 4: Survival Table Formulas**



Source: Created by the case authors.

**Exhibit 5: Cox Proportional Hazards Regression**

In the Cox proportional hazards model, the hazard function for object of interest takes the following form:

In this model, are explanatory variates for objects of interest (e.g., the patient, the customer, and the machine), are the coefficients that explain the effect the explanatory variates have on the hazard function, and is the baseline hazard function (the hazard function when all explanatory variables are zero, i.e., ).

Note that while the hazard function provides the risk that an event occurs specifically at time , it is useful to know the survival function—that is, the probability that an object of interest will survive beyond time . The survival function for the proportional hazards model,, is given below, where is the baseline survival function.

, with

Also helpful in this branch of analysis is the hazard ratio. The hazard ratio is a measure of relative risk and represents the ratio of two hazard rates corresponding to two different observations (i.e., sets of values for explanatory variates). It tells us how much more likely an object of interest is to experience the event of interest over another object at a particular time—the higher the hazard ratio, the higher the risk of experiencing the event of interest. The term “proportional hazards” refers to the fact that the hazard of any object is a fixed proportion of any other object in the study. For example, the hazard ratio comparing the second subject in a study, , to the third subject in the same study, , is

Note that there is no time component in the hazard ratio above (as the baseline hazard functions in the numerator and denominator cancel out), which means that the objects that are most at risk at a certain point in time are still the objects most at risk at any other time.

For example, suppose that the observations in question involve the study of employees’ voluntary attrition. If the hazard ratio evaluates to 1, then it means that both the second and third employees are equally likely to quit. If the hazard ratio evaluates to 2, then it means that the second employee is twice as likely to quit as the third employee. If the hazard ratio evaluates to 0.5, then the second employee is only half as likely to quit as the third employee (or, similarly, the third employee is twice as likely to quit as the second employee).

More generally, if you calculate the hazard ratio comparing object to the baseline hazard function, the result is the following:

**Exhibit 5 (Continued)**

Using this form of the hazard ratio provides a standard unit of comparison—the baseline hazard function in the denominator—and allows for uniform comparison of the hazard ratios of different observations. Note that with the Cox proportional hazards model, the hazard ratio does not depend on time.

As with logistic regression, to estimate the coefficients, a method known as maximum likelihood estimation is required. As is not specified, Cox devised an estimation method known as partial likelihood estimation that removes from the estimating equation.

The partial likelihood function, , is given by

where time is divided into periods , with representing the start of the study and representing the time of the last observed event, and and represent the covariates of the objects that have experienced the event of interest at time and , respectively. Note that the terms that would be in both the numerator and denominator have cancelled out.

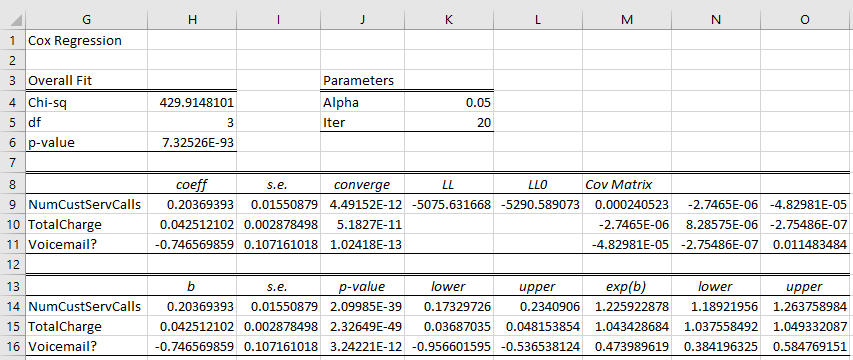
Applying the natural log to the equation to obtain the log partial likelihood, , simplifies the function and leads to the following:

Note that the coefficients that maximize the partial likelihood function also maximize the log partial likelihood function.02.

Source: Bart Baesens, *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications* (Hoboken, NJ: John Wiley & Sons, Inc., 2014), 105–117.

**Exhibit 6: Real Statistics Screenshot—Cox Regression**

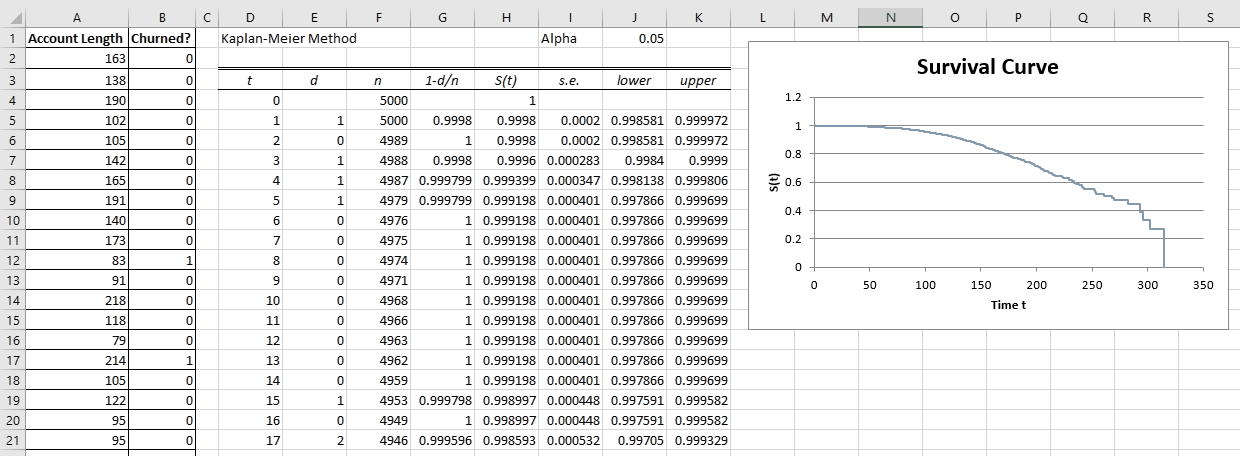
Sample output from a Cox regression analysis using the Real Statistics Resource Pack is provided below and includes estimations of coefficients, significance tests, and the model’s chi-squared test statistic.



Source: Created by the case authors.

**Exhibit 7: Real Statistics Screenshot—Kaplan-Meier Estimator**

Sample output of a Kaplan-Meier analysis using the Real Statistics Resource Pack is provided below and includes calculations of survival rate, standard error calculations, a 95 per cent confidence interval for the survival function, and a plot of the survival curve.



Source: Created by the case authors.

1. Brian T. Denton, “Unicru Uses Forecasting Methods to Estimate Employee Replacement Rates,” Interfaces 37, no. 3 (May–June 2007): 277–278. [↑](#footnote-ref-1)
2. Melissa L. McCarthy, Scott L. Zeger, Ru Ding, Scott R. Levin, Jeffrey S. Desmond, Jennifer Lee, and Dominik Aronsky, “Crowding Delays Treatment and Lengthens Emergency Department Length of Stay, Even among High-Acuity Patients,” Annals of Emergency Medicine 54, no. 4 (October 2009): 492–503, www.ncbi.nlm.nih.gov/pubmed/19423188. [↑](#footnote-ref-2)
3. “Travel & Tourism Total Contribution to GDP—Contribution of Travel and Tourism to GDP as a Share of GDP (%),” Knoema, accessed February 25, 2017, https://knoema.com/atlas/topics/Tourism/Travel-and-Tourism-Total-Contribution-to-GDP/Total-Contribution-to-GDP-percent-share. [↑](#footnote-ref-3)
4. Germán Gémar, Laura Moniche, and Antonio J. Morales, “Survival Analysis of the Spanish Hotel Industry,” Tourism Management 54 (June 2016): 428–438, www.sciencedirect.com/science/article/pii/S0261517715300637. [↑](#footnote-ref-4)
5. Hollie Shaw, “How Silently Enraged Consumers Are Hurting Canada’s Businesses: ‘We Will Just Quietly Slip Away,’” Financial Post, May 9, 2016, accessed March 8, 2017, http://business.financialpost.com/news/retail-marketing/how-silently-enraged-consumers-are-hurting-canadas-businesses-we-will-just-quietly-slip-away. [↑](#footnote-ref-5)
6. Edward L. Kaplan and Paul Meier, “Nonparametric Estimation from Incomplete Observations,” Journal of the American Statistical Association 53, no. 282 (June 1958): 457–481, www.jstor.org/stable/2281868?seq=1#page\_scan\_tab\_contents. [↑](#footnote-ref-6)
7. Julian Champkin, “Significance: The Timeline of Statistics,” Royal Statistical Society: American Statistical Association, January 24, 2014, accessed March 7, 2017, www.statslife.org.uk/history-of-stats-science/1190-the-timeline-of-statistics. [↑](#footnote-ref-7)
8. Bruce D. Meyer, “Unemployment Insurance and Unemployment Spells,” Econometrica 58, no. 4 (July 1990): 757–782, www.jstor.org/stable/2938349?seq=1#page\_scan\_tab\_contents. [↑](#footnote-ref-8)
9. “Datasets from UCI,” SGI: The Trusted Leader in High Performance Computing, accessed March 8, 2017, www.sgi.com/tech/mlc/db/. [↑](#footnote-ref-9)
10. David R. Cox, “Regression Models and Life-Tables,” Journal of the Royal Statistical Society. Series B (Methodological) 34, no. 2 (1972): 187–220, www.jstor.org/stable/2985181?seq=1#page\_scan\_tab\_contents. [↑](#footnote-ref-10)
11. “Sir David R. Cox Wins Kettering Prize,” Chance 3, no. 3 (1990): 37, www.tandfonline.com/doi/abs/10.1080/09332480.1990.10554968. [↑](#footnote-ref-11)
12. Fredrik Odegaard, Note on Logistic Regression (London, ON: Ivey Publishing, 2010). Available from Ivey Publishing, product no. 9B10E005. [↑](#footnote-ref-12)
13. Addinsoft, “Statistical Software & Data Analysis Add-On for Excel,” XLSTAT, accessed February 21, 2017, www.xlstat.com/en/. [↑](#footnote-ref-13)
14. Charles Zaiontz, “Real Statistics Using Excel,” accessed November 29, 2016, www.real-statistics.com/. [↑](#footnote-ref-14)