Introduction to Business Analytics

Lecture 11: Survival Analysis in R

Iegor Vyshnevskyi

Woosong University

May 9/13, 2023

Agenda

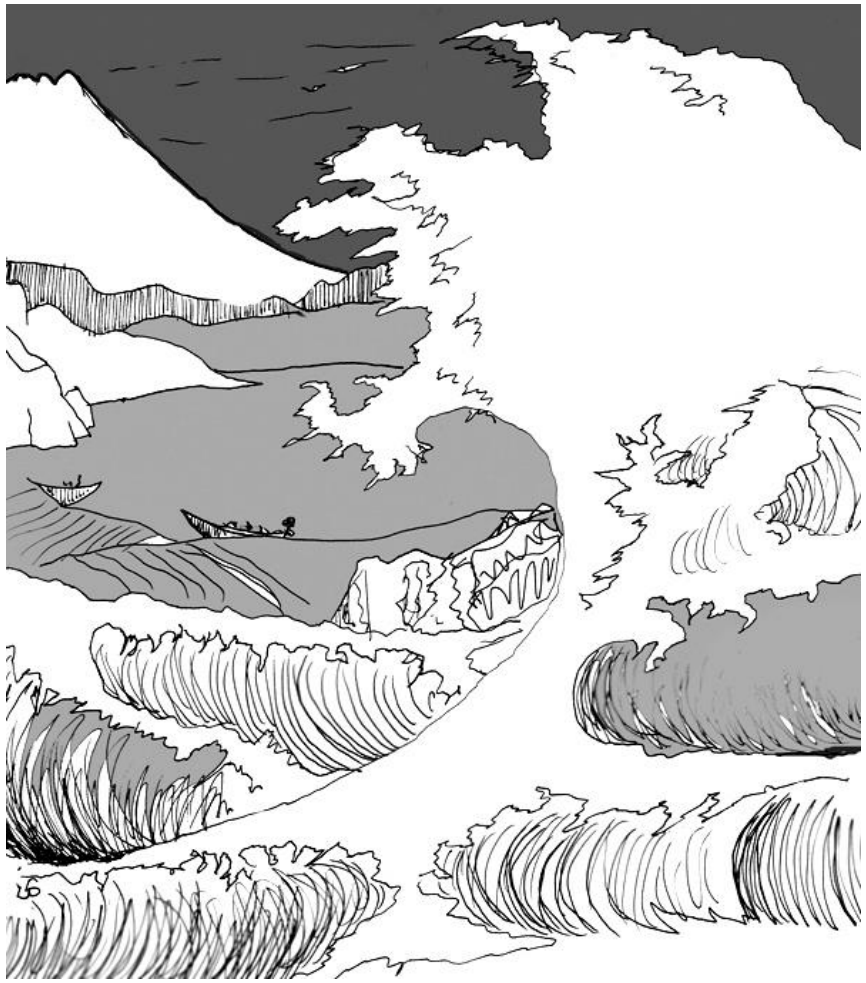
1. Intro to Survival Analysis
2. Survival Analysis Applications
3. Survival Analysis Key Feature
4. Survival Analysis Methods
5. Survival Analysis in Practice
6. In-class Assignment

Acknowledgment: Used a number of open sources and materials from the web.

1. Intro to Survival Analysis

*What is **Survival Analysis?***

Survival analysis refers to a class of statistical techniques that measure the *effect of predictors on the time until an event*, rather than the probability of an event occurring.



With roots dating back to at least 1662 when John Graunt, a London merchant, published an extensive set of inferences based on mortality records, survival analysis is one of the oldest subfields of Statistics

*What is **Survival Analysis?** (cont.)*

Survival analysis is perhaps more accurately referred to as “time-to-event analysis”.

As the name suggests, it is a statistical approach to analyzing the time until a particular event occurs, such as the failure of a mechanical component or a customer ending their subscription.

While there are other methods that can be applied to these problems (e.g. regression), survival analysis offers a *robust method* for dealing with those individuals or elements that haven’t yet experienced the event at the time of the study, so-called “right censored” data.

Survival analysis can be performed using a number of software tools, including Excel and SAS, R, etc.

Usefulness of Survival Analysis for Business Analytics

As the name indicates, this technique has roots in the field of medical research for evaluating the effect of drugs or medical procedures on time until death.

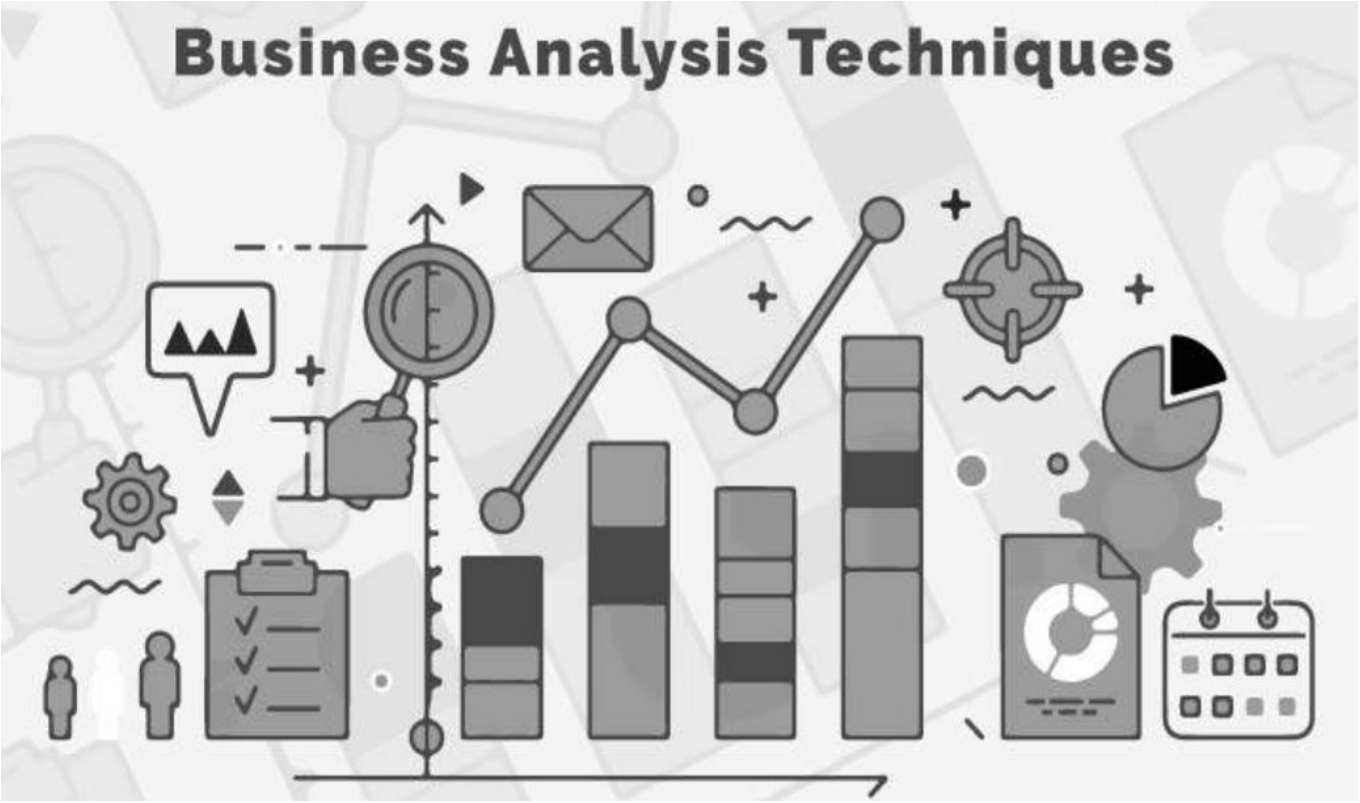
However, there are many other applications of this technique, such as the following business analytics examples:

- Time until product failure
- Time until a warranty claim
- Time until a process reaches a critical level
- Time from initial sales contact to a sale
- Time from employee hire to either termination or quit
- Time from a salesperson hire to their first sale
- How long a customer is likely to remain a customer

Usefulness of Survival Analysis for Business Analytics

Overall, Survival analysis is a great technique for performing time-to-event analysis. In many use cases, a survival analysis makes better use of the data than methods such as regression.

It can also be used for descriptive reporting and to generate hypotheses to test using other methods.



2. Survival Analysis Applications

Example of Survival Analysis

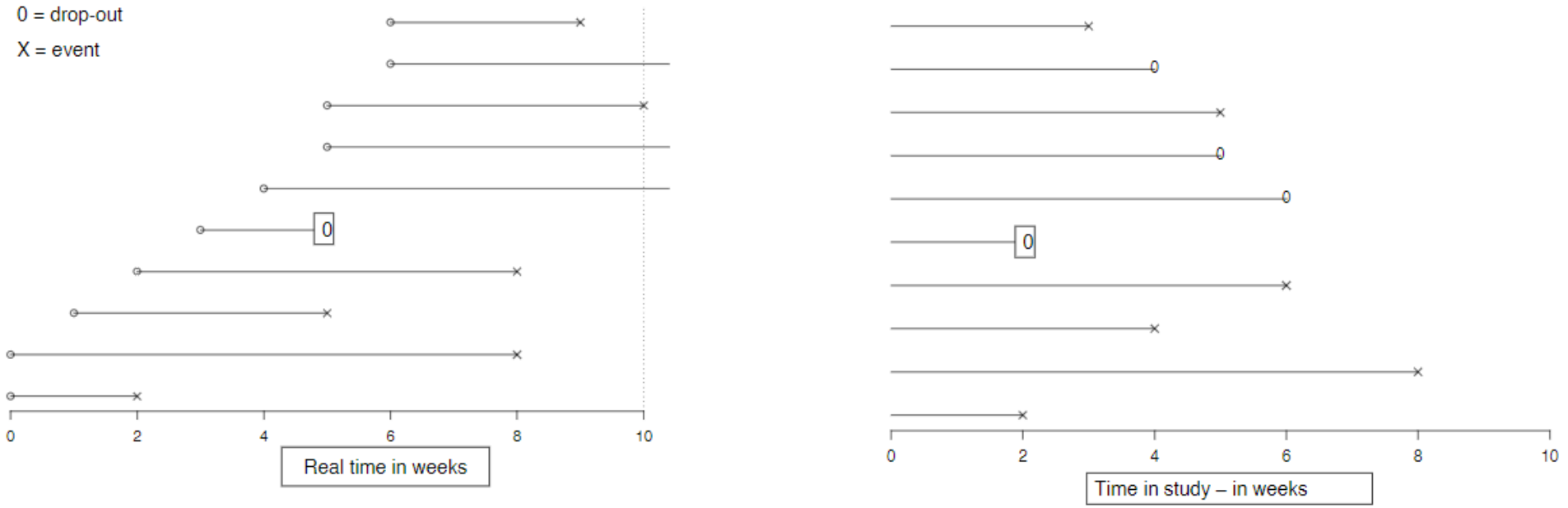
- Knowing how long a customer is likely to remain a customer is a prerequisite to effectively model lifetime value, manage resources, and calculate target costs per acquisition. Also, understanding the factors that influence customer churn can reveal priceless opportunities to optimize performance.

- There's little doubt that survival analysis is a useful tool for understanding customer retention, answering questions such as "What is the median retention time of a new customer acquired through existing customer referral?" and "When should I expect 25% of my customers to have churned?"

- Yet the true power of survival analysis is only revealed when you apply it to answer more complex questions, like "Is my customer lifespan longer in country A vs. country B?" and "Do I retain customers longer if I perform my new retention strategy?"

3. Survival Analysis Key Features

Structure of Event Time Data



Event time data as observed (L) versus to a data analyst (R)

Characteristics of Event Time Data

- 'Individuals' do not all enter the study at the same time.
  - This is referred to as staggered entry.
- When the study ends, some individuals still haven't had the event.
- Other individuals drop out or are lost during the study.
  - The last time they were still "free" of the event is all that is known.

The last two features relate to *censoring* of the failure times.

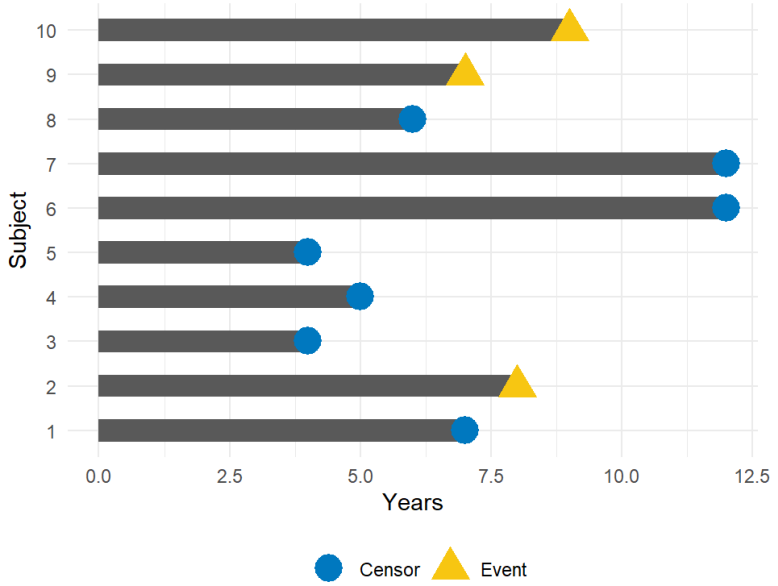
The first of the times until the study ends or the subject drops out is called a *censoring time*.

A key feature of survival data is censoring

- Censoring is a type of missing data problem unique to survival analysis. This happens when you track the sample/subject through the end of the study and the event never occurs.
- A subject may be censored due to (Specifically these are examples of right censoring.):
  - Loss to follow-up
  - Withdrawal from study
  - No event by end of fixed study period
- To illustrate the impact of censoring, suppose we have the following data:

How would we compute the proportion who are event-free at 10 years?

Subjects 6 and 7 were event-free at 10 years.
Subjects 2, 9, and 10 had the event before 10 years.
Subjects 1, 3, 4, 5, and 8 were censored before 10 years, so we don't know whether they had the event or not at 10 years. But we know something about them - that they were each followed for a certain amount of time without the event of interest prior to being censored.



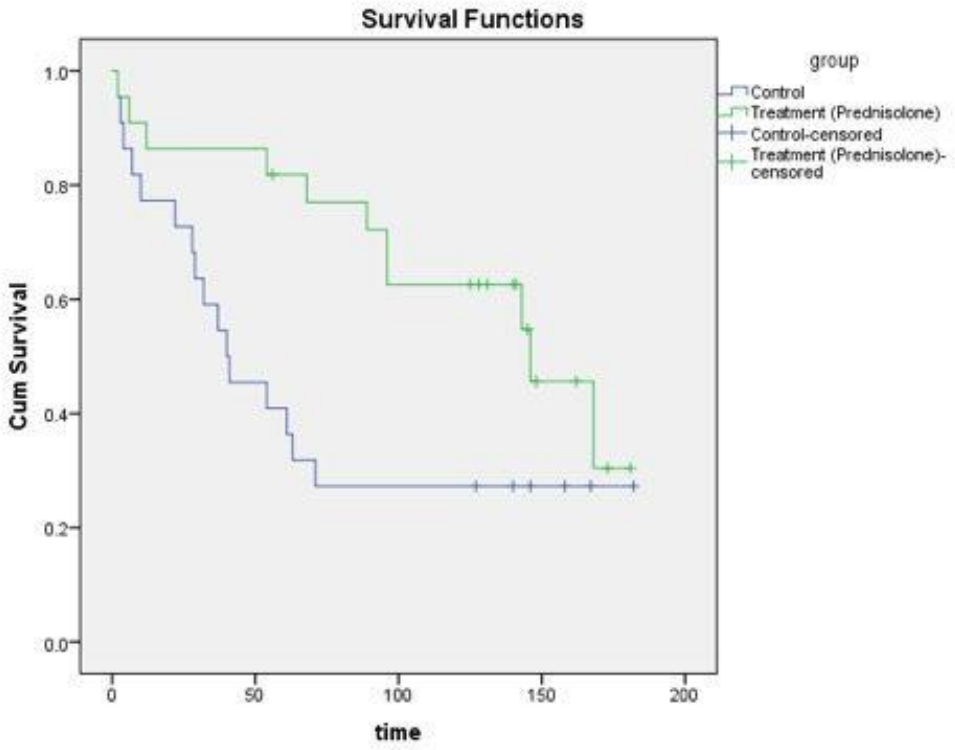
Key concept about Survival Analysis

- For survival analysis, we want to understand how long it takes for an event to occur.
  - **How long**: Tenure. This is how many months a customer has been subscribed.
  - **Survival Event**: Churn = 1. This the event that connects the time to the death of a subscriber (they didn’t really die, but just unsubscribed).

4. Survival Analysis Methods

Kaplan-Meier Method

- The Kaplan-Meier estimate is the simplest way of computing the survival over time.
- For each time interval, survival probability is calculated as the number of subjects surviving divided by the number of patients at risk.
- This type of test determines if there is a statistically significant difference between the survival time of two or more groups. For clinical drug trials, a successful test typically means that the group taking the new drug has a shorter time to recovery or death than the group taking a placebo (and determines whether the trial can move to the next stage).



Cox Proportional hazard model

- It is a regression modeling that measures the instantaneous risk of deaths and is bit more difficult to illustrate than the Kaplan-Meier estimator.
- The advantage of Cox Regression over Kaplan-Meier is that it can accommodate any number of predictors, rather than group membership only.
- As is the case for all regression techniques, there are two potential benefits of analysis using Cox Regression:
  - predictor ranking, with each predictor's effect measured above and beyond the other predictors' effects, and
  - the ability to make predictions with the regression results.
- Predictor rankings enable the analyst to identify the factors that have the most influence on time to an event, and the regression results can be used to estimate the amount of until an event for a specific profile of any subject.

Other Methods

- There are other analytical techniques that can predict time until an event, but *survival analysis techniques have the unique advantage* of including cases that have experienced the event and cases that have not.

5. Survival Analysis in Practice

Survival Analysis for Customer Churn

$$\frac{\text{USERS AT BEGINNING OF PERIOD - USERS AT END OF PERIOD}}{\text{USERS AT BEGINNING OF PERIOD}} = \text{CHURN RATE}$$

What we do…

- We have a custom dataset that has churn data.
  - This metric represents the number of customers that have stopped using your product or service during a given period of time.
- We'll get some business insights using the Survival Analysis in R.

Loading packages and the data

Install and load needed packages first.

```
# Install libraries
install.packages('tidyverse')
install.packages('janitor')
install.packages('tidyquant')
install.packages('patchwork')
install.packages('survival')
install.packages('survminer')

# Load libraries
library(tidyverse)
library(janitor)
library(tidyquant)
library(patchwork)
library(survival)
library(survminer)
```

Load data.

You need to insert your folder path

```
# Load data
# First, set your working folder
setwd("C:/Users/user/OneDrive - kdis.ac.kr/Woosong_2022/Work/2023_spring/Introduction to Business Analytics/L11_Survival")
# remember to change a path to your working folder.
```

```
customer_churn_tbl <- read_csv("customer_churn.csv") %>%
  clean_names() %>%
  mutate(churn = ifelse(churn == "Yes", 1, 0)) %>%
  mutate_if(is.character, as_factor)
```

```
Rows: 7043 Columns: 5
── Column specification ────────────────────────
Delimiter: ","
chr (4): customerID, Contract, gender, Churn
dbl (1): tenure
```

Data exploration

```
# Dataset information
customer_churn_tbl

glimpse(customer_churn_tbl)
```

```
> customer_churn_tbl
# A tibble: 7,043 × 5
   customer_id tenure contract       gender churn
   <fct>        <dbl> <fct>          <fct>  <dbl>
 1 7590-VHVEG       1 Month-to-month Female     0
 2 5575-GNVDE      34 One year       Male       0
 3 3668-QPYBK       2 Month-to-month Male       1
 4 7795-CFOCW      45 One year       Male       0
 5 9237-HQITU       2 Month-to-month Female     1
 6 9305-CDSKC       8 Month-to-month Female     1
 7 1452-KIOVK      22 Month-to-month Male       0
 8 6713-OKOMC      10 Month-to-month Female     0
 9 7892-POOKP      28 Month-to-month Female     1
10 6388-TABGU      62 One year       Male       0
# … with 7,033 more rows
```

We see 5 columns:

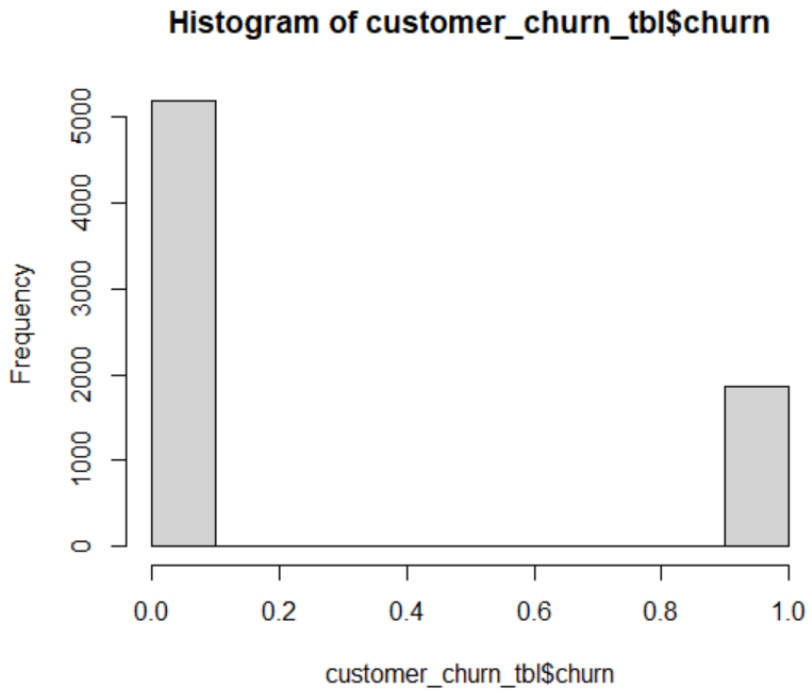
- customer_id - customer's unique identification;
- tenure - how many months a customer has been subscribed;
- contract - type of contract;
- gender - customer gender;
- churn - whether or not customers have churned/left.

```
> glimpse(customer_churn_tbl)
Rows: 7,043
Columns: 5
$ customer_id <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOCW, 9237-HQITU, 9305-CDSKC, 1452-KIOVK, 6713-OKOMC, 7892-POOKP, 638…
$ tenure      <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 25, 69, 52, 71, 10, 21, 1, 12, 1, 58, 49, 30, 47, 1, 72, 17…
$ contract    <fct> Month-to-month, One year, Month-to-month, One year, Month-to-month, Month-to-month, Month-to-month, Month-to-mo…
$ gender      <fct> Female, Male, Male, Male, Female, Female, Male, Female, Female, Male, Male, Male, Male, Male, Male, Female, Fem…
$ churn       <dbl> 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, …
```

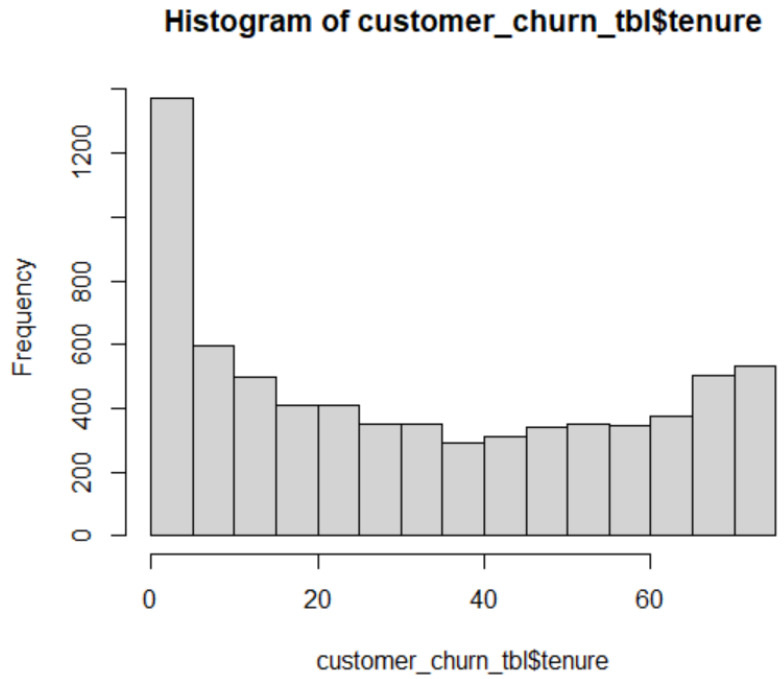
Data exploration

Visualization of our data.

```
# Data seems to be bimodal
hist(customer_churn_tbl$churn)
```




```
# Tenure distribution
hist(customer_churn_tbl$tenure)
```



Fitting the survival model: general **KM model**

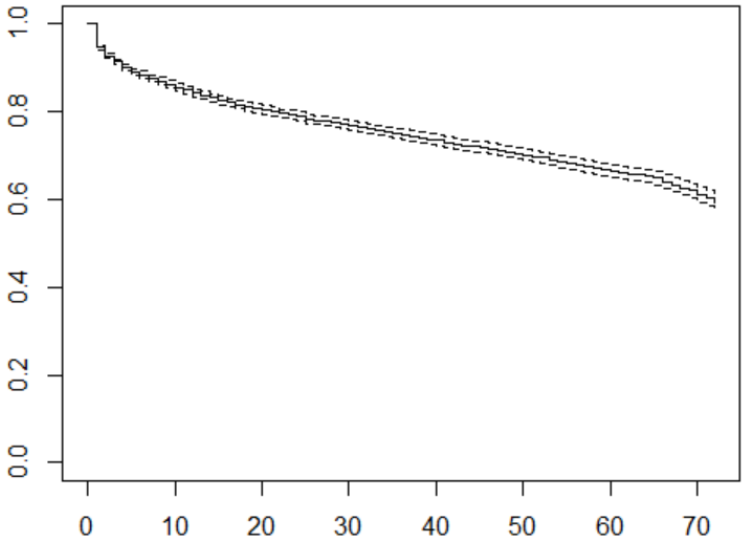
```
# Fitting the survival model:general KM
sfit_1 <- survfit(Surv(tenure, churn) ~ 1, data = customer_churn_tbl)
```

survfit() creates survival curves and prints the number of values, number of events(people suffering from cancer), the median time and 95% confidence interval.

The *Surv()* function takes two times and status as input and creates an object which serves as the input of *survfir()* function.
We pass ~1 in *survfit()* function to ensure that we are telling the function to fit the model on basis of the survival object and have an interrupt.

```
# Plotting the function KM
plot(sfit_1)
```

The plot gives the following output:



Here, the x-axis specifies "**Number of months**" and the y-axis specifies the "**probability of survival/being a client**". The dashed lines are *upper* confidence interval and *lower* confidence interval.

We see that the probability of staying for 20 months with a company is 0.82 or 82%.

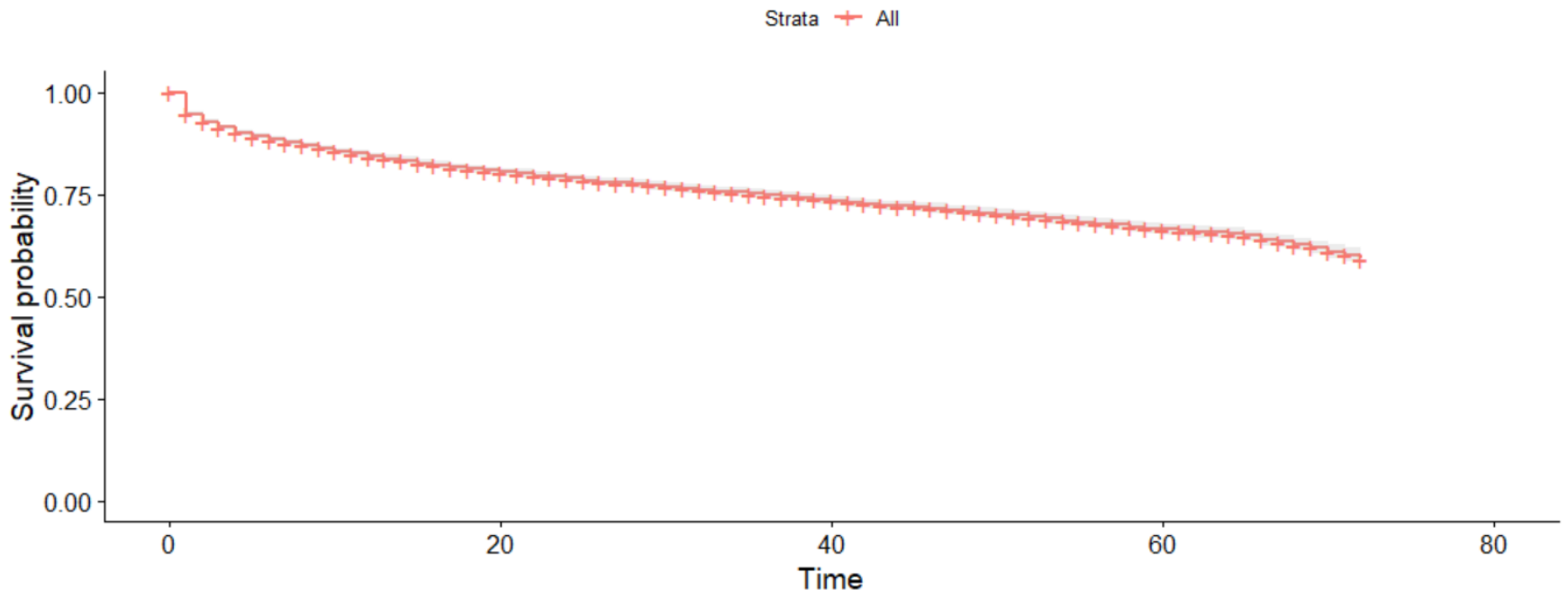
We also have the confidence interval which shows the margin of error expected i.e. In months of surviving 20 months (i.e., being a client for 20 months), upper confidence interval reaches 0.84 or 84% and then goes down to **0.80 or 80%**.

Fitting the survival model: general KM model

```r
# Plotting the function KM (better way)
g1 <- ggsurvplot(
  sfit_1,
  conf.int = TRUE,
  data     = customer_churn_tbl
)

g1
```

We can use ggsurvplot() to have better illustration.



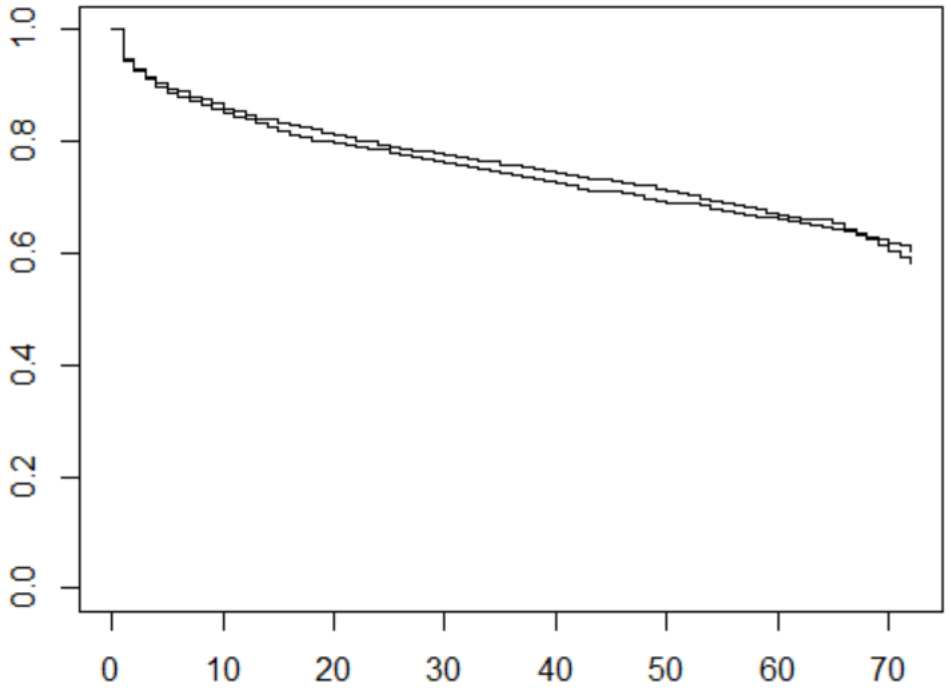
Fitting the survival model: specific **KM model** *(1)*

```
# Fitting the survival model:specific KM model (1)
sfit_2 <- survfit(Surv(tenure, churn) ~ gender, data = customer_churn_tbl)
```

We check for behavioral differences for male and female clients.

```
# Plotting the function KM
plot(sfit_2)
```

We see just a small difference between male and female clients.



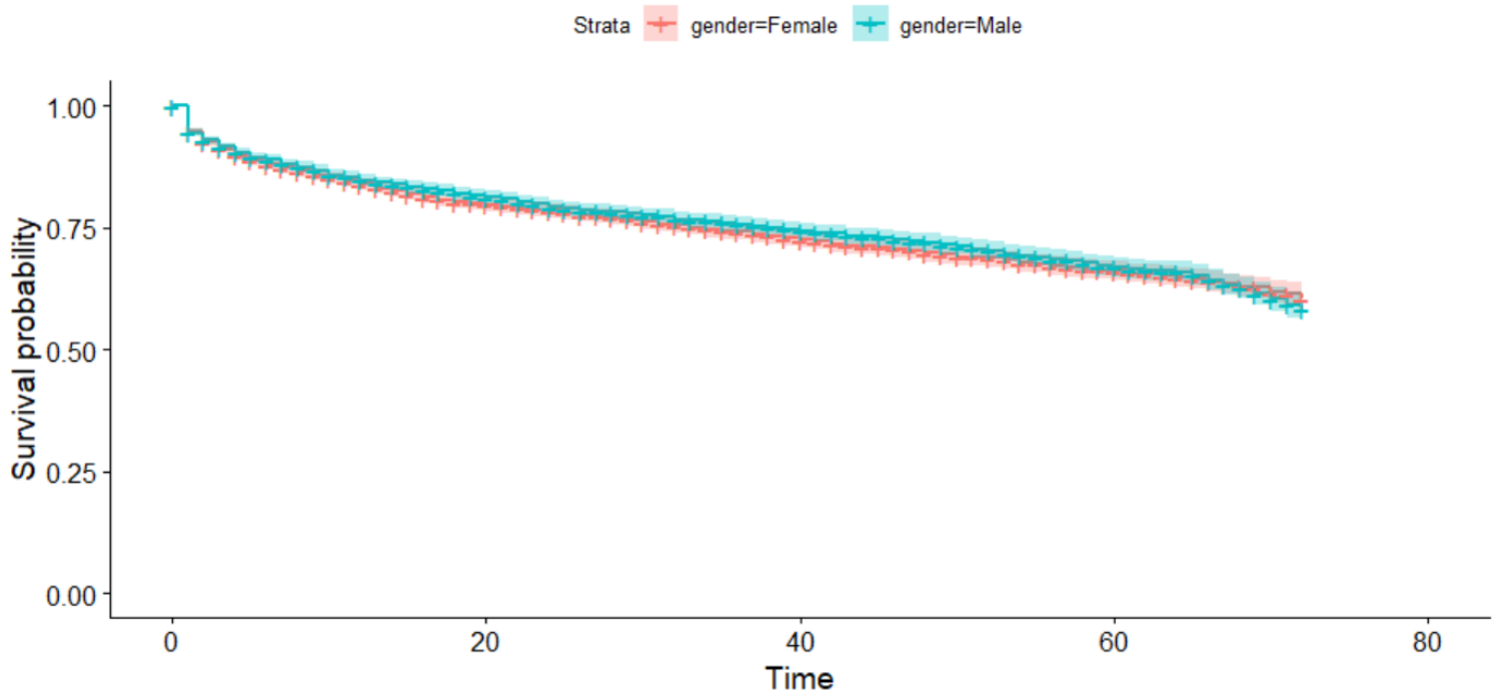
Fitting the survival model: specific KM model (1)

```
# Plotting the specific KM (better way)
g2 <- ggsurvplot(
  sfit_2,
  conf.int = TRUE,
  data     = customer_churn_tbl
)

g2
```

We can use ggsurvplot() to have better illustration.

We see just a small difference between male and female clients.



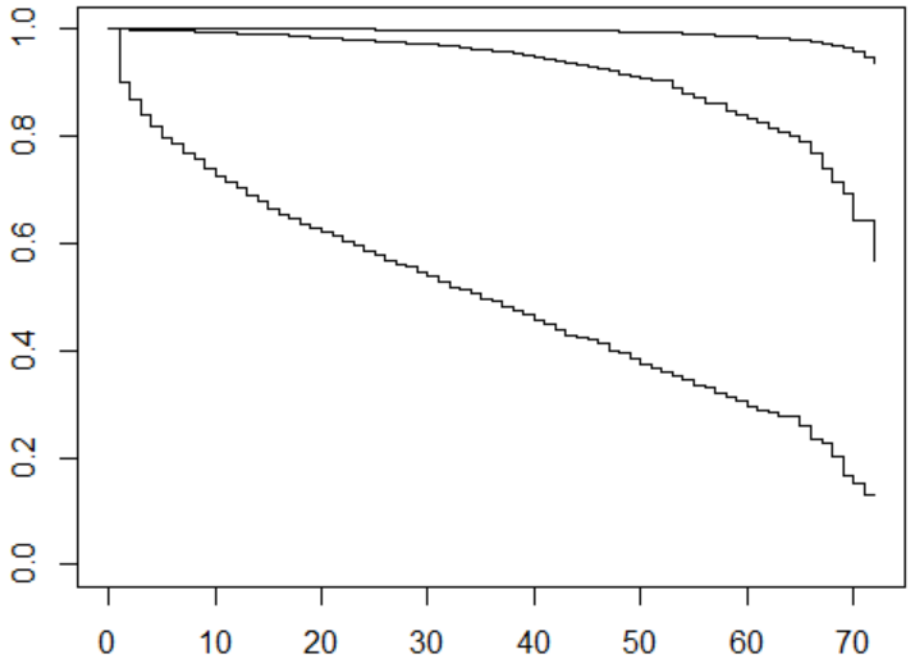
Fitting the survival model: specific **KM model** *(2)*

```
# Fitting the survival model:specific KM model (2)
sfit_3 <- survfit(Surv(tenure, churn) ~ contract, data = customer_churn_tbl)
```

We check for behavioral differences for different contract types.

Remember there are 3 types of contracts.

```
# Plotting the function
plot(sfit_3)
```



We see a huge difference between contract types.

Fitting the survival model: specific **KM model** *(2)*

```
# Plotting the specific KM (better way)
g3 <- ggsurvplot(
  sfit_3,
  conf.int = TRUE,
  data     = customer_churn_tbl
)

g3
```

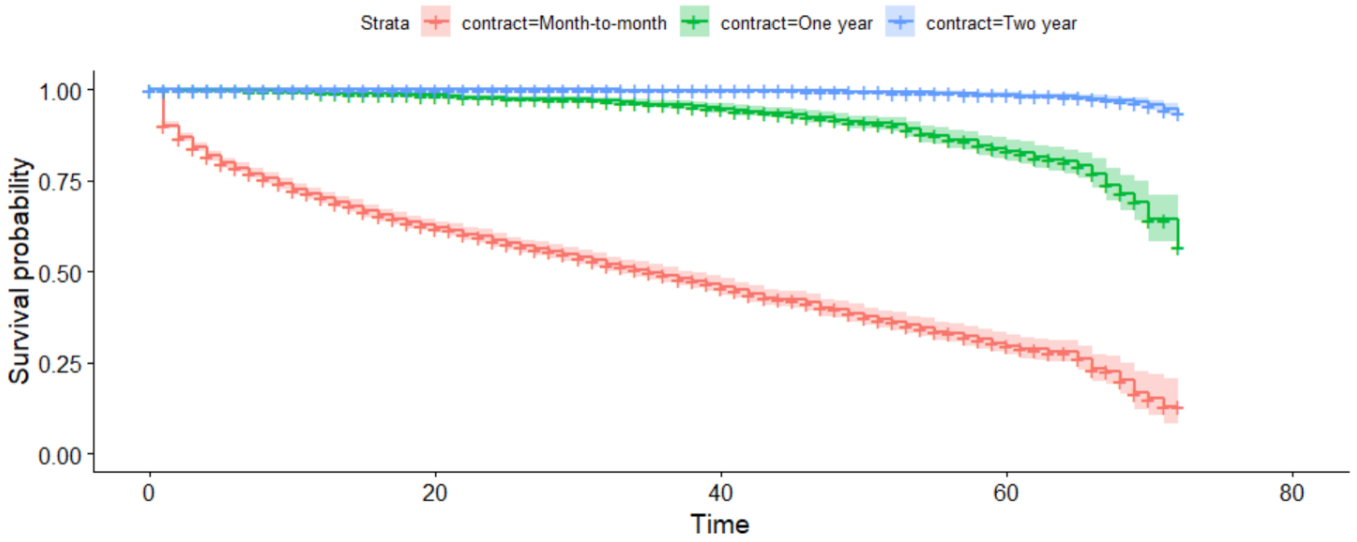
We can use ggsurvplot() to have better illustration.

We see a huge difference between contract types.

The plot has great potential for business insights!

We can see that:

- Month-to-Month Contracts are **hurting the business**. After 40-months, only 50% have survived (meaning that 50% have churned).
- Conversely, long-term contracts are really **beneficial to the business**. At 40-months, 90% of 1-year contracts have survived. These customers are more loyal!
- Also, those who have been with a company for 2 years are the **most loyal**!



Fitting the survival model: Cox proportional hazard model

```
# Fitting the Cox model ----
Cox_mod <- coxph(Surv(tenure, churn) ~ contract, data = customer_churn_tbl)
```

It is a regression modeling that measures the instantaneous risk of deaths and is bit more difficult to illustrate than the Kaplan-Meier estimator.

```
# Summarizing the model
summary(Cox_mod)
```

```
# Fitting survfit()
Cox <- survfit(Cox_mod)
# Plotting the function
plot(Cox)
```

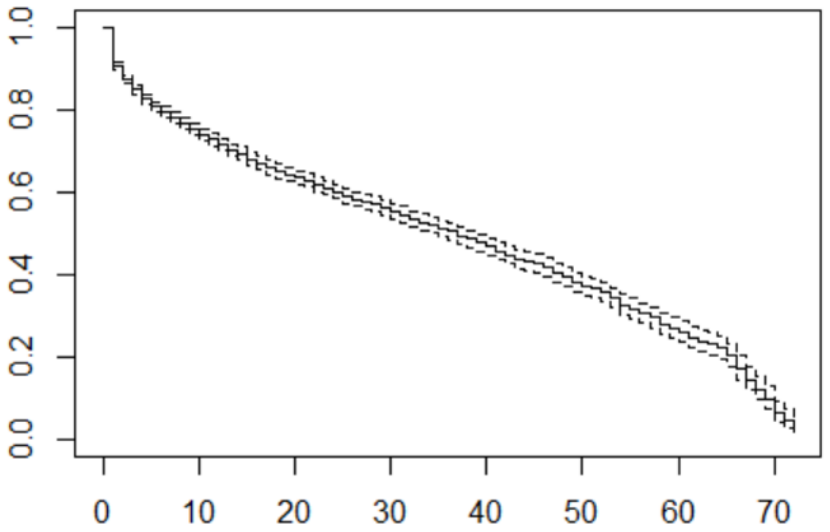
```
> summary(Cox_mod)
Call:
coxph(formula = Surv(tenure, churn) ~ contract, data = customer_churn_tbl)

  n= 7043, number of events= 1869

                    coef exp(coef) se(coef)      z Pr(>|z|)
contractOne year -2.19110   0.11179  0.08345 -26.26   <2e-16 ***
contractTwo year -4.22539   0.01462  0.15626 -27.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
contractOne year   0.11179      8.945   0.09493   0.13166
contractTwo year   0.01462     68.401   0.01076   0.01986

Concordance= 0.772  (se = 0.003 )
Likelihood ratio test= 2619  on 2 df,   p=<2e-16
Wald test            = 1281  on 2 df,   p=<2e-16
Score (logrank) test = 2347  on 2 df,   p=<2e-16
```



Fitting the survival model: Cox proportional hazard model

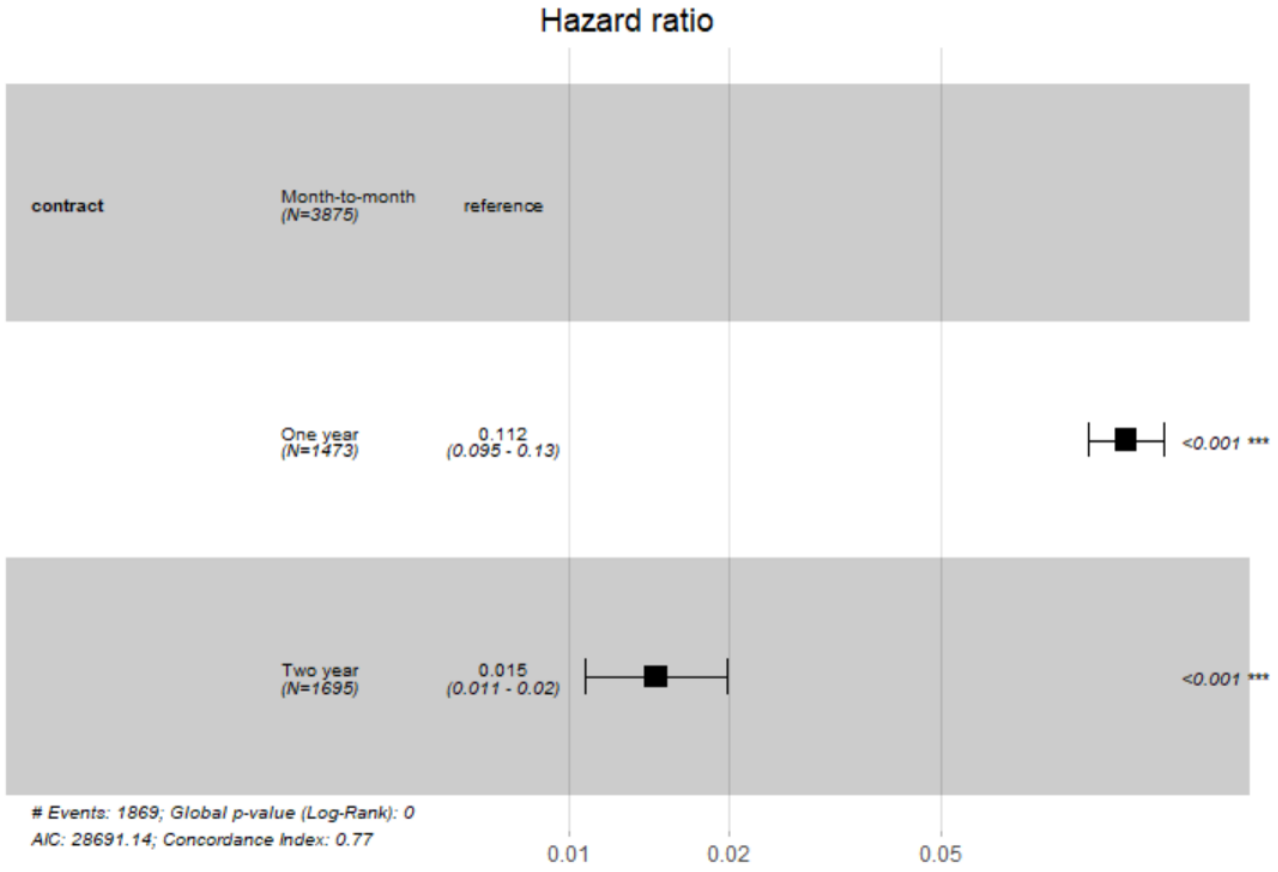
```
# Fit a Cox proportional hazards model
ggforest(Cox_mod, data = customer_churn_tbl)
```

We can use ggforest*()* to have better illustration.

The quantity of interest from a Cox regression model is a *hazard ratio* (HR). The HR represents the ratio of hazards between two groups at any particular point in time. The HR is interpreted as the instantaneous rate of occurrence of the event of interest in those who are still at risk for the event. It is not a risk, though it is commonly mis-interpreted as such.

A HR < 1 indicates reduced hazard of churn whereas a HR > 1 indicates an increased hazard of churn.

So the HR = 0.112 implies that 0.112 times as many one year clients are leaving as month-to-month clients, at any given time. Stated differently, one year clients have a significantly lower hazard of churn than month-to-month clients in these data.



5. In-class Assignment