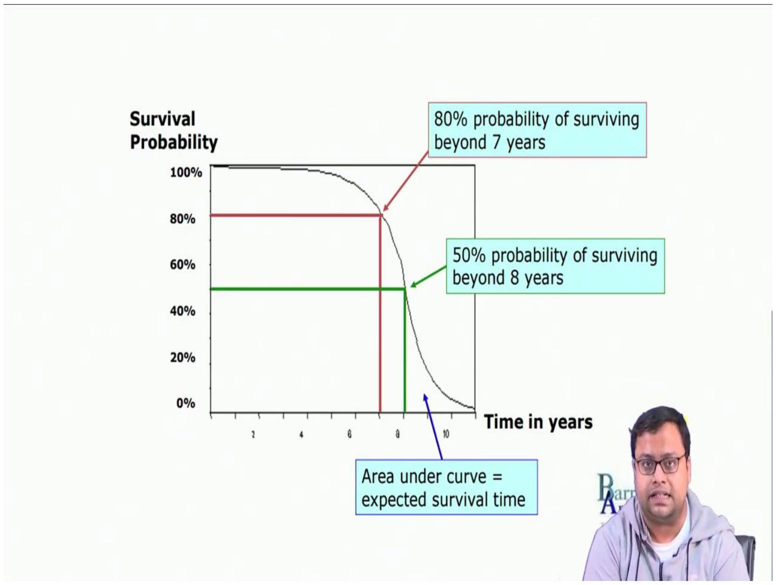
Marketing Analytics
Professor. Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture 50
Customer Churn and Customer Lifetime Value (Contd.)

Hello, everybody, welcome to marketing analytics course. This is Dr. Swagato Chatterjee from VGSOM IIT Kharagpur who is taking this course for you and this particular class, which is in week 9, last session, we will discuss about survival analysis in case of Customer Churn. So, what is survival analysis?

(Refer Slide Time: 00:34)

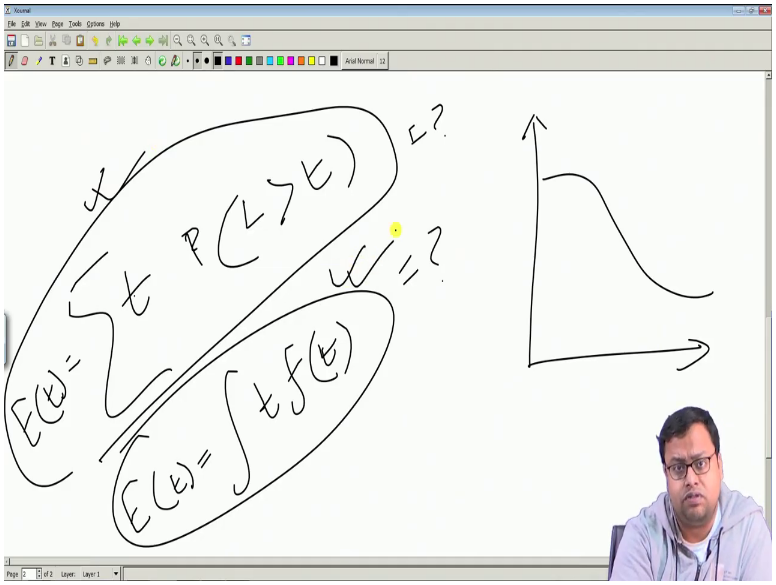


So, you will see that this is a survival probability curve. So, it is actually talking about that every time period if you are alive today what is the probability that you will be alive tomorrow? What is the probability that you will be alive day after tomorrow and so on two days later, three days later, four days later.

So, how much is the probability that over a certain period of time you will be alive and this is something which is important to know. Because certain times it is in the context of let us say, customer lifetime calculation that part, I will take as much as customer lifetime, this is something matters.

So, survival probability, you see that initially 80 percent of guys survive beyond seven years and then at this point, which is the green point and it is 50 percent of guys who is surviving after eight years. So, any end of the curve is basically the expected survival time.

(Refer Slide Time: 01:37)

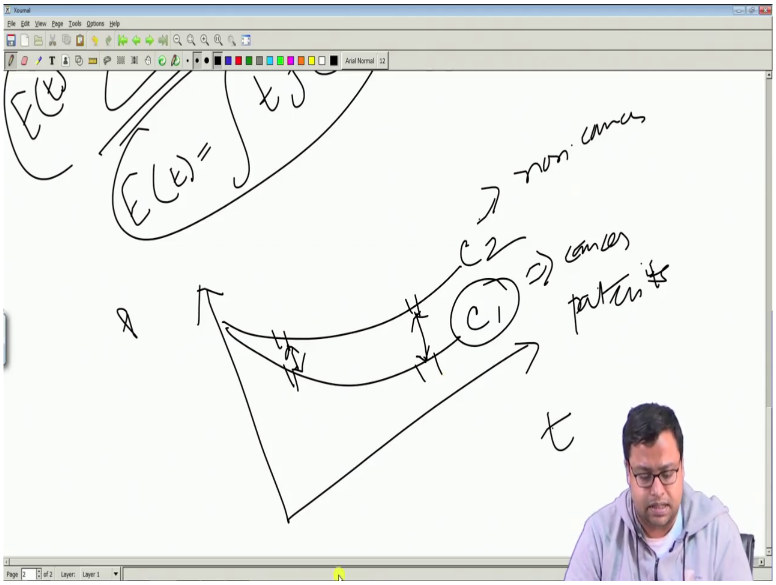


So, if I can give you information like this, that the probability, that your lifetime L is greater than t, if I can give this particular probability, that what is the probability that your lifetime will be greater than t. So, that is particularly this particular graph.

So, that is something that I am giving and then t into this thing, summation of that or in this case integration of that, if by chance if it is a continuous distribution that is basically the your expected lifetime. So, expected lifetime. So, if I can by chance, get this particular probability, if I can find out what is the expected lifetime then that is something that I can use for customer lifetime value calculation.

Because you will see that whether I can do the calculation for 5 years or 6 years or 10 years, what is the value that I will do is sometimes come from here this kind of analysis and how much you were concentrating as expected lifetime. So, there are various other case where these particular things apply.

(Refer Slide Time: 02:49)



For example, it can also be applied that how much will be the time taken before you die if you have corona virus, let us say. So, in certain cases, it is much low, in certain cases virus cases it must low, in certain viruses cases it is much high or how much time will this particular thing, what will the probability?

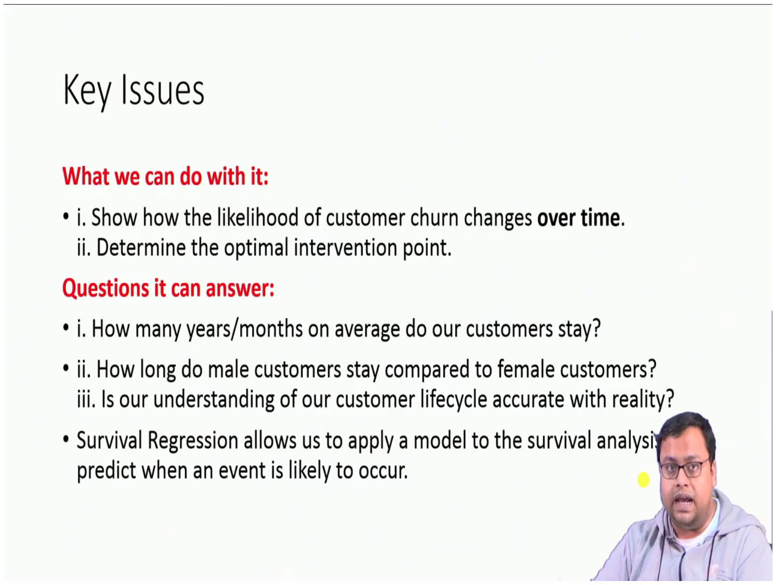
So let us say, in general, we have a probability of living up to 60 years, 70 years, 80 years. But given that you have this particular disease, what is a probability that you will live and how that changes how the probability changes over time. So, let us say if there are two curves, which looks like this, this is for probability, this is time, this is for curve one, this is for curve two, which is group one and group two, you will say that as time increases, as you gets older, your probability of living comes down, obviously.

But for C1 group of customer who is let us say, cancer patients are much lower than people who are not cancer patients and here, the ratio is the odds ratio, or not exactly odds ratio, the hazard rate, hazard ratio is much less and here it is much higher, let us say. So, once you get aged, the impact of cancer on your probability of living is much, much higher and that is something that we sometimes want to see.

Now, in this case, it is not cancer, we are talking about customers. So, given that there is some service failure, the older customers will be having higher chance or the loyal customers will have a higher chance to stay back and non-loyal customers will be having higher chances of going away.

So, if I can find out that loyalty card. But you can find out that one of the various aspects which impact your decision of staying, surviving as a customer or not surviving as a customer or staying in this particular company or not staying in a particular company then that can be analyzed using this survival analysis techniques.

(Refer Slide Time: 04:53)



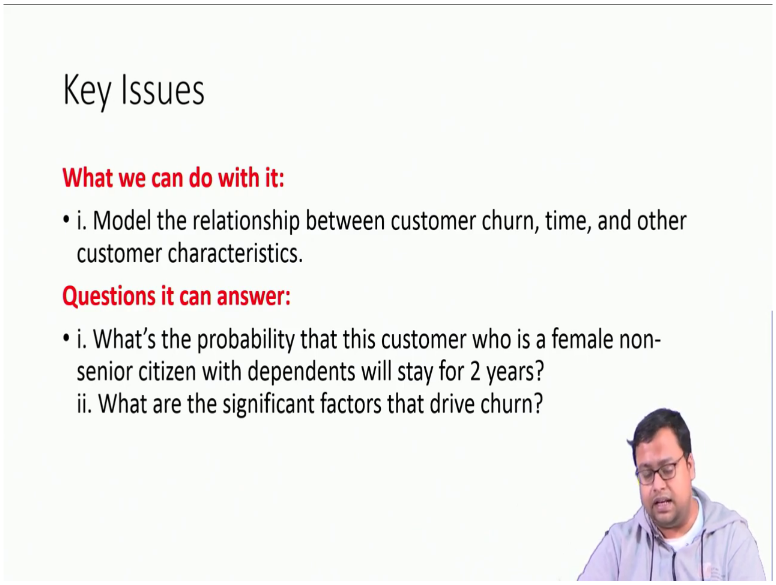
Key Issues

What we can do with it:

- i. Show how the likelihood of customer churn changes **over time**.
  ii. Determine the optimal intervention point.

Questions it can answer:

- i. How many years/months on average do our customers stay?
- ii. How long do male customers stay compared to female customers?
  iii. Is our understanding of our customer lifecycle accurate with reality?
- Survival Regression allows us to apply a model to the survival analysi
  predict when an event is likely to occur.



Key Issues

What we can do with it:

- i. Model the relationship between customer churn, time, and other customer characteristics.

Questions it can answer:

- i. What's the probability that this customer who is a female non-senior citizen with dependents will stay for 2 years?
  ii. What are the significant factors that drive churn?

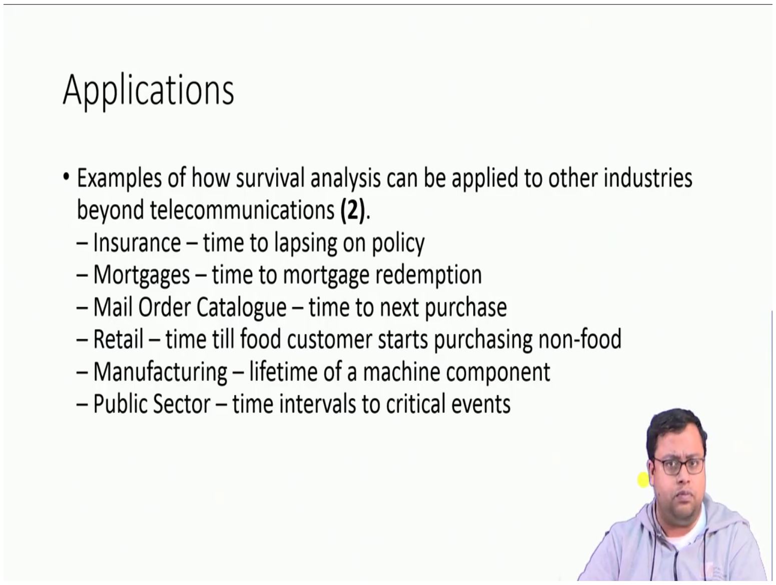
So, what are the key issues? The key issues are, so how the likelihood of customer churn changes over time and determine the potential intervention points. So, at one point you should intervene and questions if it can answer is how many years months on an average do the customer stay?

So, if I do not do anything by one time, how much will leave the customer in my service? And how long do male customers stay compared to female customers? So, is there any difference between gender? Is there very difference between nationality is there any difference between one customer group and another customer group? So, this is something that I can try to find out using this particular technique and is our understanding of customer lifecycle accurate with reality.

So, this is something that I can also try to find out that whether whatever we understand about the customer lifecycle, is this the case also. So, survival regression allows us to apply a model to the survival analysis to predict when an event is likely to occur. So, these are basic issues in which this particular analysis technique works. What are the key issues? What we can do with it? So, we can model the relationship between customer churn time and other customer characteristics.

This is something that we can do and we can also answer, this kind of question that what is the probability that this customer who is a female non-senior citizen with dependents will stay for two years? And what are the significant factors that drive churn this kind of answers I can give.

(Refer Slide Time: 06:26)

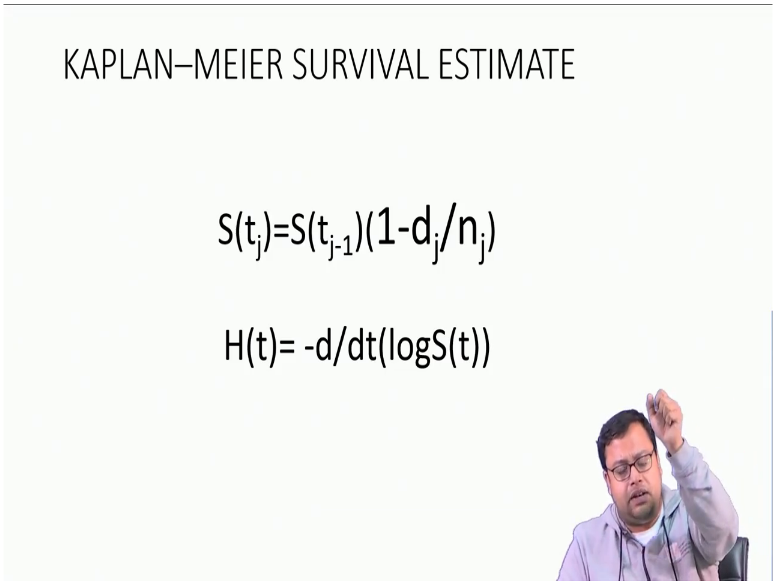


Applications

- Examples of how survival analysis can be applied to other industries beyond telecommunications **(2)**.
  - Insurance – time to lapsing on policy
  - Mortgages – time to mortgage redemption
  - Mail Order Catalogue – time to next purchase
  - Retail – time till food customer starts purchasing non-food
  - Manufacturing – lifetime of a machine component
  - Public Sector – time intervals to critical events

So, there are basically there are lots of applications. In telecommunication, there is a huge application, there is application in insurance, mortgages, mail order catalogue, retail, manufacturing and public sector. So, there are lots of places where it has a, so for example manufacturing it has a application is lifetime machine components in public sectaries, time intervals to critical events. In retail, time till food customer start purchasing nonfood and so on.

(Refer Slide Time: 06:56)



KAPLAN–MEIER SURVIVAL ESTIMATE

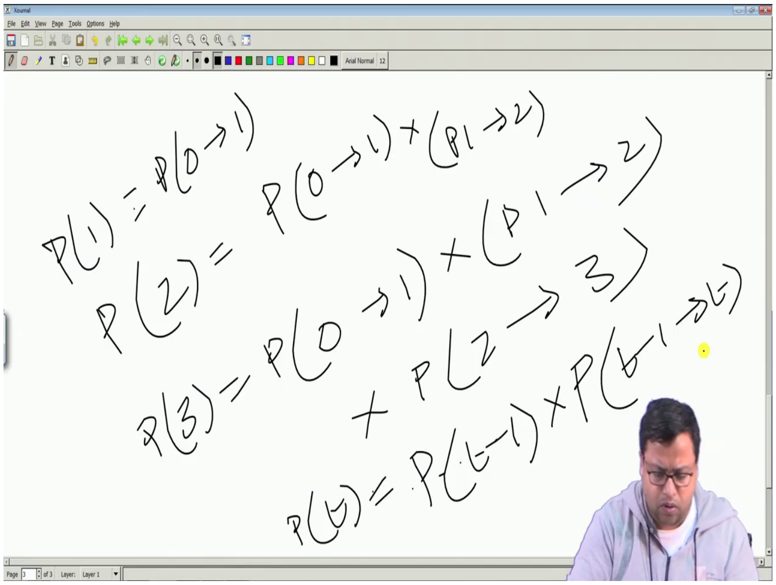
$$S(t_j)=S(t_{j-1})(1-d_j/n_j)$$

$$H(t)= -d/dt(logS(t))$$

Now, we have two estimation techniques. One is Kaplan-Meier which is very basic which says that the probability of staying back is nothing but, the probability of leaving given that you have leave the first time. So, it is saying that every time period, the factor that he will be living in time period two is independent of whatever was your probability that he will be living in time period one.

But, so switching, so from time period one to time period two this switch will not depend on the probability of time period 0 to time period 1. So, that is something that they are saying. That you will survive from time period 0 to time period 1 has no impact, that you will time period 2, given time period 1. But it is also saying the maths is also saying that the probability, if they are not independent.

(Refer Slide Time: 07:45)

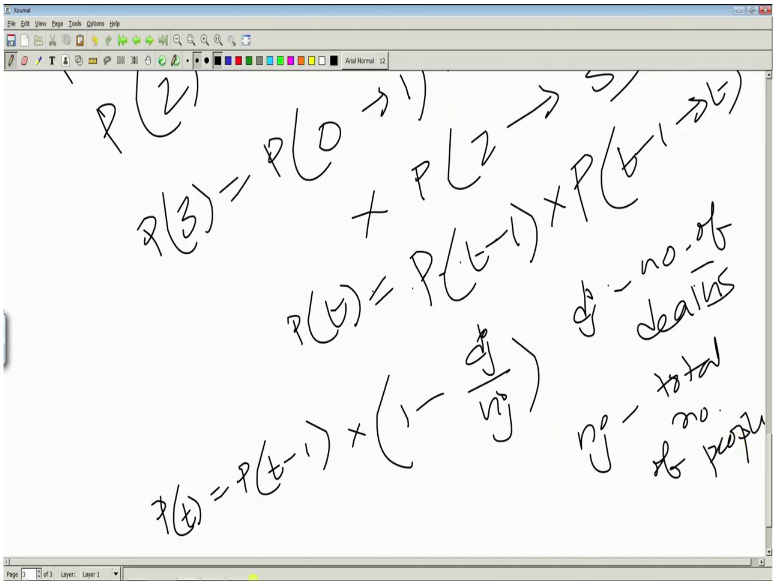

Xournal
File Edit View Page Tools Options Help
Arial Normal 12
Page 3 of 3 Layer: Layer 1

$P(1) = P(0 \to 1)$

$P(2) = P(0 \to 1) \times (P1 \to 2)$

$P(3) = P(0 \to 1) \times (P1 \to 2) \times P(2 \to 3)$

$P(t) = P(t-1) \times P(t-1 \to t)$


Xournal
File Edit View Page Tools Options Help
Arial Normal 12
Page 3 of 3 Layer: Layer 1

$P(2)$

$P(3) = P(0 \to 1)$

$\times P(2 \to 3)$

$P(t) = P(t-1) \times P(t-1 \to t)$

d_j - no. of deaths

$P(t) = P(t-1) \times \left(1 - \frac{d_j}{n_j}\right)$

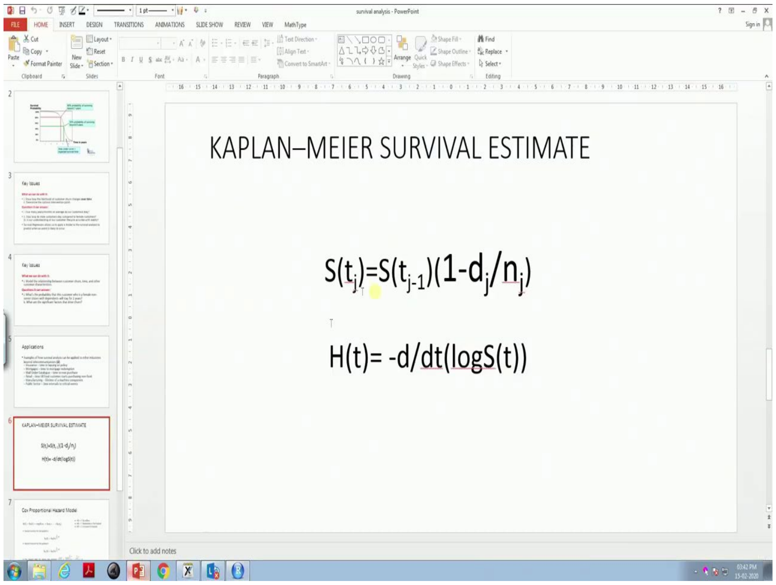
n_j - total no. of people

Then the probability that you will survive time period two is time period, 0 to 1 into time period 1 to 2. Fair enough, $P(1) = P(0 \to 1)$, $P(2) = P(0 \to 1) X P(1 \to 2)$, $P(2) = P(0 \to 1) X P(1 \to 2) XP(2 \to 3)$. So, then $P(t) = P(t \to 1) \times P(t-1 \ \to t)$

So, then if I go on doing this maths, then probability of t is nothing but $P(t) = P(t-1)(1 - \frac{d_j}{n_j})$

probability what is d_j, d_j is number of deaths, at that time period and n_j is total number of people. So, that is what this particular formula is saying.

(Refer Slide Time: 08:54)



KAPLAN–MEIER SURVIVAL ESTIMATE

$S(t_j)=S(t_{j-1})(1-d_j/n_j)$

$H(t)= -d/dt(\log S(t))$

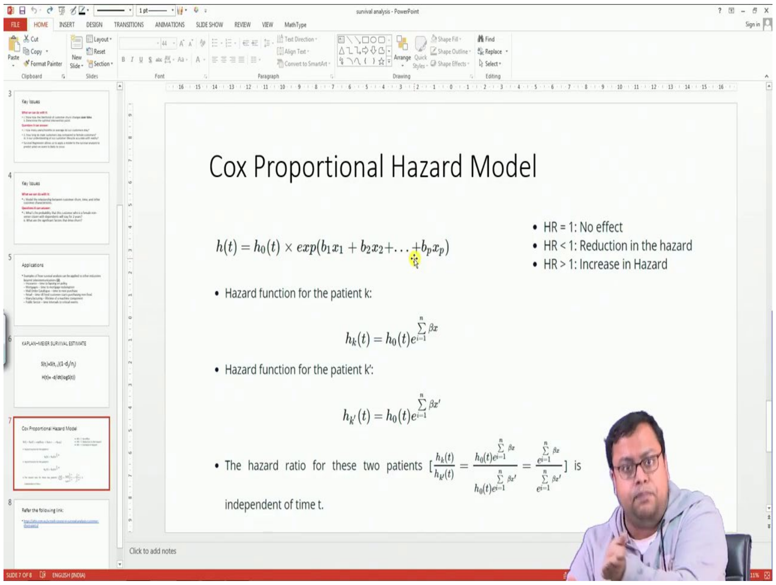
Click to add notes

If you check this formula, this formula is saying that that that your survival rate is: $S(t_j) = S(t_{j-1})(1-\frac{d_j}{n_j})$. So, that is the formula dj is number of death. So, this is a probability of being death, 1 minus that this is probability of being remaining alive. So, the probability that you will remain alive up to j-1, j minus 1th time into jth time periods probability of remaining alive is the probability that you will be alive till jth time period, fair enough and then the hazard rate is the change of this death rate.

Nothing but the change of log of this rate, why log? Because this is a probability, which is a very small number 0.00000 something if you take log, it becomes a little bit of handleable log likelihood we used to take that is why and if we you take log, this 0.001 becomes 10 to the power minus 6. So, log of 10 to the power minus 6 is basically minus 6 that minus 6 number is still handleable than 10 to power minus 6 which is a very small number.

So, this is called hazard rate, that rate of change of log of likelihood of survival. $H(t) = -\frac{d}{dt}(\log S(t))$. So, survival rate is basically probability of surviving. So, that is some things it is a nonparametric method we can calculate and then we can calculate between two groups and show that one group has higher hazard rate than other group. So, one group's rate of survival drop is much steep, other group is not so steep. If that is the case, then I can say that these two groups are different.

(Refer Slide Time: 10:46)



Cox Proportional Hazard Model

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + \ldots + b_px_p)$$

- HR = 1: No effect
- HR < 1: Reduction in the hazard
- HR > 1: Increase in Hazard

- Hazard function for the patient k:

$$h_k(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x}$$

- Hazard function for the patient k':

$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x'}$$

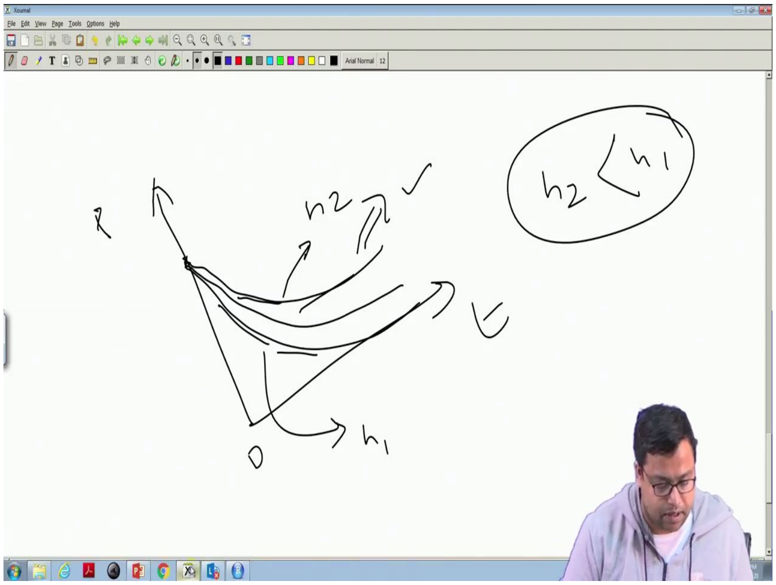
- The hazard ratio for these two patients $[\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_{i=1}^{n}\beta x}}{h_0(t)e^{\sum_{i=1}^{n}\beta x'}} = \frac{e^{\sum_{i=1}^{n}\beta x}}{e^{\sum_{i=1}^{n}\beta x'}}]$ is independent of time t.

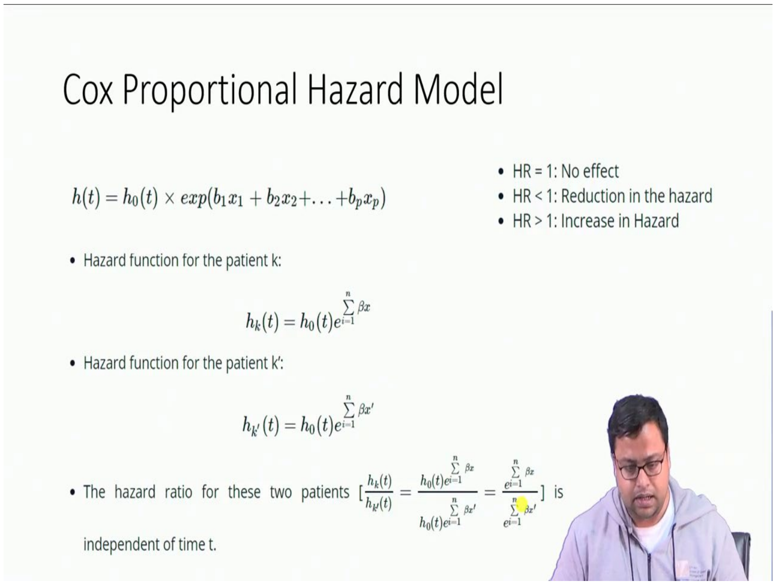
Click to add notes

Another method is basically Cox proportional hazard method which is say that this hazard rate is not only dependent on time, but also dependent on many other factors and these x1, x2 are those factors. So, instead of two groups, you can take multiple groupings, like age, gender, income together and say that all of this impacts the hazard rate.

So, all of this impacts the rate in which you will drop. So, in general, your probability of leaving third year, fourth year, fifth year is this much. But if you have from high income group, your hazard rate is much lower than a lower income group. In lower income group, the probability of surviving over some number of years drops in a much steeper way, drops in a much steeper way.

(Refer Slide Time: 11:38)





Cox Proportional Hazard Model

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + \ldots + b_px_p)$$

- HR = 1: No effect
- HR < 1: Reduction in the hazard
- HR > 1: Increase in Hazard

- Hazard function for the patient k:

$$h_k(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x}$$

- Hazard function for the patient k':

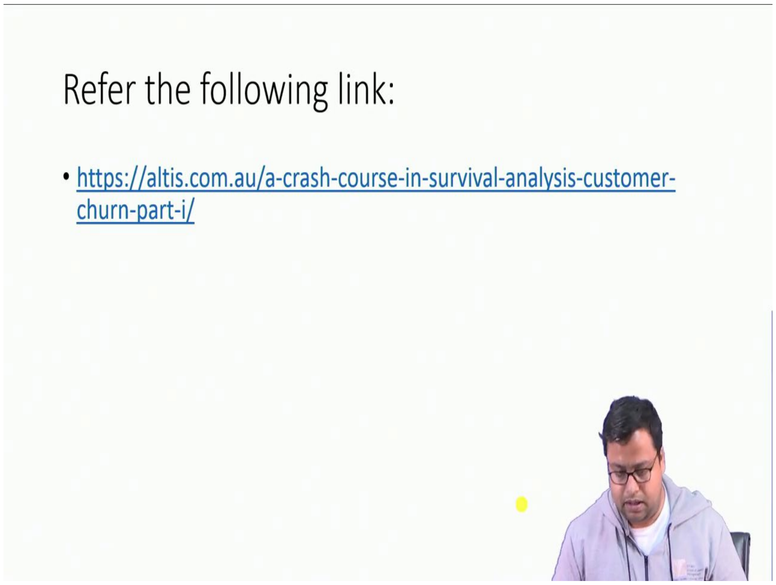
$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x'}$$

- The hazard ratio for these two patients $[\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_{i=1}^{n}\beta x}}{h_0(t)e^{\sum_{i=1}^{n}\beta x'}} = \frac{e^{\sum_{i=1}^{n}\beta x}}{e^{\sum_{i=1}^{n}\beta x'}}]$ is independent of time t.

So, which looks like this. If I normally, the probability of leaving at certain time period is high at 0 time period. Slowly as you go on aging, the probability gets dropped fair enough, this is the probability of leaving this is time, slowly it goes down, makes sense. If I, if you are a age of 95, the probability that he will he live 96 years is much low goes down. But these going down rate if it is average for rich people and educated people and informative people this is this and for poor people it might be a little bit much lower than that.

So this, this is the change the change is much steeper, here the changes much flatter. So, this case in the lower case the hazard rate if I say h1, in the upper case, If I write say h2, h2 is much lower than h1, the hazard rate is lower. So, that is something that I am trying to say here that if the hazard rate is lower or the ratio of hazards is lower than one. Then there is a reduction of hazard. If the other case there is an increase of hazard. So, there are two hazard functions that we find out and then we try to check that the hazard ratio we try to find out.

(Refer Slide Time: 12:55)

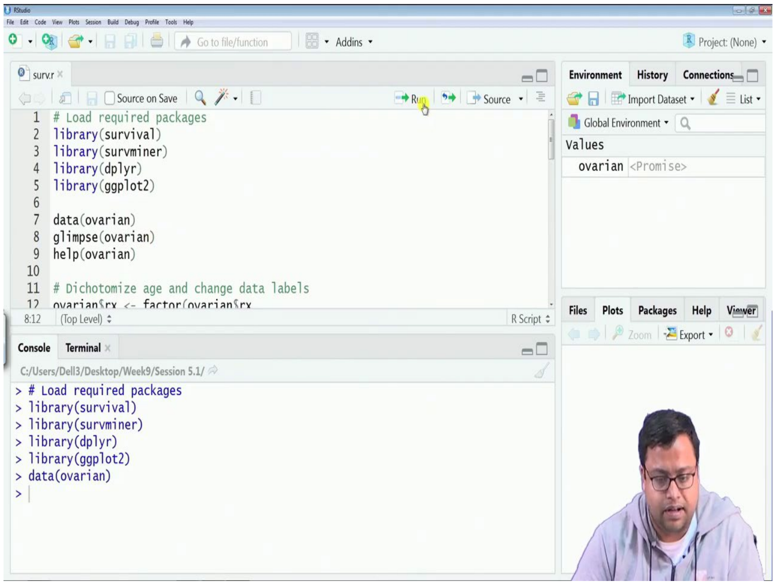


Refer the following link:

- https://altis.com.au/a-crash-course-in-survival-analysis-customer-churn-part-i/

And you can read from this particular link about more about this survival analysis.

(Refer Slide Time: 13:10)



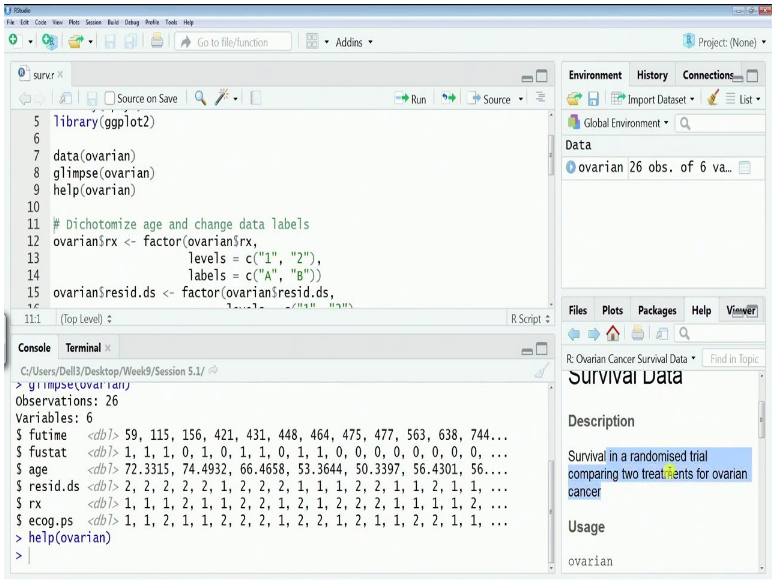
```r
# Load required packages
library(survival)
library(survminer)
library(dplyr)
library(ggplot2)

data(ovarian)
glimpse(ovarian)
help(ovarian)

# Dichotomize age and change data labels
ovarian$rx <- factor(ovarian$rx
```

Values

ovarian <Promise>

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
> # Load required packages
> library(survival)
> library(survminer)
> library(dplyr)
> library(ggplot2)
> data(ovarian)
>
```



```r
library(ggplot2)

data(ovarian)
glimpse(ovarian)
help(ovarian)

# Dichotomize age and change data labels
ovarian$rx <- factor(ovarian$rx,
                     levels = c("1", "2"),
                     labels = c("A", "B"))
ovarian$resid.ds <- factor(ovarian$resid.ds,
```

Data

ovarian 26 obs. of 6 va...

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
> glimpse(ovarian)
Observations: 26
Variables: 6
$ futime   <dbl> 59, 115, 156, 421, 431, 448, 464, 475, 477, 563, 638, 744...
$ fustat   <dbl> 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ age      <dbl> 72.3315, 74.4932, 66.4658, 53.3644, 50.3397, 56.4301, 56....
$ resid.ds <dbl> 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, ...
$ rx       <dbl> 1, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 2, ...
$ ecog.ps  <dbl> 1, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, ...
> help(ovarian)
>
```

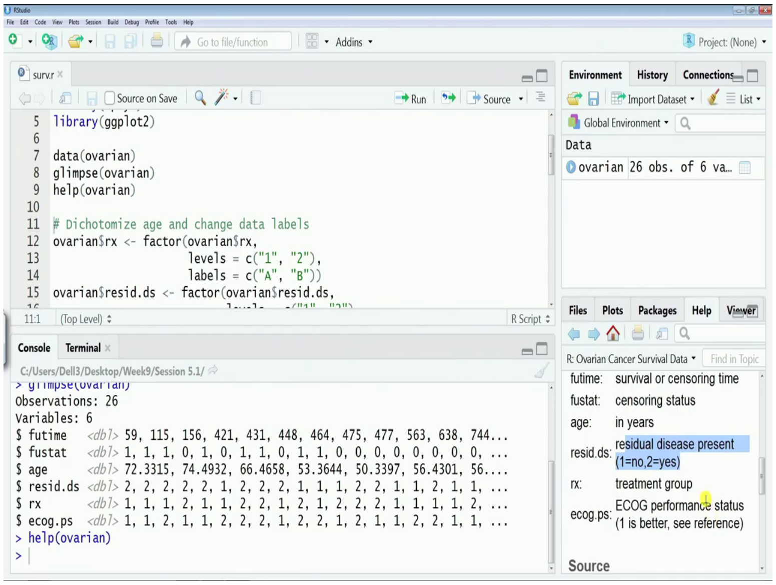
Survival Data

Description

Survival in a randomised trial comparing two treatments for ovarian cancer

Usage

ovarian



```r
library(ggplot2)

data(ovarian)
glimpse(ovarian)
help(ovarian)

# Dichotomize age and change data labels
ovarian$rx <- factor(ovarian$rx,
                     levels = c("1", "2"),
                     labels = c("A", "B"))
ovarian$resid.ds <- factor(ovarian$resid.ds,
```

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
> glimpse(ovarian)
Observations: 26
Variables: 6
$ futime   <dbl> 59, 115, 156, 421, 431, 448, 464, 475, 477, 563, 638, 744...
$ fustat   <dbl> 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ age      <dbl> 72.3315, 74.4932, 66.4658, 53.3644, 50.3397, 56.4301, 56....
$ resid.ds <dbl> 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, ...
$ rx       <dbl> 1, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 2, ...
$ ecog.ps  <dbl> 1, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, ...
> help(ovarian)
>
```

Data

ovarian 26 obs. of 6 va...

futime:	survival or censoring time
fustat:	censoring status
age:	in years
resid.ds:	residual disease present (1=no,2=yes)
rx:	treatment group
ecog.ps:	ECOG performance status (1 is better, see reference)

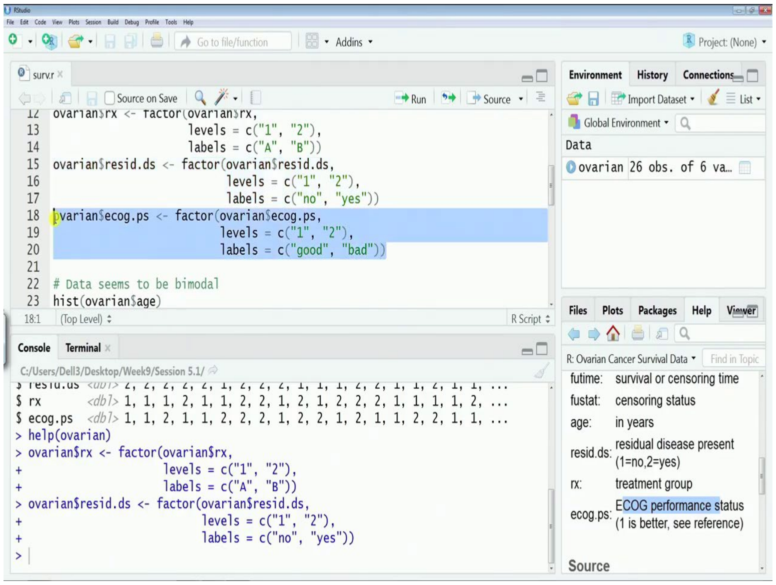
Source

But I will just quickly show a study, , for which you have to open this surv.r file. So, there are four libraries, I called the library, you have to install these libraries before if you do not have them and the data that I will be using is basically the ovarian data. It is ovarian cancer data already inbuilt in us and if you want to have a glimpse of this data, it has 26 observations of six variables and this guy is the futime fustat age, the residents rx and ecog.

So these are various kinds of factors which impact the chances of living and this is the age and this is fustat and whether this guy is living or death 1, 0 and the time period and what time span this measurement has been done is something that they are checking and the help of the description of this information will come here.

So, you can get all the details what is this? So, if time is the survival of our censoring time when you are measuring, this is the censoring status, this is age in years and then this is basically residual disease present, 1 is no 2 is yes. This treatment group or control group and ecog is the performance status. These are some of the things that they are checking.

(Refer Slide Time: 14:36)



surv.r

```
ovarian$rx <- factor(ovarian$rx,
                     levels = c("1", "2"),
                     labels = c("A", "B"))
ovarian$resid.ds <- factor(ovarian$resid.ds,
                           levels = c("1", "2"),
                           labels = c("no", "yes"))
ovarian$ecog.ps <- factor(ovarian$ecog.ps,
                          levels = c("1", "2"),
                          labels = c("good", "bad"))

# Data seems to be bimodal
hist(ovarian$age)
```

Console — C:/Users/Dell3/Desktop/Week9/Session 5.1/

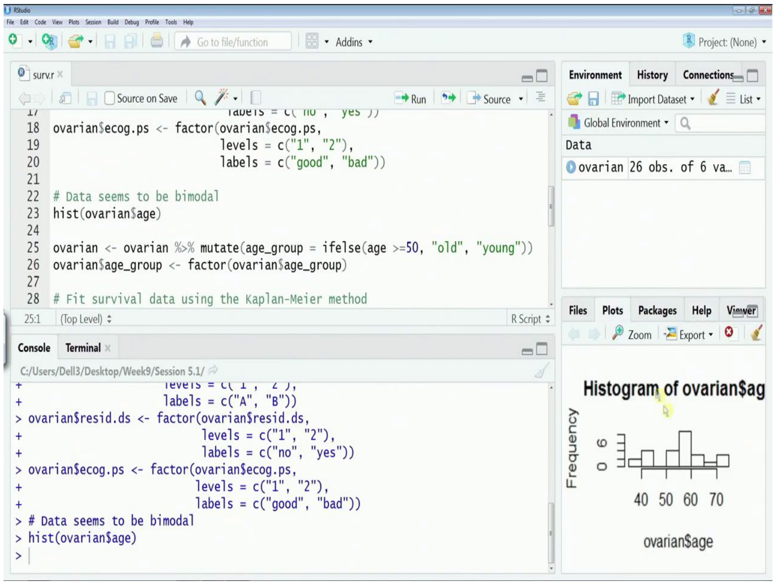
```
$ resid.ds <dbl> 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, ...
$ rx       <dbl> 1, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 2, ...
$ ecog.ps  <dbl> 1, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, ...
> help(ovarian)
> ovarian$rx <- factor(ovarian$rx,
+                      levels = c("1", "2"),
+                      labels = c("A", "B"))
> ovarian$resid.ds <- factor(ovarian$resid.ds,
+                            levels = c("1", "2"),
+                            labels = c("no", "yes"))
> 
```

Environment — Global Environment — Data: ovarian 26 obs. of 6 va...

R: Ovarian Cancer Survival Data

futime:	survival or censoring time
fustat:	censoring status
age:	in years
resid.ds:	residual disease present (1=no,2=yes)
rx:	treatment group
ecog.ps:	ECOG performance status (1 is better, see reference)

Source



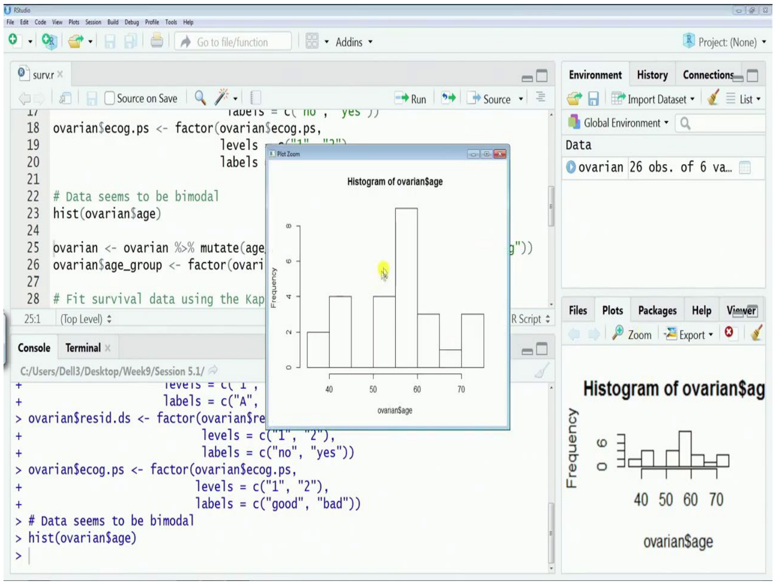
surv.r

```
                           labels = c("no", "yes"))
ovarian$ecog.ps <- factor(ovarian$ecog.ps,
                          levels = c("1", "2"),
                          labels = c("good", "bad"))

# Data seems to be bimodal
hist(ovarian$age)

ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
ovarian$age_group <- factor(ovarian$age_group)

# Fit survival data using the Kaplan-Meier method
```

Console — C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
+                      levels = c("1", "2"),
+                      labels = c("A", "B"))
> ovarian$resid.ds <- factor(ovarian$resid.ds,
+                            levels = c("1", "2"),
+                            labels = c("no", "yes"))
> ovarian$ecog.ps <- factor(ovarian$ecog.ps,
+                           levels = c("1", "2"),
+                           labels = c("good", "bad"))
> # Data seems to be bimodal
> hist(ovarian$age)
> 
```

Environment — Global Environment — Data: ovarian 26 obs. of 6 va...



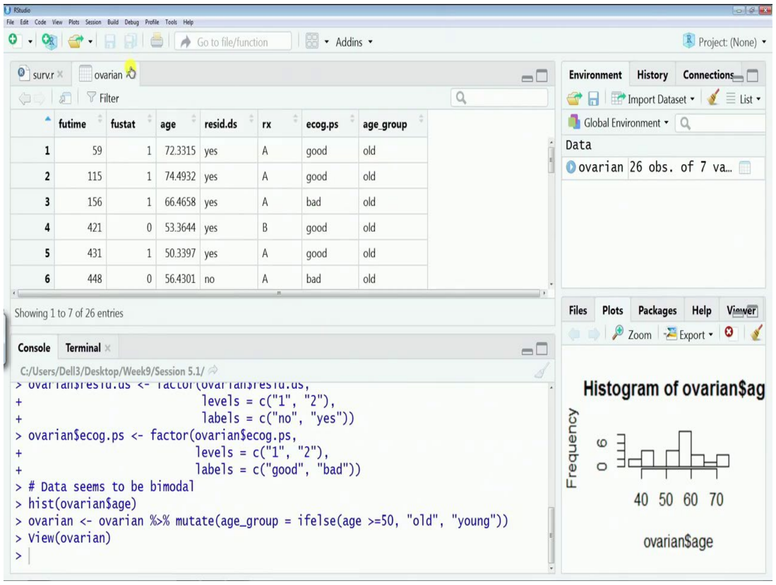
```
                         labels = c("no", "yes"))
ovarian$ecog.ps <- factor(ovarian$ecog.ps,
                          levels = c("1", "2")
                          labels

# Data seems to be bimodal
hist(ovarian$age)

ovarian <- ovarian %>% mutate(age                                    g"))
ovarian$age_group <- factor(ovari

# Fit survival data using the Kap
```

Data

ovarian 26 obs. of 6 va…

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
+                    levels = c("1",
+                    labels = c("A",
> ovarian$resid.ds <- factor(ovarian$re
+                          levels = c("1", "2"),
+                          labels = c("no", "yes"))
> ovarian$ecog.ps <- factor(ovarian$ecog.ps,
+                         levels = c("1", "2"),
+                         labels = c("good", "bad"))
> # Data seems to be bimodal
> hist(ovarian$age)
> 
```



	futime	fustat	age	resid.ds	rx	ecog.ps	age_group
1	59	1	72.3315	yes	A	good	old
2	115	1	74.4932	yes	A	good	old
3	156	1	66.4658	yes	A	bad	old
4	421	0	53.3644	yes	B	good	old
5	431	1	50.3397	yes	A	good	old
6	448	0	56.4301	no	A	bad	old

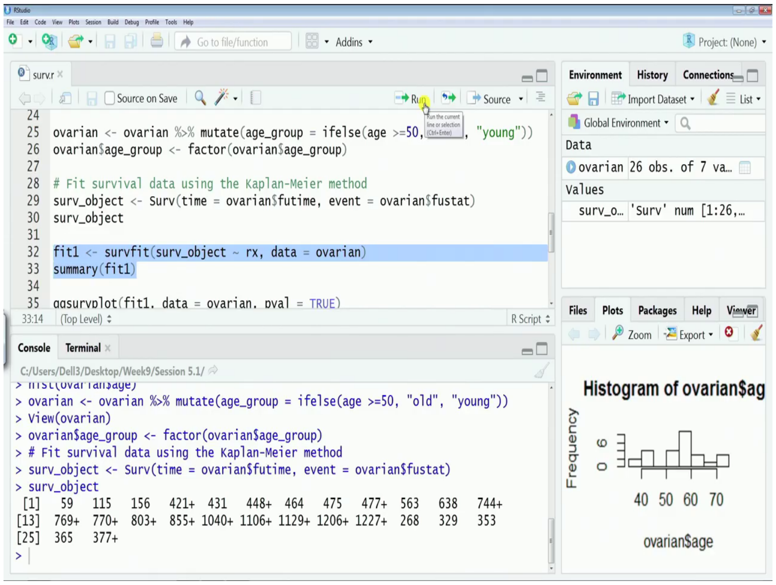
Showing 1 to 7 of 26 entries

Data

ovarian 26 obs. of 7 va…

C:/Users/Dell3/Desktop/Week9/Session 5.1/

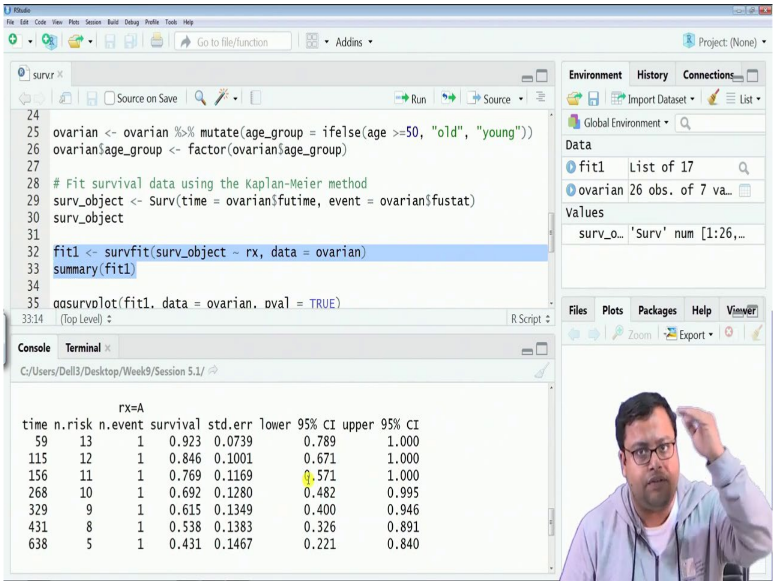
```
> ovarian$resid.ds <- factor(ovarian$resid.ds,
+                          levels = c("1", "2"),
+                          labels = c("no", "yes"))
> ovarian$ecog.ps <- factor(ovarian$ecog.ps,
+                         levels = c("1", "2"),
+                         labels = c("good", "bad"))
> # Data seems to be bimodal
> hist(ovarian$age)
> ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
> View(ovarian)
> 
```



```r

ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
ovarian$age_group <- factor(ovarian$age_group)

# Fit survival data using the Kaplan-Meier method
surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)
surv_object

fit1 <- survfit(surv_object ~ rx, data = ovarian)
summary(fit1)

ggsurvplot(fit1, data = ovarian, pval = TRUE)
```

C:/Users/Dell3/Desktop/Week9/Session 5.1/

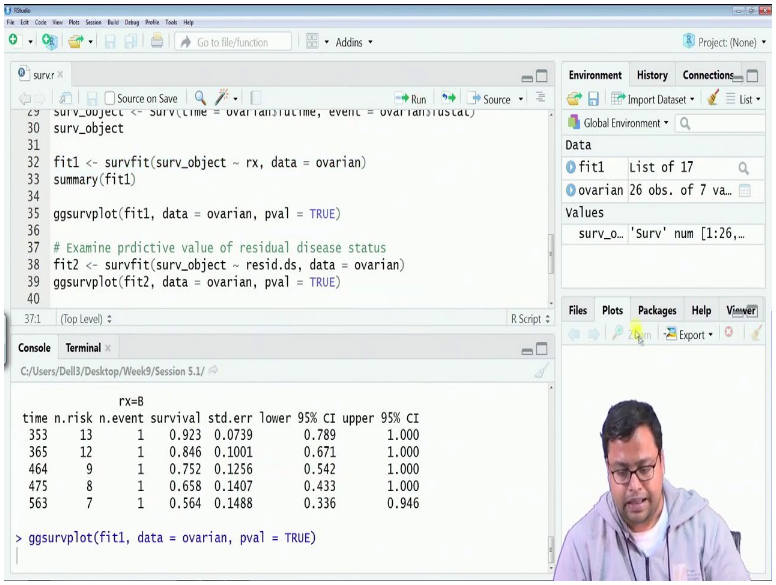
```
> hist(ovarian$age)
> ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
> View(ovarian)
> ovarian$age_group <- factor(ovarian$age_group)
> # Fit survival data using the Kaplan-Meier method
> surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)
> surv_object
 [1]   59   115   156   421+  431   448+  464   475   477+  563   638   744+
[13]  769+  770+  803+  855+ 1040+ 1106+ 1129+ 1206+ 1227+  268   329   353
[25]  365   377+
> 
```



```r

ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
ovarian$age_group <- factor(ovarian$age_group)

# Fit survival data using the Kaplan-Meier method
surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)
surv_object

fit1 <- survfit(surv_object ~ rx, data = ovarian)
summary(fit1)

ggsurvplot(fit1, data = ovarian, pval = TRUE)
```

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```

                rx=A 
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  59     13       1    0.923  0.0739        0.789        1.000
 115     12       1    0.846  0.1001        0.671        1.000
 156     11       1    0.769  0.1169        0.571        1.000
 268     10       1    0.692  0.1280        0.482        0.995
 329      9       1    0.615  0.1349        0.400        0.946
 431      8       1    0.538  0.1383        0.326        0.891
 638      5       1    0.431  0.1467        0.221        0.840
```



```r
surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)
surv_object

fit1 <- survfit(surv_object ~ rx, data = ovarian)
summary(fit1)

ggsurvplot(fit1, data = ovarian, pval = TRUE)

# Examine prdictive value of residual disease status
fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
ggsurvplot(fit2, data = ovarian, pval = TRUE)
```

Data
fit1 List of 17
ovarian 26 obs. of 7 va...
Values
surv_o... 'Surv' num [1:26,...

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
                rx=B 
time n.risk n.event survival std.err lower 95% CI upper 95% CI
 353     13       1    0.923  0.0739        0.789        1.000
 365     12       1    0.846  0.1001        0.671        1.000
 464      9       1    0.752  0.1256        0.542        1.000
 475      8       1    0.658  0.1407        0.433        1.000
 563      7       1    0.564  0.1488        0.336        0.946

> ggsurvplot(fit1, data = ovarian, pval = TRUE)
```

So, what first to do is I have put these groups rx groups into two groups A and B. So, 1 and 2, I make them factor and label them as A and B. Similarly, ovarian residual, the death status is one yes or no and then this one is good and bad. So, ecog condition is good and bad is something that we are changing. So, 1 and 2 and etc, we are recording and now if I just plot the age, this is the age, so it seems like there are bimodal, two modes are there. It might be that is something that this particular thing is being shown.

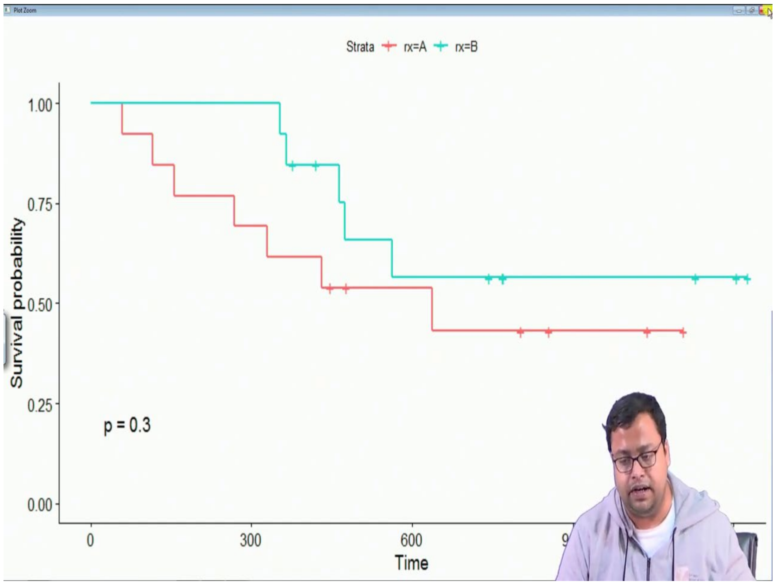
Then what I am doing is, I am changing this ovarian data set and mutate, mutate means if you change and put something here per age group is if the age group is higher than 15 than its old or otherwise it is young. So, I am creating a new variable in this. So, which is basically ovarian age group, which is old or new, I have created and make them factor. Now, what I am doing, I am creating a survival data set. So surv object is survival data, time is equal to the time when you are measuring and event is the fustat.

So, that is the death or alive that particular thing. So, I have created the surv object which is basically, a numeric variable, which is looks like this 59, 150 and plus or not plus is actually telling that whether it has survived or not and then if I just feed it with rx. So for 2 rx A and B, it will be separately creating the charts and if I put the summary what it is doing, you just carefully see.

First of all, it is making the data set for rx=A and rx is rx=B. 2 separate groups, it is creating two separate groups it is creating rx=A and rx=B and then what it is doing is that, it is saying that what are the various time periods when some event occurred? When time period was 59, there was thirteen people out there in rx A and only one event means one died and then twelve were there. So, 12 by 13 was the probability from time period 0 to 0 to time period 1. 12 by 13 comes up to be 0.923. Then in 115 time period there were 12 people another death happened. So, 11 by 12 was the death survival rate.

So, what was the S2? S2 is S1*11/12, that means 0.923*11/12, if you do the calculation it will come as 0.846 and slowly that went on calculating the survival rates, same thing they did for rx is equal to B also. Now if I just plot you will know, that this is how the plots look like with probability point three, they are different.

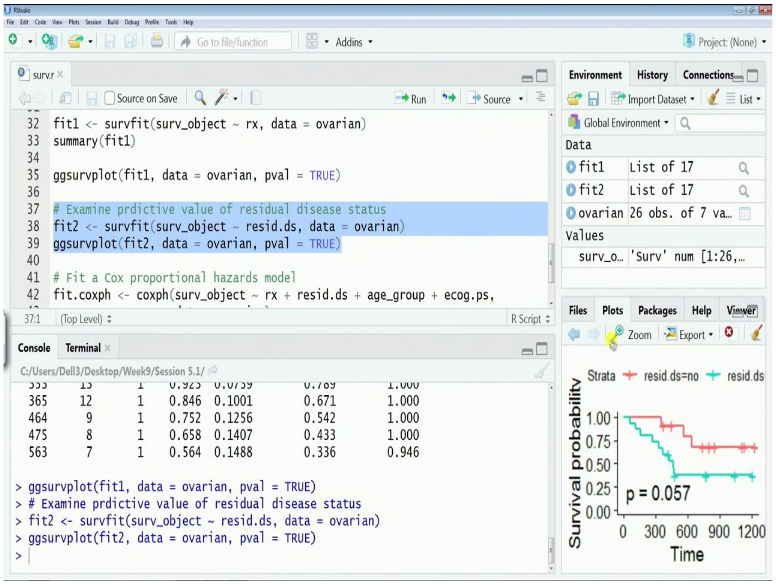
(Refer Slide Time: 17:35)



And if I just zoom it up, you will see that, rx=A looks like this , rx=B looks like this. Can you tell me which one has higher has hazard rate, obviously the graph which is at the top has the higher hazard rate. So rx=B has much less, lesser hazard rate the one at the top.

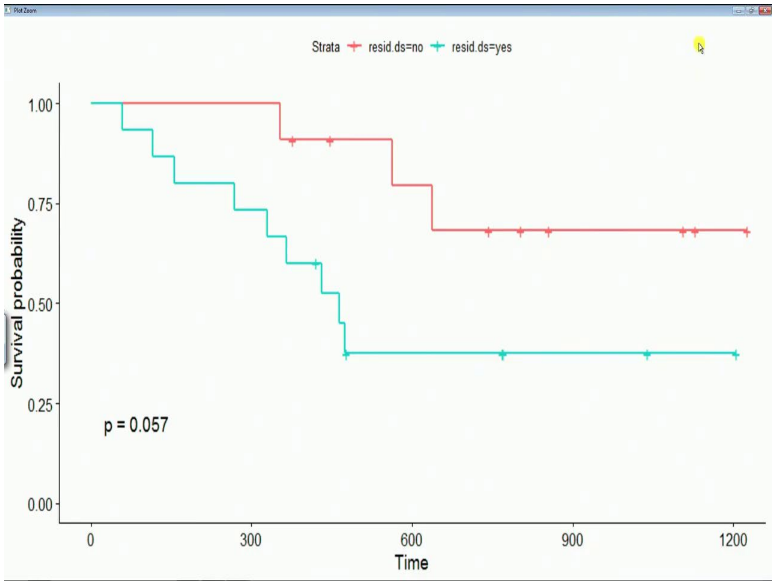
So, rx=B has much lesser hazard rate than rx=A and but on the other hand p is equal to point three these states state that whether these two are significantly different or not p lower than 0.05, they are significantly different p higher than 0.05, they are not different. So, as per my statistical analysis at least I can say that these two are not very different.

(Refer Slide Time: 18:14)



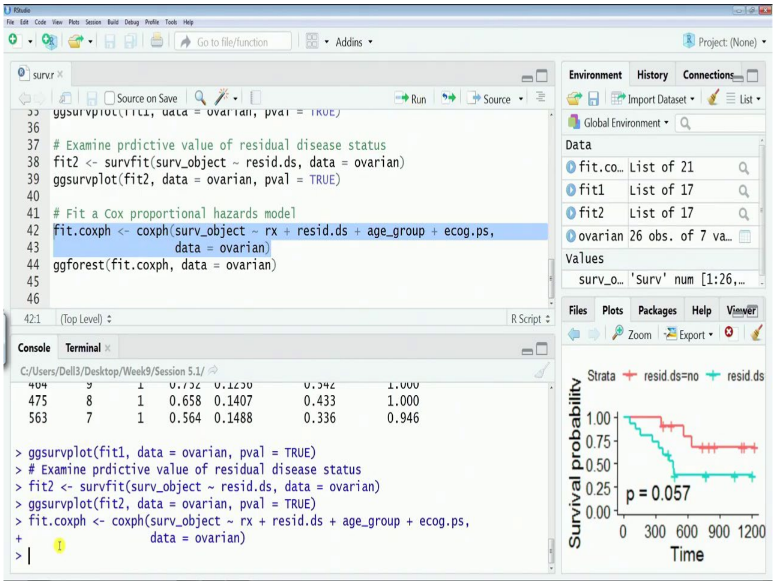
```
fit1 <- survfit(surv_object ~ rx, data = ovarian)
summary(fit1)

ggsurvplot(fit1, data = ovarian, pval = TRUE)

# Examine prdictive value of residual disease status
fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
ggsurvplot(fit2, data = ovarian, pval = TRUE)

# Fit a Cox proportional hazards model
fit.coxph <- coxph(surv_object ~ rx + resid.ds + age_group + ecog.ps,
```

```
 355     13     1    0.923  0.0739        0.789        1.000
 365     12     1    0.846  0.1001        0.671        1.000
 464      9     1    0.752  0.1256        0.542        1.000
 475      8     1    0.658  0.1407        0.433        1.000
 563      7     1    0.564  0.1488        0.336        0.946

> ggsurvplot(fit1, data = ovarian, pval = TRUE)
> # Examine prdictive value of residual disease status
> fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
> ggsurvplot(fit2, data = ovarian, pval = TRUE)
> 
```



If I do the same thing for resis.ds and then plot it. Now, it is 0.057. So, at least at 6 percent level they are different, if not 5 percent level and see, when resis.ds is no less survival rate when it is yes very high survival, very, very high hazard rate very much less survival rate is something that we can see here, at least based on the data that we have.

(Refer Slide Time: 18:44)



```
ggsurvplot(fit1, data = ovarian, pval = TRUE)

# Examine prdictive value of residual disease status
fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
ggsurvplot(fit2, data = ovarian, pval = TRUE)

# Fit a Cox proportional hazards model
fit.coxph <- coxph(surv_object ~ rx + resid.ds + age_group + ecog.ps,
                   data = ovarian)
ggforest(fit.coxph, data = ovarian)
```

Data

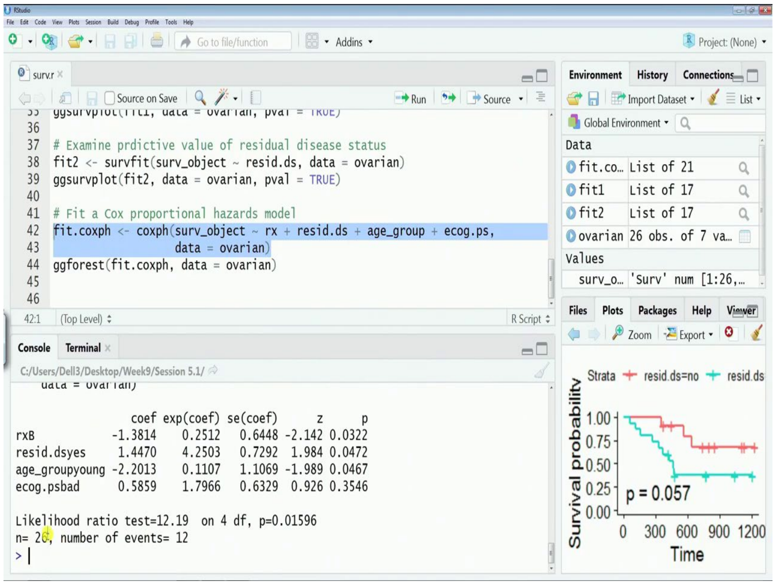
fit.co...	List of 21
fit1	List of 17
fit2	List of 17
ovarian	26 obs. of 7 va...

Values

surv_o...	'Surv' num [1:26,...

C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
 464      9      1    0.752  0.1256        0.542        1.000
 475      8      1    0.658  0.1407        0.433        1.000
 563      7      1    0.564  0.1488        0.336        0.946

> ggsurvplot(fit1, data = ovarian, pval = TRUE)
> # Examine prdictive value of residual disease status
> fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
> ggsurvplot(fit2, data = ovarian, pval = TRUE)
> fit.coxph <- coxph(surv_object ~ rx + resid.ds + age_group + ecog.ps,
+                    data = ovarian)
> |
```



```
ggsurvplot(fit1, data = ovarian, pval = TRUE)

# Examine prdictive value of residual disease status
fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
ggsurvplot(fit2, data = ovarian, pval = TRUE)

# Fit a Cox proportional hazards model
fit.coxph <- coxph(surv_object ~ rx + resid.ds + age_group + ecog.ps,
                   data = ovarian)
ggforest(fit.coxph, data = ovarian)
```

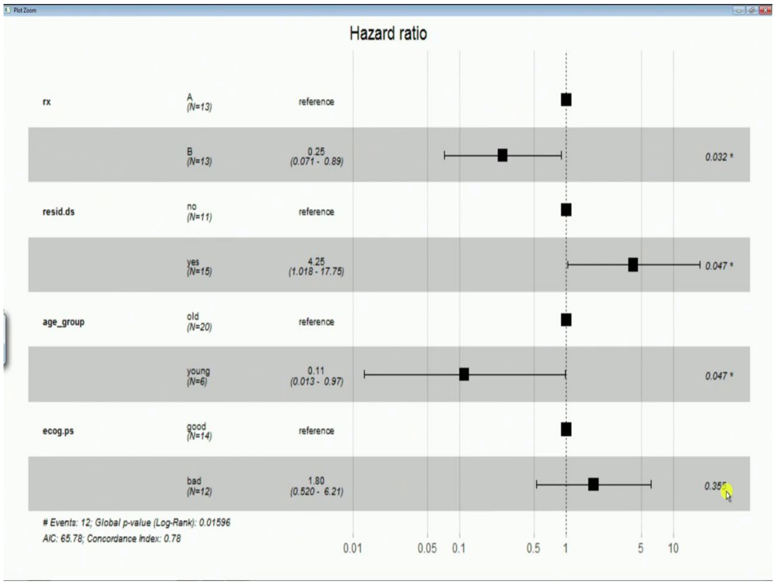
C:/Users/Dell3/Desktop/Week9/Session 5.1/

```
    data = ovarian)

                  coef exp(coef) se(coef)      z      p
rxB            -1.3814    0.2512   0.6448 -2.142 0.0322
resid.dsyes     1.4470    4.2503   0.7292  1.984 0.0472
age_groupyoung -2.2013    0.1107   1.1069 -1.989 0.0467
ecog.psbad      0.5859    1.7966   0.6329  0.926 0.3546

Likelihood ratio test=12.19  on 4 df, p=0.01596
n= 26, number of events= 12
> |
```

So, till now, we were doing the first method in the second method in that cox proportional hazard model. We can actually have multiple, so it is the summation of beta x, the equation is given in this particular in the, in our PPT and from there you can will understand that what this mean.

So, I am just plotting it. So, if I just run this much and see what this fit dot coxph say. It is giving the f coefficient beta coefficient's value there, this is the beta coefficients value for rx B and corresponding standard error of the coefficient and then corresponding Z values of the coefficient and then it is also giving the p value of the coefficient. So, these three has affect rx, resid ds and age group by young has effect, the ecological condition has no effect. So, I could have dropped that in the model. So, by two ecological condition it is not creating any different hazard rate for the groups.

(Refer Slide Time: 19:52)



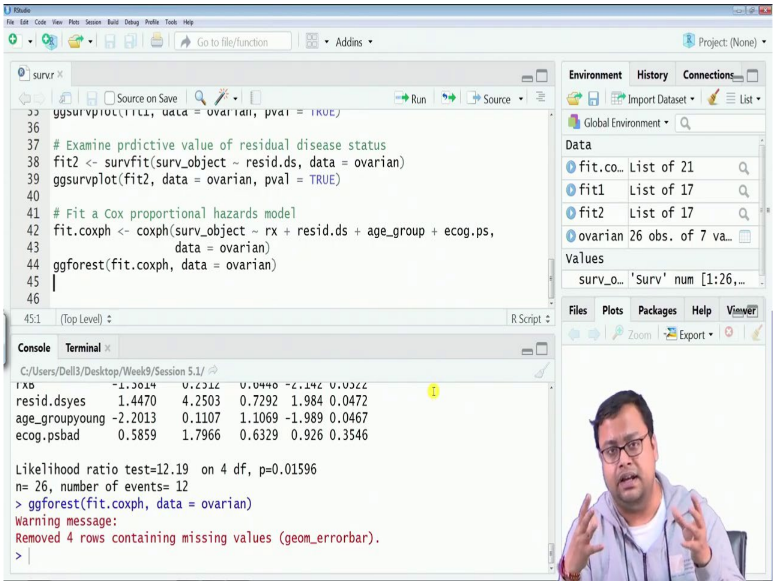
And if I just plot the ggforest, this is how it looks like. You would see that for rx A and B when I am saying rx A and rx B two groups. This guy has if keeping these as reference the hazard ratio is 0.25 and significant level is 0.032. So, this is significant. On the other hand hazard keeping resid ds as reference, no. The yes is 4.25 and the, this is upper limit and lower limit and it is also significant, very highly significant and then if I say that old this also that means here I am saying that the group B, 0.25 means hazard ratio less than 1.

Less than 1 means group B is less affected I would say the keeping everything else constant, Group B will have lower hazard ratio than Group A, how much lower four times lower I would say the ratio will be 1 by 4 and that is significant and here it is 4.25 by 1 and that is also significant.

Here it is 0.1 by 0.11 means basically, this ratio means 1 by 9. So, the young people has much less hazard ratio than old people. I am not giving this example in the context of, in the context of our customer behavior. But similar data you can create from the customer behaviour also and for ecog the bad is 1.8 means by chance this ecog is bad, you will have 1.8 times hazard ratio than when it is good, but this is not significant.

So, I can probably ignore this result, I have to consider this result, this result and the result because in all the cases this is lower than 0.05. But in this case, I can above the probably ignore. So, that is how we can create survival analysis. This is highly applicable in checking their how much time a customer will be staying with me or how much time I can when I can do the intervention which kind of intervention reduces the hazard rate.

(Refer Slide Time: 22:03)



```
ggsurvplot(fit1, data = ovarian, pval = TRUE)

# Examine prdictive value of residual disease status
fit2 <- survfit(surv_object ~ resid.ds, data = ovarian)
ggsurvplot(fit2, data = ovarian, pval = TRUE)

# Fit a Cox proportional hazards model
fit.coxph <- coxph(surv_object ~ rx + resid.ds + age_group + ecog.ps,
                   data = ovarian)
ggforest(fit.coxph, data = ovarian)
```

```
rx             -1.3814     0.2512    0.6448 -2.142 0.0322
resid.dsyes     1.4470     4.2503    0.7292  1.984 0.0472
age_groupyoung -2.2013     0.1107    1.1069 -1.989 0.0467
ecog.psbad      0.5859     1.7966    0.6329  0.926 0.3546

Likelihood ratio test=12.19  on 4 df, p=0.01596
n= 26, number of events= 12
> ggforest(fit.coxph, data = ovarian)
Warning message:
Removed 4 rows containing missing values (geom_errorbar).
>
```

Data
fit.co… List of 21
fit1 List of 17
fit2 List of 17
ovarian 26 obs. of 7 va…
Values
surv_o… 'Surv' num [1:26,…

So, like here we are saying that whether rx or age or something else is impacting the hazard rate, you could have seen that, let us say if customer lifetime is your value, that you are checking and every time period after 2 months, 3 months, 5 months, 6 months whether the customer. But they are with you 1 or 0 that was your hazard rate calculation.

Then you could have find out that what impacts your hazard rate. You did some intervention in January, the customer is still alive on September, you did another intervention on May, you this is also staying back in September, which one has an impact the intervention in January or intervention in May, which one actually made the customer to stay back? You do not know until unless you do this Cox analysis.

So Cox hazard model analysis. So that is how we can also find out that how to calculate exactly the customer lifetime and how to increase customer lifetime. One is churn management. But churn is a immediate management sometimes, you have to do something much ago, so that the lifetime increases.

So, how to do that? Which one will impact? Which one will not impact? Can be done using Cox proportional hazards model you can try out for other customer data you can try out a churn data from your customer databases and try your whether that can be applicable here. So, that is what I will stop week 9 in week 10 we will start text mining, text analytics and etc. Thank you very much for being with me. I will see you in the next week, thank you