# Marketing Analytics

**Dr. Swagato Chatterjee**

**Vinod Gupta School of Management**

**Indian Institute of Technology, Kharagpur**

## Lecture 11

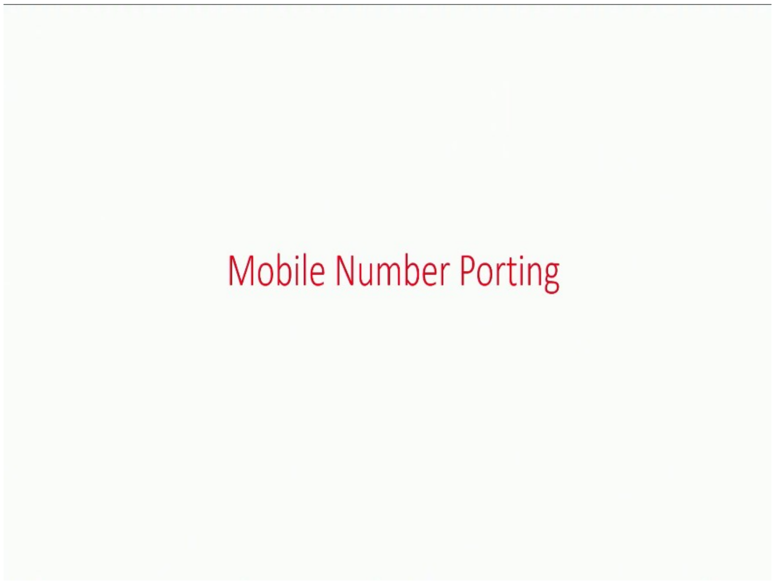
## What consumers want (contd.)

Hello everybody. Welcome to Marketing Analytics course. This is week 2, session 5. And this is Doctor Swagato Chatterjee, Vinod Gupta School of Management, IIT Kharagpur, will be taking this course. So, we are discussing about what consumers want and we will continue on that discussion in this particular video as well. Till the last session we have discussed about conjoint analysis a little bit and here in this particular video, we will actually setup the premise for choice based conjoint.

So, before we do anything on choice based conjoint, we have to do a little bit of something called choice modelling. So what is choice modelling? Choice modelling is actually trying to empirically see or trying to create an econometric or I would say quantitative model to actually find out how customers make choice. And here in this particular video we are focusing on choice which is yes, no kind of choice. So whether you will buy something or we will not buy something.

So, it is a binomial choice, but there can be multinomial choices as well, which we are keeping aside right now. So multinomial choices are something like,where you have multiple choices in your hand. So, choice A versus choice B versus choice C, choice D and you try to choose that which one out of this will, which one out of these 4 will you choose. So, here in this particular one we are focusing on binomial choice.

(Refer Slide Time: 1:47)



Mobile Number Porting

So, I have a data set where we are majorly focusing on mobile number porting. So the data set is about mobile number porting. So we have to understand what is mobile number porting? So, this is something that has happened in India probably around 7-8 years back. So, when the mobile number porting facility was brought in, that means you can keep your same mobile number, but you can switch from one service provider to another service provider; keeping the mobile number same.

So the generic procedure of this mobile number porting is like this that you send a request to your service provider that you want to port, and the service provider will mandatorily have to give you a porting code, and with that porting code you have to go to the next, new service provider, give that porting code and within some days that particular number gets transferred. But there are certain problems in that, there are certain issues that creates that probably sometimes hinders you from going from one service provider to another service provider.

So, mobile number service providers, what they do is the moment you actually try to do porting, you have sent the request, more specifically if you are a very profitable customer or you are a probably a customer who is postpaid customer. So, they think that postpaid is more profitable and probably that is true as well. So, if you are a postpaid customer they will call you, and they will ask you what happened sir?

Why do you want to switch from service provider 1 to service provider somebody else? And what kind of facilities I can give you? Is there a service problem because we can improve the service problem and so on? And then, they will also say that, okay, “So, there are new-new offers that are coming up. You can get some discounts in the next billing and etcetera.” So they will give you lots of offerings to make sure that you stay back. So that is something that they do. But when you still say that, “Ok, I will not go into all these details and etcetera, they will send you a survey. Probably sometimes they send you a survey or sometimes through that vocal voice based conversation, they take a survey and ask you what is the service quality perception, price perception, blah, blah, blah and you give ratings on that. So, now this is something like exit interview in case of HR problems.

So, it is like when somebody is leaving the service provider, why he is leaving. So, when you get this data you can analyze this data to find out what are the things that customers who are switchers or customers who are non-switchers think that is important. So, what customers value when they are switching. In this context, I want to talk about 4 switching barriers. So this is a marketing concept.

Where we say that you can also set up certain kinds of switching barriers to make sure that your customer does not go away from you. So, what are the various switching barriers possible? One is economic switching barriers. Economic switching barriers is related to this, for example, in this same context if I talk about let us say mobile number porting, the moment you try to port from one mobile service provider to another service provider all the balance that you have in your sim card goes actually gets nullified, and you do not get those facilities.

It is that the same thing is applicable for other cases also, where let us say you a have 6 months or one year contract or something like that, and you are getting some benefits because of those contract, and now you are leaving the service provider and going to some other service provider, all of your contracts benefit whatever you used to get will go off. So, that is a very basic problem, which will probably sometimes stop you from switching. So which is called economic switching barrier.

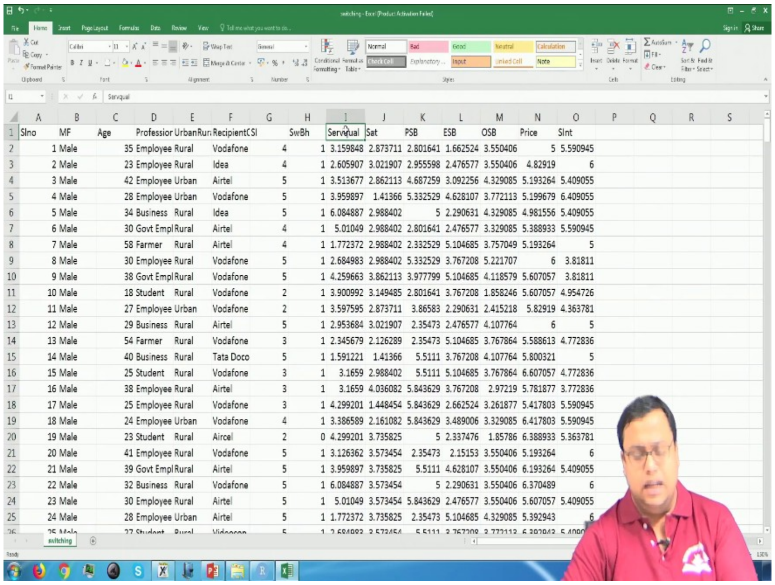
Now another is let us say social and psychological switching barrier. This happens in mobile number, mobile phones quite commonly. That I have all my people in the same network, which is let us say xyz. So, then why will I switch from xyz to somebody else? So that becomes a very major problem, and that is something called social and psychological. So, I am habituated with this mobile service provider, I do not want to switch from here to somebody else.

And then comes procedural switching barrier, procedural switching barrier is about how difficult is the procedure. For example let us say if you have to take lots of, I would say permission from here and there to switch, switch from one place to another place in terms of service provider. If there were mobile number code, then you go and give it to the new mobile number service provider, and then they give some code, and then you have to again go back to the old one and give that number to that old one.

And there is lots of paper works in between and blah blah blah. Then all of these things is a procedural problem. And if there is a huge procedural switching barrier you might not want to switch. So that is something sometimes you can create as a customer, that there is a procedural problem. So there are obviously the agencies which will take into account that how I can reduce this procedural problems so that customers get benefit. But often time marketers do that.

And then the fourth one is option related, so whether multiple options are available or not. So if you want to switch, options should be available. If this guy is the only possible option or number of options is less, then that creates a switching barrier, because people cannot switch.

(Refer Slide Time: 7:40)



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Slno | MF | Age | Professior | UrbanRur | Recipient | CSI | SwBh | Servqual | Sat | PSB | ESB | OSB | Price | SInt |
| 2 | 1 | Male | 35 | Employee | Rural | Vodafone | 4 | 1 | 3.159848 | 2.873711 | 2.801641 | 1.662524 | 3.550406 | 5 | 5.590945 |
| 3 | 2 | Male | 23 | Employee | Rural | Idea | 4 | 1 | 2.605907 | 3.021907 | 2.955598 | 2.476577 | 3.550406 | 4.82919 | 6 |
| 4 | 3 | Male | 42 | Employee | Urban | Airtel | 5 | 1 | 3.513677 | 2.862113 | 4.687259 | 3.092256 | 4.329085 | 5.193264 | 5.409055 |
| 5 | 4 | Male | 28 | Employee | Urban | Vodafone | 5 | 1 | 3.959897 | 1.41366 | 5.332529 | 4.628107 | 3.772113 | 5.199679 | 6.409055 |
| 6 | 5 | Male | 34 | Business | Rural | Idea | 5 | 1 | 6.084887 | 2.988402 | 5 | 2.290631 | 4.329085 | 4.981556 | 5.409055 |
| 7 | 6 | Male | 30 | Govt Empl | Rural | Airtel | 4 | 1 | 5.01049 | 2.988402 | 2.801641 | 2.476577 | 3.329085 | 5.388933 | 5.590945 |
| 8 | 7 | Male | 58 | Farmer | Rural | Airtel | 4 | 1 | 1.772372 | 2.988402 | 2.332529 | 5.104685 | 3.757049 | 5.193264 | 5 |
| 9 | 8 | Male | 30 | Employee | Rural | Vodafone | 5 | 1 | 2.684983 | 2.988402 | 5.332529 | 3.767208 | 5.221707 | 6 | 3.81811 |
| 10 | 9 | Male | 38 | Govt Empl | Rural | Vodafone | 5 | 1 | 4.259663 | 3.862113 | 3.977799 | 5.104685 | 4.118579 | 5.607057 | 3.81811 |
| 11 | 10 | Male | 18 | Student | Rural | Vodafone | 2 | 1 | 3.900992 | 3.149485 | 2.801641 | 3.767208 | 1.858246 | 5.607057 | 4.954726 |
| 12 | 11 | Male | 27 | Employee | Urban | Vodafone | 2 | 1 | 3.597595 | 2.873711 | 3.86583 | 2.290631 | 2.415218 | 5.82919 | 4.363781 |
| 13 | 12 | Male | 29 | Business | Rural | Airtel | 5 | 1 | 2.953684 | 3.021907 | 2.35473 | 2.476577 | 4.107764 | 6 | 5 |
| 14 | 13 | Male | 54 | Farmer | Rural | Vodafone | 3 | 1 | 2.345679 | 2.126289 | 2.35473 | 5.104685 | 3.767864 | 5.588613 | 4.772836 |
| 15 | 14 | Male | 40 | Business | Rural | Tata Doco | 5 | 1 | 1.591221 | 1.41366 | 5.5111 | 3.767208 | 4.107764 | 5.800321 | 5 |
| 16 | 15 | Male | 25 | Student | Rural | Vodafone | 3 | 1 | 3.1659 | 2.988402 | 5.5111 | 5.104685 | 3.767864 | 6.607057 | 4.772836 |
| 17 | 16 | Male | 38 | Employee | Rural | Airtel | 3 | 1 | 3.1659 | 4.036082 | 5.843629 | 3.767208 | 2.97219 | 5.781877 | 3.772836 |
| 18 | 17 | Male | 25 | Employee | Rural | Vodafone | 3 | 1 | 4.299201 | 1.448454 | 5.843629 | 2.662524 | 3.261877 | 5.417803 | 5.590945 |
| 19 | 18 | Male | 24 | Employee | Urban | Vodafone | 4 | 1 | 3.386589 | 2.161082 | 5.843629 | 3.489006 | 3.329085 | 6.417803 | 5.590945 |
| 20 | 19 | Male | 23 | Student | Rural | Aircel | 2 | 0 | 4.299201 | 3.735825 | 5 | 2.337476 | 1.85786 | 6.388933 | 5.363781 |
| 21 | 20 | Male | 41 | Employee | Rural | Vodafone | 5 | 1 | 3.126362 | 3.573454 | 2.35473 | 2.15153 | 3.550406 | 5.193264 | 6 |
| 22 | 21 | Male | 39 | Govt Empl | Rural | Airtel | 5 | 1 | 3.959897 | 3.735825 | 5.5111 | 4.628107 | 3.550406 | 6.193264 | 5.409055 |
| 23 | 22 | Male | 32 | Business | Rural | Vodafone | 5 | 1 | 6.084887 | 3.573454 | 5 | 2.290631 | 3.550406 | 6.370489 | 6 |
| 24 | 23 | Male | 30 | Employee | Rural | Airtel | 5 | 1 | 5.01049 | 3.573454 | 5.843629 | 2.476577 | 3.550406 | 5.607057 | 5.409055 |
| 25 | 24 | Male | 28 | Employee | Urban | Airtel | 5 | 1 | 1.772372 | 3.735825 | 2.35473 | 5.104685 | 4.329085 | 5.392943 | 6 |
| 26 | 25 | Male | 27 | Student | Rural | Videocon | 5 | 1 | 2.684983 | 3.573454 | 5.5111 | 3.767208 | 3.772113 | 6.392943 | 5.4090 |

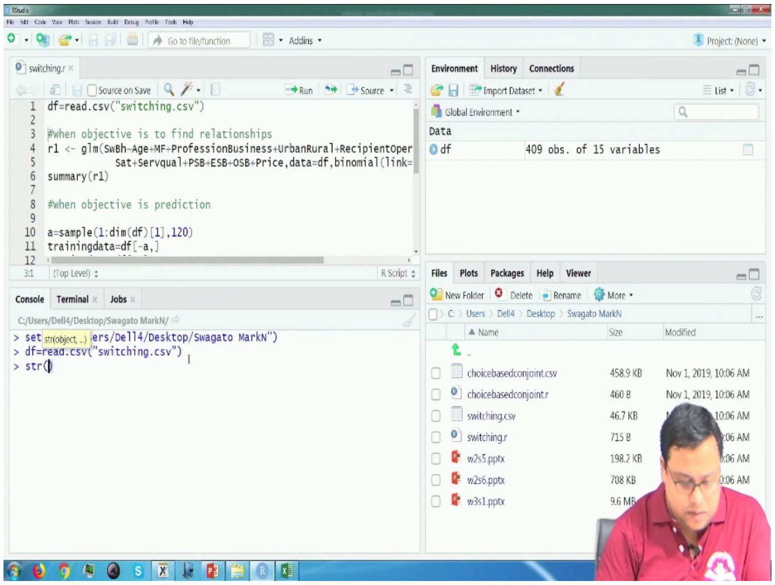
So given that as a background, I have a data set which is called switching.csv file which looks like this and this is a switching data set, where I have let us say the gender of the person, I have the age of the person who wanted to switch, these are all the people who actually wanted to switch, they expressed their intention to switch and then their profession, their whether they are urban or rural, what is the recipient. So from when they switch, what they did, so we also ask that which service provider you want to go to. So what is the recipient service provider?

And then ultimately these are certain attribute wise waiting. So, I will focus on this SInt is actually switching intention, and these are other things like service quality. The first one is servqual, that means service quality. Next one is satisfaction and then 3 switching barriers PSB, ESB, and OSB. PSB is procedural, ESB is economical, and OSB is operational. So operational will include both option related and social psychological. So, together, these and these together we have created the OSB. So that is there, and then that is a price component, price means price perception. So, if whether the price is good or bad in comparison to the other service providers.

So, we think that all of these values will impact your switching behavior, and that might change where the effect of this from column number I to column number N or column number H, might differ if you are an employee or rural or a, based on recipients and etcetera. So this is something that we will try to study that what matters, whether the switching barriers matter, if the switching barriers matter, then which switching barrier will matter, and then whether satisfaction matters or service quality matters or price matters, what is people actually value.

So, here you see the difference between the previous examples, and this example is your y variable, which is the SwBh switching behaviour, that means ultimately you switched or not, is a 1 0 variable. 1 means you switched and 0 means you did not switch or something like that. Actually, 1 means you stayed back and zero or so we will see that.

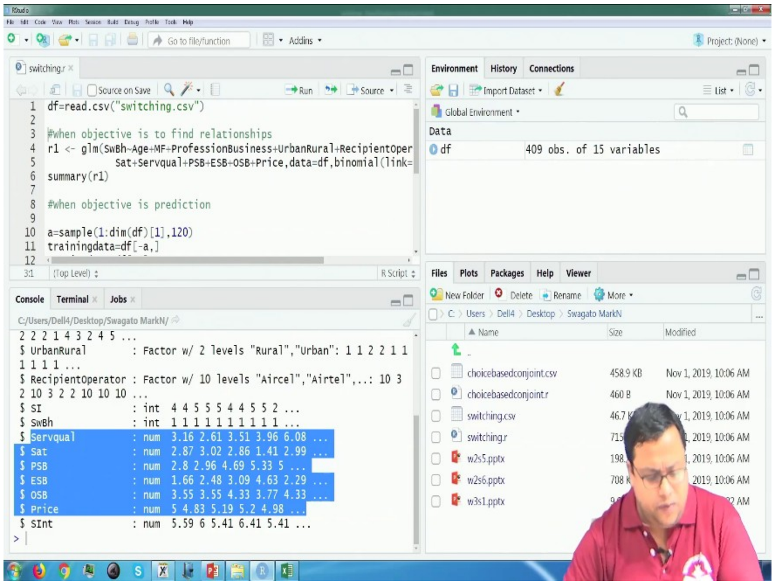
(Refer Slide Time: 10:03)



```
df=read.csv("switching.csv")

#when objective is to find relationships
r1 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOper
          Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link=
summary(r1)

#when objective is prediction

a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
```

```
df    409 obs. of 15 variables
```

```
C:/Users/Dell4/Desktop/Swagato MarkN/
> set str(object, ...) ers/Dell4/Desktop/Swagato MarkN")
> df=read.csv("switching.csv")
> str()
```

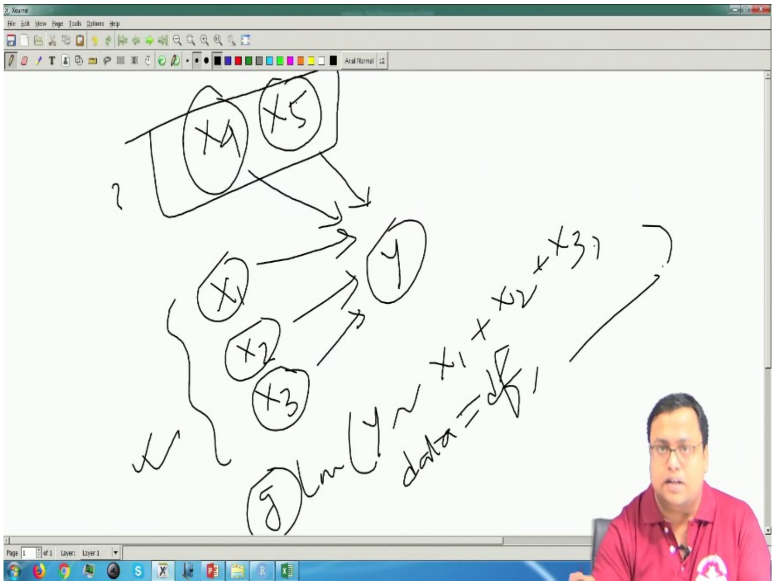


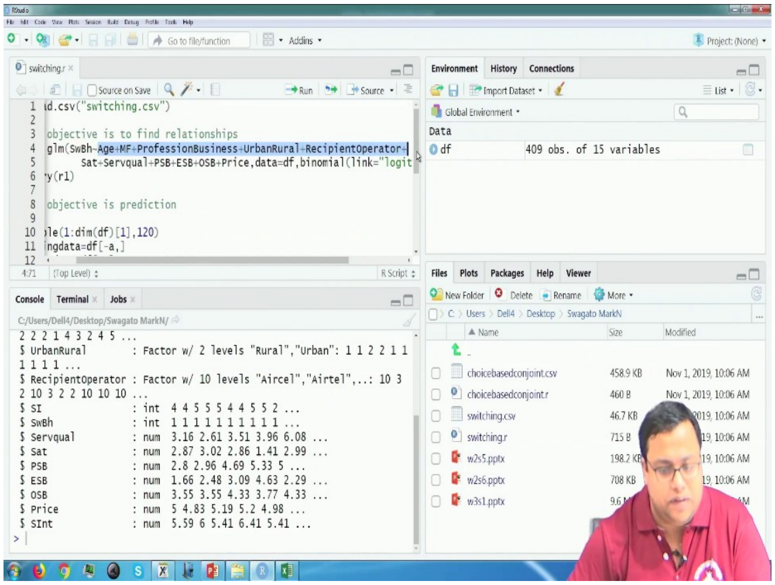
```
df=read.csv("switching.csv")

#when objective is to find relationships
r1 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOper
          Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link=
summary(r1)

#when objective is prediction

a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
```

```
C:/Users/Dell4/Desktop/Swagato MarkN/
2 2 2 1 4 3 2 4 5 ...
$ UrbanRural        : Factor w/ 2 levels "Rural","Urban": 1 1 2 2 1 1
1 1 1 1 ...
$ RecipientOperator : Factor w/ 10 levels "Aircel","Airtel",..: 10 3
2 10 3 2 2 10 10 10 ...
$ SI                : int  4 4 5 5 5 4 4 5 5 2 ...
$ SwBh              : int  1 1 1 1 1 1 1 1 1 1 ...
$ Servqual          : num  3.16 2.61 3.51 3.96 6.08 ...
$ Sat               : num  2.87 3.02 2.86 1.41 2.99 ...
$ PSB               : num  2.8 2.96 4.69 5.33 5 ...
$ ESB               : num  1.66 2.48 3.09 4.63 2.29 ...
$ OSB               : num  3.55 3.55 4.33 3.77 4.33 ...
$ Price             : num  5 4.83 5.19 5.2 4.98 ...
$ SInt              : num  5.59 6 5.41 6.41 5.41 ...
>
```

So, to do that what I will do is there is a switching behavior.r file, I will open that as usual and then set working directory to source file location and then read the data. So, my data set is more or less simplified. So, strdf will give me that, okay my SwBh is an integer function, so this is my y variable. And servqual to price is my x variable, all are numeric, so I do not have any problem right now. And I will add certain covariates as well. Covariates means certain control variables. The variables which I will not focusing be on them, I will not focus on their behaviour. So covariates is like this.

(Refer Slide Time: 10:49)





```
d.csv("switching.csv")

objective is to find relationships
glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
       Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link="logit
y(r1)

objective is prediction

le(1:dim(df)[1],120)
ngdata=df[-a,]
```

4:71 (Top Level) R Script

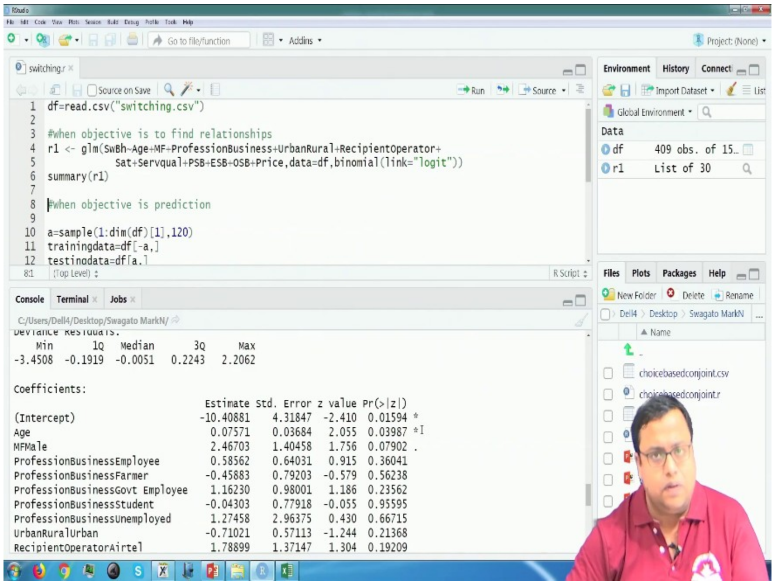
Data
df 409 obs. of 15 variables

C:/Users/Dell4/Desktop/Swagato MarkN/

```
2 2 2 1 4 3 2 4 5 ...
$ UrbanRural        : Factor w/ 2 levels "Rural","Urban": 1 1 2 2 1 1
1 1 1 1 ...
$ RecipientOperator : Factor w/ 10 levels "Aircel","Airtel",..: 10 3
2 10 3 2 2 10 10 10 ...
$ SI                : int  4 4 5 5 5 4 4 5 5 2 ...
$ SwBh              : int  1 1 1 1 1 1 1 1 1 1 ...
$ Servqual          : num  3.16 2.61 3.51 3.96 6.08 ...
$ Sat               : num  2.87 3.02 2.86 1.41 2.99 ...
$ PSB               : num  2.8 2.96 4.69 5.33 5 ...
$ ESB               : num  1.66 2.48 3.09 4.63 2.29 ...
$ OSB               : num  3.55 3.55 4.33 3.77 4.33 ...
$ Price             : num  5 4.83 5.19 5.2 4.98 ...
$ SInt              : num  5.59 6 5.41 6.41 5.41 ...
> 
```

C > Users > Dell4 > Desktop > Swagato MarkN

| Name | Size | Modified |
|---|---|---|
| .. | | |
| choicebasedconjoint.csv | 458.9 KB | Nov 1, 2019, 10:06 AM |
| choicebasedconjoint.r | 460 B | Nov 1, 2019, 10:06 AM |
| switching.csv | 46.7 KB | 2019, 10:06 AM |
| switching.r | 715 B | 19, 10:06 AM |
| w2s5.pptx | 198.2 KB | 19, 10:06 AM |
| w2s6.pptx | 708 KB | 19, 10:06 AM |
| w3s1.pptx | 9.6 | M |



```
df=read.csv("switching.csv")

#when objective is to find relationships
r1 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link="logit"))
summary(r1)

#when objective is prediction

a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]
```

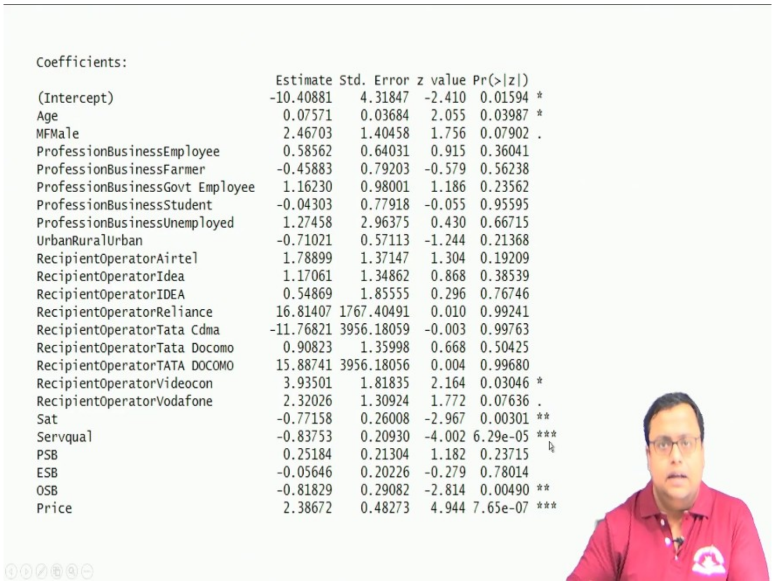
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4508  -0.1919  -0.0051   0.2243   2.2062

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -10.40881    4.31847  -2.410  0.01594 *
Age                            0.07571    0.03684   2.055  0.03987 *
MFMale                         2.46703    1.40458   1.756  0.07902 .
ProfessionBusinessEmployee     0.58562    0.64031   0.915  0.36041
ProfessionBusinessFarmer      -0.45883    0.79203  -0.579  0.56238
ProfessionBusinessGovt Employee  1.16230    0.98001   1.186  0.23562
ProfessionBusinessStudent     -0.04303    0.77918  -0.055  0.95595
ProfessionBusinessUnemployed   1.27458    2.96375   0.430  0.66715
UrbanRuralUrban               -0.71021    0.57113  -1.244  0.21368
RecipientOperatorAirtel        1.78899    1.37147   1.304  0.19209
```

So, let us say I am trying to find out the impact of X1, X2, X3 on y. But there is another guy, so let us say X4 and X5. Historically they were there, they have impact, people have said that this guy also has impact, but I am not focusing on them. My research is not focusing on them, but they should also be there because historically, they have been considered as a predictor of y. So, X4 and X5 should also be there, I just take them as covariates. So, I just put them in the regression. So these are covariates, I do not focus on them much, I focus on these guys much, but I also take the other two guys in the model.

So, here you will see that I have to write glm, glm stands for Generalized Linear Model. And here, family is equal to binomial logit. So, I am doing a logistic regression. The only thing changes from linear regression to logistic regression is instead of lm you write glm and here one more parameter you add, that is all. The other things is same, so whether you run logistic regression or linear regression, it is lm y tilda, X1 plus X2 plus X3 comma data is equal to the data set name. So, in this case df and then if there is, it is a binomial regression, then if this becomes g and then you add another thing here, that is all. So otherwise it is similar to linear regression, the code is similar.

So, I just run that and here age, gender, MF is for gender, profession and urban, rural and recipient operator, all of these things is my basically the predictor. So, here my objective is to find relationship. So, there two different objectives when we deal with 1 0 variables, I can do two different stuff, and we have to focus on that as well. So, I can find relationships, or I can do predictions. There are two jobs that I can do. So, when I am trying to find out relationships between X1, X2, X3 and y; I will include the whole data, and I will try to find out the relationships, so that is what I am doing. I am using the whole data, data is equal to df and I run this, and I get certain results. So, the result is given here. So, I will once more just run result.

(Refer Slide Time: 13:32)



```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -10.40881    4.31847  -2.410  0.01594 *
Age                               0.07571    0.03684   2.055  0.03987 *
MFMale                            2.46703    1.40458   1.756  0.07902 .
ProfessionBusinessEmployee        0.58562    0.64031   0.915  0.36041
ProfessionBusinessFarmer         -0.45883    0.79203  -0.579  0.56238
ProfessionBusinessGovt Employee   1.16230    0.98001   1.186  0.23562
ProfessionBusinessStudent        -0.04303    0.77918  -0.055  0.95595
ProfessionBusinessUnemployed      1.27458    2.96375   0.430  0.66715
UrbanRuralUrban                  -0.71021    0.57113  -1.244  0.21368
RecipientOperatorAirtel           1.78899    1.37147   1.304  0.19209
RecipientOperatorIdea             1.17061    1.34862   0.868  0.38539
RecipientOperatorIDEA             0.54869    1.85555   0.296  0.76746
RecipientOperatorReliance        16.81407 1767.40491   0.010  0.99241
RecipientOperatorTata Cdma      -11.76821 3956.18059  -0.003  0.99763
RecipientOperatorTata Docomo      0.90823    1.35998   0.668  0.50425
RecipientOperatorTATA DOCOMO     15.88741 3956.18056   0.004  0.99680
RecipientOperatorVideocon         3.93501    1.81835   2.164  0.03046 *
RecipientOperatorVodafone         2.32026    1.30924   1.772  0.07636 .
Sat                              -0.77158    0.26008  -2.967  0.00301 **
Servqual                         -0.83753    0.20930  -4.002 6.29e-05 ***
PSB                               0.25184    0.21304   1.182  0.23715
ESB                              -0.05646    0.20226  -0.279  0.78014
OSB                              -0.81829    0.29082  -2.814  0.00490 **
Price                             2.38672    0.48273   4.944 7.65e-07 ***
```

The result is given here. So. the same result I just copied and pasted it here and I want you to understand what this means. So, in this particular thing I am seeing that see, age is significant. Age is significant means that as age is significant in positive, that means as age increases, your chances of staying back increases. So, 1 means staying back here. I will just see first, one minute. So, 1 means, no 1 means actually leaving here, in this particular case that we can see that as age increases, people are more prone towards leaving and changing. So, which is which we have to explain theoretically, we will see how we can explain.

Now what I can see in terms of satisfaction, as satisfaction increases the switching probability decreases, -0.77 obvious. As service quality increases, switching probability also decreases, so this is also obvious. But in comparison to satisfaction, service quality has more impact and then another thing out of this procedural, economic and OSB, the most important thing that comes up is OSB that is operational. If you remember, operational is social psychological plus options. So operational switching barrier will reduce the switching probability.

So, if there are lots of switching barrier, you will not switch. So that something that is only significant, the other two PSB and ESB are not coming significant. So, I mean to, that means to say that if you want to make sure that the people as people should stay back in your service providers, case and remember the data set only those people who have shown interest to switch. So, they have shown the interest, they all have switching intention. But ultimately they did not switch or switch is something that we are focusing on.

So, all of these guys have switching intention means they have already thought about procedural problems, they will already think about economic problems; after considering that assuming that they are informed customers, after considering that they have shown interest or expressed their intention to switch by asking for the porting number. Now, after that, if you want to stop them, you cannot give economic benefit. Because they already know that you will give economic benefit.

And then you will not create more procedural problems because it is something that is governed by an agency, a government agency who will actually come and sue you in that case. But the only thing that matters in this particular context for a customer, for a company is that you have to create operational switching barrier which is related to the psychology of the person, the social psychology, the social and psychological issues and options. So you have to make sure that your option is better than others or the availability of options is something that you can create.

Now, sometimes availability of options is not also in your hand, then what is IN your hand, you have to make sure the network of that particular person is also in your network actually. The human beings who are connected with this particular person are also in your network. That means his family and friends, and there are his business colleagues and etcetera. If they are also in your network, then the chance that this guy will switch from your network to somebody else's network, some other service provider is very low.

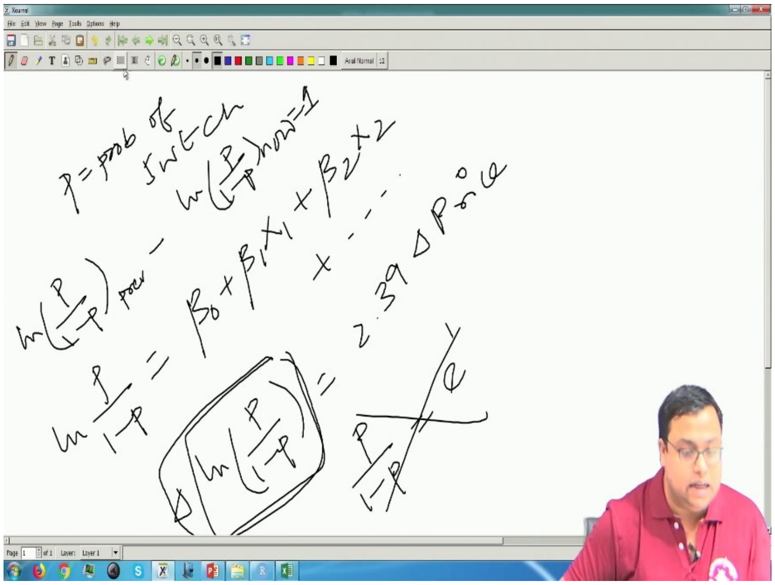
So that is something that comes up here very strongly that is what matters other than procedural switching barrier or economic switching barrier. Now price is coming positive, this is actually the higher the price, so as the price perception increases. That means I think that your price is much higher than others, then I will switch. Now, you see that is 2.38, which is much-much higher than many other factors. So that means one of the most important factor is price.

So, obviously service quality, satisfaction etcetera, etcetera, is related obviously. But persons who are trying to switch, who have shown interest in switching, price is also a very important component. So mobile number porting does not only happen at some point when used to happen. It is not only happened because of your quality of the service and etcetera, because we generally now know that most of the guys will have same level of service.

Somewhere it is better somewhere it is worse and so on. But price is something that is very important. So you have to play the price. Now this might be a different situation in a different context, probably in some other service situation price might not be an issue that much. But in this case we are finding that. This is what we are trying to find. Now, you have to also understand that this equation that you are getting and all these guys which you are getting here which are insignificant you can consider all of these things to be zero.

So, you can just drop profession from your model or probably the recipient operator, urban-rural, all of these things you can drop from your model. But you have to also understand what they mean. So, for example here, this recipient operator is coming Videocon is coming significant. So, that has some meaning. You have to actually understand that. So, I will give you a little bit of a theoretical understanding of what this result is saying to me.

(Refer Slide Time: 18:59)



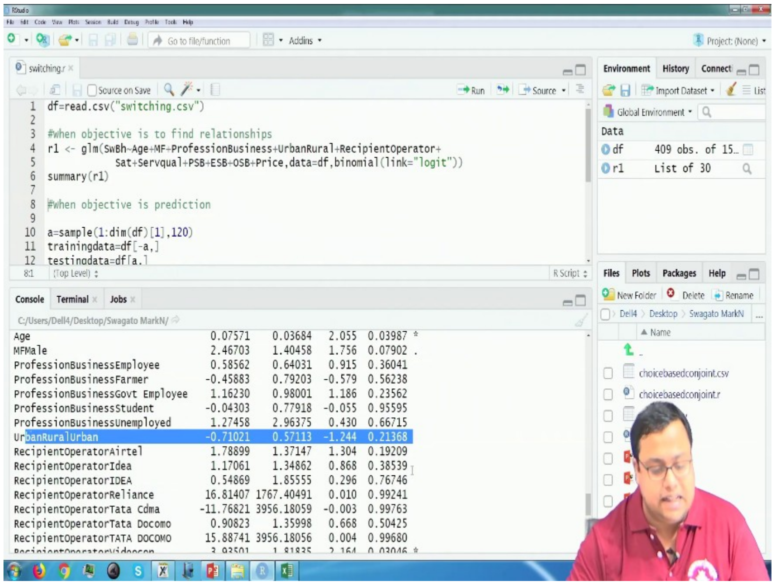
p = prob of switch
$\ln\left(\frac{p}{1-p}\right)_{\text{new}} = 1$
$\ln\left(\frac{p}{1-p}\right)_{\text{prev}} -$
$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$
$\boxed{\Delta\left[\ln\left(\frac{p}{1-p}\right)\right]} = 2.39\,\Delta\,\text{Price}$
$\frac{p}{1-p}$
y
e

So, this result is actually trying to predict where p is probability of switching, switching let us say. Then it is trying to predict basically $\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X1 + \beta_2 X2$ and so on. So, that is logistic regression. If you have an idea about that. So, that is something that I am trying to predict. So, that means that keeping everything else same, if I increase let us say price by 1 unit. Then what will change the probability of $\Delta \ln(\frac{p}{1-p})$, this will change by 2.3; if I am not wrong 2.38 or 3 9 into delta price.

So, if price increases by 1 unit, this is the guy who increases by one unit. And then the probability how much it increases you have to calculate that. If this increases by 1 unit then p/1-p increases by e^1 unit then what is the formula for p, how much p increases? What was the p before and what was the p now is something that you have to understand. So, this is not actually correct.

So this is like $\Delta \ln(\frac{p}{1-p})$ is something that is e not exactly this. So we have to check that. So, $\Delta \ln(\frac{p}{1-p})_{prev} - \Delta \ln\left(\frac{p}{1-p}\right)_{now} = 1$, and you have to solve, what is the value is coming up. So, this is something that we have to take into account. So, this is not a linear regression, this is a logistic regression and the formula is like that. The formula is in a logistic function. So that we will have to consider.

(Refer Slide Time: 21:25)

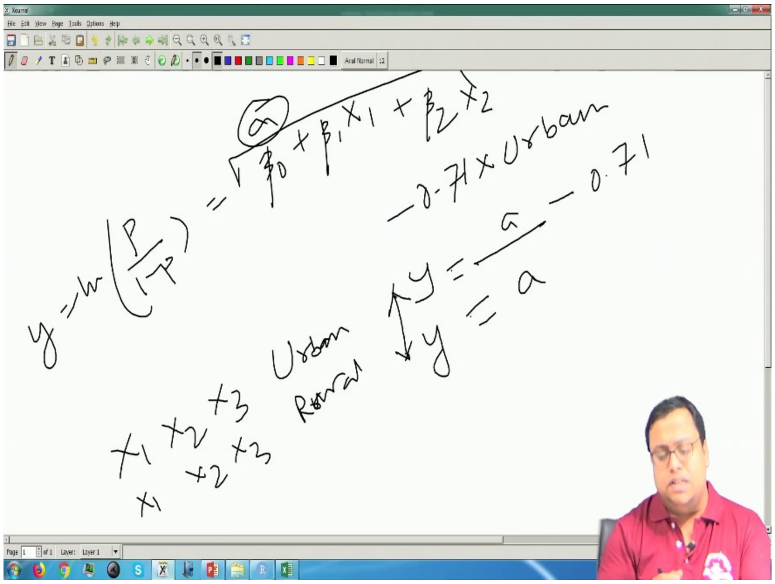


```
df=read.csv("switching.csv")

#when objective is to find relationships
r1 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
          Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link="logit"))
summary(r1)

#when objective is prediction

a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]
```

```
Age                              0.07571    0.03684   2.055  0.03987 *
MFMale                           2.46703    1.40458   1.756  0.07902 .
ProfessionBusinessEmployee       0.58562    0.64031   0.915  0.36041
ProfessionBusinessFarmer        -0.45883    0.79203  -0.579  0.56238
ProfessionBusinessGovt Employee  1.16230    0.98001   1.186  0.23562
ProfessionBusinessStudent       -0.04303    0.77918  -0.055  0.95595
ProfessionBusinessUnemployed     1.27458    2.96375   0.430  0.66715
UrbanRuralUrban                 -0.71021    0.57113  -1.244  0.21368
RecipientOperatorAirtel          1.78899    1.37147   1.304  0.19209
RecipientOperatorIdea            1.17061    1.34862   0.868  0.38539
RecipientOperatorIDEA            0.54869    1.85555   0.296  0.76746
RecipientOperatorReliance       16.81407 1767.40491   0.010  0.99241
RecipientOperatorTata Cdma     -11.76821 3956.18059  -0.003  0.99763
RecipientOperatorTata Docomo     0.90823    1.35998   0.668  0.50425
RecipientOperatorTATA DOCOMO    15.88741 3956.18056   0.004  0.99680
RecipientOperatorVideocon        3.93501    1.81835   2.164  0.03046 *
```

Now, we have to also consider what is the dummy variable? So, what this categorical variable mean? So, by chance if let us say, if urban rural is coming insignificant here, by chance if that was significant, what would have been the meaning?

(Refer Slide Time: 21:36)

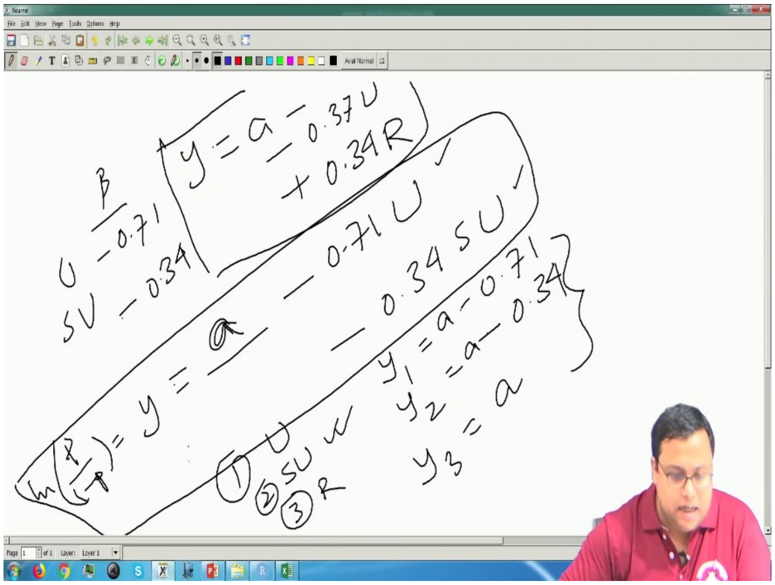


The meaning is that whatever is my y, y = $\ln(\frac{p}{1-p})$, whatever is that, that is actually something ( $\beta_0 + \beta_1 X1 + \beta_2 X2$ - 0.71)* urban. I am just assuming that all other things are remaining constant.

That means if all other things remain constant, that means that let us say X1, X2, X3 everything is same and then there is a urban guy, versus X1, X2, X3 everything same then there is a rural guy.

So for the first guy, the y will be some value a, let us say a is the summation of all of these things a- 0.71 and the second guy y will be a, because he is not urban, this part will not happen. So, the difference between the urban and rural is 0.71 and, rural, urban guy is here less proponent towards switching. So, their propensity towards switching is less than the rural guy. So, that is something that we are getting here. If by chance that minus 0.71 was significant.

(Refer Slide Time: 23:04)



Similarly, if there are more than one categories here in this particular case, let us say there is urban, semi-urban and rural. By chance let say in the result I have got urban, semi-urban and rural. And urban, so rural was the reference point which got removed and urban and semi-urban both of them are significant and this value was 0.71 and this value was 0.34. So, the model would have come urban and semi-urban.

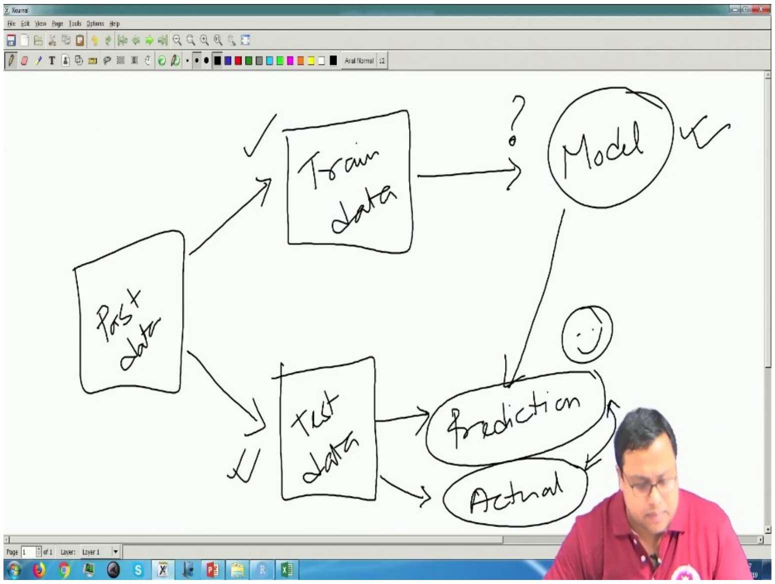
So, I would say then that y is equal to basically keeping everything else same, (a-0.71)* urban - 0.34*( semi urban), that is what you would have got in the regression equation. So y = $\ln(\frac{p}{1-p})$, and what is a? a = $\beta_0 + \beta_1 X1 + \beta_2 X2$ and so on. Now, if I have three persons, let us say one person number 1, person number 2, person number 3, person number 1, everything is same but it is urban guy, person number 2 is semi-urban guy, person number 3 is rural guy.

So, y1 will be a-0.71, y2 will be a-0.34 and y3 will be just 'a', because this guy is neither this is one nor this is one, neither this is one nor this is one for y3. So, if that is the case then I can say that, okay, Y2 who is a semi-urban guy is 0.34 less propensity of going for switching. So, log observed 0.34 less and in first case log observed 0.71 less. By chance if I have taken SU as you as the mid as the factor, SU as the reference point; then the whole equation would have changed in a different way.

Here, in this particular equation, rural was the factor which was the reference point. By chance in the regression if you have got or you have to know that how a guy is more, has more propensity of switching in comparison to the semi urban guys. So, semi-urban guys because few reference points. In that case the function will be y is equal to a minus, now remember y3 is 0.34 less than SU.

So, this will be just +0.34 into rural and what is the difference between SU and U that means semi-urban and urban. This guy is -0.71 this one is -0.34. So another 0.37 less. So, -0.37*urban, so then if you just consider that then the difference between the urban and semi-urban and semi-urban and rural is still as it is, as it was in this case. So, this is something that you have to understand that I can find out, for any variable I can find out, I would say the corresponding levels, how they are different.

(Refer Slide Time: 26:15)



Now, this is for explaining purpose. Now, I can do predictions as well. So, predictions is something that we do, also in this particular modelling. Prediction is something like how you go and talk with a fortune teller. Let us say any kind of fortune teller. If you go in, go to a fortune teller and you want to know that whether the fortune teller is good enough or not, you try to say that, “Okay, why do not you actually, so I will give you some information and why do not you actually tell me about my few past, which has already happened.

I will not disclose the past to you. And you should tell me a little bit about my past.” If you can explain it properly, then I will say that, “Okay, you are good enough.” If you cannot explain it properly, then I would say that, “Okay, I do not think you are good enough and I should go ahead and do something else.” So that is something that I would also want to do in this particular case. So, let us say, so fortune teller what do we do, so you have a past data.

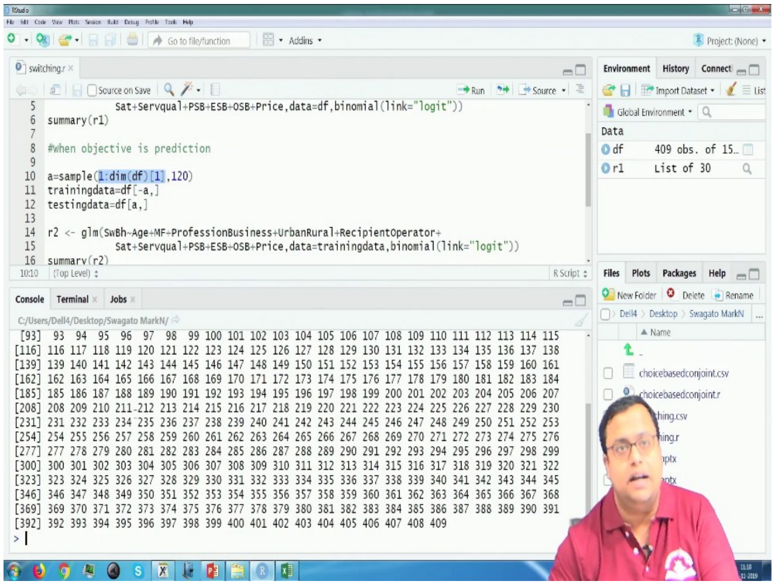
You form that past data, you break it; you say that this part of the data is something that I will give to fortune teller. So, fortune teller has his own model, every fortune teller will have his own model, own way of creating a model. So, this is some part of the data and this is that data that you do not disclose. So, this is called training data and this is called test data. This part is something that we do not disclose. This we do not disclose and this you disclose and then based on this training data he creates the model, some model.

Now in that model, then he comes up with the model and applies in the testing data a little bit and create a prediction. So, let us say he asks your date of birth, age of birth sometimes some people actually ask you to see in a cup or some people ask you to pick up a card, all of that it is training data, that is something based on which you will create the model and based on that creating the models, he gives some prediction.

So, he sees you, your behavior and etcetera and he also creates a model, together he gives the prediction. So, that prediction, you actually go and check with the actual values and if the actual value and prediction value are close to each other, then you are happy, if it is not so close to each other then you are unhappy. So, how close it is based on that you decide that which fortune teller is better, which fortune teller is worse.

So, here how he has created the model using what kind of methodology we generally do not know. So that is why we generally do not focus. In this particular case, in our case the only thing that it is different from there fortune teller’s case and our case is probably we have to say that how this model is created. We have to actually talk about that also. So, I will do something like that here in my data set.

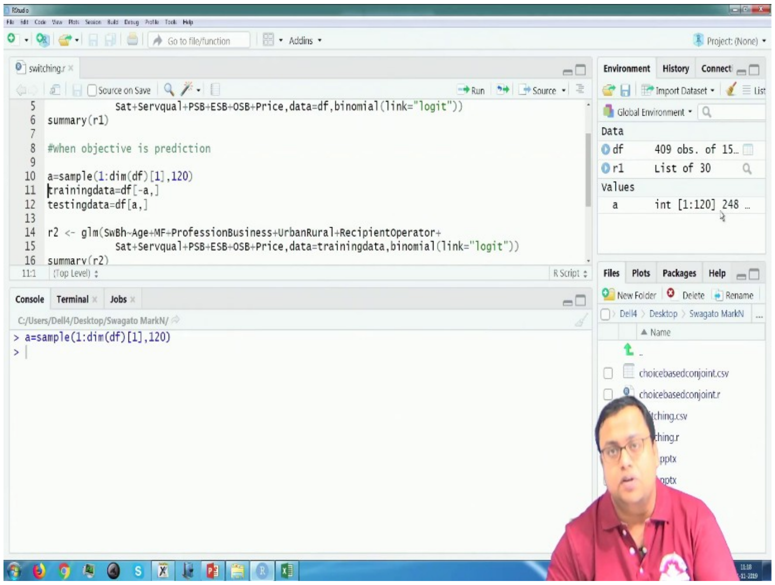
(Refer Slide Time: 29:21)



```
             Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link="logit"))
summary(r1)

#when objective is prediction

a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]

r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
             Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)
```

```
C:/Users/Dell4/Desktop/Swagato MarkN/
 [93]  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115
[116] 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
[139] 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161
[162] 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184
[185] 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207
[208] 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230
[231] 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253
[254] 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276
[277] 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299
[300] 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322
[323] 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345
[346] 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368
[369] 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391
[392] 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409
>
```

Data
df 409 obs. of 15...
r1 List of 30

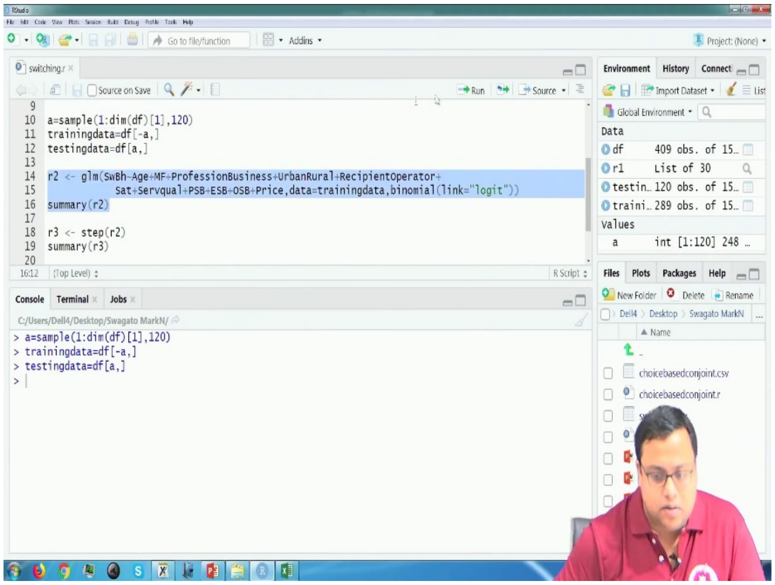


```
             Sat+Servqual+PSB+ESB+OSB+Price,data=df,binomial(link="logit"))
summary(r1)

#when objective is prediction

a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]

r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
             Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)
```

```
C:/Users/Dell4/Desktop/Swagato MarkN/
> a=sample(1:dim(df)[1],120)
>
```

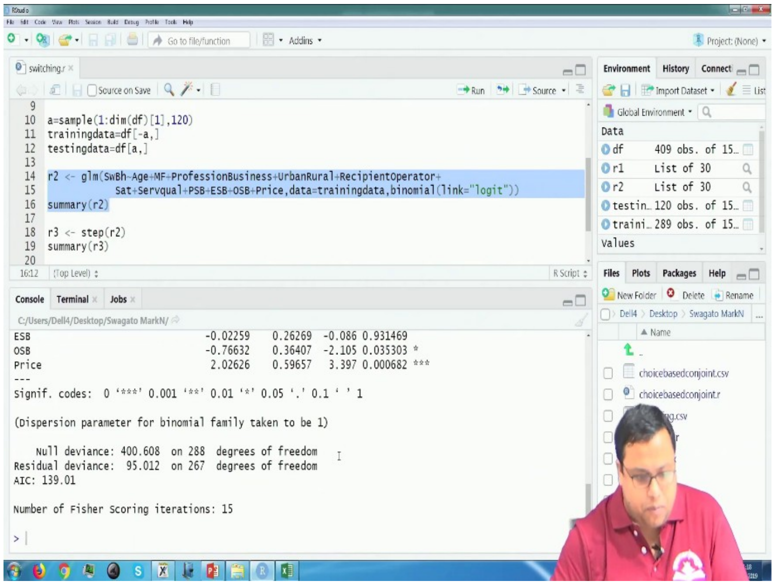
Data
df 409 obs. of 15...
r1 List of 30

Values
a int [1:120] 248 ...



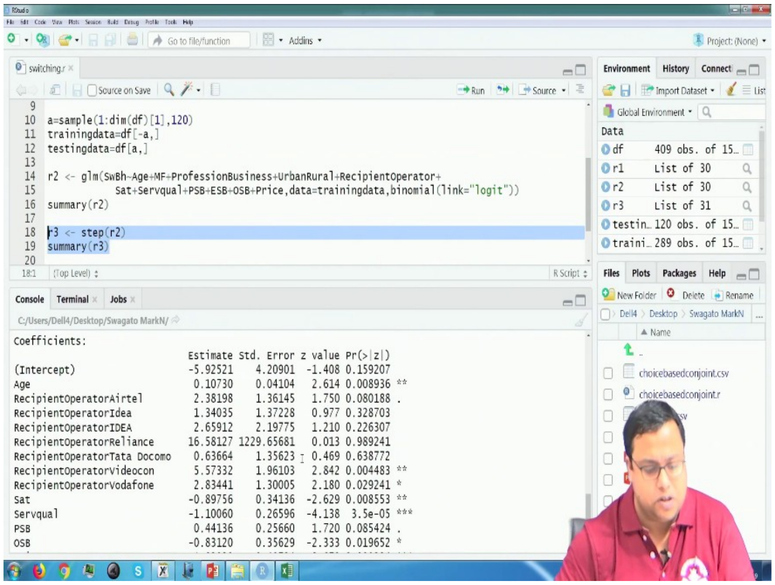
```
a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]

r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)

r3 <- step(r2)
summary(r3)
```

```
> a=sample(1:dim(df)[1],120)
> trainingdata=df[-a,]
> testingdata=df[a,]
> |
```



```
a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]

r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)

r3 <- step(r2)
summary(r3)
```

```
ESB                              -0.02259    0.26269  -0.086 0.931469
OSB                              -0.76632    0.36407  -2.105 0.035303 *
Price                             2.02626    0.59657   3.397 0.000682 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 400.608  on 288  degrees of freedom
Residual deviance:  95.012  on 267  degrees of freedom
AIC: 139.01

Number of Fisher Scoring iterations: 15

> |
```



```
a=sample(1:dim(df)[1],120)
trainingdata=df[-a,]
testingdata=df[a,]

r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)

r3 <- step(r2)
summary(r3)
```

```
Data
df        409 obs. of 15...
r1        List of 30
r2        List of 30
r3        List of 31
testin..  120 obs. of 15...
traini..  289 obs. of 15...
```

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -5.92521    4.20901  -1.408 0.159207
Age                             0.10730    0.04104   2.614 0.008936 **
RecipientOperatorAirtel         2.38198    1.36145   1.750 0.080188 .
RecipientOperatorIdea           1.34035    1.37228   0.977 0.328703
RecipientOperatorIDEA           2.65912    2.19775   1.210 0.226307
RecipientOperatorReliance      16.58127 1229.65681   0.013 0.989241
RecipientOperatorTata Docomo    0.63664    1.35623   0.469 0.638772
RecipientOperatorVideocon       5.57332    1.96103   2.842 0.004483 **
RecipientOperatorVodafone       2.83441    1.30005   2.180 0.029241 *
Sat                            -0.89756    0.34136  -2.629 0.008553 **
Servqual                       -1.10060    0.26596  -4.138  3.5e-05 ***
PSB                             0.44136    0.25660   1.720 0.085424 .
OSB                            -0.83120    0.35629  -2.333 0.019652 *
```

So, objective is prediction. So, what I create is the first thing that I create is the training data and testing data. So, I create a sample of  dim(df)[1], what is dim(df)[1] just check. Dim of df is actually dimension of df, dimension of my data set. dim(df)[1] is nothing but the number of rows. Because see there are 400 and 15 observations. So, that is coming in the dimension of df. The first entry of that is just the number of rows.

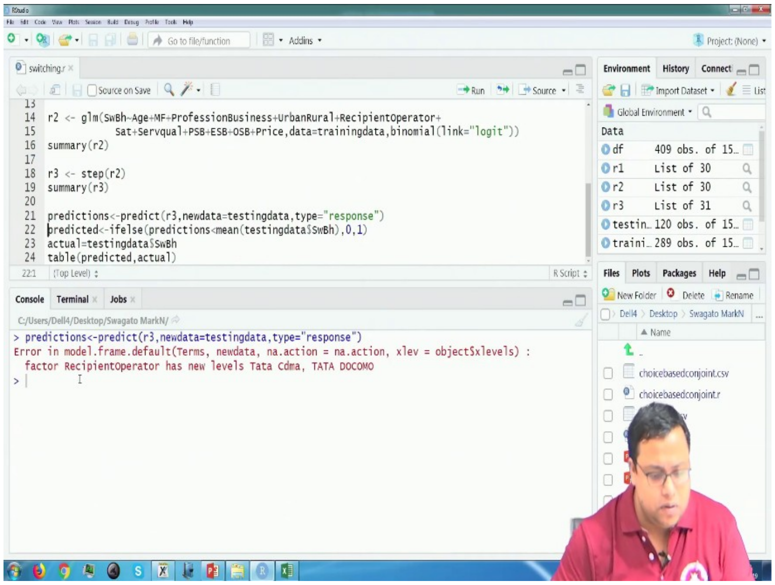
So, out of 400 rows, so out of 1 to 409, so if I just write this, this will come like this 1 to 409. Out of 1 to 409 I am choosing 120 guys, randomly I am choosing. So, I am choosing this, run, and I am getting here 120 observations which are some numbers between 1 to 409. Those numbers, those 120 I am taking. So, I am breaking it in 70 percent 30 percent. 70 percent are my training data, 30 percent is testing data. 30 percent of 409 is around 120. So that is why I am taking 120.

These 120 guys creates my testing data. So, randomly selected 120 rows and anything other than this 120 rows selects, creates my training data. So, I just create that. And then keeping everything same I create the model on my training data. And as usual the model is coming, there are various variables which are coming significant. There are some more variable which are coming not significant. To improve the model we use something called step functions.

Step function is stepwise regression, one at a time it will go on dropping based on the goodness of fit, it will start dropping the insignificant variable and after doing that, it will only show me the best model possible. So this is the best model that is possible. See recipient operator has not been dropped because some of the recipient operators are coming significant so you cannot drop. So this is the base model based on this AIC value.

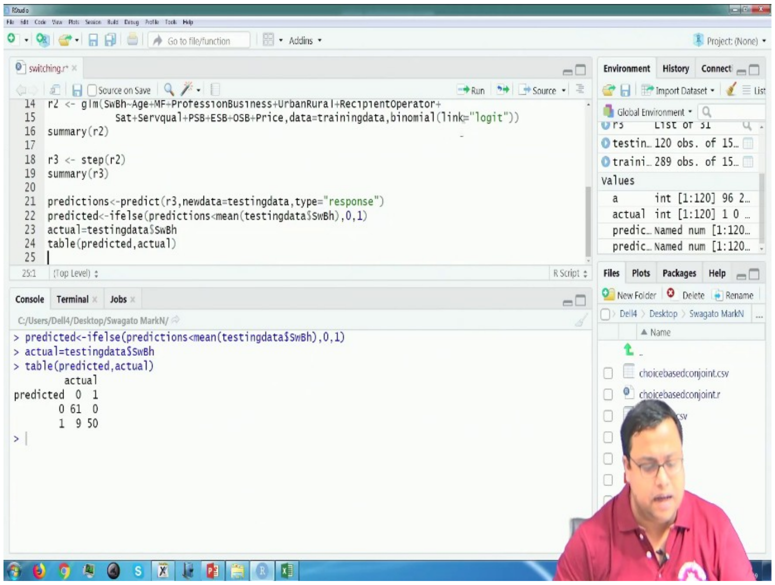
So goodness of fit value I am getting the best model currently. So, I will not see the model how the model looks like. My job is not seeing that right now, when I am doing prediction. When I am doing explanation I will do all of this thing, but when I am doing prediction my job is to do prediction only. So, now with this model I told you that this thing that with my model and with my test data I create predictions.

(Refer Slide Time: 32:03)



```
r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)

r3 <- step(r2)
summary(r3)

predictions<-predict(r3,newdata=testingdata,type="response")
predicted<-ifelse(predictions<mean(testingdata$SwBh),0,1)
actual=testingdata$SwBh
table(predicted,actual)
```

```
C:/Users/Dell4/Desktop/Swagato MarkN/
> predictions<-predict(r3,newdata=testingdata,type="response")
Error in model.frame.default(Terms, newdata, na.action = na.action, xlev = object$xlevels) :
  factor RecipientOperator has new levels Tata Cdma, TATA DOCOMO
> 
```



```r
r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)

r3 <- step(r2)
summary(r3)

predictions<-predict(r3,newdata=testingdata,type="response")
predicted<-ifelse(predictions<mean(testingdata$SwBh),0,1)
actual=testingdata$SwBh
table(predicted,actual)
```

```
C:/Users/Dell4/Desktop/Swagato MarkN/
> predicted<-ifelse(predictions<mean(testingdata$SwBh),0,1)
> actual=testingdata$SwBh
> table(predicted,actual)
         actual
predicted  0  1
        0 61  0
        1  9 50
> |
```

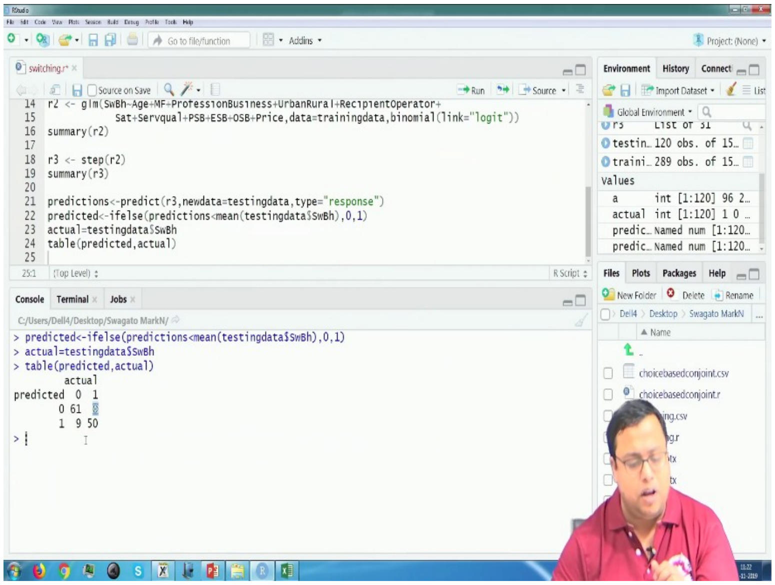
```
r3        List of 31
testin... 120 obs. of 15...
traini... 289 obs. of 15...
Values
  a       int [1:120] 96 2...
  actual  int [1:120] 1 0 ...
  predic... Named num [1:120...
  predic... Named num [1:120...
```

So, the same thing I will do here. So, predict.r3 is the model name and new data is testing data and type is equal to response means I am asking them that this is, so treat Y as response. That means Y as a probability, you give me the probability other than the exact value. That means you give me the $\ln(\frac{p}{1-p})$, solve that p and give me the value. So that is the predictions. So, if new data is equal to, just one minute, so error in model frame, recipient operator has new labels.

So, here by chance the recipient operator had new labels. So, I will run this once more. I will quickly run this part once more, yeah, so I got it. So it is better to sets it to solve this problem. Anyway so I got it and then I want to predict. So, here I am saying that my cutoff will be the switching behaviors mean. How many people are switching, so classically many people take 0.5. So, I will just change it 0.5 for currently take 0.5 as the cut off. So, anything lower than 0.5 probability is zero, any higher than 0.5 probability is 1.

What I did previously was taking the mean propensity? What is the mean propensity of switching? Not always probability, so let us say if the mean propensity very low only 30 percent guys are switching, then I will take that mean propensity as my cutoff. So, that is what let us say that is what I do. I take the mean propensity of switching as my cutoff. Anything lower than that probability is 0 anything higher than that probability is 1 and then actual is obviously the switching behavior and I create a table.

(Refer Slide Time: 34:38)



```
r2 <- glm(SwBh~Age+MF+ProfessionBusiness+UrbanRural+RecipientOperator+
            Sat+Servqual+PSB+ESB+OSB+Price,data=trainingdata,binomial(link="logit"))
summary(r2)

r3 <- step(r2)
summary(r3)

predictions<-predict(r3,newdata=testingdata,type="response")
predicted<-ifelse(predictions<mean(testingdata$SwBh),0,1)
actual=testingdata$SwBh
table(predicted,actual)
```

```
> predicted<-ifelse(predictions<mean(testingdata$SwBh),0,1)
> actual=testingdata$SwBh
> table(predicted,actual)
         actual
predicted  0  1
        0 61  0
        1  9 50
>
```

Simply I create a table here and I see that this is the confusion matrix. What is this confusion matrix? The columns are 0 1 which is the actual 0 and actual 1 and this is my predicted 0 and predicted 1 and out of actual zeros I am predicting 61 right out of 70. So, 61 plus 9. There were 70 actual cases where people did not switch, and 50 actual cases, so if you just do the column sums, 50 actual case where there was switching and 70 actual case there was not switching.

Out of those guys who are switching I am correctly predicting everybody, but out of those guys who are not switching I am correctly predicting 61 and this is the error. This is one error. So overall how much am I correct. Overall accuracy is 60+50/120, this is my overall accuracy 92.5 percent based on this particular data set. So, this is how you can do prediction of switching behavior or non-switching behavior.

So, I have shown you the basic choice modelling of how people switch, how people do not switch and we have used logistic regression for that. There are lots of other methods that we can use to improve the model, other machine learning methods that I will not go into that but whatever machine learning method you use the procedure will be same and for explaining purpose, when your job is to explain X1, X2, X3's impact on y, where was the switching behavior, you use the whole data set.

When your job is to predict you break into testing and training. Testing 30 percent of data set, training 70 percent of data set, you can change it also you can do something called set.seed which we will discuss in a different class probably, how to set seed. So, that you can produce, reproducible research. So, that the number of data set that you create, the random numbers that you generate, when you are doing sampling, and random numbers that I generate remains same. So, all of these things can be done and I get this result. So, we will continue on this in the next video and where we will discuss about choice based conjoint.