

Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture 17
Segmentation Targeting and Positioning (Contd.)

Hello everybody, welcome to marketing analytics course, this is Professor Swagato Chatterjee from Vinod Gupta School of Management, IIT Kharagpur, who is taking this course for you. So, we have been actually doing this particular course and we are right now in module 3 where we were discussing segmentation targeting and positioning and I will continue teaching on that. This is the last video under segmentation targeting and positioning and that is fair, we will start.

So, till now we have actually collected data about people while doing segmentation targeting positioning we collected data about people and what we did is we have broken the data row wise I would say means the human beings based on their preferences, based on their behavior, based on their preferences and then those who are close to each other, we brought them together and those who are away from each other, we actually put them into two different buckets and that is how the clustering is done.

Now, coming back to the problem of where the, where we can marry lots of different techniques, I told while discussing about consumer preferences that we can actually find out consumer preferences from their behavior, from their choices that they have made. For example, let us say if you have chosen consistently chosen brand A over brand B, I know that there is something in brand A that you like very strongly that is why you are choosing A over brand B.

So, we compare whatever is there in brand A, whatever is there in brand B, and whatever there is similarity, I do not think that will impact your choice. But if there is something which is different in brand A and brand B, let us say the fragrance is different, if let us say the color is different, let us say something else is, some performance quality is different, you might give lot of weightage to that. That is why you, you actually choose brand A over brand B consistently.

Now, all I am trying to say here that by focusing on the choice of the consumers then we can come to know that what is their underlying preferences, how, what are the things that within their, now this is not something that is visible. See, I have always told in the previous lectures or

previous videos that we should focus on behavior because behavior is something that is visible. Now, here in this context the only behavior is the choice, the choice that we have made. Oftentimes that is the data that we have.

Now, if I have only one data, which is your choice data then that becomes difficult, it becomes difficult to understand that what kind of other behavior you do and if you do not have lots of behavioral data about the consumers, then I cannot use behavioral segmentation in my segmentation method. So, in that context what we try to do is we try to analyze something and try to find out the underlying preferences and as I told just now, that the choice that our consumer made can actually help us to find out the underlying preferences of the consumers.

Now, if, how to do that in the choice modeling; in what consumers want module; or in the conjoint analysis model, we have discussed about that, that how I can actually consider consumer's data, how can I consider consumer's behavioral data, only one behavior; in this case, the choice and how I can find out references. Now, let us assume that there are multiple such consumers and I am able to find out the preferences of each of these consumers.

If I am able to find out that, then I should be able to actually find out the segments; segments means those consumers who have similar kind of preference. For example, let us say in a choice of B school, let us say you are choosing which B school you will go for your Management Studies. So some customers, in this case the student, there are some applicants, some students who focus on only ROI. So, whatever I am paying and whatever I am getting back in terms of the monetary figures.

Some people focus on the experience as well. Some people probably have a little bit of more preference on the atmosphere; the opportunity of having start up; opportunity of doing something extracurricular as well. So, all of these things also, for some people it might matter. More matter for those kind of people who might be the first time they are going for a hostel life. So, there can be one group of people who have already had a hostel life, a campus life and they might not focus on one another so they might not give more preference on a campus specifically in the second time after bachelors when they are doing masters.

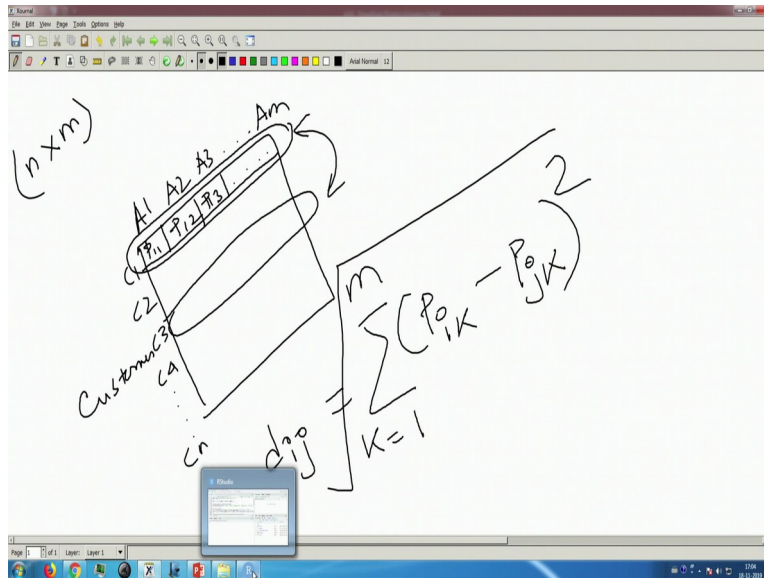
On the other hand, they have another group who probably were day scholars, they used to go to the to their respective colleges from their home. So, they were always hometown based and this is the first time they are going for a management study while they are, they will be away from their home. So, they might want a campus.

So, the preference of campus might be different for different people; preference for ROI will be different; preference for academic load will be different; preference for industry connect will be different. So, there can be different kind of people in the same segment, overall population of let us say students who are interested in doing MBA. There can be several people who might have several preferences.

Now, we cannot do a survey to all this guy and ask them that okay what is your preference in 1 to 5 point scale. You just give me that how much you prefer this, how much you prefer that. Sometimes or many a times it is not possible. So, then what we do is we use that conjoint analysis kind of technique to find out that okay given these four choices, you have chosen A, given these four choices you have chosen D, then that means the preference towards various attributes are this, this and this.

Now, if that is the case then I can also find out there will be some people who will be close to each other and some people who are different from each other. So, that is where I will focus right now.

(Refer Slide Time: 6:55)



Let us assume that from certain kind of conjoint analysis, we found that this is my dataset where these are my customers. So, customer (C1, customer 2, customer 3, customer 4, dot, dot, dot dot customer n and these are my aspects. Aspect 1, aspect 2, aspect 3, aspect m. So, this is basically a (n x m) matrix. And in this particular data set, each row is, each one of the customer's preference towards certain aspects. So, this guy's preference towards A11 is written as let us say preference 11 (P₁₁), then preference 12 (P₁₂), then preference 13 (P₁₃) and so on.

This is the, actually the preference, how much weightage you give to this aspect over some other aspect. Now, what is important to understand is that I am trying to find out the distance between these two group of people or multiple customers and we are creating segments by checking how the preference of this customer is close to this customer. So, remember we created a distance matrix, a Euclidean distance matrix in our, so, here I can also find out a Euclidean distance between these two.

So, how will be the Euclidean distance? The distance between the ith customer and jth customer will be

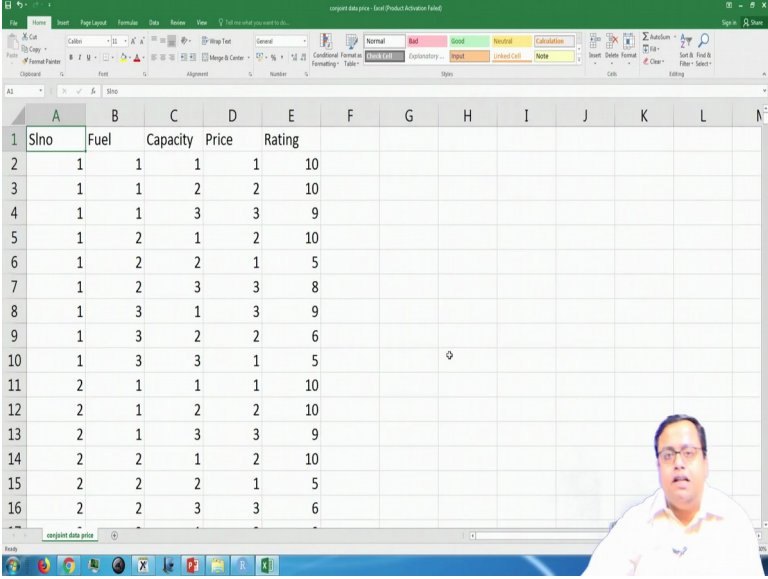
$$d_{ij} = \sqrt{\sum (P_{ik} - P_{jk})^2} \text{ [for } k=1 \text{ to } k=m \text{]}$$

summation of the preference of ith customer on kth attribute minus preference of jth customer minus in kth attributes square of that and K varies from 1 to m. So, this is the Euclidean and actually square root of that is the Euclidean distance.

So, I can find out euclidean distance for each of the customers, each pair and then whoever is closer I will join them and I can do a hierarchical clustering by that. We can also do a K means kind of clustering where if there are n number of customers, I can randomly choose 2 or 3 and see that who falls closer to that particular centroid, two or three centroids I will find out and if somebody is close to such centroid, I will put them in that particular I would say cluster.

And the moment I change the cluster little bit if the customers also change, if the customer's cluster representation also changes, then that is a unstable clustering if I go on and doing and when I reach stability, that means, each customer is assigned to a particular cluster or a particular segment and he is not shifting from that segment. Even if I move the cluster node he is not shifting from the segment. That is something is called k means clustering. We can use that also. So, here in this in this course, in the next probably 10-15 minutes, we will do that.

(Refer Slide Time: 10:10)



| Slno | Fuel | Capacity | Price | Rating |
|------|------|----------|-------|--------|
| 1 | 1 | 1 | 1 | 10 |
| 2 | 1 | 1 | 2 | 10 |
| 3 | 1 | 1 | 3 | 9 |
| 4 | 1 | 2 | 1 | 10 |
| 5 | 1 | 2 | 2 | 5 |
| 6 | 1 | 2 | 3 | 8 |
| 7 | 1 | 3 | 1 | 9 |
| 8 | 1 | 3 | 2 | 6 |
| 9 | 1 | 3 | 3 | 5 |
| 10 | 2 | 1 | 1 | 10 |
| 11 | 2 | 1 | 2 | 10 |
| 12 | 2 | 1 | 3 | 9 |
| 13 | 2 | 2 | 1 | 10 |
| 14 | 2 | 2 | 2 | 5 |
| 15 | 2 | 2 | 3 | 6 |
| 16 | 2 | 2 | 3 | 6 |

So, if you remember, we have a data which look like this. So, this data was part of our conjoint analysis data, but I changed the E column, the column number E a little bit, the rest of the things remain same. If you remember the data set had fuel, fuel 1, 2, 3. There were three types of fuel. Fuel 1 was diesel, fuel 2 was petrol and fuel 3 was CNG. Similarly, I had capacity 1, 2, 3. Capacity 1 was if I am not wrong, 8 seater, capacity 2 is 6 seater and capacity 3 is 4 seater.

And then price 1, 2, 3. Price 1 was if I am not wrong, I think 12 lakhs and 8 lakhs and then 4 lakhs or 6 lakhs, something like that. So, you can go back to the, so this was on conjoint analysis, rating based or ranking based conjoint analysis. And here the ratings are given into 1 to 10 point scale. 1 means do not like this model at all, 10 means, I like the model a lot.

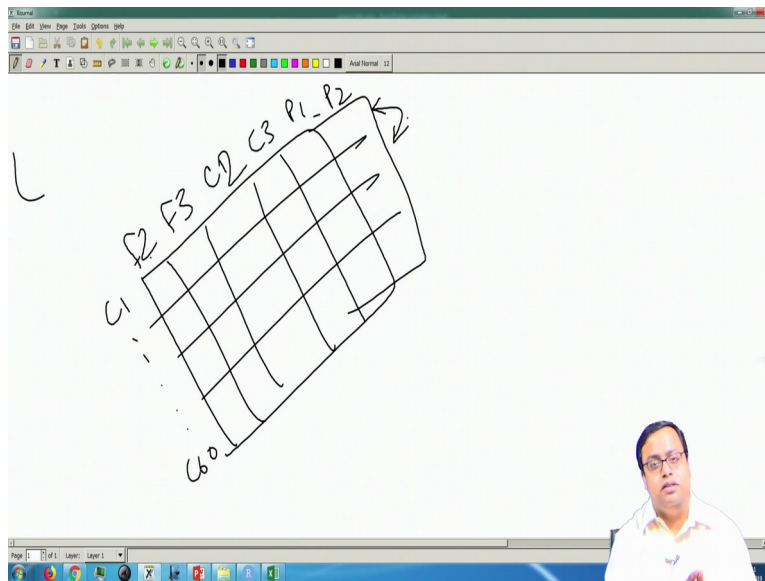
Now, there in that particular problem, we did a regression over the whole data because we thought our underlying assumption was that our customer base is homogeneous. What is homogeneous? That means, they are of similar kind of quality, their focus is same, their preferences are same. So, there is not much difference between customer 1 and customer 2. That is why we have run the analysis on the whole data.

But here in this context I am trying to do something else? What I am trying to do? I am picking up each of the customers' data. If you check, that each customer this serial number is actually the

customer ID. Each customer will have 9 data points. So, each of the customer's data I am picking up and running a regression on rating with fuel, capacity and price as my X variables.

So, each small small regression there are, if you check, there are 30, no there are 60 customers in total. So, then I will actually run 60 regressions. So, if I run 60 regressions, and if I do something like this, then what will I get?

(Refer Slide Time: 12:46)



I will get from customer 1 up to customer 60, I will get their preference for fuel, their preference for let us diesel and their preference for price. And not only that, actually not only that, I will get their preference, so, this is actually wrong. I will get their preference for fuel 2 and 3 in comparison to fuel 1. Remember that it is a categorical data, fuel 1, 2, 3 is actually categories. So, keeping fuel 1 as my, I would say as my categorical reference point, we will find out that how much the customer prefers fuel 2 and fuel 3.

So, keeping diesel as reference point how much they prefer petrol or CNG. So, fuel 2 and fuel 3. Then similarly, I will get let us say capacity 2 and capacities 3 and price 2 and price 3. These are the 6 things that I will get for each of the customer, for each customer. Each row is 1 customer. So, I will get for each of the customer. And then I will find out how customer 1 and customer 2 are distant from each other. If they are close, I will put them in the same cell; if there are away, I will put them in the other cell if I use hierarchical clustering.

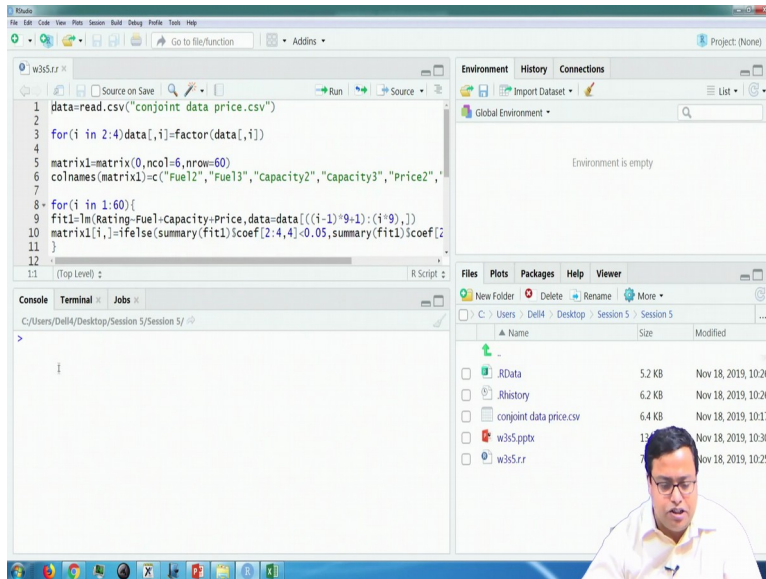
If I use k means based clustering, I already told you what kind of algorithm you will use. But this is something that I will create first. If I create this, the rest of the job is absolutely picking up the code from the previous codes and just run this thing, but this part is something that will be that is new and that is something that I will do. So, in other words, what I am trying to do here is I am adding conjoint analysis or regression analysis whatever you want to name. Conjoint is the mathematical, is the marketing name or mythological name; on the other underlying methodology is econometric methodology is regression.

So, I am marrying regression analysis or conjoint analysis with clustering in this particular problem. So, I told in the first presentation, or in the introductory presentation of this particular course that marketing analytics is often a smart job, its often you have to know in your mind that what are the various things that you have in your arsenal and for a particular problem, what kind of such tools you can bring in and work them parallelly or sequentially so that you can solve the problem.

In this particular context, I am using regression and cluster analysis sequentially. In the previous problem, when I was doing previous thing, there was a segmentation targeting positioning using cluster analysis, followed by targeting for which we used basically multinomial regression. You can also use linear discriminant analysis for that in the previous slide, in previous video.

And there, there was only 6 behaviors, if you remember carefully, there were only 6 behaviours and that is why we only went ahead with factor analysis. But if there are lots of behavioural data you get then to make those behaviour, less number of behaviour, you can also do factor analysis before doing cluster analysis. So, we can, we will do give you that as a do it yourself kind of a project that you can do it on your own, try out on your own and see that what it comes. Now, there we married basically cluster analysis with multinomial logistic regression. Here, we are mining simple linear regression with cluster analysis. So, let us see what do I have.

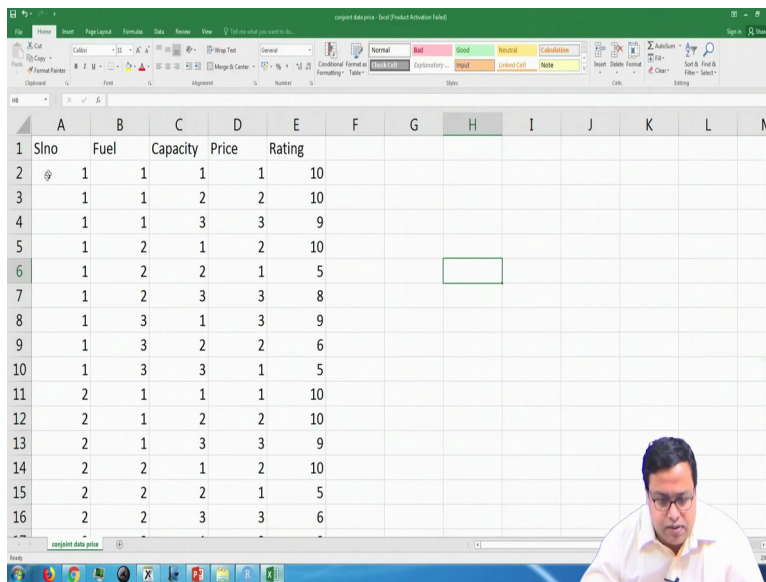
(Refer Slide Time: 16:35)



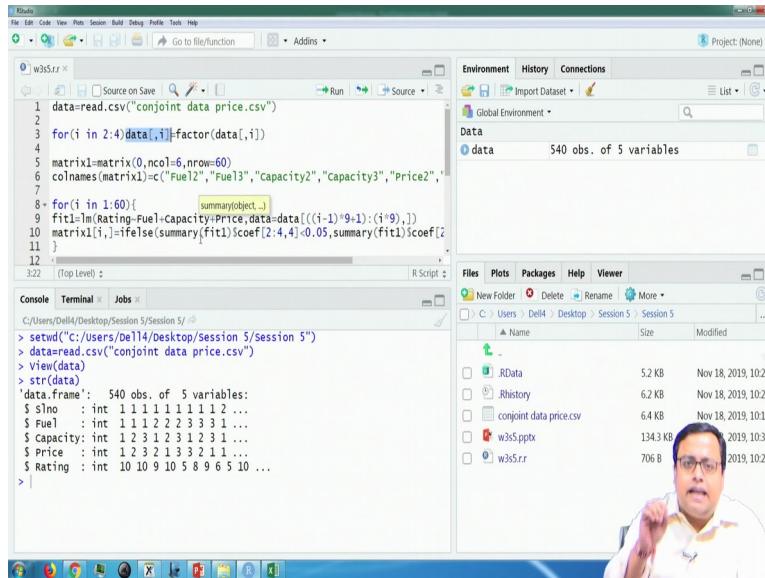
```
1 data=read.csv("conjoint data price.csv")
2
3 for(i in 2:4)data[,i]=factor(data[,i])
4
5 matrix1=matrix(0,ncol=6,nrow=60)
6 colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2","Price3")
7
8 for(i in 1:60){
9   fit1=lm(Rating~Fuel+Capacity+Price,data=data[((i-1)*9+1):(i*9),])
10  matrix1[i,]=ifelse(summary(fit1)$coef[2:4,4]<0.05,summary(fit1)$coef[2:4,4],0)
11 }
12
13 (Top Level)
```

Environment: Global Environment (empty)

Files: RData (5.2 KB), Rhistory (6.2 KB), conjoint data price.csv (6.4 KB), w3s5.pptx (13 KB), w3s5.r (13 KB)



| Slno | Fuel | Capacity | Price | Rating | |
|------|------|----------|-------|--------|----|
| 1 | 1 | 1 | 1 | 10 | |
| 2 | 1 | 1 | 2 | 10 | |
| 3 | 1 | 1 | 3 | 9 | |
| 4 | 1 | 2 | 1 | 10 | |
| 5 | 1 | 2 | 2 | 5 | |
| 6 | 1 | 2 | 3 | 8 | |
| 7 | 1 | 3 | 1 | 9 | |
| 8 | 1 | 3 | 2 | 6 | |
| 9 | 1 | 3 | 3 | 1 | 5 |
| 10 | 2 | 1 | 1 | 1 | 10 |
| 11 | 2 | 1 | 2 | 2 | 10 |
| 12 | 2 | 1 | 3 | 3 | 9 |
| 13 | 2 | 2 | 1 | 2 | 10 |
| 14 | 2 | 2 | 2 | 1 | 5 |
| 15 | 2 | 2 | 3 | 3 | 6 |



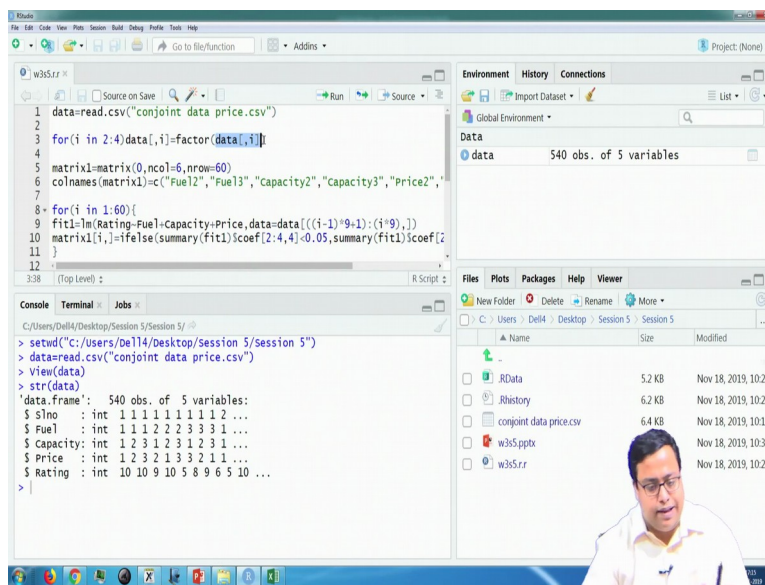
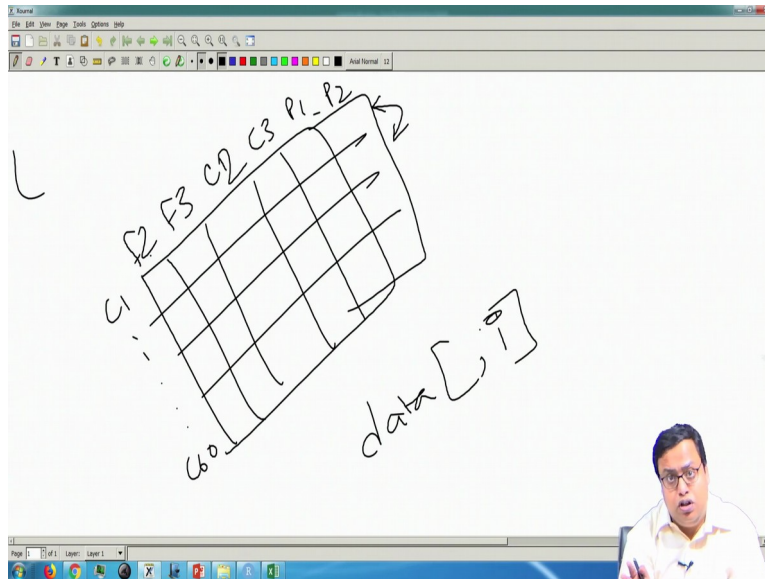
So, I have opened this week 3 session 5 dot R file which is there with you and the corresponding data file looks like this I have already shared the data file, the data file looks like this. And then what I do is, the first things first. I said my working directory to source file location, I actually have kept these 2 files in the same place, the R file. My global environment and console is clean, and I reading the data. So the data looks like this, as you have already seen and in line number 3, what I am doing?

See if I just see the structure of the data, the structure of the data says that this fuel capacity and price are integer variables. Are they integer variables? No, they are not integer values. Why? Because they are categories. 1, 2 and 3 is actually 3 different types of fuels. You cannot say that diesel is 3 times of CNG or petrol is 2 times that of CNG, you cannot say that. So 3, 2, 1 has no meaning. They are just 3 categories, right.

So, we will, what will you do then if they are 3 categories and here it is shown that they are integer variables as for the structure of the data, what will you do? You will change it to factor variable, very good. So, we will change it to factor variable by running the line number 3. See what did it say? For i in 2 to 4. Why 2 to 4? Because second column is fuel, third column is capacity and 4th column is price.

So, i will value from 2 to 4, second column to fourth column. What will I say? I will say that data[,i] the moment I put a third bracket after data, what it does?

(Refer Slide Time: 18:39)

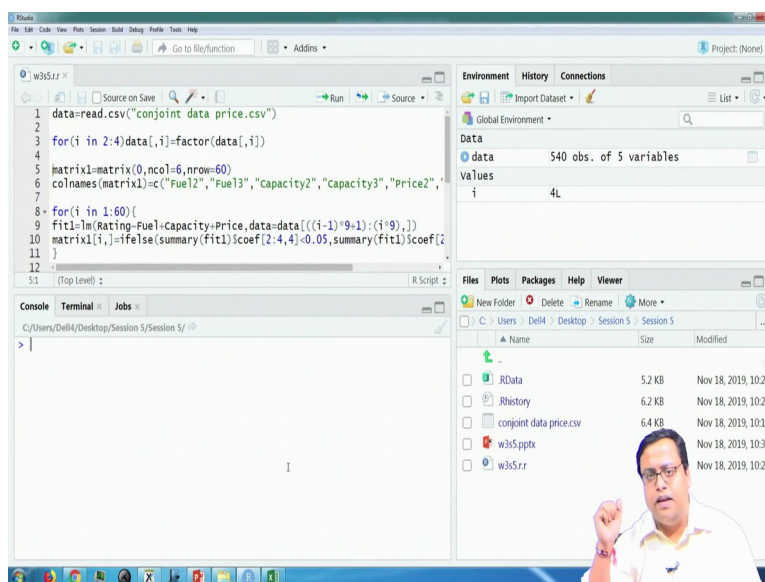


The moment I write some data file name and then I write third bracket, it knows that I am doing a sub setting. I wrote a comma means whatever I wrote before the coma, this was the first module remember. Whatever I wrote before the comma are what? The row names, the row numbers that I would use to subset. And whatever I will put after the comma is what? The column numbers, very good, the column numbers that I will use to subset. Now, if I write nothing after the comma, that means I will take all the columns.

And if I take, if I write something before the comma then I will take such rows. So, the moment I write something like I this means that there is nothing before the comma. That means, I will take all the rows and ith column. What is ith column? When i is 2, it is second column, when i is 3 it is third column, when i is 4, it is 4th column. So that is what I am doing. So, data i is equal to, is equal to what? The factor representation of data i.

So, you pick up data's ith column, change it to its factor form and save it back to data comma i means ith column. So, pick up the fuel column, which is second column when i is equal to 2, it will pick up the fuel column, change it to its factor form. Factor form means 1, 2 and 3 will now be used as categories and save it back. And that is what I am doing for all the 3 columns. So, that is what I am doing, running this, it is changing my 3 things, fuel, capacity and price to their factor forms.

(Refer Slide Time: 20:21)



```
1 data=read.csv("conjoint data price.csv")
2
3 for(i in 2:4)data[,i]=factor(data[,i])
4
5 matrix1=matrix(0,ncol=6,nrow=60)
6 colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2",
7
8 for(i in 1:60){
9 fit1=lm(Rating~Fuel+Capacity+Price,data=data[((i-1)*9+1):(i*9),])
10 matrix1[i,]=ifelse(summary(fit1)$coef[2,4]<0.05,summary(fit1)$coef[2,
11
12
```

The screenshot shows the RStudio interface. The script editor contains the code above. The Environment pane shows a data frame 'data' with 540 observations and 5 variables. The Files pane shows the project files, including 'conjoint data price.csv' and 'w3s5.r'. A small video inset of a man is visible in the bottom right corner of the RStudio window.

See, these are your now factor variables. So, if they are now factor variables, I can run the regression. Now remember, each of these factor variable has 3 categories. Fuel has 3 categories, capacity has 3 categories and price has 3 categories. When I run a regression, because of multicollinearity issues, at least once, so two dummies will be created for each categorical variable let us say for fuel, there are 3 categories.

Two dummy variables will be created, one will be dropped. And, why? Please go back and study your linear regressions, details about some a little bit of econometric background is needed. Why? Because there is multicollinearity issue. I cannot use more, than I cannot use all the categories of a categorical variable in my model because if I use that, they are absolutely multi coordinate, they have a multi correlation, 100 percent multi correlation.

So, the R square value will probably come as I do not know infinite, VIF score will come probably like infinite or something like that. So, that is something that we have to avoid. So, how to avoid that? We drop any one of them. So, in this case, what I will do automatically is R will drop alphabetically whichever is the, so fuel 1, 2, 3, 1 is alphabetically first then 2 and 3, so it will drop that and fuel 2 and 3 it will give you. So, this is the similar result we have got when we run the conjoint analysis.

So, there are 3 categorical variables. Each has 3 levels. If on each of the categorical variable 1 level gets dropped that means there will be 6 columns, 6 categorical variables, 6 dummy variables that will be created from these 3 categorical variables. Since dummy, what is this dummy? For fuel, it will be fuel 2 and fuel 3. For capacity, it will be capacity 2 and capacity 3 and for price it will be price 2 and price 3.

(Refer Slide Time: 22:40)

A hand-drawn diagram of a data matrix. The matrix is a 6x6 grid. The top row is labeled with variables: F_2 , F_3 , C_2 , C_3 , P_1 , P_2 . The left side is labeled with C_1 and P_0 . The bottom side is labeled with C_0 . A double-headed arrow on the top indicates a width of 6, and a double-headed arrow on the left indicates a height of 6. To the right of the matrix, the text $data[i, j]$ is written. The diagram is drawn on a white background within a software window.

A screenshot of the R Studio interface. The main editor window contains the following R code:

```
1 data=read.csv("conjoint data price.csv")
2
3 for(i in 2:4) data[,i]=factor(data[,i])
4
5 matrix1=matrix(0, ncol=6, nrow=60)
6 colnames(matrix1)=c("Fuel2", "Fuel3", "Capacity2", "Capacity3", "Price2",
7
8 for(i in 1:60){
9 fit1=lm(Rating~Fuel+Capacity+Price, data=data[((i-1)*9+1):(i*9),])
10 matrix1[i,]=ifelse(summary(fit1)$coef[2:4,4]<0.05, summary(fit1)$coef[2:4,4], 0)
11 }
12
```

The Environment pane on the right shows the following objects:

- data: 540 obs. of 5 variables
- matrix1: num [1:60, 1:6] 0 0 0 0 0 0 0 0
- Values: 4L

The Console pane at the bottom shows the command:

```
> matrix1=matrix(0, ncol=6, nrow=60)
>
```

The Files pane on the right shows a list of files in the current directory:

| Name | Size | Modified |
|-------------------------|--------|---------------------|
| .. | | |
| ..Data | 5.2 KB | Nov 18, 2019, 10:26 |
| ..History | 6.2 KB | Nov 18, 2019, 10:26 |
| conjoint data price.csv | 6.4 KB | Nov 18, 2019, 10:17 |
| w3s5.pptx | 1.1 MB | Nov 18, 2019, 10:30 |
| w3s5.r | 1.1 KB | Nov 18, 2019, 10:25 |

The screenshot shows the RStudio interface. The main editor displays a data frame with 60 rows and 6 columns, all containing zeros. The console shows the following code:

```

> matrix(0, ncol=6, nrow=60)
> view(matrix1)
>

```

The Environment pane on the right shows a data frame named 'matrix1' with 60 observations and 6 variables, all with values of 0. The Files pane shows the project files, including 'w3s5.pptx' and 'w3s5.r'. A small video inset of a person is visible in the bottom right corner.

The screenshot shows the RStudio interface. The main editor displays the following code:

```

1 csv("conjoint data price.csv")
2
3 data[,i]=factor(data[,i])
4
5 matrix(0, ncol=6, nrow=60)
6 matrix1=c("Fuel2", "Fuel3", "Capacity2", "Capacity3", "Price2", "Price3")
7
8 1:60{
9   rating=Fuel+Capacity+Price, data=data[((1-1)*9+1):(1*9), ]
10  }=ifelse(summary(fit1)$coef[2:4,4]<0.05, summary(fit1)$coef[2:4,1], 0)
11
12

```

The console shows the following code:

```

> matrix(0, ncol=6, nrow=60)
> view(matrix1)
>

```

The Environment pane on the right shows a data frame named 'matrix1' with 60 observations and 6 variables, all with values of 0. The Files pane shows the project files, including 'w3s5.pptx' and 'w3s5.r'. A small video inset of a person is visible in the bottom right corner.

So, I am, here I am creating this kind of a column which has 6 columns and 60 rows and I am creating that, it was in this line where all the entries as a starting point I am writing a 0. So 0, n col is equal to 6, n row is equal to 60, I am creating a matrix. So this is the matrix, a matrix looks like this. There are 60 rows, all of them are 0. And I am naming the column names. Col names of matrix 1 I am writing as fuel 2, fuel 3; capacity 2, capacity 3; and price 2 and price 3.

(Refer Slide Time: 23:19)

The screenshot shows the RStudio interface with the following code in the editor:

```
1 data=read.csv("conjoint data price.csv")
2
3 for(i in 2:4)data[,i]=factor(data[,i])
4
5 matrix1=matrix(0,ncol=6,nrow=60)
6 colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2",
7
8
9 for(i in 1:60){
10 fit=lm(Rating=Fuel+Capacity+Price,data=data[((i-1)*9+1):(i*9),])
11 matrix1[i,]=ifelse(summary(fit)$coef[2,4,4]<0.05,summary(fit)$coef[2,
12
13
14
15
16
```

The console shows the execution of the first few lines of code:

```
> matrix1=matrix(0,ncol=6,nrow=60)
> View(matrix1)
> colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2","Price3")
> View(matrix1)
> |
```

The Environment pane shows the following objects:

| Object | Class | Attributes |
|---------|------------|----------------------------------|
| data | data.frame | 540 obs. of 5 variables |
| matrix1 | matrix | num [1:60, 1:6] 0 0 0 0 0 0 0... |
| i | integer | 4L |

The Files pane shows the following files:

| Name | Size | Modified |
|-------------------------|----------|---------------------|
| .. | | |
| ..RData | 5.2 KB | Nov 18, 2019, 10:26 |
| ..Rhistory | 6.2 KB | Nov 18, 2019, 10:26 |
| conjoint data price.csv | 6.4 KB | Nov 18, 2019, 10:17 |
| w3S5.pptx | 134.3 KB | Nov 18, 2019, 10:30 |
| w3S5.rrr | 706 B | Nov 18, 2019, 10:25 |

The screenshot shows the RStudio interface with the following code in the editor:

```
5 matrix1=matrix(0,ncol=6,nrow=60)
6 colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2",
7
8
9 for(i in 1:60){
10 fit=lm(Rating=Fuel+Capacity+Price,data=data[((i-1)*9+1):(i*9),])
11 matrix1[i,]=ifelse(summary(fit)$coef[2,4,4]<0.05,summary(fit)$coef[2,
12
13
14 df=data.frame(matrix1)
15 wss <- (nrow(df)-1)*sum(apply(df,2,var))
16
```

The console shows the execution of the first few lines of code:

```
> matrix1=matrix(0,ncol=6,nrow=60)
> View(matrix1)
> colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2","Price3")
> View(matrix1)
> View(data)
> |
```

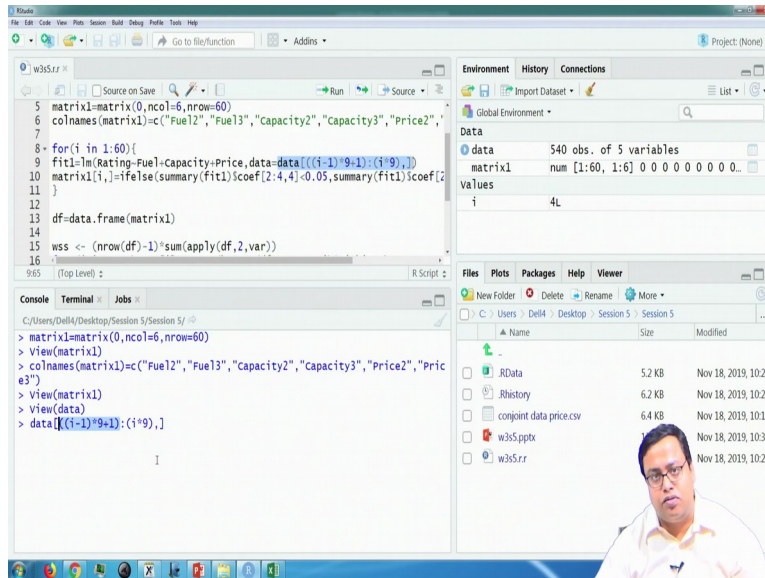
The Environment pane shows the following objects:

| Object | Class | Attributes |
|---------|------------|----------------------------------|
| data | data.frame | 540 obs. of 5 variables |
| matrix1 | matrix | num [1:60, 1:6] 0 0 0 0 0 0 0... |
| i | integer | 4L |

The Files pane shows the following files:

| Name | Size | Modified |
|-------------------------|----------|---------------------|
| .. | | |
| ..RData | 5.2 KB | Nov 18, 2019, 10:26 |
| ..Rhistory | 6.2 KB | Nov 18, 2019, 10:26 |
| conjoint data price.csv | 6.4 KB | Nov 18, 2019, 10:17 |
| w3S5.pptx | 134.3 KB | Nov 18, 2019, 10:30 |
| w3S5.rrr | 706 B | Nov 18, 2019, 10:25 |

A video inset in the bottom right corner shows a presenter speaking.



So, now if I running and after that if I see that name of the columns got changed. Now, what do I do? I run the regression 60 times. So, each time with the data of the corresponding i . So, let us say for the i th guy remember in this data set in this data set if I am the i th guy, then what is the starting point? Each guy will have 9 observations. So, if you remember, the i th guy will start $(i-1)*9$, that many observations will be before him.

So, i th guy is let us say 2, is a second guy. That means 9 observations will be before second guy, that means $(2-1)*9$, i.e. $(1*9)$. If I am fifth guy, then 4 observations, 4 customers observations is already there. That means 36 observations is already there, which is $4*9$ or in other words $(5-1)*9$. So, $(i-1)*9$, these many observations will be, these many rows will be before i th guy's observation starts.

So his starting point that is why will be if you see carefully what I wrote, here I wrote $data[(i-1)*9+1:i*9]$, that is the first observation of i th guy. And what is the last observation of i th guy? $i*9$. So I run the regression, simple regression simple linear regression Lm rating fuel capacity, price but data is equal to a subset of the data, not the whole data, a subset of the data. How to write a subset of the data? Check it carefully what did I write? I just I am just copying this part and pasting it here.

Check it carefully what have I written. Now, I am copying this part and removing it. So, $data[record,]$ that means a subset of data. Nothing written after comma means all the columns.

Now, what did I write before comma? I wrote this. So, I wrote this value, which is if i is equal to 2, that value is 10; if i is equal to 3, that value is 18 plus 1, 19 up to how much? i into 9. So, the first row number and the last row number that much rows will be picked up for every i .

And that will be going on, as i changes these values will change. So, I am writing for i in 1 to 60 for each of the customer, you first run a regression with the corresponding data set of the customer and save it in fit 1, that is the first job. What is the second job? Now, if let us say I am running it for i is equal to 1. Okay, for i is equal to 1 if I run this, this is how the fit 1 looks like.

(Refer Slide Time: 26:37)

```

5 matrix1=matrix(0, ncol=6, nrow=60)
6 colnames(matrix1)=c("Fuel2", "Fuel3", "Capacity2", "Capacity3", "Price2", "Price3")
7
8 for(i in 1:60){
9   fit1=lm(Rating~Fuel+Capacity+Price, data=data[((i-1)*9+1):(i*9),])
10  matrix1[i,]=ifelse(summary(fit1)$coef[2:4,4]<0.05, summary(fit1)$coef[2:4,4], 0)
11 }
12
13 df=data.frame(matrix1)
14
15 wss <- (nrow(df)-1)*sum(apply(df, 2, var))
16

```

Environment: Global Environment
Data: data (540 obs. of 5 variables), fit1 (List of 13), matrix1 (num [1:60, 1:6] 0 0 0 0 0 0 0...), Values: i (1)

Files: RData (5.2 KB), Rhistory (6.2 KB), conjoint data price.csv (6.4 KB), w3s.pptx (134 KB), w3s.r (70 KB)

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.0000    0.6831  14.639  0.00463 **
Fuel2        -2.0000    0.6325  -3.162  0.08713 .
Fuel3        -3.0000    0.6325  -4.743  0.04169 *
Capacity2    -2.8000    0.6928  -4.041  0.05612 .
Capacity3    -2.2000    0.6928  -3.175  0.08650 .
Price2        2.2000    0.6928  3.175  0.08650 .
Price3        1.8000    0.6928  2.598  0.12169
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7746 on 2 degrees of freedom
Multiple R-squared:  0.9667,    Adjusted R-squared:  0.8667
F-statistic: 9.667 on 6 and 2 DF,  p-value: 0.0967

```

Environment: Global Environment
Data: data (540 obs. of 5 variables), fit1 (List of 13), matrix1 (num [1:60, 1:6] 0 0 0 0 0 0 0...), Values: i (1)

Files: RData (5.2 KB), Rhistory (6.2 KB), conjoint data price.csv (6.4 KB), w3s.pptx (134 KB), w3s.r (70 KB)

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for a linear regression model. The code defines a matrix, fits a model, and calculates the variance-covariance matrix of the coefficients.
- Environment Pane:** Shows the global environment with objects: 'data' (540 observations, 5 variables), 'fit1' (a list of 13 elements), and 'matrix1' (a numeric matrix of size 1x6).
- Console:** Displays the output of the R script, including model fit statistics and coefficient estimates.
- Files Pane:** Lists files in the current session, including 'RData', 'Rhistory', 'conjoint data price.csv', 'w3s5.pptx', and 'w3s5.rrr'.

```
5 fix(0, ncol=6, nrow=60)
6 rix1=c("Fuel2", "Fuel3", "Capacity2", "Capacity3", "Price2", "Price3")
7
8 }{
9   ing=Fuel+Capacity+Price, data=data[[(i-1)*9+1):(i*9),])
10   ifelse(summary(fit1)$coef[2:6,4]<0.05, summary(fit1)$coef[2:6,1], 0)
11
12
13 se(matrix1)
14
15 ((df-1)*sum(apply(df, 2, var)))
16
1045 (Top Level) z
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7746 on 2 degrees of freedom
Multiple R-squared: 0.9667, Adjusted R-squared: 0.8667
F-statistic: 9.667 on 6 and 2 DF, p-value: 0.0967

```
> summary(fit1)$coef[2:4, 4]
Fuel2 Fuel3 Capacity2
0.08712907 0.04168515 0.05612019
> (summary(fit1)$coef[2:6, 4]
+)
Fuel2 Fuel3 Capacity2 Capacity3 Price2
0.08712907 0.04168515 0.05612019 0.08649972 0.08649972
> |
```

| Name | Size | Modified |
|-------------------------|----------|---------------------|
| .. | | |
| RData | 5.2 KB | Nov 18, 2019, 10:26 |
| Rhistory | 6.2 KB | Nov 18, 2019, 10:26 |
| conjoint data price.csv | 6.4 KB | Nov 18, 2019, 10:17 |
| w3s5.pptx | 134.3 KB | Nov 18, 2019, 10:30 |
| w3s5.rrr | 706 B | Nov 18, 2019, 10:25 |

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
w3s5.rrr
5 fix(0, ncol=6, nrow=60)
6 price=c("Fuel2", "Fuel3", "Capacity2", "Capacity3", "Price2", "Price3")
7
8
10:45 (Top Level) R Script
Environment History Connections
Global Environment
Data
data 540 obs. of 5 variables
fit1 List of 13
matrix1 num [1:60, 1:6] 0 0 0 0 0 0 0
Values
i 1
Files Plots Packages Help Viewer
New Folder Delete Rename More
C:\Users\Delia\Desktop\Session 5 Session 5
Name Size Modified
.RData 5.2 KB Nov 18, 2019, 10:26
.Rhistory 6.2 KB Nov 18, 2019, 10:26
conjoint data price.csv 6.4 KB Nov 18, 2019, 10:17
w3s5.pptx 134 KB Nov 18, 2019, 10:30
w3s5.rrr 700 KB Nov 18, 2019, 10:25
C:\Users\Delia\Desktop\Session 5 Session 5
Name Size Modified
.RData 5.2 KB Nov 18, 2019, 10:26
.Rhistory 6.2 KB Nov 18, 2019, 10:26
conjoint data price.csv 6.4 KB Nov 18, 2019, 10:17
w3s5.pptx 134 KB Nov 18, 2019, 10:30
w3s5.rrr 700 KB Nov 18, 2019, 10:25
C:\Users\Delia\Desktop\Session 5 Session 5
>
(Intercept) 10.0000 0.6831 14.639 0.00463 **
Fuel2 -2.0000 0.6325 -3.162 0.08713 .
Fuel3 -3.0000 0.6325 -4.743 0.04169 *
Capacity2 -2.8000 0.6928 -4.041 0.05612 .
Capacity3 -2.2000 0.6928 -3.175 0.08650 .
Price2 2.2000 0.6928 3.175 0.08650 .
Price3 1.8000 0.6928 2.598 0.12169
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7746 on 2 degrees of freedom
Multiple R-squared: 0.9667, Adjusted R-squared: 0.8667
F-statistic: 9.667 on 6 and 2 DF, p-value: 0.0967

> summary(fit1)$coef[2:4, 4]
Fuel2 Fuel3 Capacity2
0.08712907 0.04168515 0.05612019
> (summary(fit1)$coef[2:6, 4])
+ )
Fuel2 Fuel3 Capacity2 Capacity3 Price2
0.08712907 0.04168515 0.05612019 0.08649972 0.08649972
>

```

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
w3s5.rrr
matrix1
Filter
Fuel2 Fuel3 Capacity2 Capacity3 Price2 Price3
5 -2.333333 -4.333333 -1.933333 -4.733333 2.066667 3.266667
6 -2.666667 -4.333333 -2.400000 -4.600000 2.600000 3.400000
7 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
8 -2.666667 -4.333333 -2.400000 -4.600000 2.600000 3.400000
9 0.000000 -5.666667 0.000000 0.000000 0.000000 0.000000
10 -2.666667 -4.333333 -2.400000 -4.600000 2.600000 3.400000
11 -2.666667 -4.666667 -2.666667 -4.666667 2.333333 3.333333
Showing 5 to 12 of 60 entries, 6 total columns
Environment History Connections
Global Environment
Data
data 540 obs. of 5 variables
fit1 List of 13
matrix1 num [1:60, 1:6] 0 0 0 -2.67 -2.33...
Values
i 60L
Files Plots Packages Help Viewer
New Folder Delete Rename More
C:\Users\Delia\Desktop\Session 5 Session 5
Name Size Modified
.RData 5.2 KB Nov 18, 2019, 10:26
.Rhistory 6.2 KB Nov 18, 2019, 10:26
conjoint data price.csv 6.4 KB Nov 18, 2019, 10:17
w3s5.pptx 134 KB Nov 18, 2019, 10:30
w3s5.rrr 700 KB Nov 18, 2019, 5:23
C:\Users\Delia\Desktop\Session 5 Session 5
>
Fuel2 Fuel3 Capacity2
0.08712907 0.04168515 0.05612019
> (summary(fit1)$coef[2:6, 4])
+ )
Fuel2 Fuel3 Capacity2 Capacity3 Price2
0.08712907 0.04168515 0.05612019 0.08649972 0.08649972
> for(i in 1:60){
+ fit1=lm(Rating~Fuel+Capacity+Price, data=data[((i-1)*9+1):(i*9), ])
+ matrix1[i,]=ifelse(summary(fit1)$coef[2:7, 4]<0.05, summary(fit1)$coef[2:7,
+ ], 0)
+ }
> View(matrix1)
>

```

This is how fit 1 looks like. So I am saying that now remember there are some values which I consider to be 0 because these guys are not significant, like this is not lower than 0.05. This is not lower than 0.05. So I can consider these to be 0, this to be 0, this to be 0, this to be 0. So, I am saying that if else if summary feed coefficients is less than 5, what is this? This is nothing but the P values, see the P values.

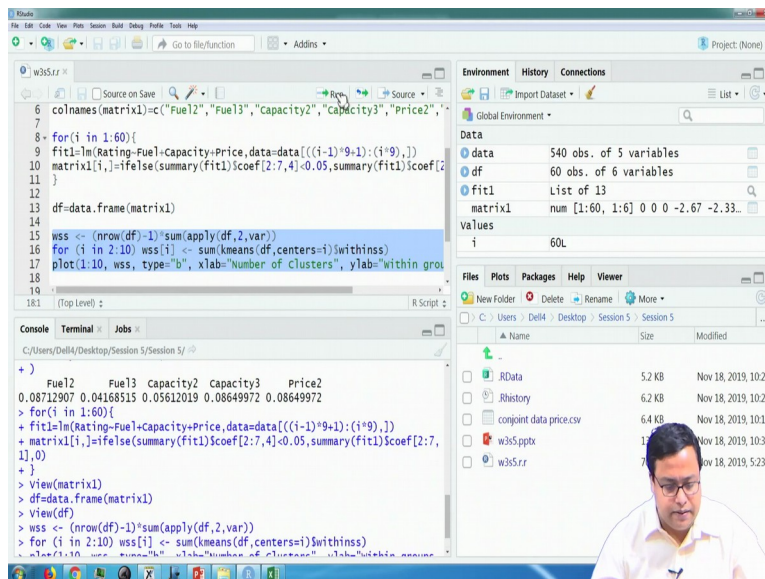
The P value of fuel 2, fuel 3 are okay sorry it is it should not be 2 to 4, it should be 2 to 6 and here also it should be 2 to 6. So, I will just run once more. So, what is this thing? This is nothing but, this is nothing but the P value. See, check the P values, the P values are 0.8713, 0.04169,

0.056. Similarly, 0.08713, 0.04, 0.05, 0.08. So, I will only consider such values, I will only consider such values where these values are lower than 0.05.

If they are lower than 0.05, consider the coefficients as it is. If they are not lower than 0.05 consider the coefficients to be 0. So 2 to 7. 2, 3, 4, 5, 6, 7 that means, total 7 things, I am running this now, so, 2 to 7. So, in the actual file that will be given to you, this will be edited 2 to 7 and then I run this. So, I run this. So, once I run this, the matrix looks like this, this is the matrix that I got.

All the zeros means for customer 2 nothing was significant, for customer 3 nothing was significant, but there are some customers for whom some things were significant and corresponding weightages were written here for each of the customers, it got populated. Now, rest of the things is simple. I will use this matrix to run a cluster analysis. So, first I will convert this to a data frame called DF. DF looks like this, same thing. Now, I will use this DF to run a K means kind of clustering.

(Refer Slide Time: 29:05)



The screenshot displays the RStudio interface. The script editor contains the following R code:

```
colnames(matrix1) = c("Fuel2", "Fuel3", "Capacity2", "Capacity3", "Price2", "Price3")
for(i in 1:60){
  fit1 = lm(Rating ~ Fuel + Capacity + Price, data = data[((i-1)*9+1):(i*9),])
  matrix1[i,] = ifelse(summary(fit1)$coef[2:7,4] < 0.05, summary(fit1)$coef[2:7,4], 0)
}
df = data.frame(matrix1)
wss <- (nrow(df)-1) * sum(apply(df, 2, var))
for(i in 2:10) wss[i] <- sum(kmeans(df, centers=i)$withinss)
plot(1:10, wss, type="b", xlab="Number of clusters", ylab="within group sum of squares")
```

The console shows the output of the first row of the matrix:

```
      Fuel2 Fuel3 Capacity2 Capacity3 Price2
0.08712907 0.04168515 0.05612019 0.08649972 0.08649972
```

The Environment pane on the right shows the following objects:

- data: 540 obs. of 5 variables
- df: 60 obs. of 6 variables
- fit1: List of 13
- matrix1: num [1:60, 1:6] 0 0 0 -2.67 -2.33...
- values: 60L

The Files pane shows a file explorer view of the Desktop/Session 5 directory, listing files such as RData, Rhistory, conjoint data price.csv, w3s5.pptx, and w3s5.r.

RStudio interface showing R code for calculating the Within Groups Sum of Squares (WSS) for different numbers of clusters. The code includes a loop to fit linear models and calculate WSS for 2 to 10 clusters, followed by a plot of WSS vs. Number of Clusters.

```

6 colnames(matrix1)=c("Fuel2","Fuel3","Capacity2","Capacity3","Price2",
7
8 for(i in 1:60){
9 fit1=lm(Rating~Fuel+Capacit
10 matrix1[i,]=ifelse(summary(
11
12
13 df=data.frame(matrix1)
14
15 wss <- (nrow(df)-1)*sum(apply
16 for (i in 2:10) wss[i] <-
17 plot(1:10, wss, type="b",
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

```

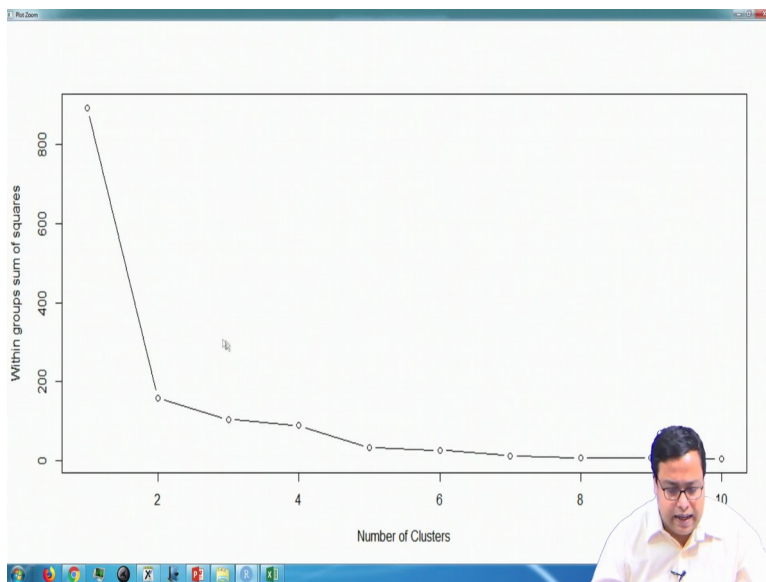
The plot shows the Within groups sum of squares decreasing as the number of clusters increases from 1 to 10. The y-axis is labeled 'Within groups sum of squares' and ranges from 0 to 800. The x-axis is labeled 'Number of Clusters' and ranges from 1 to 10.

Console output:

```

> for(i in 1:60){
+ fit1=lm(Rating~Fuel+Capacity+Pr
+ matrix1[i,]=ifelse(summary(Fit
11),0)
+ }
> View(matrix1)
> df=data.frame(matrix1)
> View(df)
> wss <- (nrow(df)-1)*sum(apply(df,2,var))
> for (i in 2:10) wss[i] <- sum(kmeans(df,centers=i)$withinss)
> plot(1:10, wss, type="b", xlab="Number of clusters", ylab="within groups
sum of squares")
>

```



RStudio interface showing R code for k-means clustering and a plot of within groups sum of squares.

```

11 }
12
13 df=data.frame(matrix1)
14
15 wss <- (nrow(df)-1)*sum(apply(df,2,var))
16 for (i in 2:10) wss[i] <- sum(kmeans(df,centers=i)$withinss)
17 plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="within groups
18
19 fit <- kmeans(df, 10) # k1 cluster solution
20
21 # get cluster means
22 aggregate(df,by=list(fit$cluster),FUN=mean)
23

```

Environment:

- data: 540 obs. of 5 variables
- df: 60 obs. of 6 variables
- fit: List of 13
- matrix1: num [1:60, 1:6] 0 0 0 -2.67 -2.33...

Values:

- i: 10L
- wss: num [1:10] 892.4 158.8 104.2 88.6 3...

RStudio interface showing R code for k-means clustering, a plot of within groups sum of squares, and the resulting cluster means.

```

11 }
12
13 df=data.frame(matrix1)
14
15 wss <- (nrow(df)-1)*sum(apply(df,2,var))
16 for (i in 2:10) wss[i] <- sum(kmeans(df,centers=i)$withinss)
17 plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="within groups
18
19 fit <- kmeans(df, 2) # k1 cluster solution
20
21 # get cluster means
22 aggregate(df,by=list(fit$cluster),FUN=mean)
23

```

Environment:

- data: 540 obs. of 5 variables
- df: 60 obs. of 6 variables
- fit: List of 9
- fit1: List of 13
- matrix1: num [1:60, 1:6] 0 0 0 -2.67 -2.33...

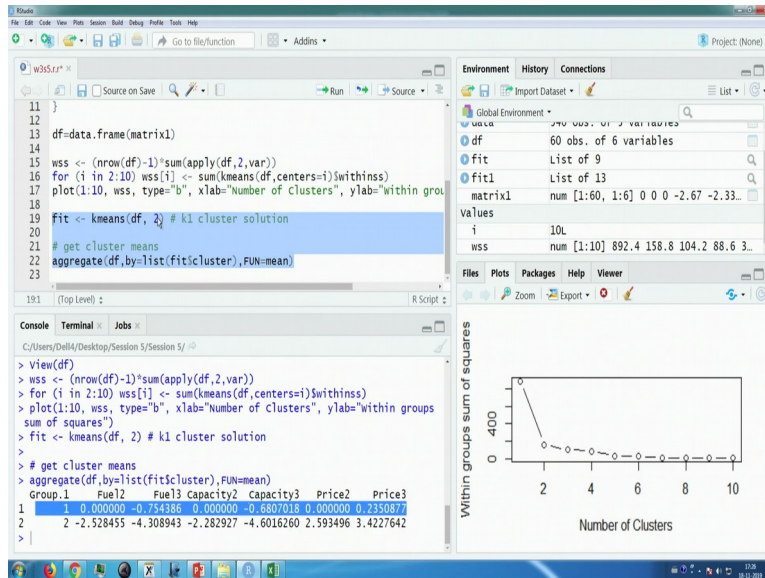
Values:

- i: 10L
- wss: num [1:10] 892.4 158.8 104.2 88.6 3...

```

> View(df)
> wss <- (nrow(df)-1)*sum(apply(df,2,var))
> for (i in 2:10) wss[i] <- sum(kmeans(df,centers=i)$withinss)
> plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="within groups
sum of squares")
> fit <- kmeans(df, 2) # k1 cluster solution
>
> # get cluster means
> aggregate(df,by=list(fit$cluster),FUN=mean)
  Group.1  Fuel2  Fuel3 capacity2 capacity3  Price2  Price3
1      1  0.000000 -0.754386  0.000000 -0.6807018 0.000000 0.2350877
2      2 -2.528455 -4.308943 -2.282927 -4.6016260 2.593496 3.4227642
>

```



So, first I will plot the scree plot. So, this is the code that I have used there also, I will just run it and this is how the plot comes up and I can see that the kink is in 2. Kink is properly coming at 2. You can probably also use 3 or something like that. So, first I will use clusters is 2, if there are only 2 clusters, what is the meaning I am getting? I will run that and I will see that. There are 2 clusters.

The first cluster looks like this. These are the average prevalence of first cluster and second cluster, this is the average preference. So, first cluster guys, they do not think petrol to be much more attractive than diesel. Diesel is the reference point, fuel 1 is diesel which got dropped. So, these guys do not find petrol to be more attractive than diesel but they definitely find CNG to be less attractive. Now, this guy is highly sensitive towards fuel. They think they are very much I would say diesel conscious.

Anything which is petrol they do not like, anything which is CNG they absolutely do not like. So these guys are very much focused on diesel. These guys are not much focused on diesel; for them diesel and petrol are not different. The coefficient is 0 and the coefficient for fuel 3 which is CNG is negative but not that very big. On the other hand, this guy is also capacity sensitive. So they want big car, 8 seater. So, if it is 6 seater they do not like, if it is 4 seater, they do not like at all.

On the other hand, this guy are also not much sensitive, 8 and 6 seater is okay. And for 4 seater, they do not prefer 4 seater, but that preference difference is not much. Similarly, these guys are price sensitive also. So, in other words, I can say group 2 is heavily sensitive on all the 3 aspects and group 1 is not so much sensitive in any of the 3 aspects. I can also find out if what if there were 3 clusters?

Let us check, if there are 3 clusters, then I can find out that okay this guy is absolutely not fuel sensitive, they are not fuel sensitive at all, but heavily capacity sensitive. So, the moment capacity drops below a level, they become very sensitive and the moment it becomes price goes up, they become very sensitive. This guy is like segment 2 in the previous one, they are sensitive to everything. On the other hand group 3 is highly fuel sensitive or moderately fuel sensitive and not sensitive to anything else.

So, I can find out different groups who have different kind of sensitivity towards different kind of things. And then if I have some demographic data, I can try to see that what is the demographic data that predicts that whether you will be group 1, or group 2 or group 3. So, is it some specific type of income group, a specific type of gender or specific type of cultural background that leads to your price sensitivity or fuel sensitivity or everything sensitivity? We can try to analyze that.

For that you have to use again kind of regression or LDA that we have the name the last video. So, that is all for segmentation targeting positioning. We have done quite a lot of thing. Thank you for being with me. I will come back with a new module in the next video. Thank you.