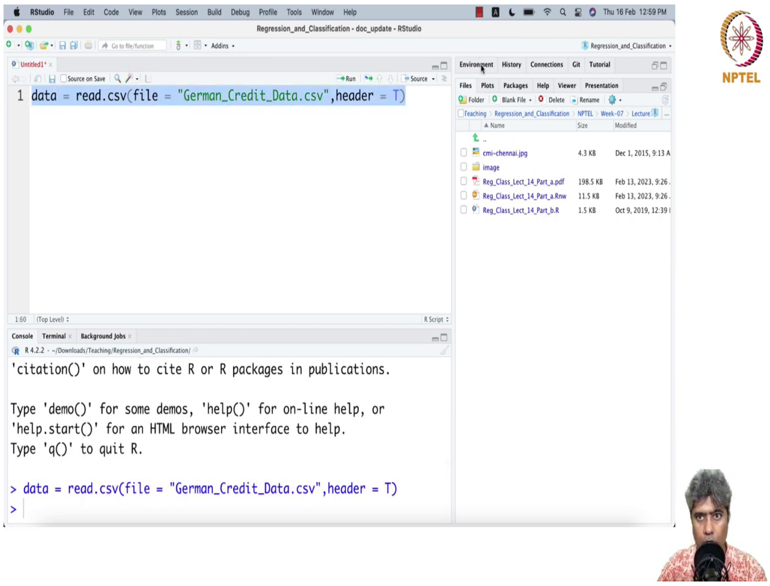
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 38
Hands on with R for Logistic Regression

Hello all, welcome back to the part C of lecture 11. In this part, we are going to do some Hands-on.

(Refer Slide Time: 00:23)

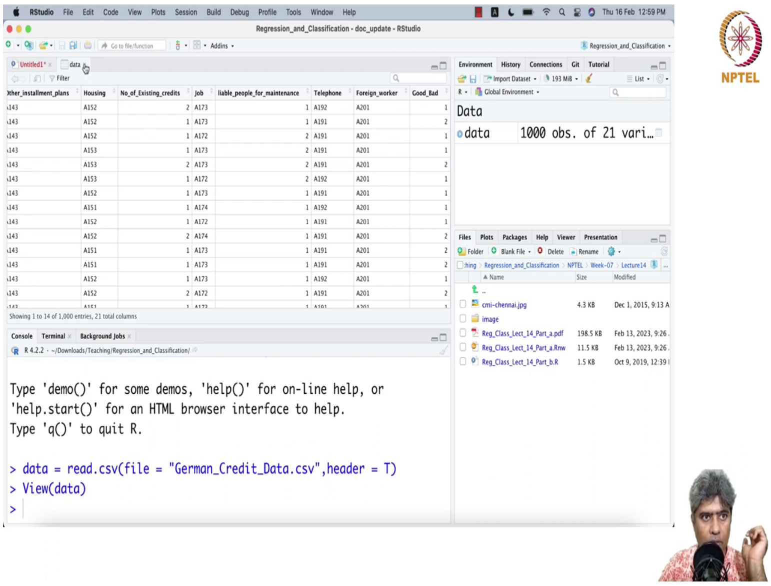


```
data = read.csv(file = "German_Credit_Data.csv",header = T)
```

```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv(file = "German_Credit_Data.csv",header = T)
>
```

So, we are going to call the and this hands-on will be based on R. So, first we are going to read the CSV file. So, we are going to read the German credit code data. So, I have already shared this on the NPTEL platform, Credit also if you do search German credit code data on internet, you will get this data same data set its a slightly old, but reasonably good data to get your hands dirty with, you know, the first logistic regression model kind of thing.

(Refer Slide Time: 01:36)


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 12:59 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data
Filter
Environment History Connections Git Tutorial
Import Dataset
193 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
NPTEL

Other_installment_plans	Housing	No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad
A143	A152	2	A173	1	A192	A201	1
A143	A152	1	A173	1	A191	A201	2
A143	A152	1	A172	2	A191	A201	1
A143	A153	1	A173	2	A191	A201	1
A143	A153	2	A173	2	A191	A201	2
A143	A153	1	A172	2	A192	A201	1
A143	A152	1	A173	1	A191	A201	1
A143	A151	1	A174	1	A192	A201	1
A143	A152	1	A172	1	A191	A201	1
A143	A152	2	A174	1	A191	A201	2
A143	A151	1	A173	1	A191	A201	2
A143	A151	1	A173	1	A191	A201	2
A143	A152	1	A173	1	A192	A201	1
A143	A152	2	A172	1	A191	A201	2
A143	A151	1	A173	1	A191	A201	1

Showing 1 to 14 of 1,000 entries, 21 total columns

Data

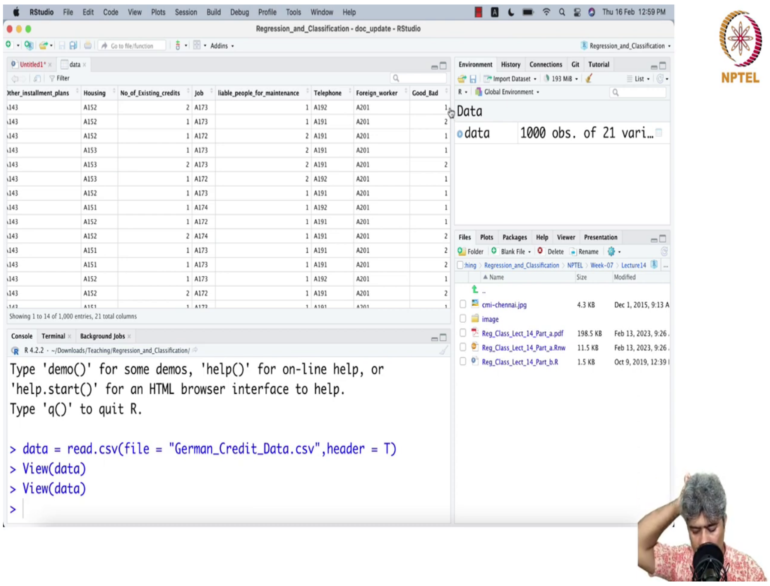
data	1000 obs. of 21 vari…

:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv(file = "German_Credit_Data.csv",header = T)
> View(data)
>
```

(Refer Slide Time: 01:54)


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 12:59 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data
Filter
Environment History Connections Git Tutorial
Import Dataset
193 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
NPTEL

)ther_installment_plans	Housing	No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad
\143	A152	2	A173	1	A192	A201	1
\143	A152	1	A173	1	A191	A201	2
\143	A152	1	A172	2	A191	A201	1
\143	A153	1	A173	2	A191	A201	1
\143	A153	2	A173	2	A191	A201	2
\143	A153	1	A172	2	A192	A201	1
\143	A152	1	A173	1	A191	A201	1
\143	A151	1	A174	1	A192	A201	1
\143	A152	1	A172	1	A191	A201	1
\143	A152	2	A174	1	A191	A201	2
\143	A151	1	A173	1	A191	A201	2
\143	A151	1	A173	1	A191	A201	2
\143	A152	1	A173	1	A192	A201	1
\143	A152	2	A172	1	A191	A201	2
\143	A151	1	A173	1	A191	A201	1

Showing 1 to 14 of 1,000 entries, 21 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv(file = "German_Credit_Data.csv",header = T)
> View(data)
> View(data)
>
```

Data

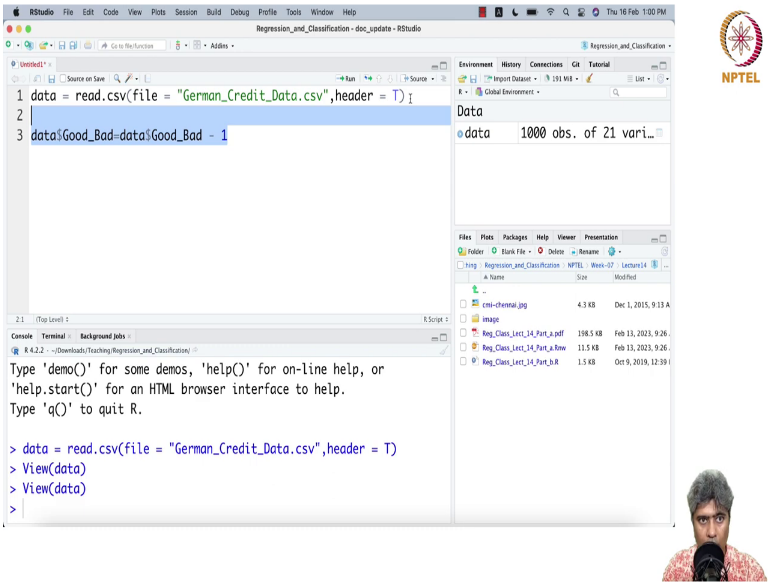
data	1000 obs. of 21 vari…

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

So, yeah, if you just come see there are about 1000 values and there are some status are given and if you look into the status, if you go to and then the target here is good or bad, good loan or bad loan and then essentially I, we have to, it is recorded as 1 or 2.

(Refer Slide Time: 02:02)



```
data = read.csv(file = "German_Credit_Data.csv",header = T)

data$Good_Bad=data$Good_Bad - 1
```

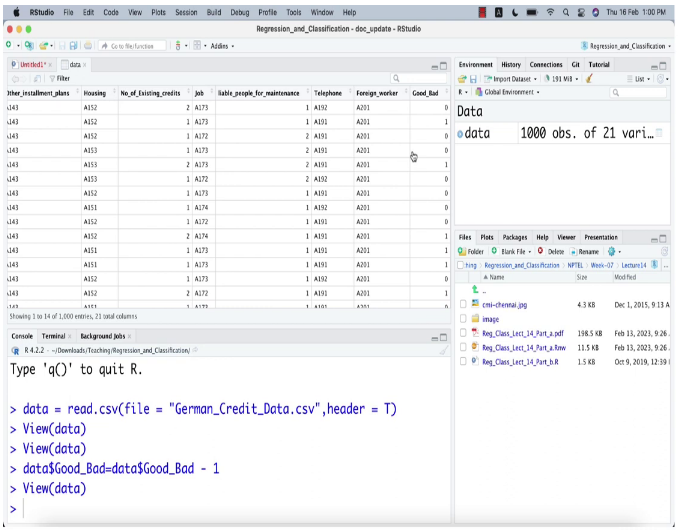
```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv(file = "German_Credit_Data.csv",header = T)
> View(data)
> View(data)
>
```

Data
data 1000 obs. of 21 vari…

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

Regression_and_Classification - doc_update - RStudio
Untitled1*
Environment History Connections Git Tutorial
Global Environment
Files Plots Packages Help Viewer Presentation
Regression_and_Classification > NPTEL > Week-07 > Lecture14
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
NPTEL

(Refer Slide Time: 02:29)



ther_installment_plans	Housing	No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad
143	A152	2	A173	1	A192	A201	0
143	A152	1	A173	1	A191	A201	1
143	A152	1	A172	2	A191	A201	0
143	A153	1	A173	2	A191	A201	0
143	A153	2	A173	2	A191	A201	1
143	A153	1	A172	2	A192	A201	0
143	A152	1	A173	1	A191	A201	0
143	A151	1	A174	1	A192	A201	0
143	A152	1	A172	1	A191	A201	0
143	A152	2	A174	1	A191	A201	1
143	A151	1	A173	1	A191	A201	1
143	A151	1	A173	1	A191	A201	1
143	A152	1	A173	1	A192	A201	0
143	A152	2	A172	1	A191	A201	1
143	A151	1	A173	1	A191	A201	0

Showing 1 to 14 of 1,000 entries, 21 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Type 'q()' to quit R.

> data = read.csv(file = "German_Credit_Data.csv",header = T)
> View(data)
> View(data)
> data$Good_Bad=data$Good_Bad - 1
> View(data)
>
```

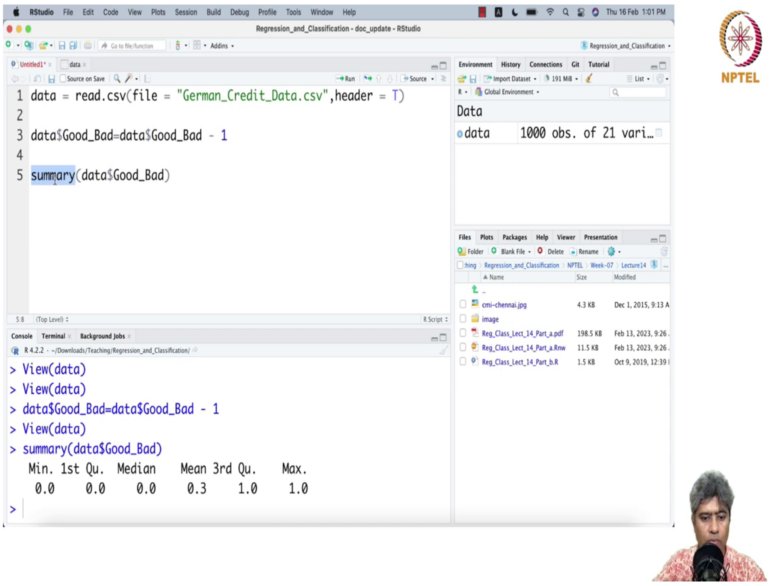
Data
data 1000 obs. of 21 vari…

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:00 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data
Filter
Console
Terminal
Background Jobs
Environment History Connections Git Tutorial
Import Dataset
191 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14

So, what we will do is data dollar Good Bad equal to, if we just say data dollar Good Bad minus 1. So, that will give us, you see good or bad, if it is 1, then it is bad, 0 means good. So, we want to predict accordingly.

(Refer Slide Time: 02:40)



```
data = read.csv(file = "German_Credit_Data.csv",header = T)

data$Good_Bad=data$Good_Bad - 1

summary(data$Good_Bad)
```

```
> View(data)
> View(data)
> data$Good_Bad=data$Good_Bad - 1
> View(data)
> summary(data$Good_Bad)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0     0.0     0.3     1.0     1.0
>
```

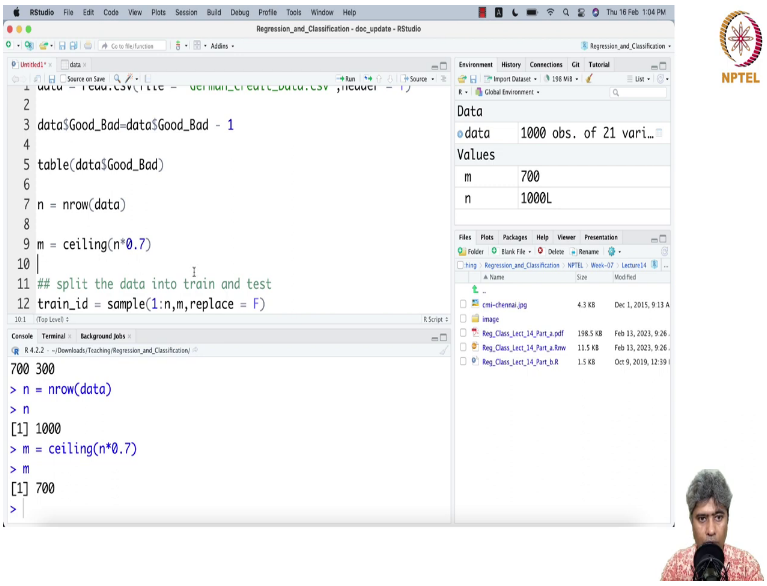
Data

data	1000 obs. of 21 vari…

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

(Refer Slide Time: 03:07)



```
data = read.csv(file = "german_credit_data.csv",header = T)

data$Good_Bad=data$Good_Bad - 1

table(data$Good_Bad)

n = nrow(data)

m = ceiling(n*0.7)


## split the data into train and test
train_id = sample(1:n,m,replace = F)
```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
700 300
> n = nrow(data)
> n
[1] 1000
> m = ceiling(n*0.7)
> m
[1] 700
>
```

Data	
data	1000 obs. of 21 vari…
Values	
m	700
n	1000L

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
Untitled1*
data
Source on Save
Run
Source
Console
Terminal
Background Jobs
Environment History Connections Git Tutorial
Files Plots Packages Help Viewer Presentation
cmi-chennai.jpg
image
Reg_Class_Lect_14_Part_a.pdf
Reg_Class_Lect_14_Part_a.Rnw
Reg_Class_Lect_14_Part_b.R
NPTEL

Now, if we just do summary, summary of good or bad. So, about 30 percent is bad actually. So, just a minute, I think if I just do table instead of summary, the if I just do a table. Yeah. So, 300 out of 1000 cases, they have 300, they have given as bad cases and 700 given as a good cases. Now, typically in real life, you will not have default more than 2, 3 percent of the entire credit customers.

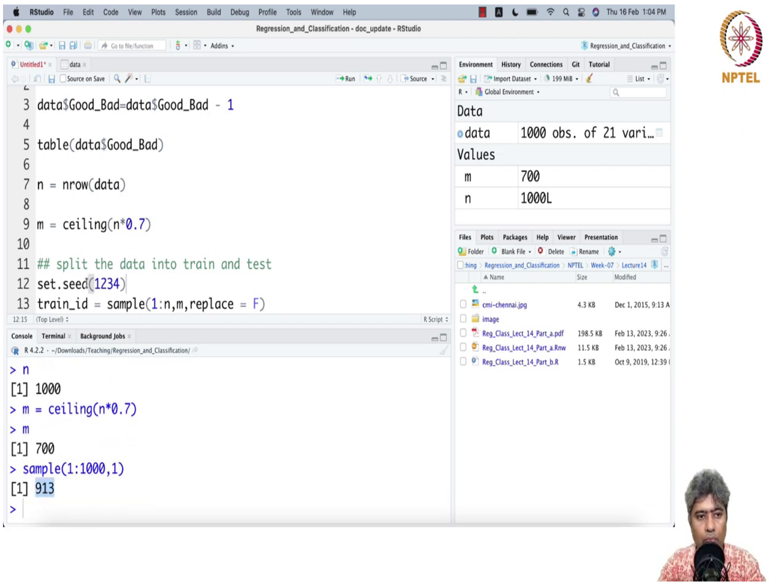
So, this is slightly a artificial in that sense because you cannot, if you are, if your 30 percent of customer is this defaulting, Then you will not be able to you know survive. But anyway, but if you have in your data set, if you have only 2 percent ones and 98 percent zeros, those kind of data sets are called imbalance data sets and in those kind of data sets doing prediction is extremely difficult.

And because most of the time, what happens that any standard model will just say everybody is kind of, you know, good, then no matter what, what is your, what happens 98 percent of the time you are correct. So, and in only 2 percent time you are wrong. So, a naive model will do correct, will be correct 98 percent of time. So, in real.

So, those kind of naive model, we have to be very careful about. So, what I will do first, I will just say number of rows in the data is 1000, I believe yeah. And then I will take m is the testing number of testing, like, you know, n into 0.7, maybe 70 percent of the data we will consider for the testing.

So, 700 out of 1000 700 data points will be considering for testing. So, for we will do a split the data into train and test. First, what we will do? we will choose the train id's sample 1, is to, n out m many samples you will we will draw and replace equal to false. Ok, we have to set up a seed.

(Refer Slide Time: 06:06)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:04 PM
Regression_and_Classification - doc_update - RStudio
Untitled1* data
Source on Save
Run Source
Environment History Connections Git Tutorial
Import Dataset 199 MiB List
Global Environment
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Regression_and_Classification > NPTEL > Week-07 > Lecture14
NPTEL

```r
data$Good_Bad=data$Good_Bad - 1

table(data$Good_Bad)

n = nrow(data)

m = ceiling(n*0.7)

## split the data into train and test
set.seed(1234)
train_id = sample(1:n,m,replace = F)
```

```
> n
[1] 1000
> m = ceiling(n*0.7)
> m
[1] 700
> sample(1:1000,1)
[1] 913
>
```

Data

data	1000 obs. of 21 vari…

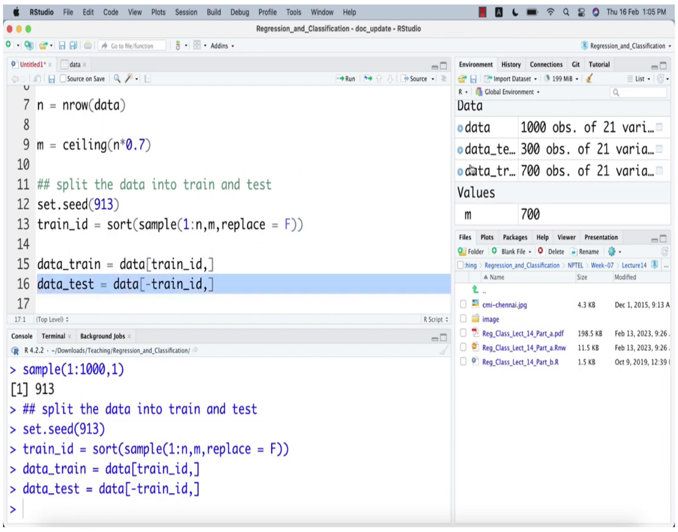
Values

m	700
n	1000L

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

Set dot, seed is 1234. So, the advantage of using seed is if you set up a seed and if you use exactly this number 1234 actually you can do any number, I can just draw a sample between 1, is to 1000, one sample and 913, I am let me just use 913 as my seed.

(Refer Slide Time: 06:34)



```r
n = nrow(data)

m = ceiling(n*0.7)

## split the data into train and test
set.seed(913)
train_id = sort(sample(1:n,m,replace = F))

data_train = data[train_id,]
data_test = data[-train_id,]
```

```
> sample(1:1000,1)
[1] 913
> ## split the data into train and test
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
>
```

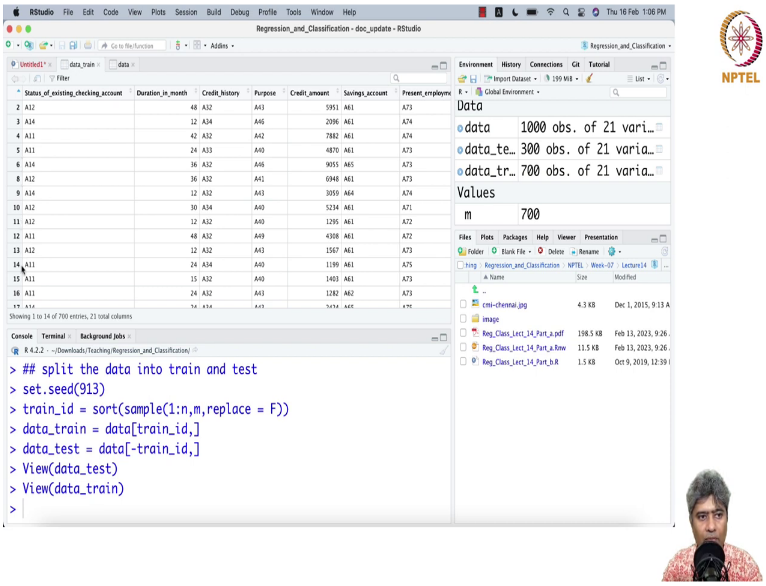
```
Data
data        1000 obs. of 21 vari…
data_te…    300 obs. of 21 varia…
data_tr…    700 obs. of 21 varia…
Values
m           700
```

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

Now, that means, if you use this seed number as your seed, then what will happen is you, you my result and your result will exactly match ok. So, this ensures because we are going to split the data set randomly 70 percent data we will take as a training data, ok. So, data underscore train equal to what I am going to do data. Just I will provide the training id comma maybe I will just do a sorting here ok.

Let me just take this and then I will just data train and data test equal to what I will do. I will just copy this and instead of train id, I will just say minus train id. So, just take all the ids, but do not take the training ids.

(Refer Slide Time: 07:40)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:06 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data_train
data
Filter
Environment History Connections Git Tutorial
Import Dataset
199 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
NPTEL

	Status_of_existing_checking_account	Duration_in_month	Credit_history	Purpose	Credit_amount	Savings_account	Present_employme
2	A12	48	A32	A43	5951	A61	A73
3	A14	12	A34	A46	2096	A61	A74
4	A11	42	A32	A42	7882	A61	A74
5	A11	24	A33	A40	4870	A61	A73
6	A14	36	A32	A46	9055	A65	A73
8	A12	36	A32	A41	6948	A61	A73
9	A14	12	A32	A43	3059	A64	A74
10	A12	30	A34	A40	5234	A61	A71
11	A12	12	A32	A40	1295	A61	A72
12	A11	48	A32	A49	4308	A61	A72
13	A12	12	A32	A43	1567	A61	A73
14	A11	24	A34	A40	1199	A61	A75
15	A11	15	A32	A40	1403	A61	A73
16	A11	24	A32	A43	1282	A62	A73
17	A14	24	A34	A43	2424	A65	A75

Showing 1 to 14 of 700 entries, 21 total columns

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…

Values

m	700

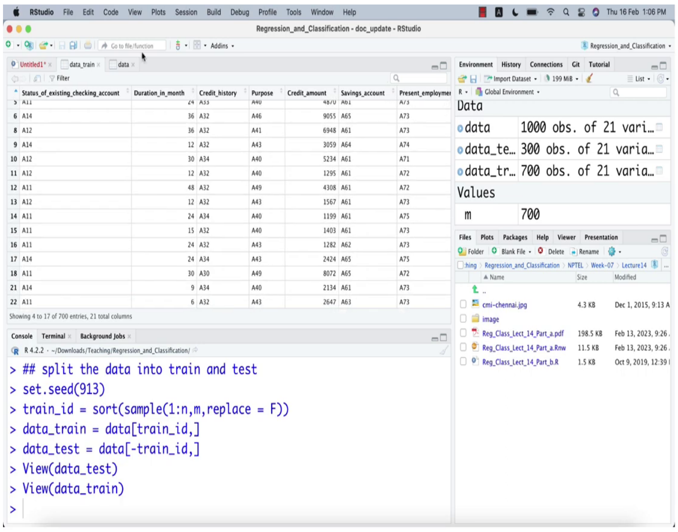
:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
> ## split the data into train and test
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
>
```

(Refer Slide Time: 07:57)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
Untitled1*
data_train
data
Filter
Environment History Connections Git Tutorial
Import Dataset
Global Environment
Files Plots Packages Help Viewer Presentation
Console Terminal Background Jobs

	Status_of_existing_checking_account	Duration_in_month	Credit_history	Purpose	Credit_amount	Savings_account	Present_employmer
5	A11	24	A33	A40	4870	A61	A73
6	A14	36	A32	A46	9055	A65	A73
8	A12	36	A32	A41	6948	A61	A73
9	A14	12	A32	A43	3059	A64	A74
10	A12	30	A34	A40	5234	A61	A71
11	A12	12	A32	A40	1295	A61	A72
12	A11	48	A32	A49	4308	A61	A72
13	A12	12	A32	A43	1567	A61	A73
14	A11	24	A34	A40	1199	A61	A75
15	A11	15	A32	A40	1403	A61	A73
16	A11	24	A32	A43	1282	A62	A73
17	A14	24	A34	A43	2424	A65	A75
18	A11	30	A30	A49	8072	A65	A72
21	A14	9	A34	A40	2134	A61	A73
22	A11	6	A32	A43	2647	A63	A73

Showing 4 to 17 of 700 entries, 21 total columns

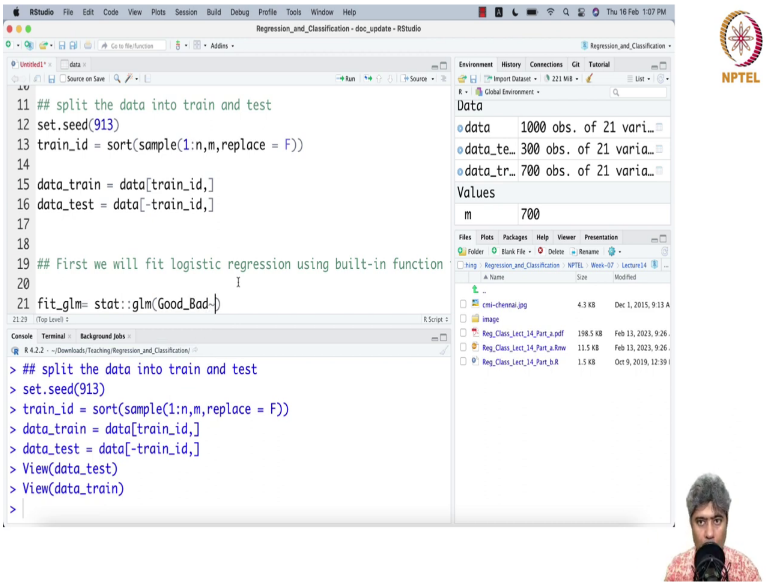
```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
> ## split the data into train and test
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
>
```

```
Data
data       1000 obs. of 21 vari…
data_te…   300 obs. of 21 varia…
data_tr…   700 obs. of 21 varia…
Values
m          700
```

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

So, that will automatically give me the testing data. And so, row number 1, 7, 19, 20, 26, this came to the test data. And if you go to the training data, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, then 19 missing 21. So, those are the train data, ok. So, now we have splitted the data. So, first we will fit a logistic regression using built-in function in R ok.

(Refer Slide Time: 08:19)



```
## split the data into train and test
set.seed(913)
train_id = sort(sample(1:n,m,replace = F))

data_train = data[train_id,]
data_test = data[-train_id,]


## First we will fit logistic regression using built-in function

fit_glm= stat::glm(Good_Bad~)
```

```
> ## split the data into train and test
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
>
```

Data

data	1000 obs. of 21 vari...
data_te...	300 obs. of 21 varia...
data_tr...	700 obs. of 21 varia...

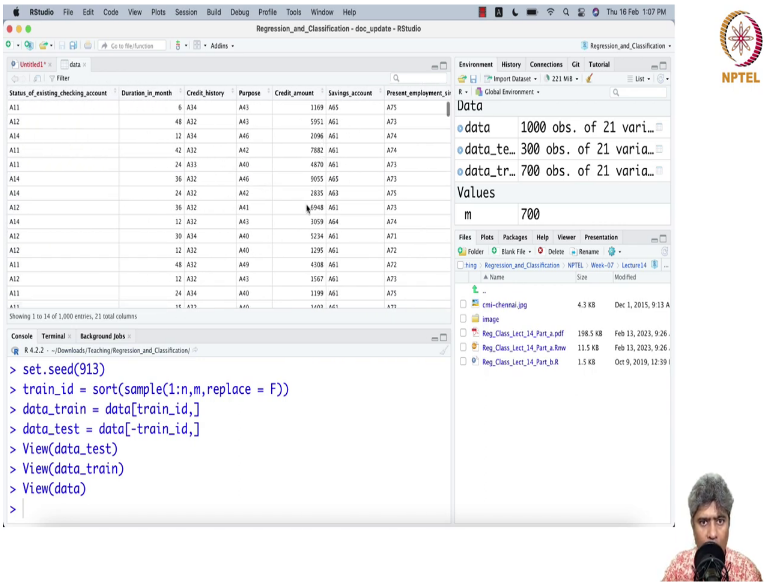
Values

m	700

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

First we will fit logistic regression using built-in function built-in function in R ok. Now, first fit glm. So, from the stats package, you extract the glm function, you call the glm function. So, you just say Good Bad and dollar.

(Refer Slide Time: 09:14)


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:07 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data
Filter
Environment History Connections Git Tutorial
Import Dataset
221 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
NPTEL

Status_of_existing_checking_account	Duration_in_month	Credit_history	Purpose	Credit_amount	Savings_account	Present_employment_si
A11	6	A34	A43	1169	A65	A75
A12	48	A32	A43	5951	A61	A73
A14	12	A34	A46	2096	A61	A74
A11	42	A32	A42	7882	A61	A74
A11	24	A33	A40	4870	A61	A73
A14	36	A32	A46	9055	A65	A73
A14	24	A32	A42	2835	A63	A75
A12	36	A32	A41	6948	A61	A73
A14	12	A32	A43	3059	A64	A74
A12	30	A34	A40	5234	A61	A71
A12	12	A32	A40	1295	A61	A72
A11	48	A32	A49	4308	A61	A72
A12	12	A32	A43	1567	A61	A73
A11	24	A34	A40	1199	A61	A75
A11	15	A32	A40	1403	A61	A73

Showing 1 to 14 of 1,000 entries, 21 total columns

Console Terminal Background Jobs

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
> View(data)
>
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…

Values

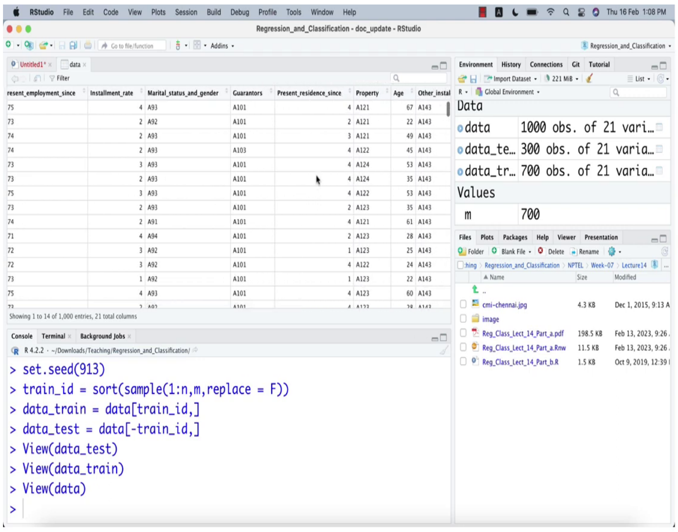
m	700

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

And I am going to take two in this data. There are most of these data are categorical, but I just want to keep it little simple. So, that, you know, you understand it more care more easily.

(Refer Slide Time: 09:40)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:08 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data
Filter
Environment History Connections Git Tutorial
Import Dataset
221 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs

resent_employment_since	Installment_rate	Marital_status_and_gender	Guarantors	Present_residence_since	Property	Age	Other_instal
75	4	A93	A101	4	A121	67	A143
73	2	A92	A101	2	A121	22	A143
74	2	A93	A101	3	A121	49	A143
74	2	A93	A103	4	A122	45	A143
73	3	A93	A101	4	A124	53	A143
73	2	A93	A101	4	A124	35	A143
75	3	A93	A101	4	A122	53	A143
73	2	A93	A101	2	A123	35	A143
74	2	A91	A101	4	A121	61	A143
71	4	A94	A101	2	A123	28	A143
72	3	A92	A101	1	A123	25	A143
72	3	A92	A101	4	A122	24	A143
73	1	A92	A101	1	A123	22	A143
75	4	A93	A101	4	A123	60	A143
73	2	A92	A101	4	A123	28	A143

Showing 1 to 14 of 1,000 entries, 21 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
> View(data)
> 
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…

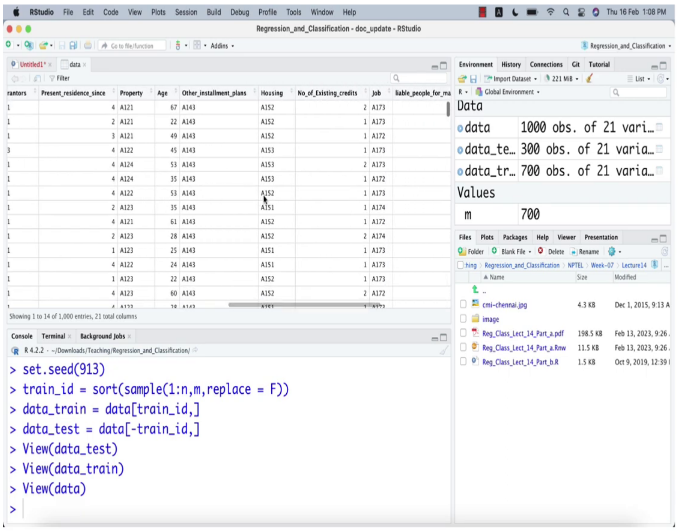
Values

m	700

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

(Refer Slide Time: 09:44)


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:08 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Untitled1*
data
Filter
Environment History Connections Git Tutorial
Import Dataset
221 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs

rantors	Present_residence_since	Property	Age	Other_installment_plans	Housing	No_of_Existing_credits	Job	liable_people_for_ma
1	4	A121	67	A143	A152	2	A173	
1	2	A121	22	A143	A152	1	A173	
1	3	A121	49	A143	A152	1	A172	
3	4	A122	45	A143	A153	1	A173	
1	4	A124	53	A143	A153	2	A173	
1	4	A124	35	A143	A153	1	A172	
1	4	A122	53	A143	A152	1	A173	
1	2	A123	35	A143	A151	1	A174	
1	4	A121	61	A143	A152	1	A172	
1	2	A123	28	A143	A152	2	A174	
1	1	A123	25	A143	A151	1	A173	
1	4	A122	24	A143	A151	1	A173	
1	1	A123	22	A143	A152	1	A173	
1	4	A123	60	A143	A152	2	A172	
1	4	A123	28	A143	A151	1	A173	

Showing 1 to 14 of 1,000 entries, 21 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
> View(data)
>
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…

Values

m	700

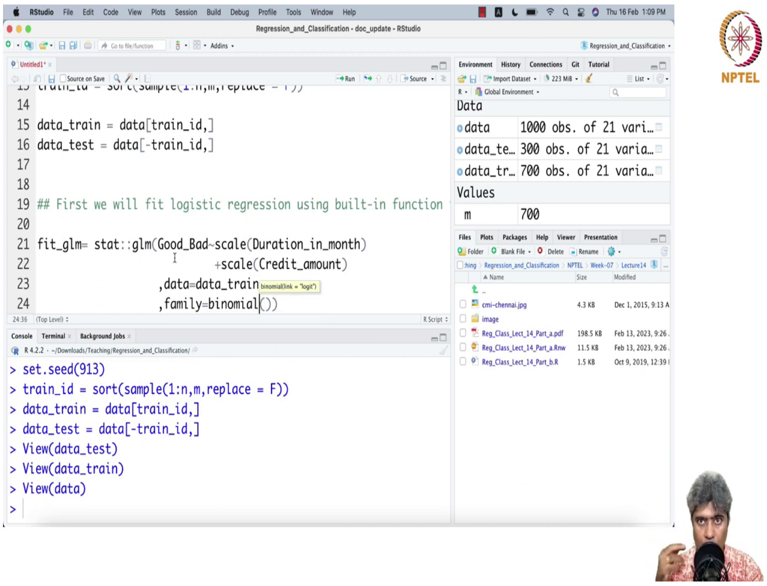
:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

The duration in month is a sort of a numeric value, you know, yeah, credit amount and duration in month. These are the 2 columns, which are in numeric. There are some other numeric also, installment rate, present, residence, since age. So, these also can be, I can take, but just to be, my purpose of this demo is to give you how the logistic regression works, ok.

You can add as many predictors as you want. There are about 20, 19 or 20 predictors are there in this data set. But you can add as many predictors as you want. My purpose is to just give you how the whole logistic regression works in R.

(Refer Slide Time: 10:20)



```
train_id = sort(sample(1:n,m,replace = F))

data_train = data[train_id,]
data_test = data[-train_id,]


## First we will fit logistic regression using built-in function

fit_glm= stat::glm(Good_Bad~scale(Duration_in_month)
                         +scale(Credit_amount)
                   ,data=data_train
                   ,family=binomial())
```

```
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
> View(data)
>
```

Data

```
data      1000 obs. of 21 vari…
data_te…  300 obs. of 21 varia…
data_tr…  700 obs. of 21 varia…
```

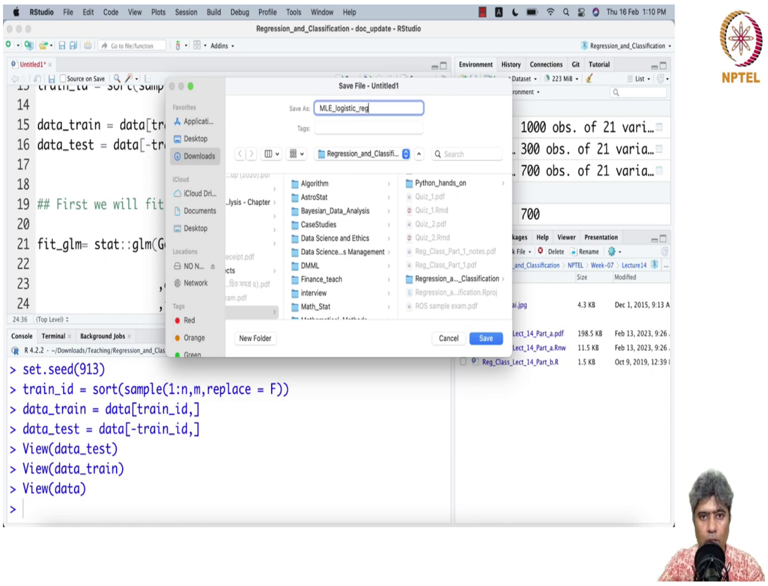
Values

```
m         700
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:09 PM
Regression_and_Classification - doc_update - RStudio
Untitled1*
Source on Save
Run
Source
binomial(link = "logit")
24:36 (Top Level)
R Script
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment History Connections Git Tutorial
Import Dataset
223 MiB
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
cmi-chennai.jpg
image
Reg_Class_Lect_14_Part_a.pdf
Reg_Class_Lect_14_Part_a.Rnw
Reg_Class_Lect_14_Part_b.R

And what I am going to do, I am just going to take 2 predictor, which are numeric predictor. And what I am going to do, I am going to just scale. This scale always helps Duration in month and plus scale Credit amount credit amount and then data equal to data train.

(Refer Slide Time: 11:26)



```
data_train = data[tr
data_test = data[-tr

## First we will fit

fit_glm= stat::glm(G
```

Save File - Untitled1

Save As: MLE_logistic_reg

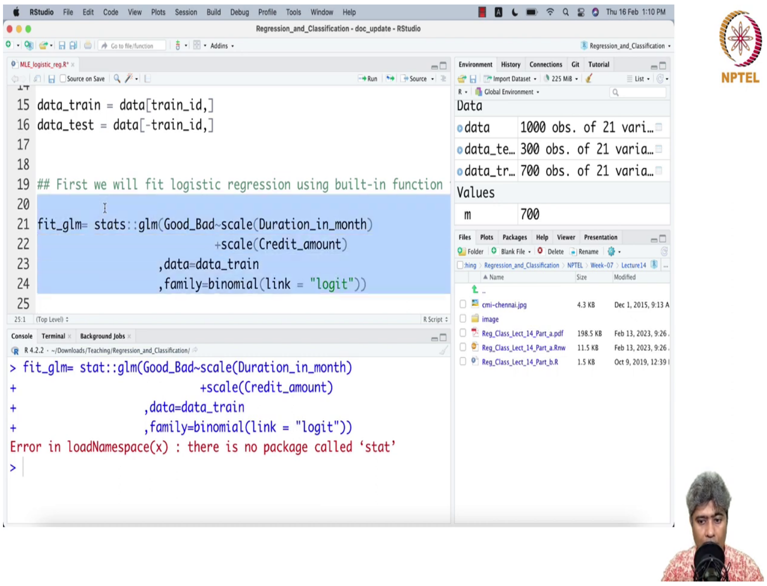
```
1000 obs. of 21 vari…
300 obs. of 21 varia…
700 obs. of 21 varia…
700
```

```
> set.seed(913)
> train_id = sort(sample(1:n,m,replace = F))
> data_train = data[train_id,]
> data_test = data[-train_id,]
> View(data_test)
> View(data_train)
> View(data)
>
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:10 PM
Regression_and_Classification - doc_update - RStudio
Untitled1*
Source on Save
Environment History Connections Git Tutorial
Favorites
Applicati…
Desktop
Downloads
iCloud
iCloud Dri…
Documents
Desktop
Locations
Network
Tags
Red
Orange
Green
Regression_and_Classifi…
Algorithm
AstroStat
Bayesian_Data_Analysis
CaseStudies
Data Science and Ethics
Data Science…s Management
DMML
Finance_teach
interview
Math_Stat
Python_hands_on
Quiz_1.pdf
Quiz_1.Rmd
Quiz_2.pdf
Quiz_2.Rmd
Reg_Class_Part_1_notes.pdf
Reg_Class_Part_1.pdf
Regression_a…_Classification
Regression_a…ification.Rproj
ROS sample exam.pdf
New Folder
Cancel
Save
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Clas
Help Viewer Presentation
Delete Rename
NPTEL Week-07 Lecture14
Size Modified
Lect_14_Part_a.pdf 198.5 KB Feb 13, 2023, 9:26
Lect_14_Part_a.Rnw 11.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R 1.5 KB Oct 9, 2019, 12:39
NPTEL

And now here is the important part that you have to mention family, family equal to binomial, because remember that it is a binary classification model. So, it raised into binomial I have to save it somewhere. Let me just call it. So, I will just call it maybe, MLE logistic regression logistic regression, ok binomial. So, since it is a binary classification problem, in the binary classification problem, the model automatically raise up what it comes is a binomial model or Bernoulli model.

(Refer Slide Time: 12:07)



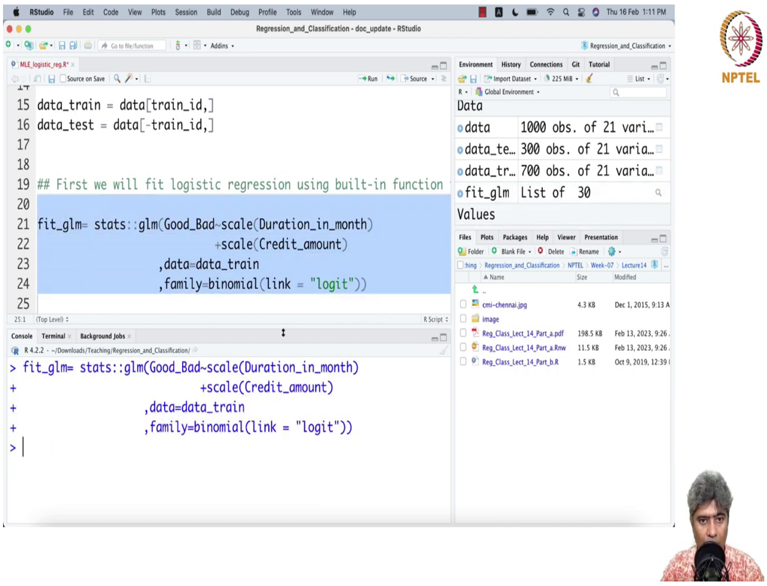
```
data_train = data[train_id,]
data_test = data[-train_id,]


## First we will fit logistic regression using built-in function

fit_glm= stats::glm(Good_Bad~scale(Duration_in_month)
                          +scale(Credit_amount)
                    ,data=data_train
                    ,family=binomial(link = "logit"))

```

```
> fit_glm= stat::glm(Good_Bad~scale(Duration_in_month)
+                           +scale(Credit_amount)
+                     ,data=data_train
+                     ,family=binomial(link = "logit"))
Error in loadNamespace(x) : there is no package called ‘stat’
> 
```

```
Data
data      1000 obs. of 21 vari…
data_te…  300 obs. of 21 varia…
data_tr…  700 obs. of 21 varia…
Values
m         700
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Console
Terminal
Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment
History
Connections
Git
Tutorial
Global Environment
Files
Plots
Packages
Help
Viewer
Presentation
cmi-chennai.jpg
image
Reg_Class_Lect_14_Part_a.pdf
Reg_Class_Lect_14_Part_a.Rnw
Reg_Class_Lect_14_Part_b.R
NPTEL

(Refer Slide Time: 12:35)



```
data_train = data[train_id,]
data_test = data[-train_id,]


## First we will fit logistic regression using built-in function

fit_glm= stats::glm(Good_Bad~scale(Duration_in_month)
                              +scale(Credit_amount)
                     ,data=data_train
                     ,family=binomial(link = "logit"))

```

```
> fit_glm= stats::glm(Good_Bad~scale(Duration_in_month)
+                              +scale(Credit_amount)
+                     ,data=data_train
+                     ,family=binomial(link = "logit"))
> 
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30

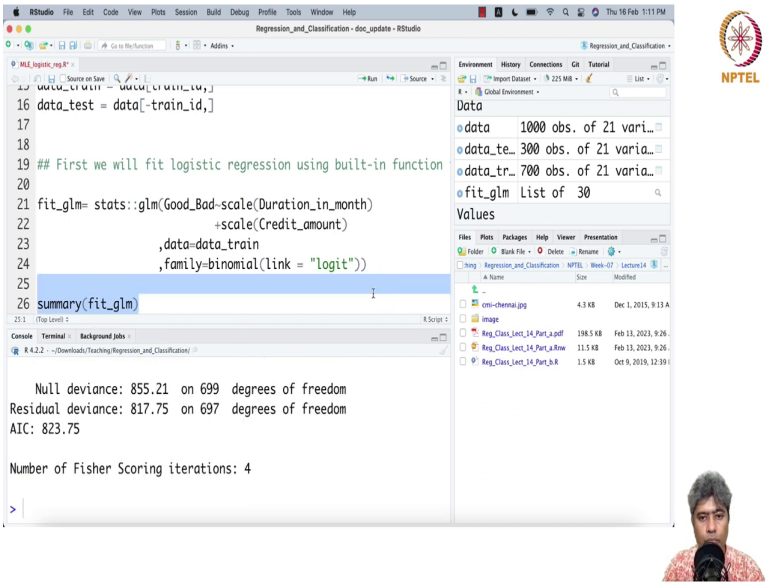
Values

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:11 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
225 MiB
Global Environment
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Files Plots Packages Help Viewer Presentation
Regression_and_Classification > NPTEL > Week-07 > Lecture14
NPTEL

So, and we have to give a link, link equal to, we will say, logit. You can use probit also. You can use probit also. And let me run. Let me run. It should be stats, not stat. Yeah, ok no error means probably it has run well.

(Refer Slide Time: 12:39)



```r
data_train = data[train_id,]
data_test = data[-train_id,]


## First we will fit logistic regression using built-in function

fit_glm= stats::glm(Good_Bad~scale(Duration_in_month)
                         +scale(Credit_amount)
                 ,data=data_train
                 ,family=binomial(link = "logit"))


summary(fit_glm)
```

```
    Null deviance: 855.21  on 699  degrees of freedom
Residual deviance: 817.75  on 697  degrees of freedom
AIC: 823.75

Number of Fisher Scoring iterations: 4

>
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30

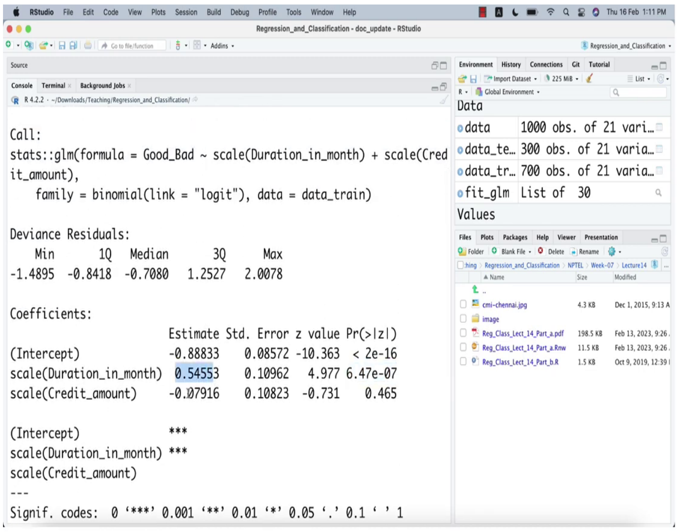
Values

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:11 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
225 MiB
List
Global Environment
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
R Script
NPTEL

(Refer Slide Time: 12:49)



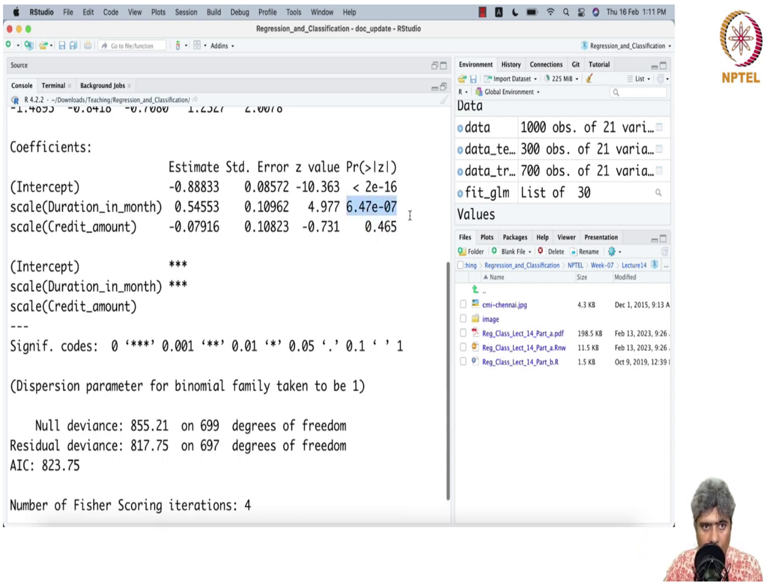
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:11 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
Source
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment History Connections Git Tutorial
Import Dataset
225 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14
Name Size Modified

```
Call:
stats::glm(formula = Good_Bad ~ scale(Duration_in_month) + scale(Cred
it_amount),
    family = binomial(link = "logit"), data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4895  -0.8418  -0.7080   1.2527   2.0078

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.88833    0.08572 -10.363  < 2e-16
scale(Duration_in_month)  0.54553    0.10962   4.977 6.47e-07
scale(Credit_amount)     -0.07916    0.10823  -0.731    0.465

(Intercept)              ***
scale(Duration_in_month) ***
scale(Credit_amount)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Data
data       1000 obs. of 21 vari…
data_te…   300 obs. of 21 varia…
data_tr…   700 obs. of 21 varia…
fit_glm    List of  30
Values
```

```
..
cmi-chennai.jpg                4.3 KB     Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf   198.5 KB   Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw   11.5 KB    Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R     1.5 KB     Oct 9, 2019, 12:39 I
```

(Refer Slide Time: 12:52)



```
-1.4895  -0.8418  -0.7080   1.2527   2.0078

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.88833    0.08572 -10.363  < 2e-16
scale(Duration_in_month) 0.54553    0.10962   4.977 6.47e-07
scale(Credit_amount)    -0.07916    0.10823  -0.731    0.465

(Intercept)             ***
scale(Duration_in_month) ***
scale(Credit_amount)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 855.21  on 699  degrees of freedom
Residual deviance: 817.75  on 697  degrees of freedom
AIC: 823.75

Number of Fisher Scoring iterations: 4
```

Data	
data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30

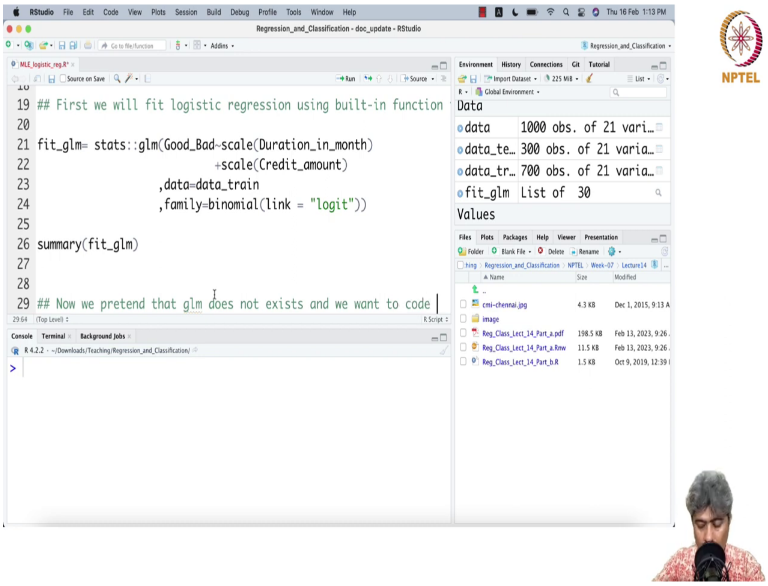
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
Source
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment History Connections Git Tutorial
Values
Files Plots Packages Help Viewer Presentation
NPTEL

Now, I am going to say summary, fit glm ok. So, it what it says that duration in month and credit amount, both of duration in month has a statistically significant m effect in the, in the whether a probe, customer is going to be a good customer or bad customer.

And turns out credit amount is not that important. You see, p value is very large, but credit amount is important. So, this is the estimate, the coefficient. These are the standard error and z value. So, you can have these, you can create these information’s. And now what I am going to do is now we are going to pretend that as if glm function in our does not exist and we want to code it from scratch ok. So, we are going to write down the negative log likelihood function.

And then we will use the optimization subroutine technique and we will find what estimates of the coefficients are. And what these estimates are going to be match going to match with built in glm function ok. So, we will first write negative log likelihood function ok, alright.

(Refer Slide Time: 14:21)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:13 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
225 MiB
List
Global Environment
NPTEL

```r
## First we will fit logistic regression using built-in function

fit_glm= stats::glm(Good_Bad~scale(Duration_in_month)
                              +scale(Credit_amount)
                    ,data=data_train
                    ,family=binomial(link = "logit"))

summary(fit_glm)


## Now we pretend that glm does not exists and we want to code
```

29:64 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
>
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30

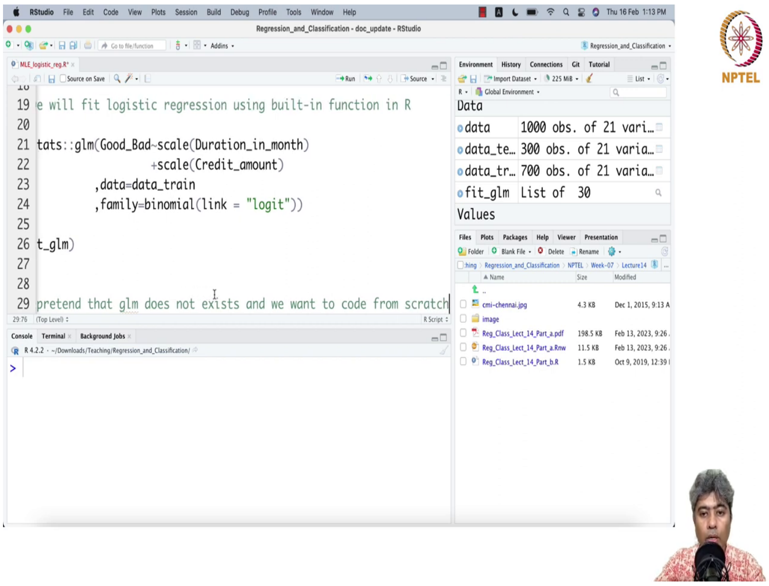
Values

Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename

:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

(Refer Slide Time: 14:54)



```
e will fit logistic regression using built-in function in R

tats::glm(Good_Bad~scale(Duration_in_month)
                    +scale(Credit_amount)
          ,data=data_train
          ,family=binomial(link = "logit"))

t_glm)


pretend that glm does not exists and we want to code from scratch
```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
>
```

Data

data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30

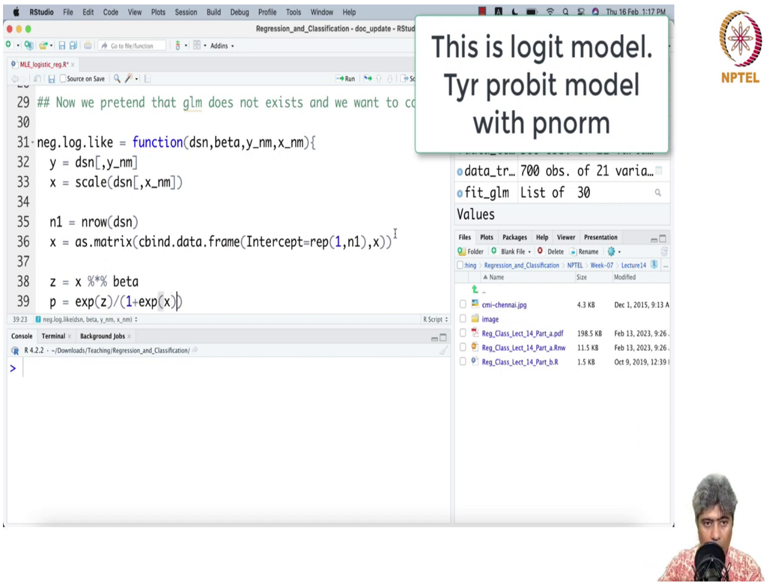
Values

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

Let me first, let me just write a point here that: Now we will we pretend that glm does not exist exists and we want to code from scratch ok.

(Refer Slide Time: 15:10)



This is logit model.
Tyr probit model
with pnorm

```
## Now we pretend that glm does not exists and we want to co

neg.log.like = function(dsn,beta,y_nm,x_nm){
  y = dsn[,y_nm]
  x = scale(dsn[,x_nm])

  n1 = nrow(dsn)
  x = as.matrix(cbind.data.frame(Intercept=rep(1,n1),x))

  z = x %*% beta
  p = exp(z)/(1+exp(x))
```

```
data_tr… 700 obs. of 21 varia…
fit_glm  List of  30
Values
```

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
>
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStud
MLE_logistic_reg.R*
Source on Save
Run
39:23 neg.log.like(dsn, beta, y_nm, x_nm)
R Script
Console Terminal Background Jobs
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14
NPTEL

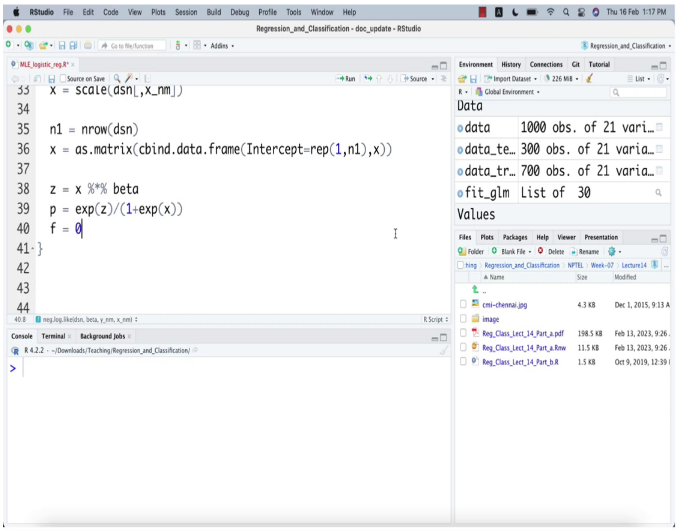
Now, we want to code from scratch. Now, first, how do you do that? Negative log likelihood you have to write the negative look likelihood function. So, there will be all the statement will be inside and then everything is first dataset name we have to give then beta values we have to give then y name we have to give and the x names we have to give correct, alright. So, the first from the we have to extract y from the dsn y name and keep it in x and then x equal to scale dsn x name.

So, the way I am doing it at this moment it will not work for categorical variable. So, it will work only for the numerical variable when writing the full blown code is not my purpose. And my purpose is to give you a demo that how if require you can write your own model you can develop your own model write down the likelihood function, pass it through the optimization sub routine and feed your own model and hey, here is your own model isn't it fun?

So, what I am going to do is first is n1 is nrow of dsn and then I am going to define it as remember that most likely dsn is going to be a data frame not most likely we are going to pass it as a data frame, but we need to do a matrix operation. Since we need to do a matrix operation what we need to do we have to call it as matrix this x as dot matrix. And we have to do this cbind dot data dot frame with x and before that Intercept equal to replicate of 1 comma n1.

So, we have to create a x matrix with all the x columns that has a name of the x columns with all the predictors name and a column, which will be intercept right. Now then once I have the x what I have to do I have to calculate for z remember that in the theory part we have a z value latent variable z. So, this is the latent variable I am going to introduce here. Percentage star percentage beta and then p equal to e to the power z divided by 1 plus e to the power z.

(Refer Slide Time: 18:50)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:17 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
226 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
R Script
40:8

```
  x = scale(dsn[,x_nm])

  n1 = nrow(dsn)
  x = as.matrix(cbind.data.frame(Intercept=rep(1,n1),x))

  z = x %*% beta
  p = exp(z)/(1+exp(x))
  f = 0
}
```

neg.log.like(dsn, beta, y_nm, x_nm)

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
>
```

Data

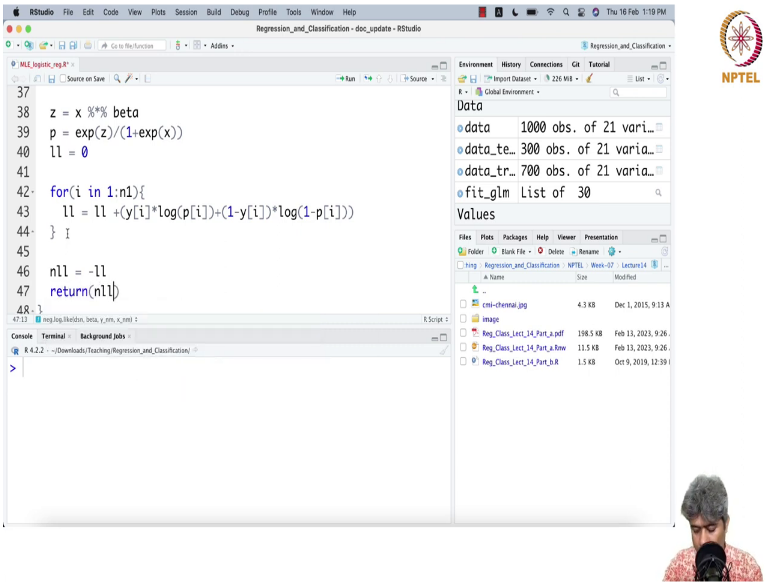
data	1000 obs. of 21 vari…
data_te…	300 obs. of 21 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30

Values

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

(Refer Slide Time: 18:55)



```
z = x %*% beta
p = exp(z)/(1+exp(x))
ll = 0

for(i in 1:n1){
  ll = ll +(y[i]*log(p[i])+(1-y[i])*log(1-p[i]))
}

nll = -ll
return(nll)
}
```

Data

data	1000 obs. of 21 vari...
data_te...	300 obs. of 21 varia...
data_tr...	700 obs. of 21 varia...
fit_glm	List of 30

Values

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

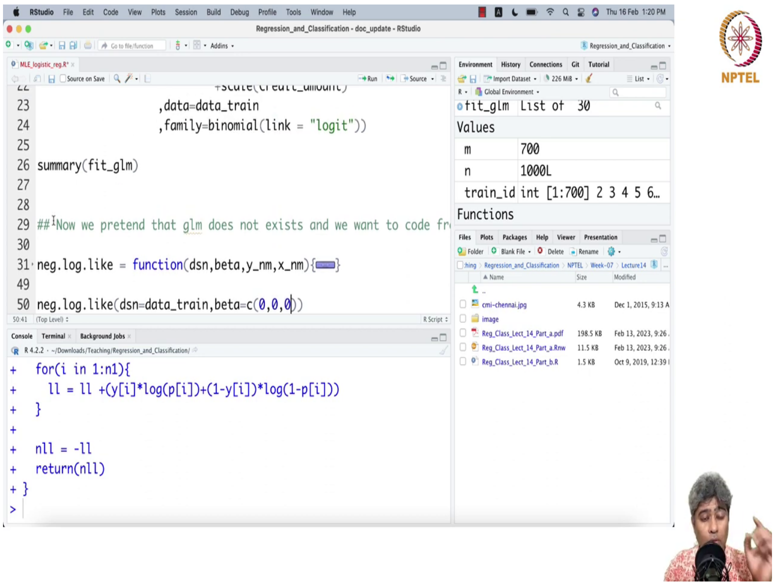
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:19 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Run
Source
Environment History Connections Git Tutorial
Global Environment
Files Plots Packages Help Viewer Presentation
Regression_and_Classification > NPTEL > Week-07 > Lecture14
47:13 neg.log.like(dsn, beta, y_nm, x_nm)
R Script
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
NPTEL

Now, I have the p wonderful now what I have to do is f equal to 0 I have to calculate the log likelihood actually I should say log likelihood ll equal to 0. For I could have done it slightly better way in a some kind of you know more computed efficient way, but my purpose is to show you not a write a write an extremely efficient way of writing the code when my purpose to show you how the concept of writing negative log likelihood.

And because the best thing is if I could have write this three lines of code in a vectorized form, but instead of writing it in a vectorized form I am purposefully writing it as in a loop so, that you can understand what I am doing. So, plus now y i times log p i plus 1 minus y i into log of 1 minus p i ok.

Now, here is a here is a task for you ok. Write this code three lines of code without for loop, can you do that and check if do the, if you do that whether your overall optimization time is going to be reduced or not. So, let me just let me just come finally, write negative log likelihood will be just minus of log likelihood. So, lll this is my log likelihood and just and then return negative log likelihood ok.

(Refer Slide Time: 21:34)



Regression_and_Classification - doc_update - RStudio

MLE_logistic_reg.R*

```
                  +scale(credit_amount)
                  ,data=data_train
                  ,family=binomial(link = "logit"))

summary(fit_glm)


##Now we pretend that glm does not exists and we want to code fr

neg.log.like = function(dsn,beta,y_nm,x_nm){…}

neg.log.like(dsn=data_train,beta=c(0,0,0))
```

Console

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
+   for(i in 1:n1){
+     ll = ll +(y[i]*log(p[i])+(1-y[i])*log(1-p[i]))
+   }
+
+   nll = -ll
+   return(nll)
+ }
>
```

Environment

fit_glm List of 30

Values

m	700
n	1000L
train_id	int [1:700] 2 3 4 5 6…

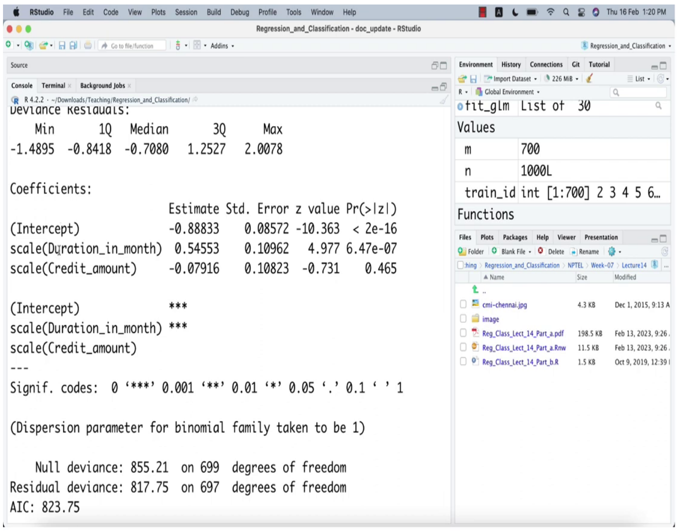
Functions

Files: Regression_and_Classification › NPTEL › Week-07 › Lecture14

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

Now, once I have written it once I have written the negative log likelihood at least we have to check whether the value the function works for some data set or not. So, some value some initial value dsn will be data underscore train then beta will be 0 comma 0 comma 0. So, why is 0 comma 0 comma 0? Because you see I have if I have in the model that I am trying to fit this model I am going to fit.

(Refer Slide Time: 22:08)



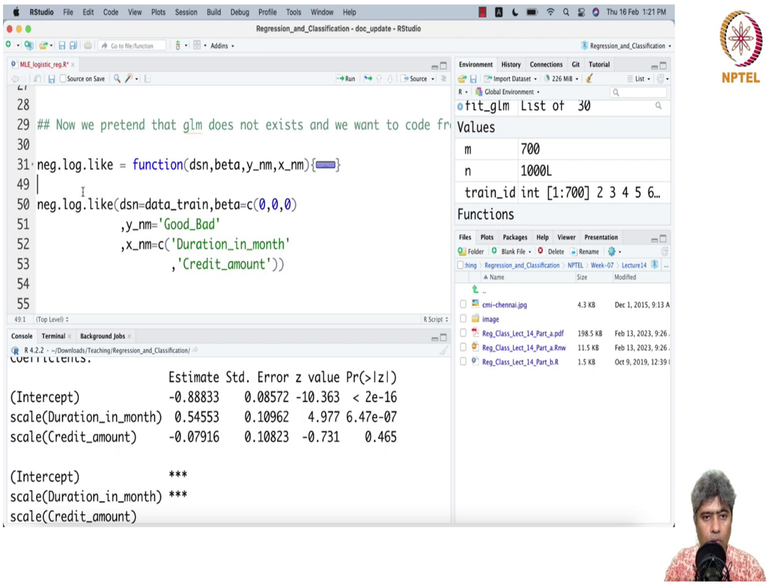
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4895  -0.8418  -0.7080   1.2527   2.0078

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.88833    0.08572 -10.363  < 2e-16
scale(Duration_in_month)  0.54553    0.10962   4.977 6.47e-07
scale(Credit_amount)    -0.07916    0.10823  -0.731    0.465

(Intercept)             ***
scale(Duration_in_month) ***
scale(Credit_amount)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 855.21  on 699  degrees of freedom
Residual deviance: 817.75  on 697  degrees of freedom
AIC: 823.75
```

fit_glm	List of 30
Values	
m	700
n	1000L
train_id	int [1:700] 2 3 4 5 6…
Functions	

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:20 PM
Regression_and_Classification - doc_update - RStudio
Regression_and_Classification
Source
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment History Connections Git Tutorial
Import Dataset 226 MiB List
R Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

(Refer Slide Time: 22:24)



```
## Now we pretend that glm does not exists and we want to code fr

neg.log.like = function(dsn,beta,y_nm,x_nm){}

neg.log.like(dsn=data_train,beta=c(0,0,0)
             ,y_nm='Good_Bad'
             ,x_nm=c('Duration_in_month'
                     ,'Credit_amount'))
```

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.88833    0.08572 -10.363  < 2e-16
scale(Duration_in_month)  0.54553    0.10962   4.977 6.47e-07
scale(Credit_amount)    -0.07916    0.10823  -0.731    0.465

(Intercept)             ***
scale(Duration_in_month) ***
scale(Credit_amount)
```

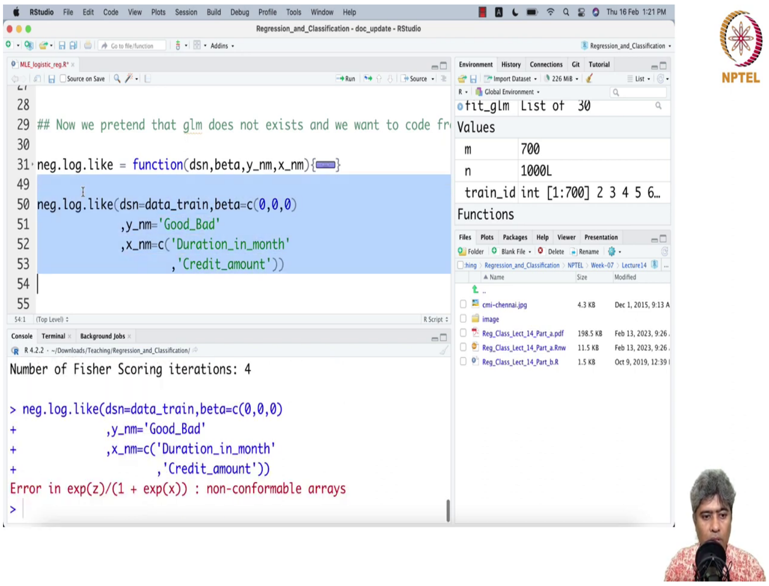
```
fit_glm   List of  30
Values
m         700
n         1000L
train_id  int [1:700] 2 3 4 5 6…
Functions
```

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

So, intercept one coefficient for Duration in month another coefficient Credit in amount. So, these are three coefficients for all three coefficients I am going to give initial value of 0. And then I have to give y name y name y name is in a vector form sorry y name is good or bad Good Bad and x name actually I can just copy these names Duration in month. Because that is the number exactly what we have and Credit amount ok.

(Refer Slide Time: 23:09)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:21 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
226 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Regression_and_Classification > NPTEL > Week-07 > Lecture14
Name Size Modified
cmi-chennai.jpg 4.3 KB Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf 198.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw 11.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R 1.5 KB Oct 9, 2019, 12:39
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
NPTEL

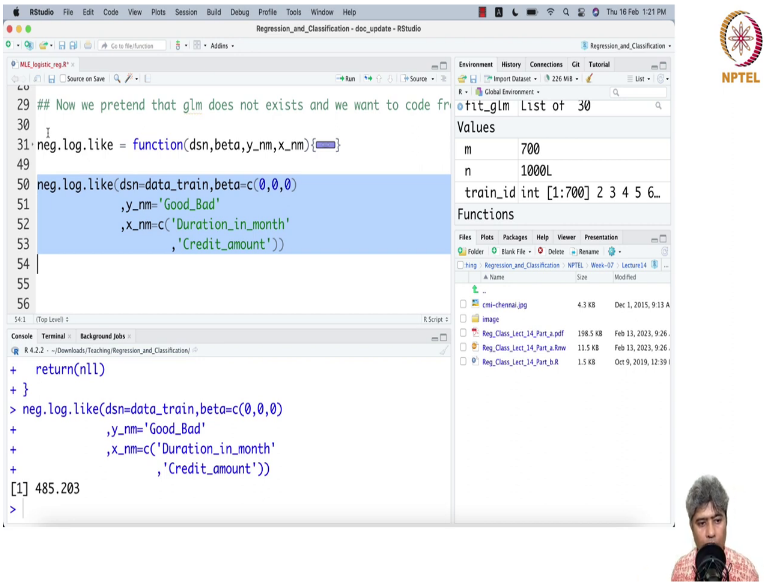
```
## Now we pretend that glm does not exists and we want to code fr

neg.log.like = function(dsn,beta,y_nm,x_nm){...}

neg.log.like(dsn=data_train,beta=c(0,0,0)
             ,y_nm='Good_Bad'
             ,x_nm=c('Duration_in_month'
                     ,'Credit_amount'))
```

```
fit_glm   List of  30
Values
  m       700
  n       1000L
  train_id int [1:700] 2 3 4 5 6…
Functions
```

```
Number of Fisher Scoring iterations: 4

> neg.log.like(dsn=data_train,beta=c(0,0,0)
+              ,y_nm='Good_Bad'
+              ,x_nm=c('Duration_in_month'
+                      ,'Credit_amount'))
Error in exp(z)/(1 + exp(x)) : non-conformable arrays
>
```

So, if I just run this these are the initial values this are called non-conformable let me see of course, because it should be z. So, there was a mistake alright.

(Refer Slide Time: 23:27)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:21 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
226 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
NPTEL

```
## Now we pretend that glm does not exists and we want to code fr

neg.log.like = function(dsn,beta,y_nm,x_nm){▭}

neg.log.like(dsn=data_train,beta=c(0,0,0)
             ,y_nm='Good_Bad'
             ,x_nm=c('Duration_in_month'
                     ,'Credit_amount'))
```

54:1 (Top Level) R Script

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

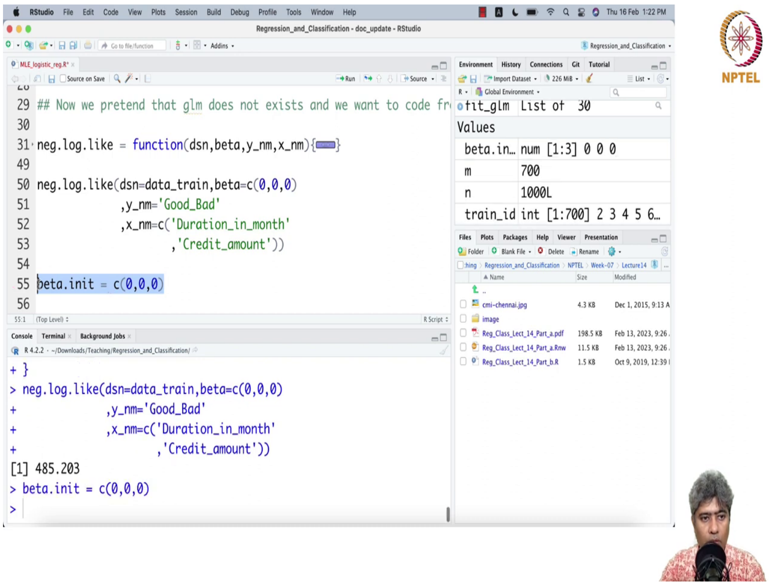
```
+   return(nll)
+ }
> neg.log.like(dsn=data_train,beta=c(0,0,0)
+              ,y_nm='Good_Bad'
+              ,x_nm=c('Duration_in_month'
+                      ,'Credit_amount'))
[1] 485.203
>
```

```
fit_glm    List of 30
Values
m          700
n          1000L
train_id   int [1:700] 2 3 4 5 6…
Functions
```

:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

(Refer Slide Time: 23:44)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R
Environment History Connections Git Tutorial
Files Plots Packages Help Viewer Presentation
Console Terminal Background Jobs
NPTEL

```
## Now we pretend that glm does not exists and we want to code fr
neg.log.like = function(dsn,beta,y_nm,x_nm){...}

neg.log.like(dsn=data_train,beta=c(0,0,0)
             ,y_nm='Good_Bad'
             ,x_nm=c('Duration_in_month'
                     ,'Credit_amount'))

beta.init = c(0,0,0)
```

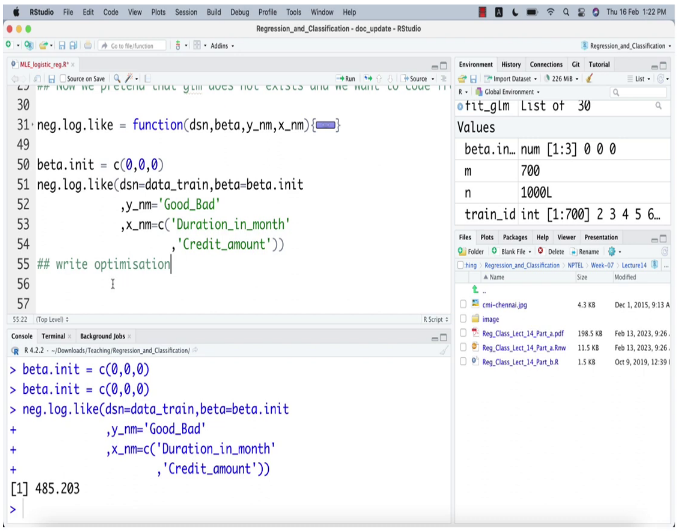
```
+ }
> neg.log.like(dsn=data_train,beta=c(0,0,0)
+              ,y_nm='Good_Bad'
+              ,x_nm=c('Duration_in_month'
+                      ,'Credit_amount'))
[1] 485.203
> beta.init = c(0,0,0)
>
```

```
fit_glm     List of 30
Values
beta.in..   num [1:3] 0 0 0
m           700
n           1000L
train_id    int [1:700] 2 3 4 5 6…
```

Name	Size	Modified
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

Let me run it once more yup. So, in the first initial value the negative log likelihood function runs and it gives me a value of the log negative log likelihood now I am going to initiate the initial value of beta I have to give initial value as 0, 0, 0 I have given here maybe I will just run it here and I will do it there.

(Refer Slide Time: 24:00)



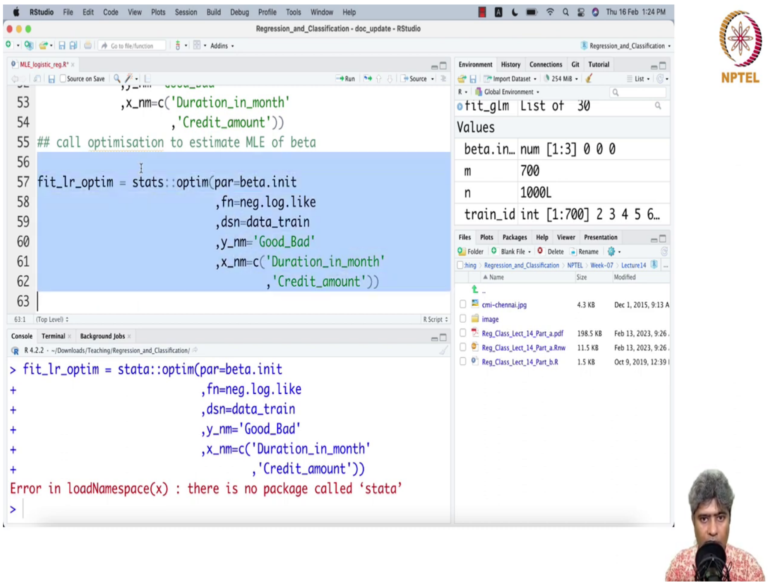
```r
## Now we pretend that glm does not exists and we want to code i

neg.log.like = function(dsn,beta,y_nm,x_nm){...}

beta.init = c(0,0,0)
neg.log.like(dsn=data_train,beta=beta.init
             ,y_nm='Good_Bad'
             ,x_nm=c('Duration_in_month'
                     ,'Credit_amount'))
## write optimisation


```

```
> beta.init = c(0,0,0)
> beta.init = c(0,0,0)
> neg.log.like(dsn=data_train,beta=beta.init
+              ,y_nm='Good_Bad'
+              ,x_nm=c('Duration_in_month'
+                      ,'Credit_amount'))
[1] 485.203
>
```

```
fit_glm     List of 30
Values
beta.in…    num [1:3] 0 0 0
m           700
n           1000L
train_id    int [1:700] 2 3 4 5 6…
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:22 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
226 MiB
Global Environment
List
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Regression_and_Classification > NPTEL > Week-07 > Lecture14
Name Size Modified
cmi-chennai.jpg 4.3 KB Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf 198.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw 11.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R 1.5 KB Oct 9, 2019, 12:39
55:22 (Top Level) R Script

(Refer Slide Time: 24:27)



```
                 ,x_nm=c('Duration_in_month'
                         ,'Credit_amount'))
## call optimisation to estimate MLE of beta

fit_lr_optim = stats::optim(par=beta.init
                            ,fn=neg.log.like
                            ,dsn=data_train
                            ,y_nm='Good_Bad'
                            ,x_nm=c('Duration_in_month'
                                    ,'Credit_amount'))
```

```
> fit_lr_optim = stata::optim(par=beta.init
+                             ,fn=neg.log.like
+                             ,dsn=data_train
+                             ,y_nm='Good_Bad'
+                             ,x_nm=c('Duration_in_month'
+                                     ,'Credit_amount'))
Error in loadNamespace(x) : there is no package called ‘stata’
>
```

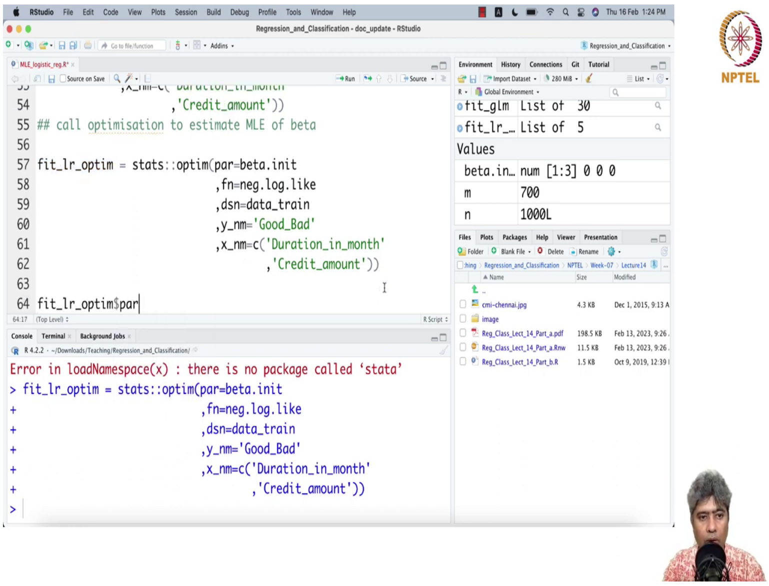
```
fit_glm      List of  30
Values
beta.in…  num [1:3] 0 0 0
m         700
n         1000L
train_id  int [1:700] 2 3 4 5 6…
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
R Script
Console
Terminal
Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment
History
Connections
Git
Tutorial
Global Environment
Files
Plots
Packages
Help
Viewer
Presentation
cmi-chennai.jpg
image
Reg_Class_Lect_14_Part_a.pdf
Reg_Class_Lect_14_Part_a.Rnw
Reg_Class_Lect_14_Part_b.R
NPTEL

And if I just run it yeah it is working perfectly fine and then I am going to write optimization sub routine to get the maximum likelihood estimate write optimization or just call optimization sub routine optimization to estimate MLE of beta ok. So, what I will do, fit lr logistic regression with optim and from the stats package you just call optim ok the parameter first is you have to give the parameter which is initialize beta dot init.

Then you have to give the objective function on what is your objective function? This is your objective function negative log likelihood function. So, you have to give the objective function then you have to give some additional in the objective function you have dsn data dot train. Then in the objective function you have this y name and x name that you have to provide ok.

(Refer Slide Time: 25:55)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:24 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
280 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
64:17 (Top Level)
R Script
NPTEL

```
                    ,'Credit_amount'))
## call optimisation to estimate MLE of beta

fit_lr_optim = stats::optim(par=beta.init
                            ,fn=neg.log.like
                            ,dsn=data_train
                            ,y_nm='Good_Bad'
                            ,x_nm=c('Duration_in_month'
                                    ,'Credit_amount'))

fit_lr_optim$par
```

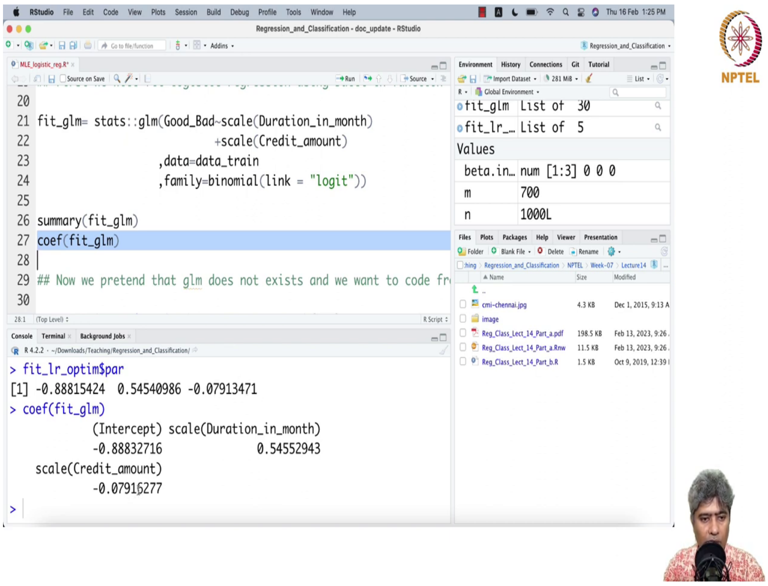
```
Error in loadNamespace(x) : there is no package called ‘stata’
> fit_lr_optim = stats::optim(par=beta.init
+                            ,fn=neg.log.like
+                            ,dsn=data_train
+                            ,y_nm='Good_Bad'
+                            ,x_nm=c('Duration_in_month'
+                                    ,'Credit_amount'))
>
```

```
fit_glm    List of  30
fit_lr_…   List of  5
Values
beta.in…   num [1:3] 0 0 0
m          700
n          1000L
```

```
:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14
Name                          Size      Modified
..
cmi-chennai.jpg               4.3 KB    Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf  198.5 KB  Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw  11.5 KB   Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R    1.5 KB    Oct 9, 2019, 12:39 I
```

So, and if you run this, I have made again a mistake it should be stats not stata and it has run and without any error. So, this is a good news.

(Refer Slide Time: 26:26)



```
fit_glm= stats::glm(Good_Bad~scale(Duration_in_month)
                             +scale(Credit_amount)
                    ,data=data_train
                    ,family=binomial(link = "logit"))

summary(fit_glm)
coef(fit_glm)

## Now we pretend that glm does not exists and we want to code fr
```

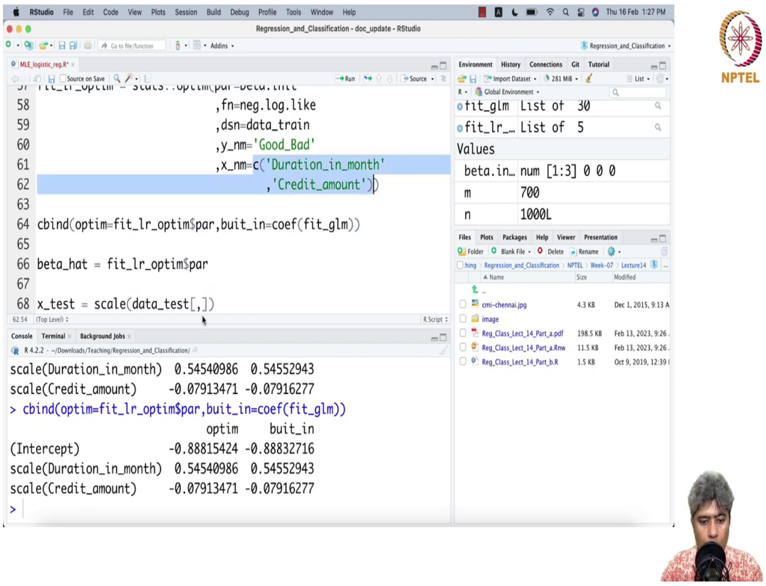
```
> fit_lr_optim$par
[1] -0.88815424  0.54540986 -0.07913471
> coef(fit_glm)
         (Intercept) scale(Duration_in_month)
         -0.88832716               0.54552943
   scale(Credit_amount)
         -0.07916277
>
```

```
fit_glm   List of 30
fit_lr_…  List of 5
Values
beta.in…  num [1:3] 0 0 0
m         700
n         1000L
```

And let me run from this let me call the parameter estimated parameter value. So, these are my estimated parameter values. So, these are my parameter values from manually estimated whereas, when we use the summary dot fit or maybe coefficients we just call coefficient of fit glm this is looks like these are exactly not exactly.

But at least up to three decimal places they are matching looks like 0791, 0791. So, up to 3 to 4 decimal places it is matching three 3 to 4 decimal places it is matching. There are ways to match it if I increase the tolerance level and all these things then it will match maybe up to a great extent.

(Refer Slide Time: 27:04)



```
fit_lr_optim = stats::optim(par=beta.init
                            ,fn=neg.log.like
                            ,dsn=data_train
                            ,y_nm='Good_Bad'
                            ,x_nm=c('Duration_in_month'
                                    ,'Credit_amount'))

cbind(optim=fit_lr_optim$par,buit_in=coef(fit_glm))

beta_hat = fit_lr_optim$par

x_test = scale(data_test[,])
```

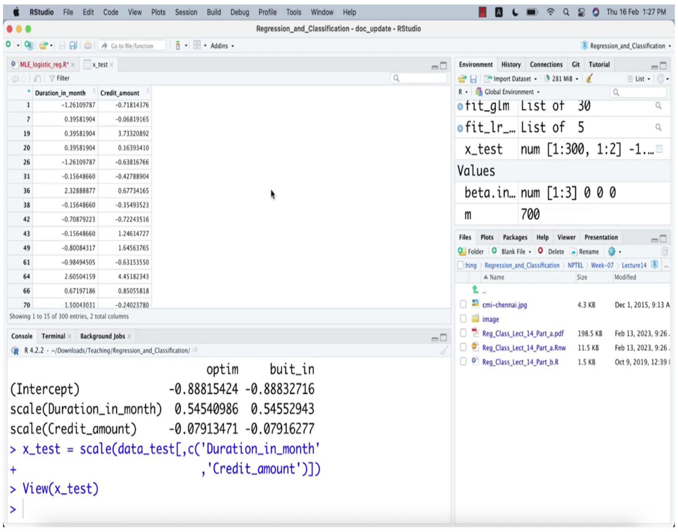
```
scale(Duration_in_month)  0.54540986  0.54552943
scale(Credit_amount)     -0.07913471 -0.07916277
> cbind(optim=fit_lr_optim$par,buit_in=coef(fit_glm))
                               optim     buit_in
(Intercept)              -0.88815424 -0.88832716
scale(Duration_in_month)  0.54540986  0.54552943
scale(Credit_amount)     -0.07913471 -0.07916277
>
```

```
fit_glm    List of 30
fit_lr_…   List of 5
Values
beta.in…   num [1:3] 0 0 0
m          700
n          1000L
```

So, what I will do if you. In fact, I can just say cbind and to just compare we can see. So, first one is from optim and second one is from the built in yeah. So, this is the when we use optim this is the value and when we use the built in r function glm this is what happens. So, they are almost matching up to the three decimal places and then they are deviating this one has matched up to four decimal places three to four decimal places they are matching.

If we increase the tolerance level probably, we will match even more. So, now, we have the let us write the beta hat so, since ok. So, beta hat equal to fit lr par and then x test x test we have to take from we have to scale it remember that we have to scale it data test comma these two value.

(Refer Slide Time: 28:41)



Regression_and_Classification - doc_update - RStudio

MLE_logistic_reg.R* | x_test

	Duration_in_month	Credit_amount
1	-1.26109787	-0.71814376
7	0.39581904	-0.06819165
19	0.39581904	3.73320892
20	0.39581904	0.16393410
26	-1.26109787	-0.63816766
31	-0.15648660	-0.42788904
36	2.32888877	0.67734165
38	-0.15648660	-0.35493523
42	-0.70879223	-0.72243516
43	-0.15648660	1.24614727
49	-0.80084317	1.64563765
61	-0.98494505	-0.63153550
64	2.60504159	4.45182343
66	0.67197186	0.85055818
70	1.50043031	-0.24023780

Showing 1 to 15 of 300 entries, 2 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
                              optim     buit_in
(Intercept)             -0.88815424 -0.88832716
scale(Duration_in_month)  0.54540986  0.54552943
scale(Credit_amount)    -0.07913471 -0.07916277
> x_test = scale(data_test[,c('Duration_in_month'
+                            ,'Credit_amount')])
> View(x_test)
> 
```

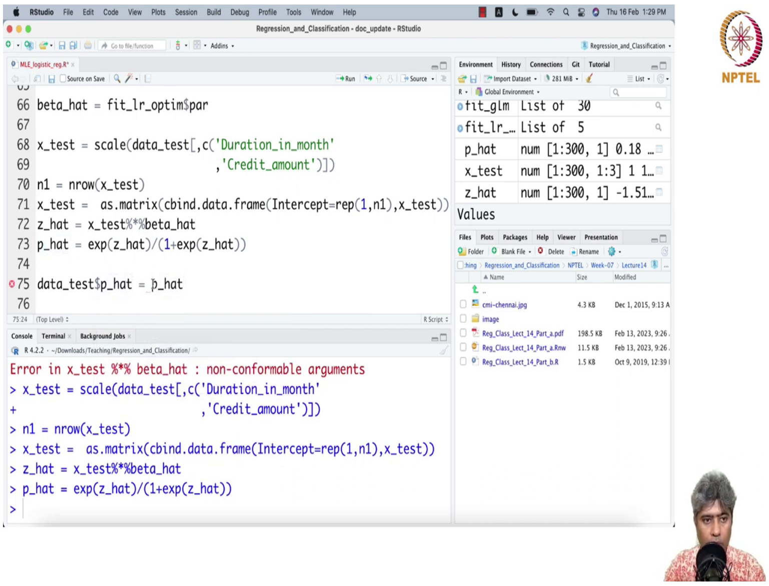
Global Environment

```
fit_glm     List of  30
fit_lr_…    List of  5
x_test      num [1:300, 1:2] -1.…
Values
beta.in…    num [1:3] 0 0 0
m           700
```

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

(Refer Slide Time: 28:58)



```r
beta_hat = fit_lr_optim$par

x_test = scale(data_test[,c('Duration_in_month'
                          ,'Credit_amount')])
n1 = nrow(x_test)
x_test =  as.matrix(cbind.data.frame(Intercept=rep(1,n1),x_test))
z_hat = x_test%*%beta_hat
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = p_hat

```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Error in x_test %*% beta_hat : non-conformable arguments
> x_test = scale(data_test[,c('Duration_in_month'
+                           ,'Credit_amount')])
> n1 = nrow(x_test)
> x_test =  as.matrix(cbind.data.frame(Intercept=rep(1,n1),x_test))
> z_hat = x_test%*%beta_hat
> p_hat = exp(z_hat)/(1+exp(z_hat))
>
```

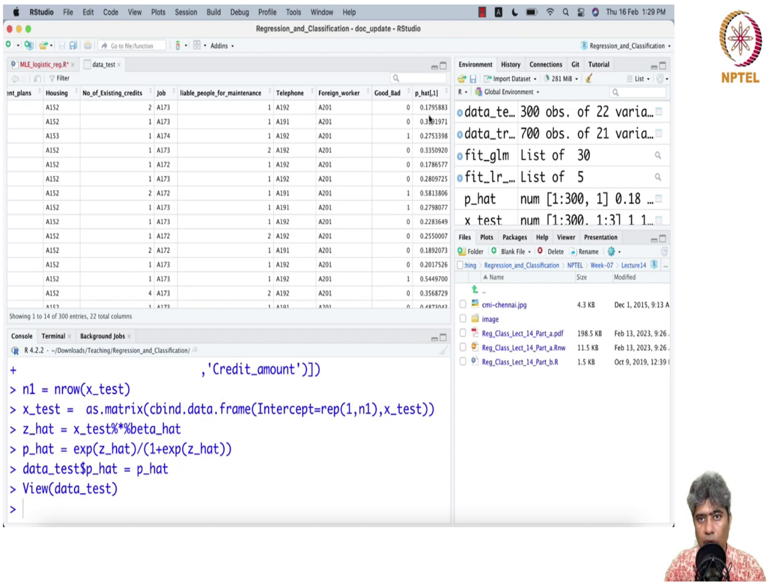
```
fit_glm    List of  30
fit_lr_…   List of  5
p_hat      num [1:300, 1] 0.18 …
x_test     num [1:300, 1:3] 1 1…
z_hat      num [1:300, 1] -1.51…
Values
```

```
:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14
..
cmi-chennai.jpg                 4.3 KB     Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf    198.5 KB   Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw    11.5 KB    Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R      1.5 KB     Oct 9, 2019, 12:39
```

So, we have to take the test ok we got the test values and then what I am going to do is we are going to calculate this first the latent variable z hat which is x test percentage star percentage beta hat. Sorry I have to call this guy and then yeah it is not confirmable. So, of course, because I have to do few more work in row of x test and then I have to call x test equal to you see I have done this line right.

So, I have to do the same kind of lining arrangement here also because that is not x test directly is not my you know my call. So, let me just run this and now probably it will be fine, yes now it is fine now it is fine. And now once I have the set test let me just have p hat p hat equals to e to the power z hat divided by 1 plus e to the power z hat ok. So, this is my p hat. Now, now in the data test data test I am going to p hat equals to p hat this is going to be a new column in the data test. See I have created a new column here alright.

(Refer Slide Time: 31:17)


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:29 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R*
data_test
Filter
NPTEL

ent_plans	Housing	No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad	p_hat[,1]
	A152	2	A173	1	A192	A201	0	0.1795883
	A152	1	A173	1	A191	A201	0	0.3391971
	A153	1	A174	1	A192	A201	1	0.2753398
	A152	1	A173	2	A192	A201	0	0.3350920
	A152	1	A172	1	A192	A201	0	0.1786577
	A152	1	A173	1	A192	A201	0	0.2809725
	A152	2	A172	1	A191	A201	1	0.5813806
	A152	1	A173	1	A191	A201	1	0.2798077
	A152	1	A173	1	A192	A201	0	0.2283649
	A152	1	A172	2	A192	A201	0	0.2550007
	A152	2	A172	1	A191	A201	0	0.1892073
	A152	1	A173	1	A192	A201	0	0.2017526
	A152	1	A173	1	A192	A201	1	0.5449700
	A152	4	A173	2	A192	A201	0	0.3568729
	A152	1	A173	1	A191	A201	0	0.4873043

Showing 1 to 14 of 300 entries, 22 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
+                              ,'Credit_amount')])
> n1 = nrow(x_test)
> x_test =  as.matrix(cbind.data.frame(Intercept=rep(1,n1),x_test))
> z_hat = x_test%*%beta_hat
> p_hat = exp(z_hat)/(1+exp(z_hat))
> data_test$p_hat = p_hat
> View(data_test)
> 
```

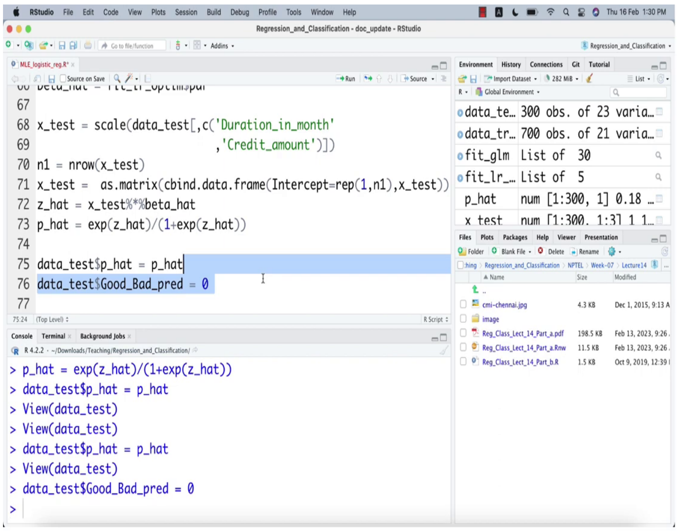
```
data_te…   300 obs. of 22 varia…
data_tr…   700 obs. of 21 varia…
fit_glm    List of  30
fit_lr_…   List of  5
p_hat      num [1:300, 1] 0.18 …
x_test     num [1:300, 1:3] 1 1…
```

:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

Next what I am going to do I am not sure why it shows a p hat column 1 ok, we will see what can be done alright. Next is we have to predict.

(Refer Slide Time: 32:08)



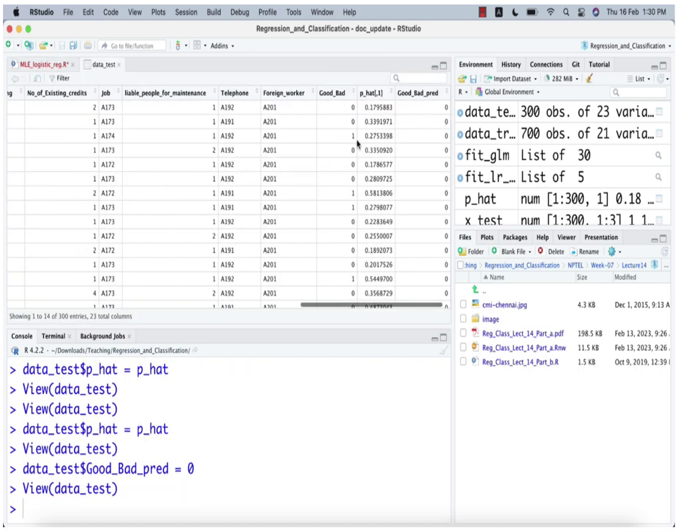
```r
beta_hat = fit_lr_optim$par

x_test = scale(data_test[,c('Duration_in_month'
                            ,'Credit_amount')])
n1 = nrow(x_test)
x_test =  as.matrix(cbind.data.frame(Intercept=rep(1,n1),x_test))
z_hat = x_test%*%beta_hat
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = p_hat
data_test$Good_Bad_pred = 0

```

```
> p_hat = exp(z_hat)/(1+exp(z_hat))
> data_test$p_hat = p_hat
> View(data_test)
> View(data_test)
> data_test$p_hat = p_hat
> View(data_test)
> data_test$Good_Bad_pred = 0
> 
```

```
data_te…  300 obs. of 23 varia…
data_tr…  700 obs. of 21 varia…
fit_glm   List of  30
fit_lr_…  List of  5
p_hat     num [1:300, 1] 0.18 …
x_test    num [1:300, 1:3] 1 1…
```

	Name	Size	Modified
	..		
□	cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
□	image		
□	Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26
□	Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26
□	Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

(Refer Slide Time: 32:24)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:30 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R*
data_test
Filter
Environment History Connections Git Tutorial
Import Dataset
282 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs

No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad	p_hat[,1]	Good_Bad_pred
2	A173	1	A192	A201	0	0.1795883	0
1	A173	1	A191	A201	0	0.3391971	0
1	A174	1	A192	A201	1	0.2753398	0
1	A173	2	A192	A201	0	0.3350920	0
1	A172	1	A192	A201	0	0.1786577	0
1	A173	1	A192	A201	0	0.2809725	0
2	A172	1	A191	A201	1	0.5813806	0
1	A173	1	A191	A201	1	0.2798077	0
1	A173	1	A192	A201	0	0.2283649	0
1	A172	2	A192	A201	0	0.2550007	0
2	A172	1	A191	A201	0	0.1892073	0
1	A173	1	A192	A201	0	0.2017526	0
1	A173	1	A192	A201	1	0.5449700	0
4	A173	2	A192	A201	0	0.3568729	0

Showing 1 to 14 of 300 entries, 23 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

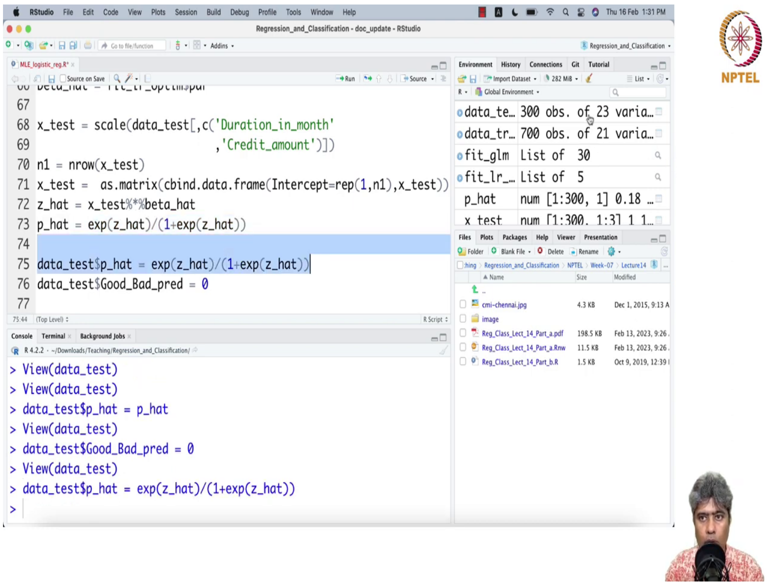
```
> data_test$p_hat = p_hat
> View(data_test)
> View(data_test)
> data_test$p_hat = p_hat
> View(data_test)
> data_test$Good_Bad_pred = 0
> View(data_test)
>
```

data_te…	300 obs. of 23 varia…	
data_tr…	700 obs. of 21 varia…	
fit_glm	List of 30	
fit_lr_…	List of 5	
p_hat	num [1:300, 1] 0.18 …	
x_test	num [1:300, 1:3] 1 1…	

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39

(Refer Slide Time: 32:32)



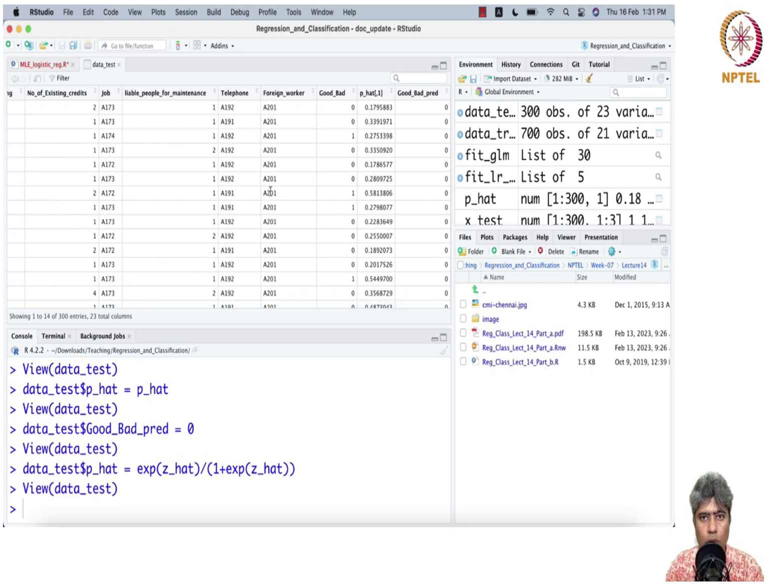
```
beta_hat = fit_lr_optim$par

x_test = scale(data_test[,c('Duration_in_month'
                            ,'Credit_amount')])
n1 = nrow(x_test)
x_test =  as.matrix(cbind.data.frame(Intercept=rep(1,n1),x_test))
z_hat = x_test%*%beta_hat
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
data_test$Good_Bad_pred = 0

```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
> View(data_test)
> View(data_test)
> data_test$p_hat = p_hat
> View(data_test)
> data_test$Good_Bad_pred = 0
> View(data_test)
> data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
>
```

```
data_te…  300 obs. of 23 varia…
data_tr…  700 obs. of 21 varia…
fit_glm   List of  30
fit_lr_…  List of  5
p_hat     num [1:300, 1] 0.18 …
x_test    num [1:300, 1:3] 1 1…
```

```
:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14
Name                          Size      Modified
..
cmi-chennai.jpg               4.3 KB    Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf  198.5 KB  Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw  11.5 KB   Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R    1.5 KB    Oct 9, 2019, 12:39
```

(Refer Slide Time: 32:38)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:31 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R*
data_test
Filter
Environment History Connections Git Tutorial
Import Dataset
282 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
NPTEL

ng	No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad	p_hat[,1]	Good_Bad_pred
	2	A173	1	A192	A201	0	0.1795883	0
	1	A173	1	A191	A201	0	0.3391971	0
	1	A174	1	A192	A201	1	0.2753398	0
	1	A173	2	A192	A201	0	0.3350920	0
	1	A172	1	A192	A201	0	0.1786577	0
	1	A173	1	A192	A201	0	0.2809725	0
	2	A172	1	A191	A201	1	0.5813806	0
	1	A173	1	A191	A201	1	0.2798077	0
	1	A173	1	A192	A201	0	0.2283649	0
	1	A172	2	A192	A201	0	0.2550007	0
	2	A172	1	A191	A201	0	0.1892073	0
	1	A173	1	A192	A201	0	0.2017526	0
	1	A173	1	A192	A201	1	0.5449700	0
	4	A173	2	A192	A201	0	0.3568729	0
	1	A173	1	A191	A201	0	0.4873043	0

Showing 1 to 14 of 300 entries, 23 total columns

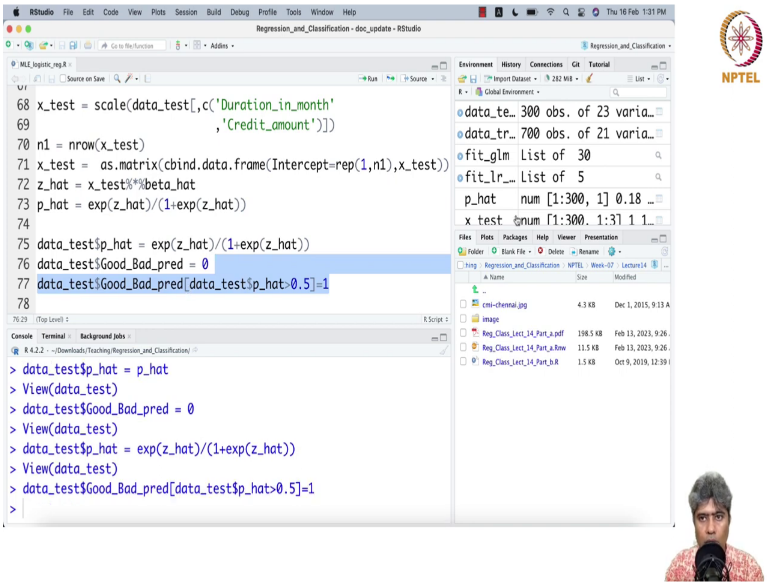
```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
> View(data_test)
> data_test$p_hat = p_hat
> View(data_test)
> data_test$Good_Bad_pred = 0
> View(data_test)
> data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
> View(data_test)
>
```

data_te…	300 obs. of 23 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30
fit_lr_…	List of 5
p_hat	num [1:300, 1] 0.18 …
x_test	num [1:300, 1:3] 1 1…

:hing > Regression_and_Classification > NPTEL > Week-07 > Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

(Refer Slide Time: 32:45)



```r
x_test = scale(data_test[,c('Duration_in_month'
                            ,'Credit_amount')])
n1 = nrow(x_test)
x_test =  as.matrix(cbind.data.frame(Intercept=rep(1,n1),x_test))
z_hat = x_test%*%beta_hat
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
data_test$Good_Bad_pred = 0
data_test$Good_Bad_pred[data_test$p_hat>0.5]=1

```

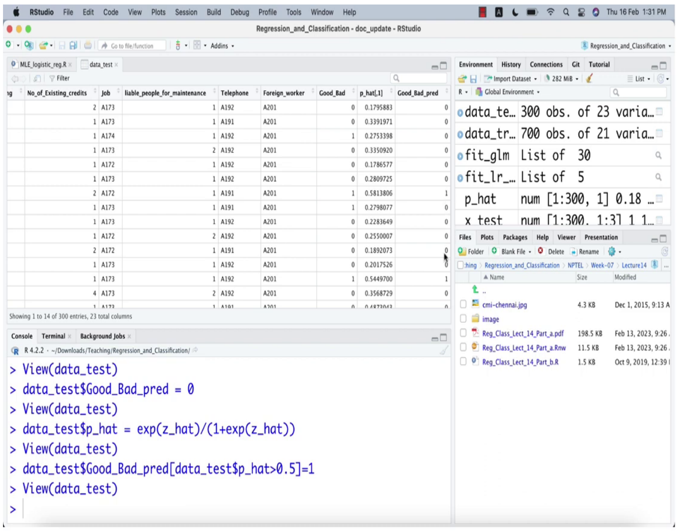
```
> data_test$p_hat = p_hat
> View(data_test)
> data_test$Good_Bad_pred = 0
> View(data_test)
> data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
> View(data_test)
> data_test$Good_Bad_pred[data_test$p_hat>0.5]=1
>
```

```
data_te… 300 obs. of 23 varia…
data_tr… 700 obs. of 21 varia…
fit_glm  List of  30
fit_lr_… List of  5
p_hat    num [1:300, 1] 0.18 …
x_test   num [1:300, 1:3] 1 1…
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:31 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R
Source on Save
Run
Source
76:29 (Top Level)
R Script
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment History Connections Git Tutorial
Import Dataset
282 MiB
Global Environment
Files Plots Packages Help Viewer Presentation
Regression_and_Classification › NPTEL › Week-07 › Lecture14
cmi-chennai.jpg 4.3 KB Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf 198.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw 11.5 KB Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R 1.5 KB Oct 9, 2019, 12:39
NPTEL

So, data test dollar Good Bad, but we have to do first prediction the prediction equal to first I am saying 0 and then first I am taking all the values of sort of a 0. Actually, what I can do? I can just take this yeah, I will see what (Refer Time: 32:44) can do and then I can just say that if p hat is greater than 1 greater than half. Then most likely its a going to be a bad customer.

(Refer Slide Time: 33:09)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:31 PM
Regression_and_Classification - doc_update - RStudio
Go to file/function
Addins
Regression_and_Classification
MLE_logistic_reg.R
data_test
Filter
Environment History Connections Git Tutorial
Import Dataset
282 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs

ıg	No_of_Existing_credits	Job	liable_people_for_maintenance	Telephone	Foreign_worker	Good_Bad	p_hat[,1]	Good_Bad_pred
	2	A173	1	A192	A201	0	0.1795883	0
	1	A173	1	A191	A201	0	0.3391971	0
	1	A174	1	A192	A201	1	0.2753398	0
	1	A173	2	A192	A201	0	0.3350920	0
	1	A172	1	A192	A201	0	0.1786577	0
	1	A173	1	A192	A201	0	0.2809725	0
	2	A172	1	A191	A201	1	0.5813806	1
	1	A173	1	A191	A201	1	0.2798077	0
	1	A173	1	A192	A201	0	0.2283649	0
	1	A172	2	A192	A201	0	0.2550007	0
	2	A172	1	A191	A201	0	0.1892073	0
	1	A173	1	A192	A201	0	0.2017526	0
	1	A173	1	A192	A201	1	0.5449700	1
	4	A173	2	A192	A201	0	0.3568729	0
	1	A173	1	A191	A201	0	0.4873043	0

Showing 1 to 14 of 300 entries, 23 total columns

data_te…	300 obs. of 23 varia…
data_tr…	700 obs. of 21 varia…
fit_glm	List of 30
fit_lr_…	List of 5
p_hat	num [1:300, 1] 0.18 …
x_test	num [1:300, 1:3] 1 1…

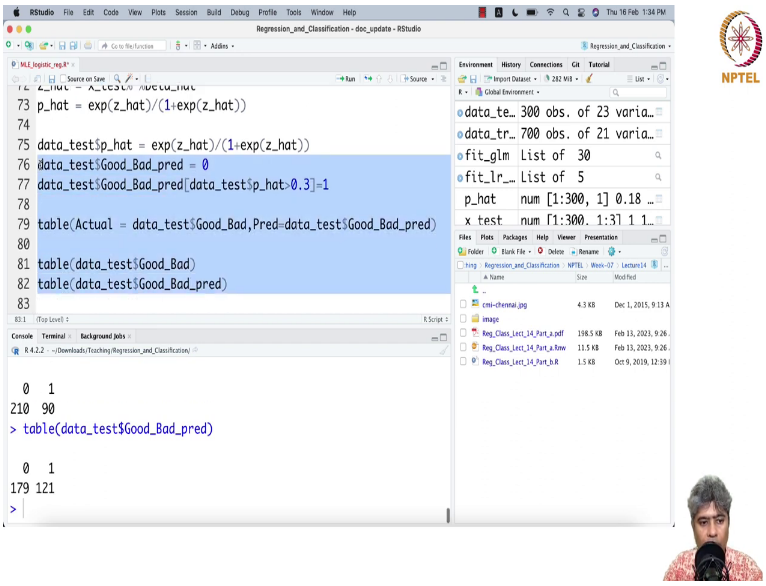
:hing › Regression_and_Classification › NPTEL › Week-07 › Lecture14

Name	Size	Modified
..		
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 A
image		
Reg_Class_Lect_14_Part_a.pdf	198.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_a.Rnw	11.5 KB	Feb 13, 2023, 9:26 .
Reg_Class_Lect_14_Part_b.R	1.5 KB	Oct 9, 2019, 12:39 I

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

```
> View(data_test)
> data_test$Good_Bad_pred = 0
> View(data_test)
> data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
> View(data_test)
> data_test$Good_Bad_pred[data_test$p_hat>0.5]=1
> View(data_test)
>
```

(Refer Slide Time: 33:21)



```
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
data_test$Good_Bad_pred = 0
data_test$Good_Bad_pred[data_test$p_hat>0.3]=1

table(Actual = data_test$Good_Bad,Pred=data_test$Good_Bad_pred)

table(data_test$Good_Bad)
table(data_test$Good_Bad_pred)
```

```
  0   1
210  90
> table(data_test$Good_Bad_pred)

  0   1
179 121
>
```

data_te...	300 obs. of 23 varia...
data_tr...	700 obs. of 21 varia...
fit_glm	List of 30
fit_lr_...	List of 5
p_hat	num [1:300, 1] 0.18 ...
x_test	num [1:300, 1:3] 1 1...

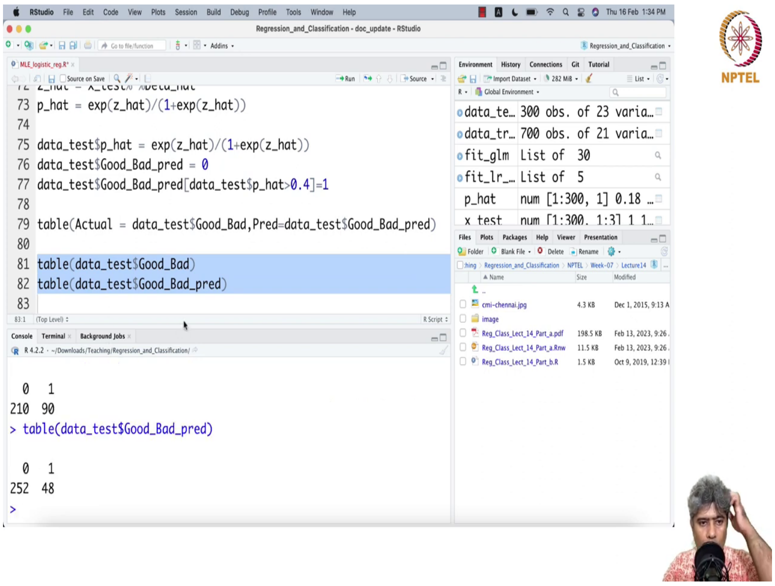
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:34 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
282 MiB
Global Environment
Console Terminal Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Files Plots Packages Help Viewer Presentation
Regression_and_Classification > NPTEL > Week-07 > Lecture14
cmi-chennai.jpg
image
Reg_Class_Lect_14_Part_a.pdf
Reg_Class_Lect_14_Part_a.Rnw
Reg_Class_Lect_14_Part_b.R
NPTEL

So, there are some prediction where it has said it is going to be a bad customer ok. Now I am going to do a table I call accuracy actual equals to data test dollar Good Bad and prediction predict equal to data test dollar Good Bad predicted values and this is the case. So, there are out of about 300, 212 cases where 202 plus 210 there are too many like bad customer is being missed and said that ok these are Good.

So, it is looks like the model is predicting most of the cases are as good 2000 out of 300 test cases. So, in the test cases test data set data test if I just do a simple table now data test comma good bad they are about 210 cases are good and 90 cases are bad. Now on the other hand if I just do a table of data test Good Bad pred. So, 282 cases are saying that they are Good and 18 cases they are saying Bad. So, this model is actually failing to you know failing to say who is Good who is Bad.

So, one thing people do try to understand that if I in case of 0.5 if I increase it to say 0.3 and will it be helpful. So, if you do that say probably yes, instead of 0.5 if you just reduce it now you have just too many probably saying or maybe 0.4. So, you have these many cases.

(Refer Slide Time: 36:07)



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 16 Feb 1:34 PM
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
Environment History Connections Git Tutorial
Import Dataset
282 MiB
Global Environment
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Console Terminal Background Jobs
NPTEL

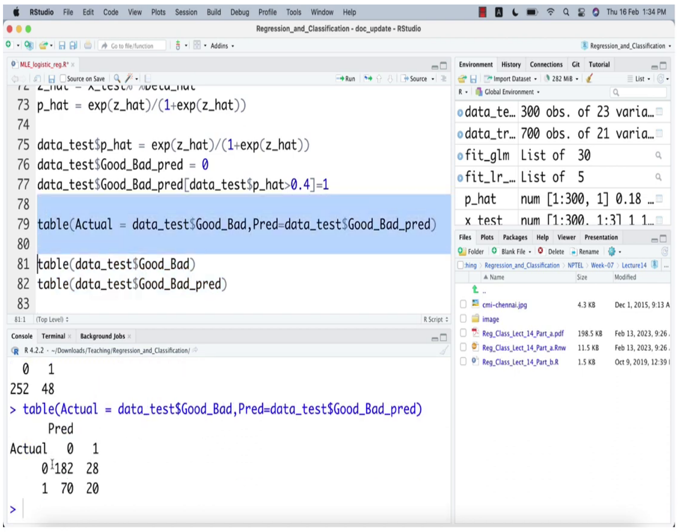
```
z_hat = x_test%*%beta_hat
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
data_test$Good_Bad_pred = 0
data_test$Good_Bad_pred[data_test$p_hat>0.4]=1

table(Actual = data_test$Good_Bad,Pred=data_test$Good_Bad_pred)

table(data_test$Good_Bad)
table(data_test$Good_Bad_pred)
```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/

  0   1
210  90
> table(data_test$Good_Bad_pred)

  0   1
252  48
>
```

```
data_te… 300 obs. of 23 varia…
data_tr… 700 obs. of 21 varia…
fit_glm  List of  30
fit_lr_… List of  5
p_hat    num [1:300, 1] 0.18 …
x_test   num [1:300, 1:3] 1 1…
```

```
Name                         Size      Modified
cmi-chennai.jpg              4.3 KB    Dec 1, 2015, 9:13 A
image
Reg_Class_Lect_14_Part_a.pdf 198.5 KB  Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_a.Rnw 11.5 KB   Feb 13, 2023, 9:26
Reg_Class_Lect_14_Part_b.R   1.5 KB    Oct 9, 2019, 12:39
```

(Refer Slide Time: 36:30)



```
p_hat = exp(z_hat)/(1+exp(z_hat))

data_test$p_hat = exp(z_hat)/(1+exp(z_hat))
data_test$Good_Bad_pred = 0
data_test$Good_Bad_pred[data_test$p_hat>0.4]=1

table(Actual = data_test$Good_Bad,Pred=data_test$Good_Bad_pred)

table(data_test$Good_Bad)
table(data_test$Good_Bad_pred)
```

```
  0   1
252  48
> table(Actual = data_test$Good_Bad,Pred=data_test$Good_Bad_pred)
      Pred
Actual   0   1
     0 182  28
     1  70  20
>
```

```
data_te… 300 obs. of 23 varia…
data_tr… 700 obs. of 21 varia…
fit_glm  List of  30
fit_lr_… List of  5
p_hat    num [1:300, 1] 0.18 …
x_test   num [1:300, 1:3] 1 1…
```

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Regression_and_Classification - doc_update - RStudio
MLE_logistic_reg.R*
Source on Save
Run
Source
R Script
Console
Terminal
Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/
Environment
History
Connections
Git
Tutorial
Global Environment
Files
Plots
Packages
Help
Viewer
Presentation
cmi-chennai.jpg
image
Reg_Class_Lect_14_Part_a.pdf
Reg_Class_Lect_14_Part_a.Rnw
Reg_Class_Lect_14_Part_b.R

So, if you have that now you have kind of 90 cases in actual cases these are the actual. So, 90 cases good and now it is better slightly better you can try out with this cut off marks. But the problem is another level now problem is there are two type of error one is truly there is a Good cases, which is being predicted as bad and there are Bad cases actually bad cases which is predicted as Good.

So, this is like actual bad cases which is predicted as good, which is actually bad thing you do not want this to happen if you as a bank manager you are giving loan to somebody and he or she is not returning in the money that is a really bad thing that is a real loss. This error on the other hand for a bank where actually the person is good, but bank thought ok if they will be bad customer and you are not giving loan.

So, in that case it is going to be a real problem. So, there are we are seeing in binary classification there are two kinds of problems. So, and one problem is typically called false positive another called false negative in statistical language it is called type 1 error and type 2 error. So, we are going to talk about that which one is good and which one is bad in the next lecture we are going to talk about that so, for now.

Thank you very much.