

# Study of Trends Using Logistic Regression Model for Congenital Heart Diseases in Adults

Anshul Srivastava\*, Anupam Sharma

1\*, 2) Dr Akhilesh Das Gupta Institute of Technology and Management, GGSIPU, New Delhi

Electronic Mail 1\*: anshul\_sriv@rediffmail.com

Electronic Mail 2: aanupam21@gmail.com

## Article Info

**Page Number:** 6340-6350

**Publication Issue:**

**Vol. 71 No. 4 (2022)**

## Article History

**Article Received:** 25 March 2022

**Revised:** 30 April 2022

**Accepted:** 15 June 2022

## Abstract

Congenital heart disease or defect is a heart abnormality present at birth. In this paper, we will be building logistic regression model on a sample of 40 adults collected from survey of various hospitals. We will be using R programming language for building the model.

---

## 1.Introduction

Congenital heart disease, also commonly known as CHD is a heart abnormality present since birth. The problem can affect heart walls, valves, blood vessels.

There are several types of congenital heart defects. They can range from simple conditions to life threatening conditions.

There are currently more than 1 million adults in India living with congenital heart defects. However over the past few decades, due to improvement in treatments and follow-up care for defects so nearly all children with heart defects survive into adulthood. Some of them need continuous care for their heart defect throughout their lives. However, many others go on to have active and productive lives despite their heart condition.

Here, in this paper we have collected sample of 40 adults and tried to predict various results using logistic regression model. Sample raw data is given in table 1.1

**Table 1.1**

S.NO.	Age	CHD	S.NO.	Age	CHD	S.NO.	Age	CHD
1	25	No	21	37	No	41	57	No
2	25	No	22	38	Yes	42	57	Yes
3	26	Yes	23	38	Yes	43	57	No
4	26	Yes	24	39	Yes	44	58	No
5	26	No	25	39	No	45	58	No
6	28	Yes	26	40	No	46	58	Yes
7	28	No	27	40	No	47	59	No
8	28	No	28	40	No	48	59	No
9	30	No	29	41	No	49	60	Yes
10	31	No	30	41	Yes	50	60	No
11	31	Yes	31	46	Yes	51	61	No
12	31	No	32	50	Yes	52	61	No
13	32	Yes	33	52	No	53	61	Yes
14	32	No	34	54	Yes	54	62	No
15	33	Yes	35	54	Yes	55	63	No
16	34	Yes	36	55	Yes	56	63	Yes
17	35	No	37	55	No	57	63	No
18	36	Yes	38	56	Yes	58	64	Yes
19	37	Yes	39	56	No	59	64	No
20	37	No	40	57	Yes	60	65	No

We will be studying above sample and build logistic regression model on it and infer various results.

## 2. Building logistic regression model

Here, we assign CHD data as Y variable and age as X variable. Since, Y is a categorical variable here, we will assign Y a dummy variable 0 and 1 and build logistic regression model here using R Studio.

### R Code:

#### # Importing and reading excel file

```
library(readxl)
```

```
chd_data <- read_excel("C:/Users/dell/Downloads/predictive analytics.xlsx",
```

```
sheet = "Sheet3")
```

```
View(chd_data)
```

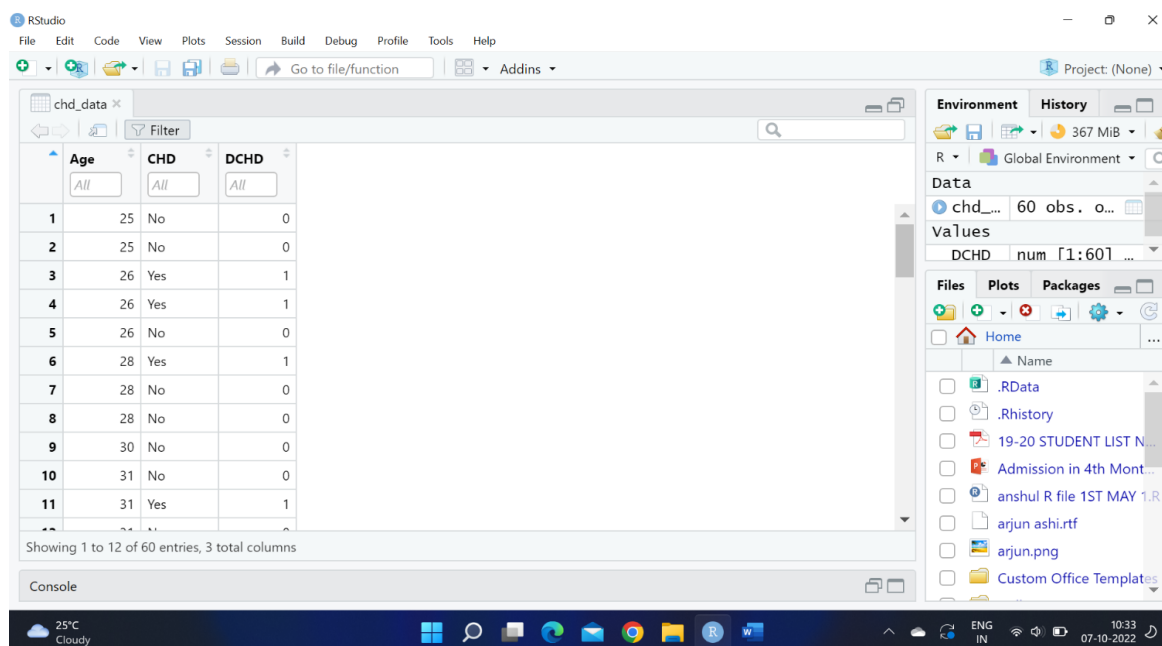
```
attach(chd_data)
```

### # Assigning dummy variable to categorical data CHD

```
DCHD=ifelse(chd_data$CHD=="Yes",1,0)
```

```
chd_data$DCHD=DCHD
```

**Fig. 2.1**



**Fig. 2.2**

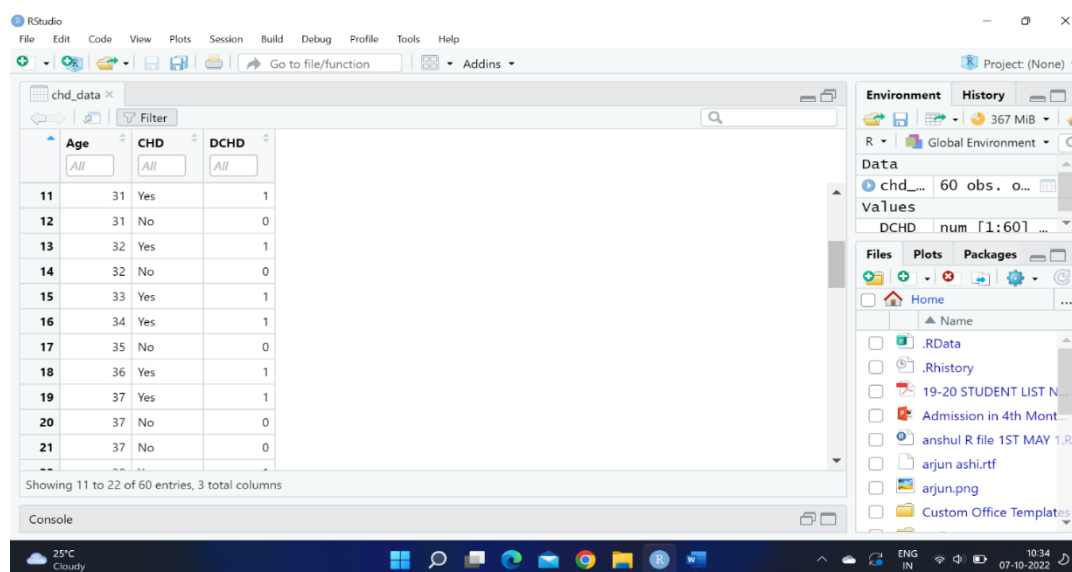


Fig. 2.3

RStudio interface showing a data table with columns Age, CHD, and DCHD. The table displays rows 21 to 32 of 60 entries. The right sidebar shows the Environment pane with 'chd\_data' selected, indicating 60 observations and 3 variables.

	Age	CHD	DCHD
21	37	No	0
22	38	Yes	1
23	38	Yes	1
24	39	Yes	1
25	39	No	0
26	40	No	0
27	40	No	0
28	40	No	0
29	41	No	0
30	41	Yes	1
31	46	Yes	1

Fig. 2.4

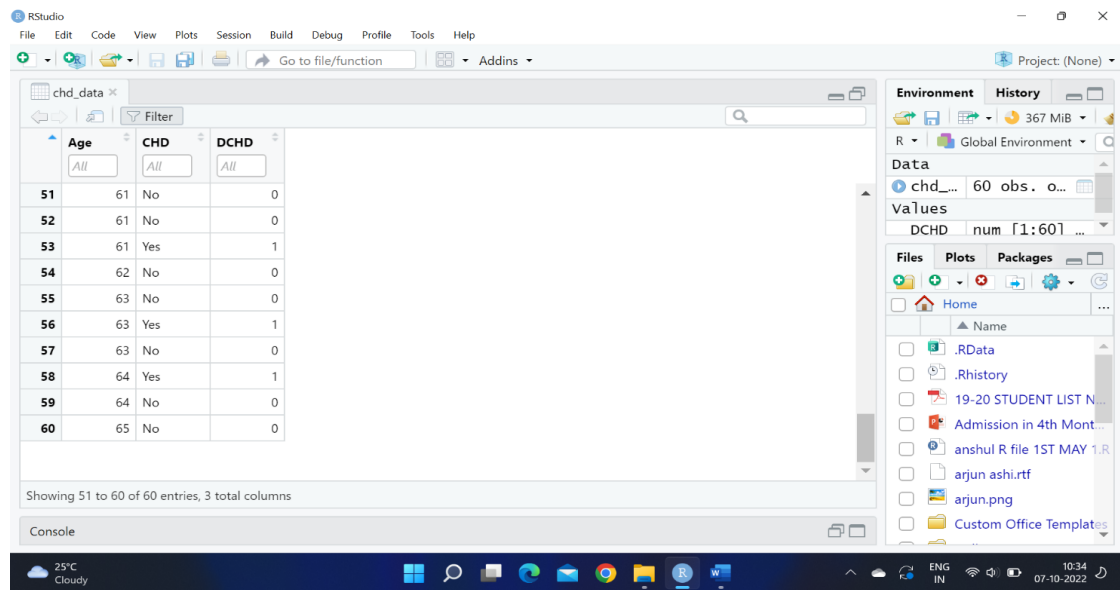
RStudio interface showing a data table with columns Age, CHD, and DCHD. The table displays rows 31 to 42 of 60 entries. The right sidebar shows the Environment pane with 'chd\_data' selected, indicating 60 observations and 3 variables.

	Age	CHD	DCHD
31	46	Yes	1
32	50	Yes	1
33	52	No	0
34	54	Yes	1
35	54	Yes	1
36	55	Yes	1
37	55	No	0
38	56	Yes	1
39	56	No	0
40	57	Yes	1
41	57	No	0

Fig.2.5

RStudio interface showing a data table with columns Age, CHD, and DCHD. The table displays rows 41 to 52 of 60 entries. The right sidebar shows the Environment pane with 'chd\_data' selected, indicating 60 observations and 3 variables.

	Age	CHD	DCHD
41	57	No	0
42	57	Yes	1
43	57	No	0
44	58	No	0
45	58	No	0
46	58	Yes	1
47	59	No	0
48	59	No	0
49	60	Yes	1
50	60	No	0
51	61	No	0

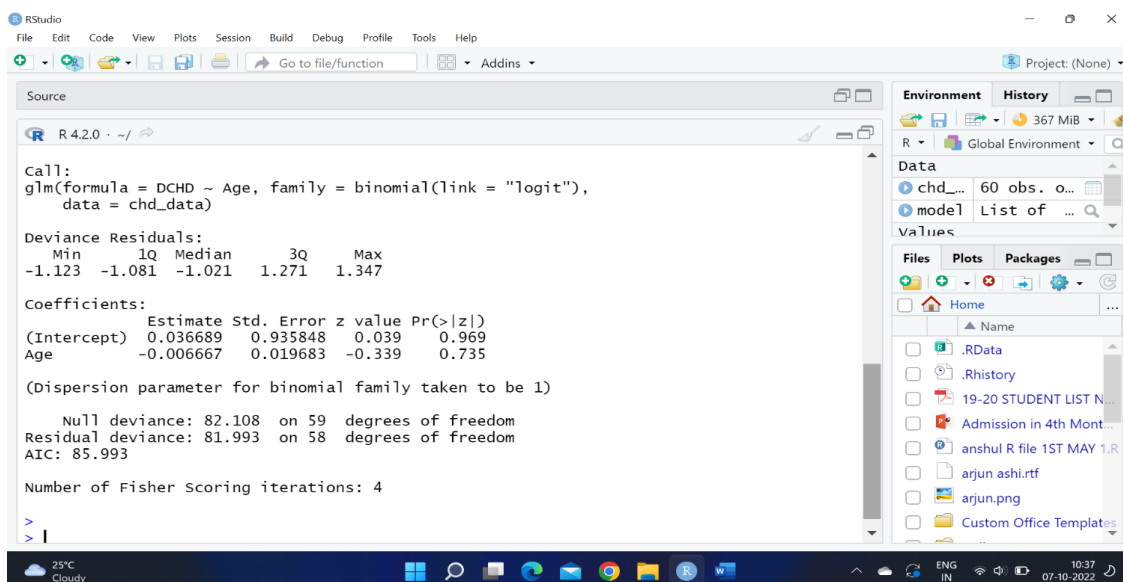
**Fig. 2.6**

### # Removing second column and building logistic regression model

```
chd_data=chd_data[,-2]
```

```
model=glm(DCHD ~ Age, family = binomial(link = "logit"),data=chd_data)
```

```
summary(model)
```

**Fig. 2.7**

### # Predicting probabilities

```
probs=predict(model,chd_data,type="response")
```

```
chd_data$probs=probs
```

Fig. 2.8

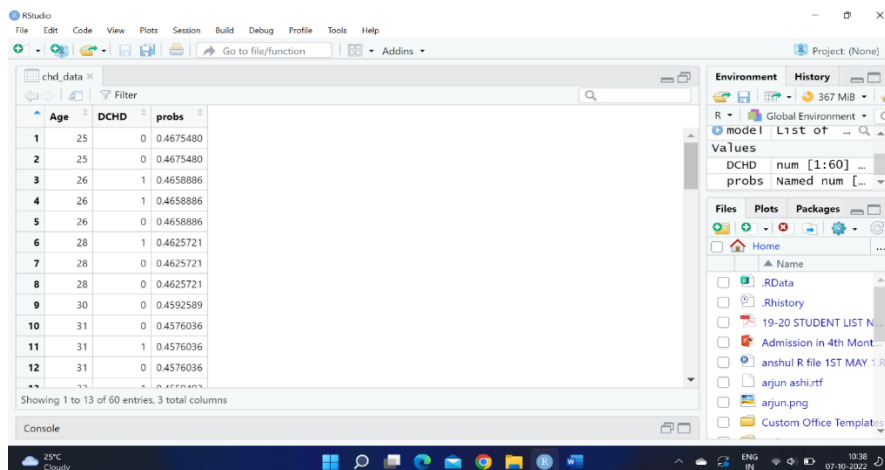


Fig. 2.9

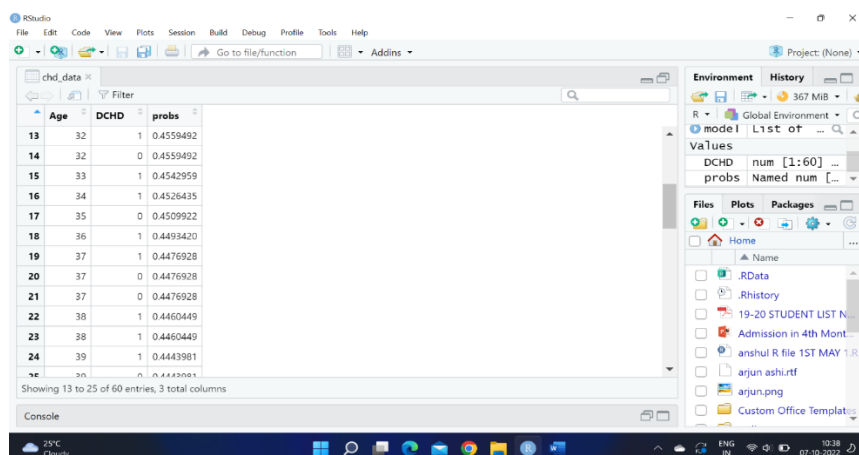


Fig. 2.10

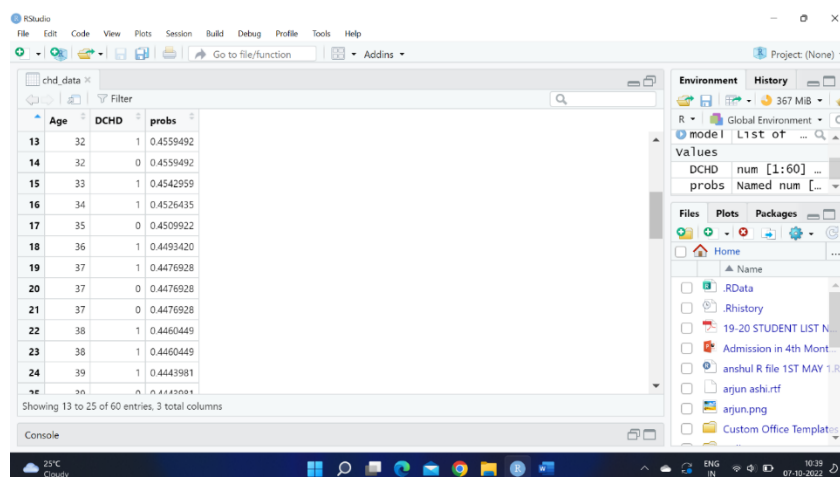
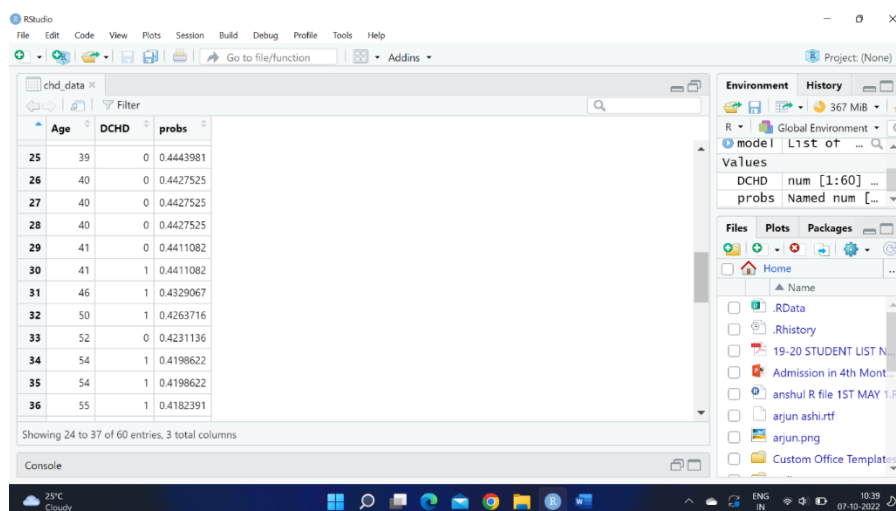


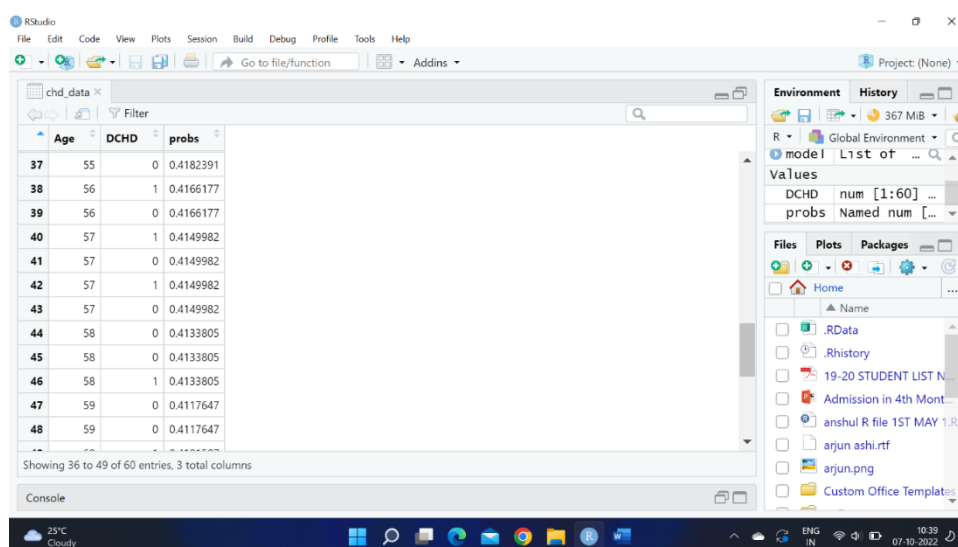
Fig. 2.11



	Age	DCHD	probs
25	39	0	0.4443981
26	40	0	0.4427525
27	40	0	0.4427525
28	40	0	0.4427525
29	41	0	0.4411082
30	41	1	0.4411082
31	46	1	0.4329067
32	50	1	0.4263716
33	52	0	0.4231136
34	54	1	0.4198622
35	54	1	0.4198622
36	55	1	0.4182391

Showing 24 to 37 of 60 entries, 3 total columns

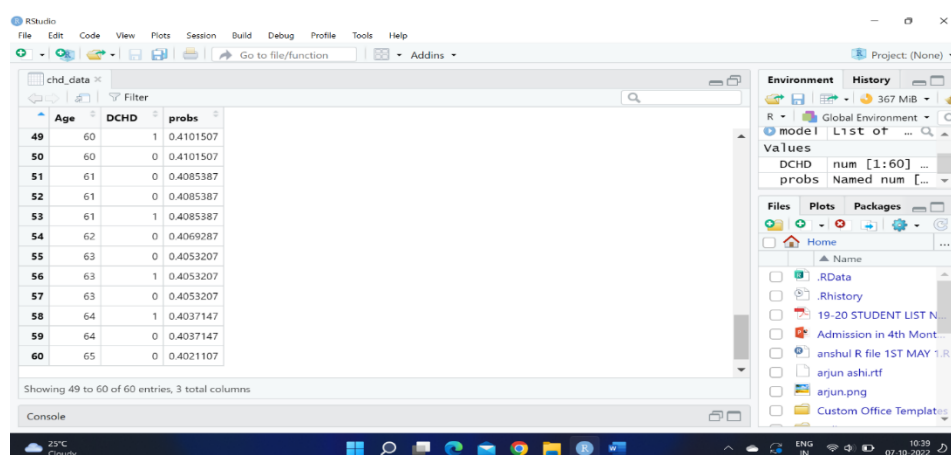
Fig 2.12



	Age	DCHD	probs
37	55	0	0.4182391
38	56	1	0.4166177
39	56	0	0.4166177
40	57	1	0.4149982
41	57	0	0.4149982
42	57	1	0.4149982
43	57	0	0.4149982
44	58	0	0.4133805
45	58	0	0.4133805
46	58	1	0.4133805
47	59	0	0.4117647
48	59	0	0.4117647

Showing 36 to 49 of 60 entries, 3 total columns

Fig. 2.13



	Age	DCHD	probs
49	60	1	0.4101507
50	60	0	0.4101507
51	61	0	0.4085387
52	61	0	0.4085387
53	61	1	0.4085387
54	62	0	0.4069287
55	63	0	0.4053207
56	63	1	0.4053207
57	63	0	0.4053207
58	64	1	0.4037147
59	64	0	0.4037147
60	65	0	0.4021107

Showing 49 to 60 of 60 entries, 3 total columns

**# Installing package InformationValue and loading library informationValue**

```
install.packages("InformationValue")
```

```
library(InformationValue)
```

```
# finding threshold probability
```

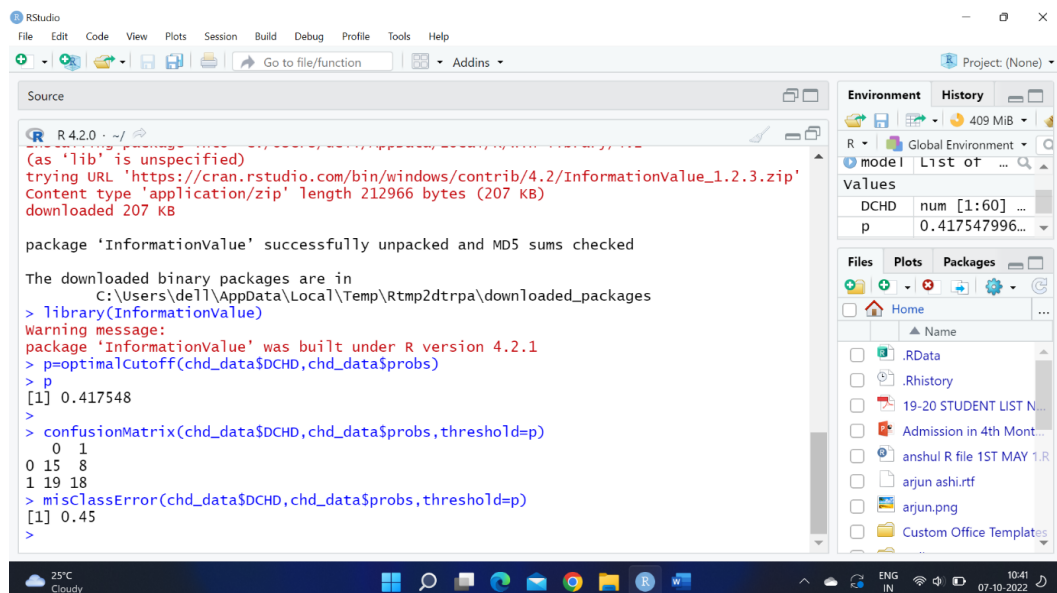
```
p=optimalCutoff(chd_data$DCHD,chd_data$probs)
```

```
p
```

**# Finding confusion matrix**

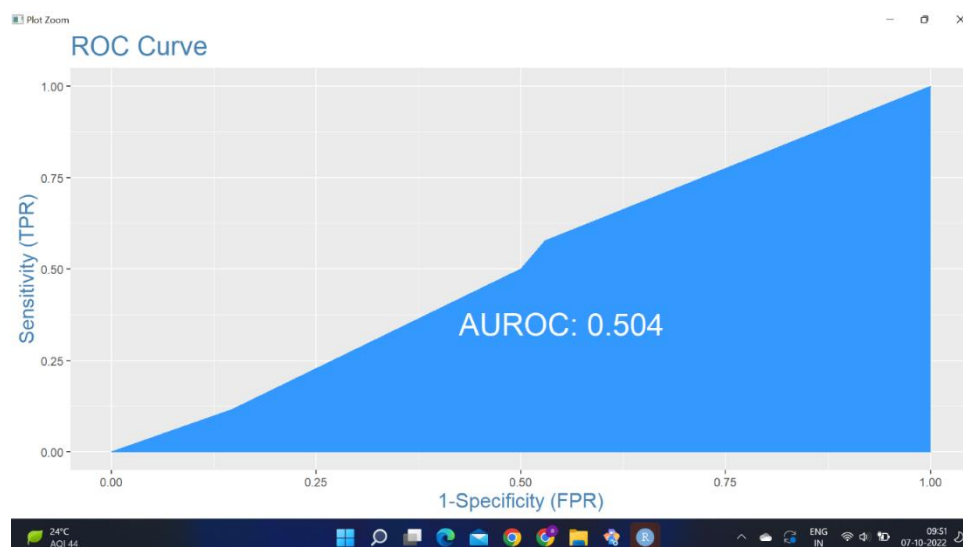
```
confusionMatrix(chd_data$DCHD,chd_data$probs,threshold=p)
```

```
misClassError(chd_data$DCHD,chd_data$probs,threshold=p)
```

**Fig. 2.14****# Plotting ROC Curve**

```
plotROC(chd_data$DCHD,chd_data$probs)
```



**Fig. 2.15**

### # Predicted Dummy Variables

```
pred_class=ifelse(chd_data$probs>p,1,0)
```

```
chd_data$pred_class=pred_class
```

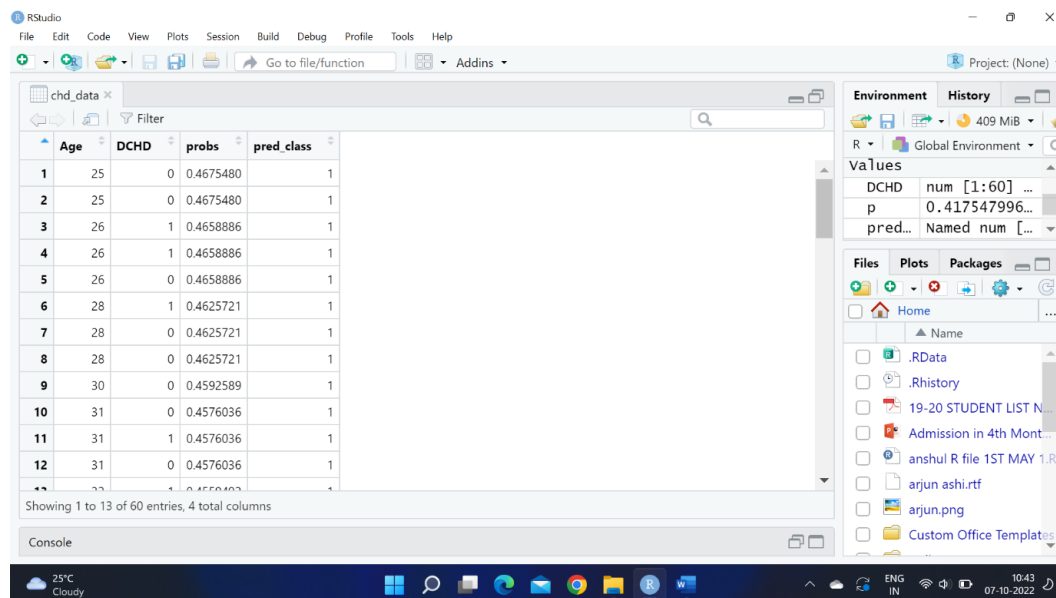
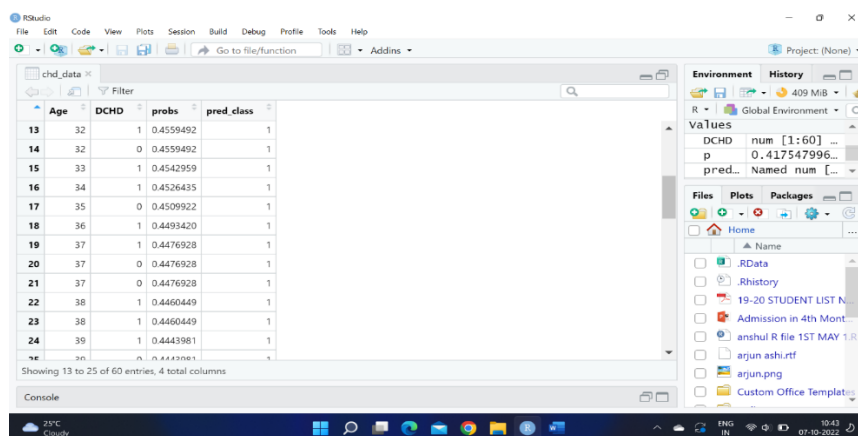
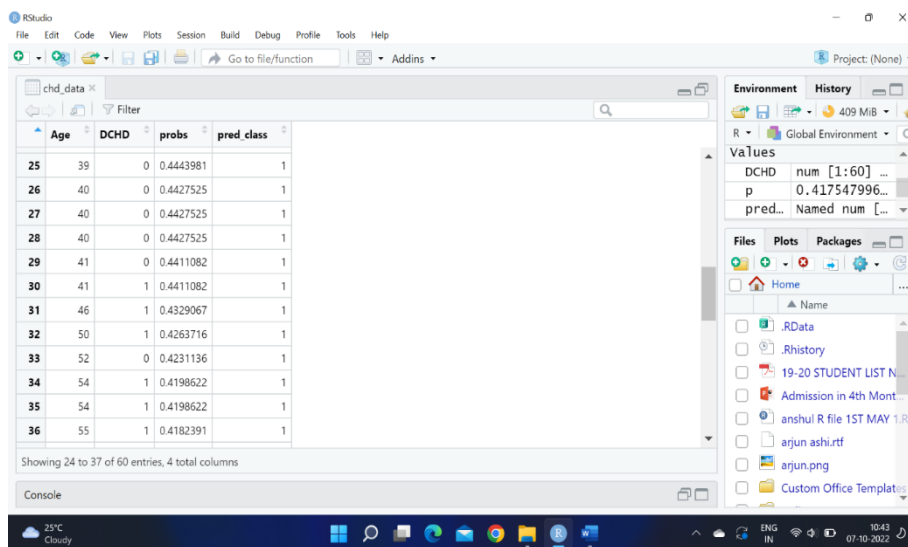
**Fig. 2.16**

Fig. 2.17



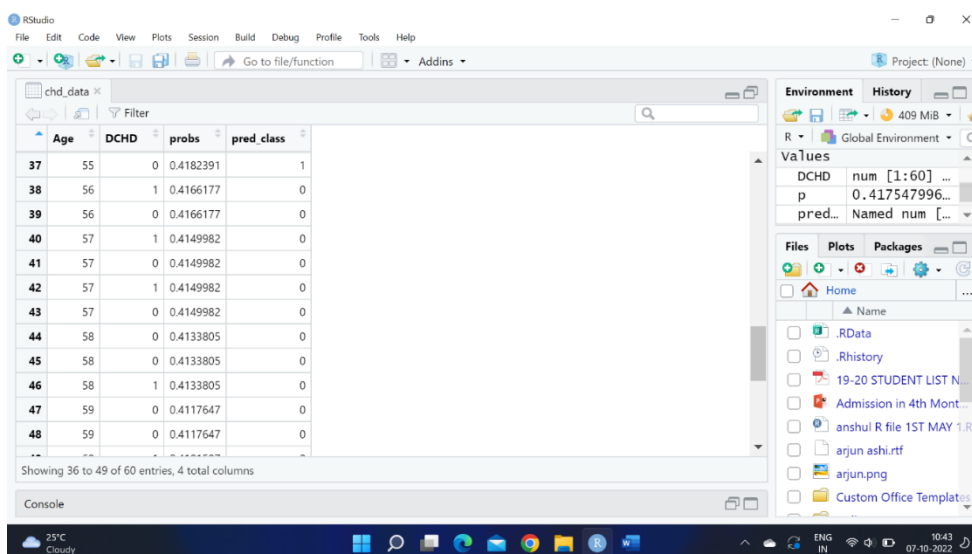
	Age	DCHD	probs	pred_class
13	32	1	0.4559492	1
14	32	0	0.4559492	1
15	33	1	0.4542959	1
16	34	1	0.4526435	1
17	35	0	0.4509922	1
18	36	1	0.4493420	1
19	37	1	0.4476928	1
20	37	0	0.4476928	1
21	37	0	0.4476928	1
22	38	1	0.4460449	1
23	38	1	0.4460449	1
24	39	1	0.4443981	1

Fig. 2.18



	Age	DCHD	probs	pred_class
25	39	0	0.4443981	1
26	40	0	0.4427525	1
27	40	0	0.4427525	1
28	40	0	0.4427525	1
29	41	0	0.4411082	1
30	41	1	0.4411082	1
31	46	1	0.4329067	1
32	50	1	0.4263716	1
33	52	0	0.4231136	1
34	54	1	0.4198622	1
35	54	1	0.4198622	1
36	55	1	0.4182391	1

Fig. 2.19



	Age	DCHD	probs	pred_class
37	55	0	0.4182391	1
38	56	1	0.4166177	0
39	56	0	0.4166177	0
40	57	1	0.4149982	0
41	57	0	0.4149982	0
42	57	1	0.4149982	0
43	57	0	0.4149982	0
44	58	0	0.4133805	0
45	58	0	0.4133805	0
46	58	1	0.4133805	0
47	59	0	0.4117647	0
48	59	0	0.4117647	0

**Fig. 2.20**

The screenshot shows the RStudio interface. The main window displays a data frame named 'chd\_data' with the following data:

	Age	DCHD	probs	pred_class
49	60	1	0.4101507	0
50	60	0	0.4101507	0
51	61	0	0.4085387	0
52	61	0	0.4085387	0
53	61	1	0.4085387	0
54	62	0	0.4069287	0
55	63	0	0.4053207	0
56	63	1	0.4053207	0
57	63	0	0.4053207	0
58	64	1	0.4037147	0
59	64	0	0.4037147	0
60	65	0	0.4021107	0

The Environment pane on the right shows the data frame's structure:

```

R | Global Environment | 409 MiB
Values
DCHD num [1:60] ...
p 0.417547996...
pred... Named num [...]
```

#### 4.Results and Conclusions

Based on above results, we infer that above logistic regression model is a poor model and we should reject this model. Reasons for rejecting above model (Fig. 2.7, 2.14, 2.15) are:

- Large AIC value and p value.
- Large misclassification error.
- Small AUROC value

#### References:

1. Maalouf, Maher. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335.
2. Niu, Lian. (2018). A review of the application of logistic regression in educational research: common issues, implications, and suggestions. Educational Review. 72. 1-27. 10.1080/00131911.2018.1483892.
3. Schober, Patrick MD, PhD, MMedStat\*; Vetter, Thomas R. MD, MPH†. Logistic Regression in Medical Research. Anesthesia & Analgesia: February 2021 - Volume 132 - Issue 2 - p 365-366, doi: 10.1213/ANE.00000000000005247
4. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):12-8. doi: 10.11613/BM.2014.003. PMID: 24627710; PMCID: PMC3936971.
5. X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.