# Sheth L.U.J. & Sir M.V. College

**Aim :-** To perform Logistic Regression using the glm() function in R and to predict whether a student will Pass or Fail based on the number of hours studied.

---

## Dataset Description

The dataset used for this practical is hours_scores.csv, which contains 25 observations.
The dataset includes the following variables:

- Hours – Number of hours studied (Independent Variable)
- Scores – Marks obtained by the student
- Pass – Binary dependent variable, created as:
  - Pass = 1, if Scores ≥ 50
  - Fail = 0, if Scores < 50

This transformation converts the problem into a binary classification task.

---

## Theory (Logistic Regression using glm)

Logistic Regression is a statistical technique used when the dependent variable is binary in nature.
In R, Logistic Regression is implemented using the Generalized Linear Model (glm) function with:

**Name :- Priya Gupta**
**Roll No :- S081**

- family = binomial
- link = logit

The mathematical model of Logistic Regression is given by:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

where:

- $p$ represents the probability of passing
- $X$ represents the number of hours studied

This model estimates the relationship between study hours and the probability of passing the exam.

---

## Procedure

### Step 1: Loading the Dataset

The dataset was loaded into R using the read.csv() function, and the column names were verified to ensure correct data structure.

---

### Step 2: Creating the Binary Target Variable

A new variable Pass was created based on student scores.
Students scoring 50 or more marks were labeled as Pass (1), while those scoring below 50 were labeled as Fail (0).

---

```
> head(data)
  Hours Scores Pass
1   2.5     21    0
2   5.1     47    0
3   3.2     27    0
4   8.5     75    1
5   3.5     30    0
6   1.5     20    0
>
```

---

### Step 3: Fitting the Logistic Regression Model

Logistic Regression was applied using the following model:

glm(Pass ~ Hours, family = binomial(link = "logit"))

This model estimates the probability of passing based on the number of hours studied.

```
> summary(model)

Call:
glm(formula = Pass ~ Hours, family = binomial(link = "logit"),
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.052     13.277  -1.435    0.151
Hours          3.842      2.662   1.443    0.149

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.6173  on 24  degrees of freedom
Residual deviance:  4.7584  on 23  degrees of freedom
AIC: 8.7584

Number of Fisher Scoring iterations: 9
```

---

## Model Summary Analysis

From the model summary:

- Intercept = −19.052
- Coefficient for Hours = 3.842

These values indicate that as the number of study hours increases, the log-odds of passing also increase.
The positive coefficient for hours confirms that study time has a positive effect on passing probability.

Model Fit Statistics

- Null Deviance: 34.617
- Residual Deviance: 4.758
- AIC: 8.758

The large reduction from null deviance to residual deviance shows that the model provides a good fit to the data.

---

**Step 4: Prediction of Probabilities**

The fitted model was used to predict the probability of passing for each student.
For example:

- A student studying 1.5 hours has a very low probability of passing (approximately 0.000001).

- A student studying 5.1 hours has a moderate probability of passing (approximately 0.63).

- A student studying 8.5 hours has a very high probability of passing (approximately 0.9999).

This clearly shows the increasing trend in passing probability with more study hours.

```
> head(data)
  Hours Scores Pass Predicted_Probability
1   2.5     21    0          7.896622e-05
2   5.1     47    0          6.326846e-01
3   3.2     27    0          1.161649e-03
4   8.5     75    1          9.999988e-01
5   3.5     30    0          3.669436e-03
6   1.5     20    0          1.693381e-06
>
```

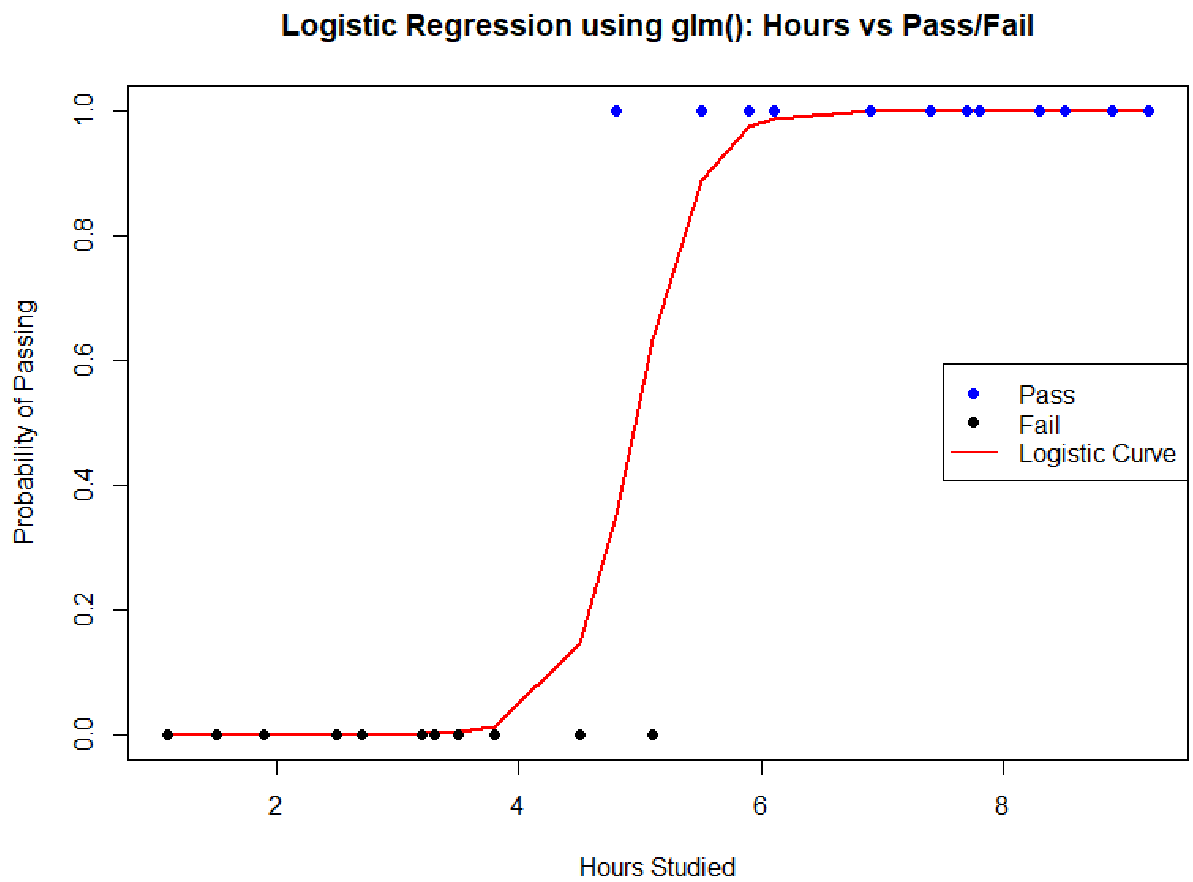
---

## Graphical Representation

A Logistic Regression curve was plotted to visualize the relationship between hours studied and the probability of passing.

### Logistic Regression Curve

# Sheth L.U.J. & Sir M.V. College




## Graph Analysis

The graph obtained from the model is correct and appropriate for Logistic Regression.

- The X-axis represents Hours Studied
- The Y-axis represents Probability of Passing
- The red curve represents the logistic (sigmoid) regression curve
- Blue points indicate students who passed

**Name :- Priya Gupta**
**Roll No :- S081**

- Black points indicate students who failed

Key Observations

- For study hours below approximately 4 hours, the probability of passing is close to zero.
- Between 5 to 6 hours, there is a sharp increase in the probability of passing.
- Beyond 6 hours, the probability approaches one, indicating a very high chance of passing.

The characteristic S-shaped curve confirms that the model behaves as expected. The predicted probabilities lie between 0 and 1, and the decision boundary is clearly visible.

## Result

Logistic Regression using the glm() function successfully predicted whether a student would pass or fail based on the number of hours studied.

---

## Conclusion

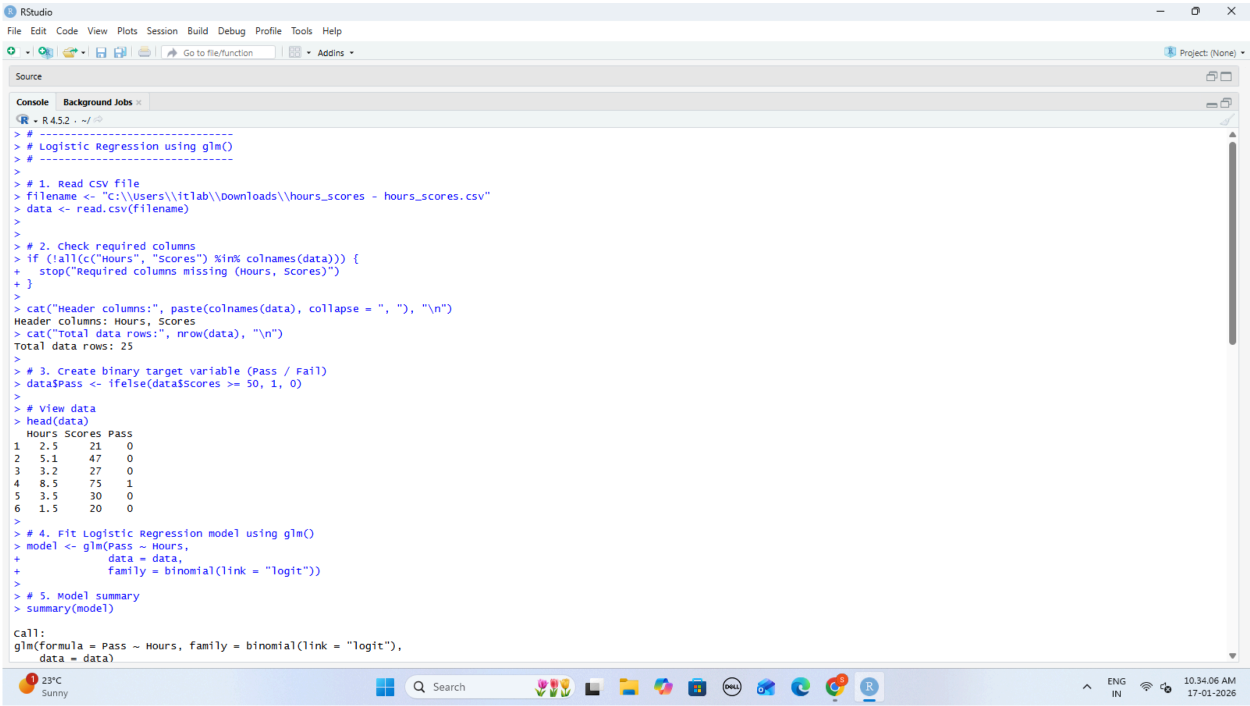
The Logistic Regression model built using glm() demonstrates that study hours have a strong positive impact on a student's probability of passing an exam. As the number of hours studied increases, the likelihood of passing increases significantly. Therefore, the objective of the practical was successfully achieved.

---

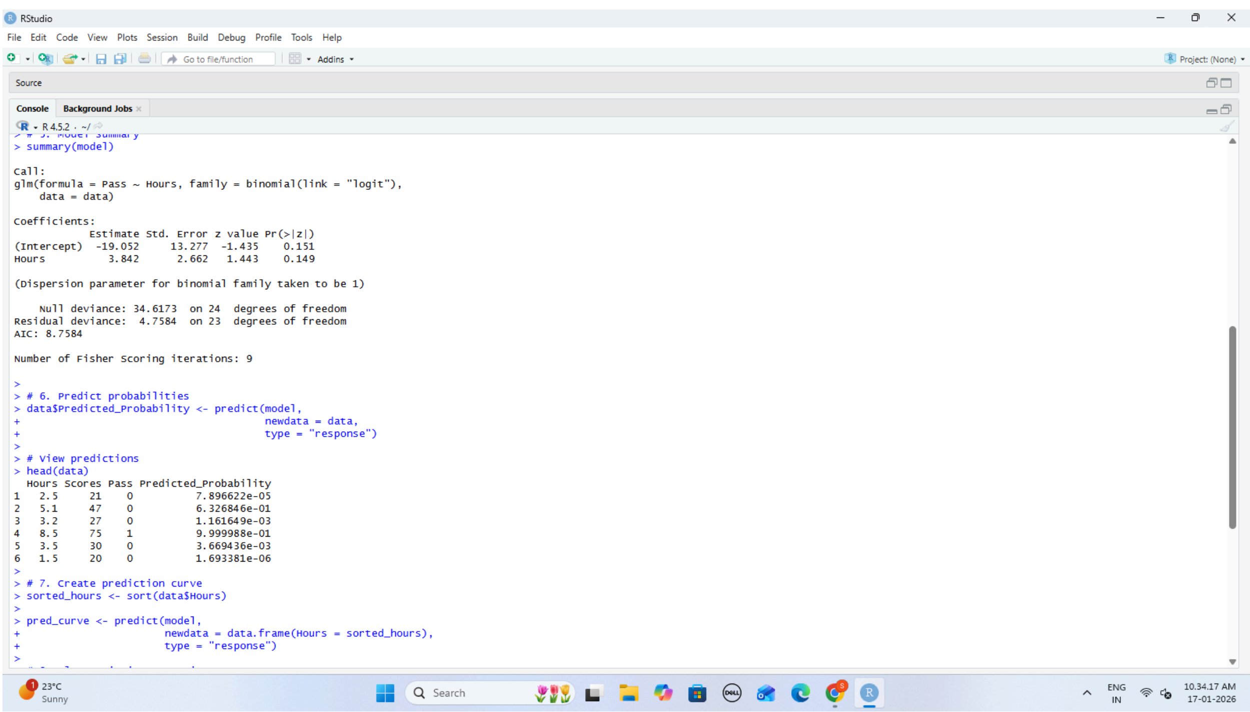
**Name :- Priya Gupta**
**Roll No :- S081**

# Sheth L.U.J. & Sir M.V. College

## ***Screenshots***



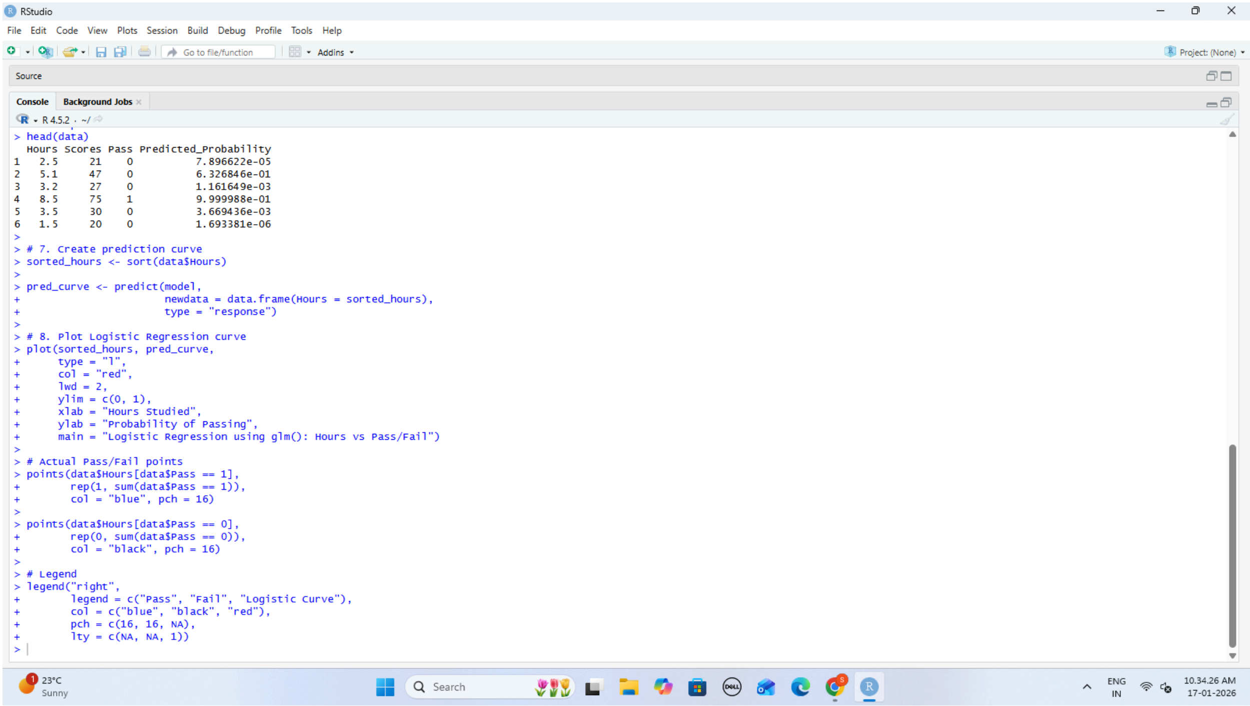
```
> # ------------------------------
> # Logistic Regression using glm()
> # ------------------------------
>
> # 1. Read CSV file
> filename <- "C:\\Users\\itlab\\Downloads\\hours_scores - hours_scores.csv"
> data <- read.csv(filename)
>
>
> # 2. Check required columns
> if (!all(c("Hours", "Scores") %in% colnames(data))) {
+   stop("Required columns missing (Hours, Scores)")
+ }
>
> cat("Header columns:", paste(colnames(data), collapse = ", "), "\n")
Header columns: Hours, Scores
> cat("Total data rows:", nrow(data), "\n")
Total data rows: 25
>
> # 3. Create binary target variable (Pass / Fail)
> data$Pass <- ifelse(data$Scores >= 50, 1, 0)
>
> # View data
> head(data)
  Hours Scores Pass
1   2.5     21    0
2   5.1     47    0
3   3.2     27    0
4   8.5     75    1
5   3.5     30    0
6   1.5     20    0
>
> # 4. Fit Logistic Regression model using glm()
> model <- glm(Pass ~ Hours,
+              data = data,
+              family = binomial(link = "logit"))
>
> # 5. Model summary
> summary(model)

Call:
glm(formula = Pass ~ Hours, family = binomial(link = "logit"),
    data = data)
```



```
> summary(model)

Call:
glm(formula = Pass ~ Hours, family = binomial(link = "logit"),
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -19.052     13.277  -1.435    0.151
Hours          3.842      2.662   1.443    0.149

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.6173  on 24  degrees of freedom
Residual deviance:  4.7584  on 23  degrees of freedom
AIC: 8.7584

Number of Fisher Scoring iterations: 9

>
> # 6. Predict probabilities
> data$Predicted_Probability <- predict(model,
+                                       newdata = data,
+                                       type = "response")
>
> # View predictions
> head(data)
  Hours Scores Pass Predicted_Probability
1   2.5     21    0          7.896622e-05
2   5.1     47    0          6.326846e-01
3   3.2     27    0          1.161649e-03
4   8.5     75    1          9.999988e-01
5   3.5     30    0          3.669436e-03
6   1.5     20    0          1.693381e-06
>
> # 7. Create prediction curve
> sorted_hours <- sort(data$Hours)
>
> pred_curve <- predict(model,
+                       newdata = data.frame(Hours = sorted_hours),
+                       type = "response")
>
```

**Name :- Priya Gupta**
**Roll No :- S081**

# Sheth L.U.J. & Sir M.V. College



```
> head(data)
  Hours Scores Pass Predicted_Probability
1   2.5     21    0          7.896622e-05
2   5.1     47    0          6.326846e-01
3   3.2     27    0          1.161649e-03
4   8.5     75    1          9.999988e-01
5   3.5     30    0          3.669436e-03
6   1.5     20    0          1.693381e-06
>
> # 7. Create prediction curve
> sorted_hours <- sort(data$Hours)
>
> pred_curve <- predict(model,
+                       newdata = data.frame(Hours = sorted_hours),
+                       type = "response")
>
> # 8. Plot Logistic Regression curve
> plot(sorted_hours, pred_curve,
+      type = "l",
+      col = "red",
+      lwd = 2,
+      ylim = c(0, 1),
+      xlab = "Hours Studied",
+      ylab = "Probability of Passing",
+      main = "Logistic Regression using glm(): Hours vs Pass/Fail")
>
> # Actual Pass/Fail points
> points(data$Hours[data$Pass == 1],
+        rep(1, sum(data$Pass == 1)),
+        col = "blue", pch = 16)
>
> points(data$Hours[data$Pass == 0],
+        rep(0, sum(data$Pass == 0)),
+        col = "black", pch = 16)
>
> # Legend
> legend("right",
+        legend = c("Pass", "Fail", "Logistic Curve"),
+        col = c("blue", "black", "red"),
+        pch = c(16, 16, NA),
+        lty = c(NA, NA, 1))
> |
```

**Name :- Priya Gupta**
**Roll No :- S081**