Modeling Binary Outcomes: Logistic Regression in R

Sherman Centre for Digital Scholarship

Thursday, November 20, 2025

4:00pm - 5:00pm **(Online)**

Modeling Binary Outcomes: Logistic Regression in R

Sahar Khademioore
PhD Candidate in Health Research Methodology

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct/

Certificate Programs

The Sherman Centre for Digital Scholarship Certificate of Attendance

The Sherman Centre's certificate program recognizes attendance at our workshops. It complements degree training, supports the development of critical competencies in data analysis, research data management, and digital scholarship, and formalizes core skills fostered by our workshops.

Participants are invited to attend seven workshops and receive a certificate of attendance. To verify your participation in today's workshop, we will provide a code and additional instructions at the end of the session.

You can learn more about the certificate program at **scds.ca/certificate-program**

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events: u.mcmaster.ca/scds-events

Nov 27, 2025: “Creating and Sharing Maps using ArcGIS”

Nov 27, 2025: “Microdata Analysis with Python using Statistics Canada Data”

Jan 15, 2025: “Introduction to R Programming”

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: **https://library.mcmaster.ca/services/dash**

Logistic regression using R

Objectives of the workshop:

- Review basics of Logistic regression
- How to fit a Logistic regression in R
- Interpret model output and coefficients
- Assess model assumptions
- Evaluate the model’s fit

What is Logistic Regression?

Logistic regression is a statistical method for analyzing datasets where the outcome variable is **binary** (0 or 1, Yes or No, Success or Failure).

Use Cases:

- Medical: Disease diagnosis (Diseased vs Healthy)
- Marketing: Customer churn (Will leave vs Will stay)
- Finance: Loan default (Default vs Non-default)

When to Use Logistic Regression

Dependent variable:

- Continues (e.g., blood sugar) → **Linear regression**
- Binary or categorical (e.g., Dead/alive) → **Logistic regression**

Introduction to Logistic Regression

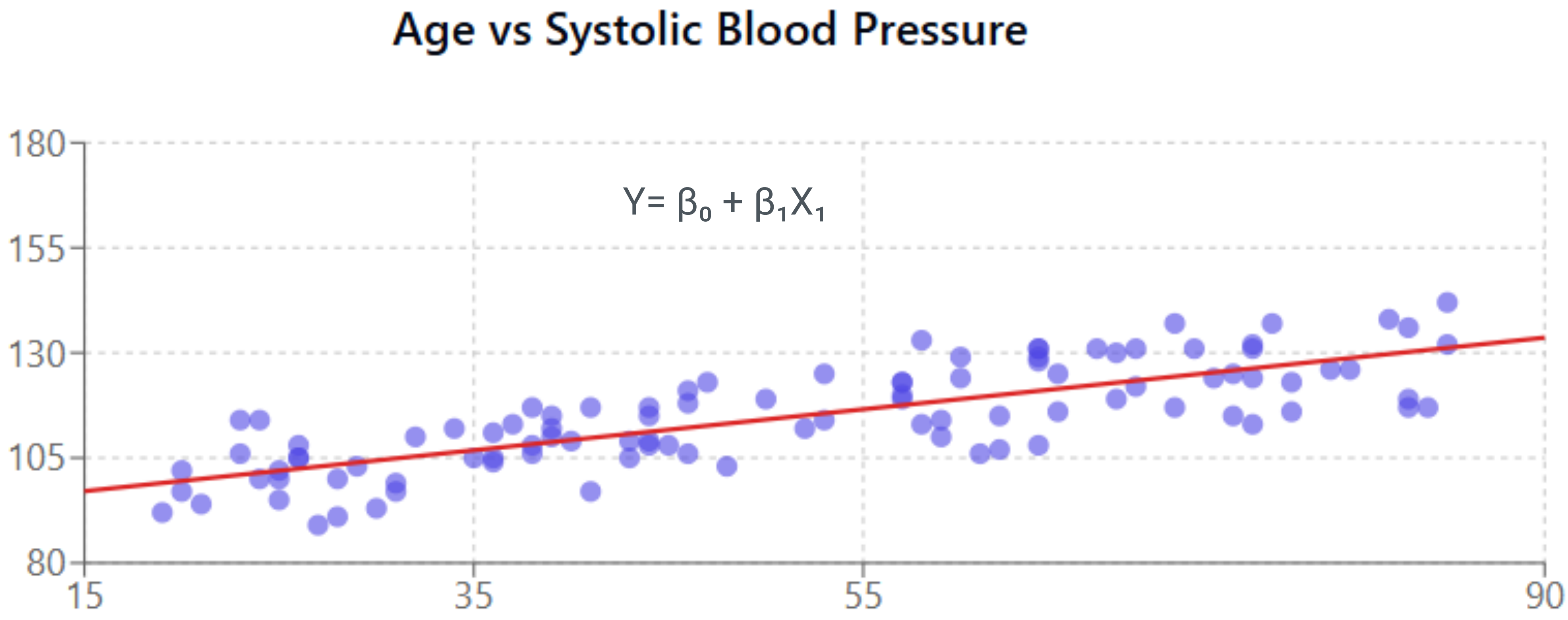
- What is the association of a binary **outcome** variable (Y) with one or more predictor variables (X's)?

- Continues predictor?
- Several covariates/confounders?

→ **Logistic regression**

Linear regression recap!

Find the fitted line and use this line to predict the blood pressure given age



The Mathematical Heart of Logistic Regression

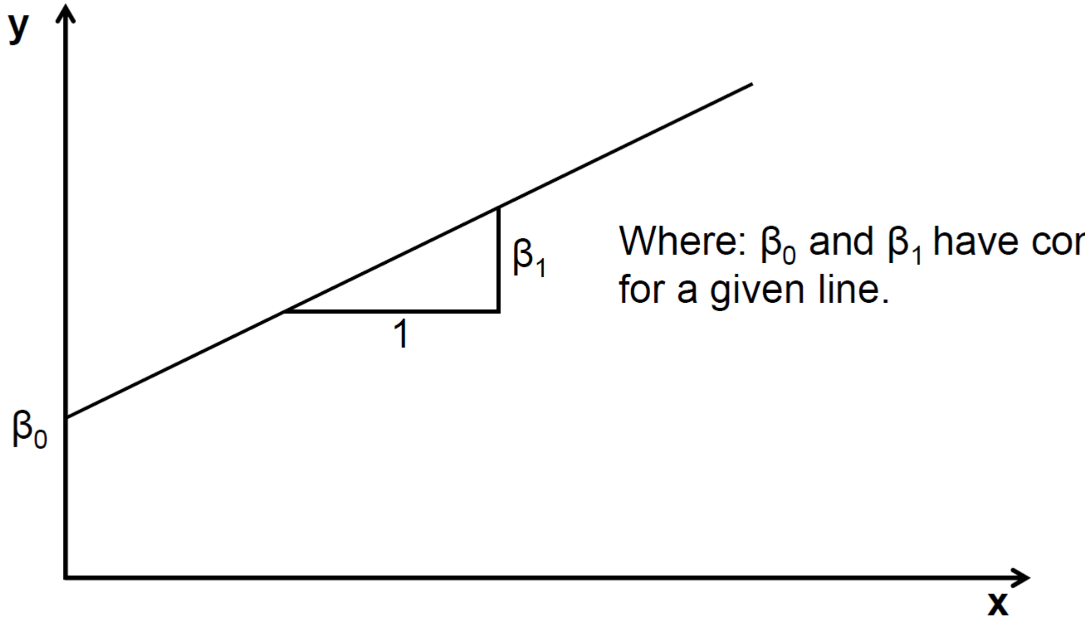
The Linear Predictor

We start with a familiar concept from linear regression – a weighted sum of our predictors:

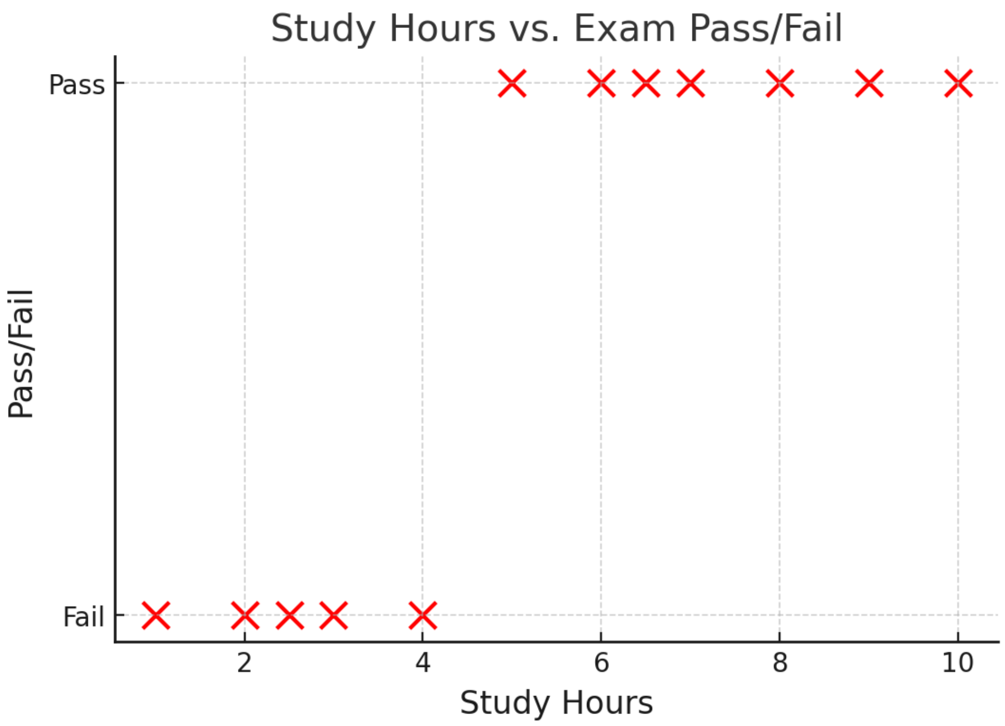
$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

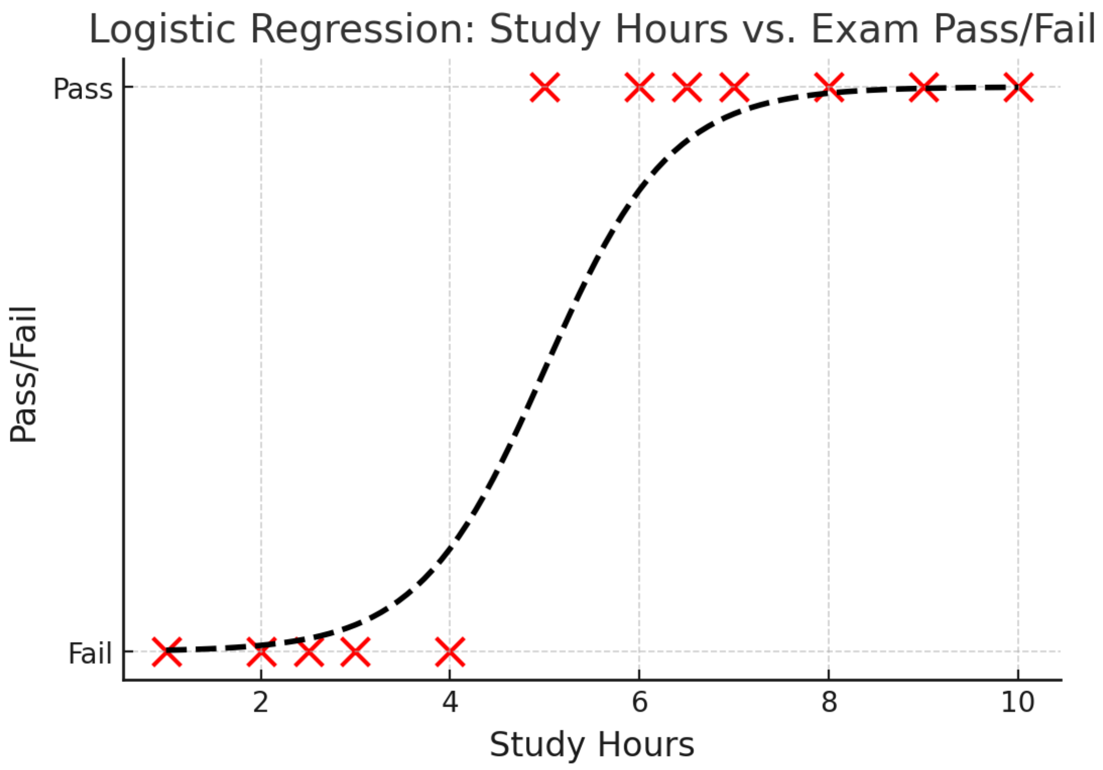
Where:

- z is called the linear predictor or "log odds" (more on this later)
- β_0 is the intercept (baseline value when all predictors are zero)
- $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients that represent the impact of each predictor
- $X_1, X_2, \ldots, X_n$ are our predictor variables



Why not just use linear regression?





Unlike linear regression, logistic regression predicts probabilities and classifies data points into discrete categories.

Understanding Odds and Odds Ratios

Probability

The chance of an event occurring.

$$P(\text{Survival}) = \frac{\text{Number Survived}}{\text{Total}}$$

Odds

The ratio of success to failure.

$$\text{Odds}(\text{Survival}) = \frac{P(\text{Survival})}{1 - P(\text{Survival})} = \frac{P(\text{Survival})}{P(\text{Death})}$$

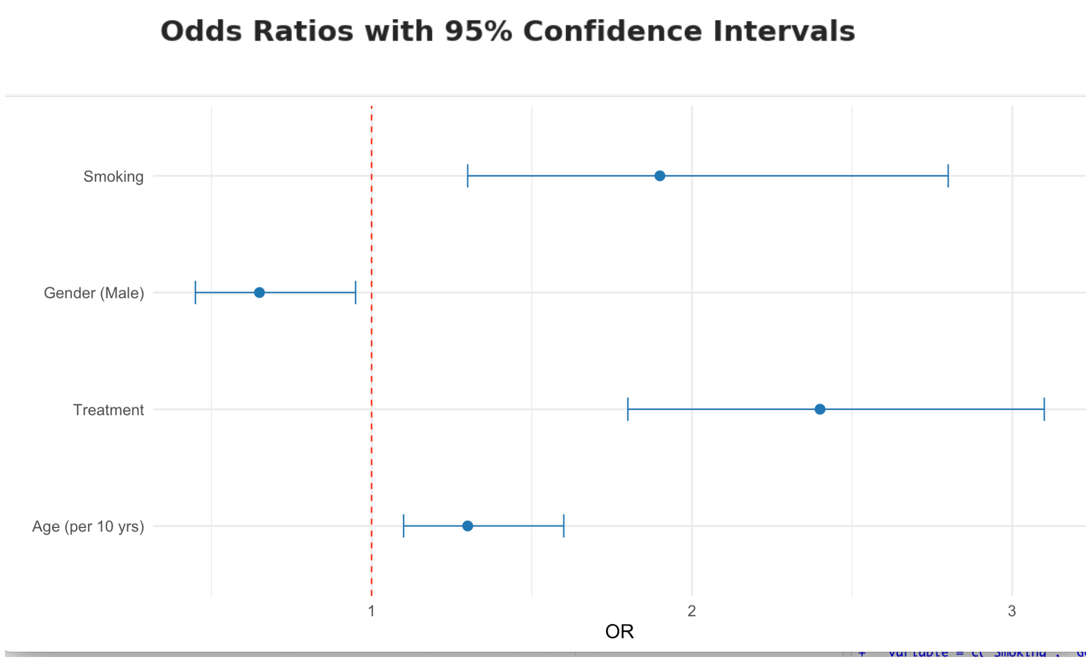
Understanding Odds Ratios

Odds Ratio (OR):

$$OR = e\beta$$

Interpretation:

- OR > 1: Increased odds
- OR = 1: No effect
- OR < 1: Decreased odds

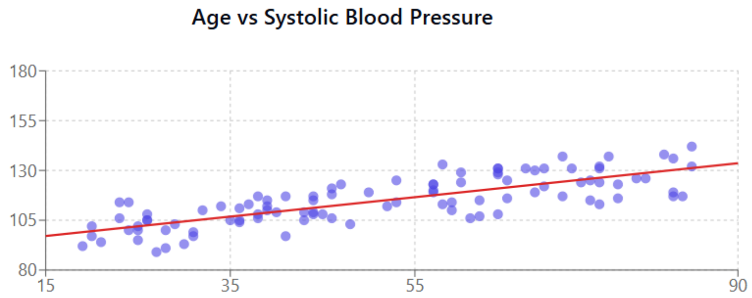


Example: OR = 2.45 for treatment means treated patients have 2.45 times higher odds of success compared to control group.

What Does Logistic Regression Do?

- Logistic regression is specifically designed to model the **probability** of a binary outcome.
- It uses a mathematical transformation (the **logistic function**) to ensure predictions always fall between 0 and 1, and it allows for non-linear relationships between predictors and outcomes.

From Linear to Logistic



Linear regression predicts the outcome directly:

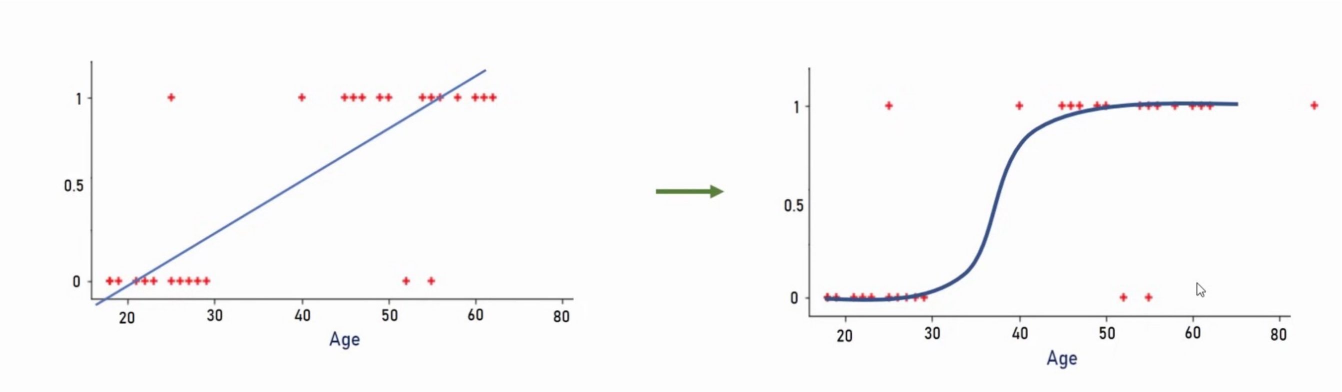
$$Y = \beta_0 + \beta_1 X$$

Logistic regression predicts the log-odds (logit) of the outcome:

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

Then we convert the log-odds back into a probability using the logistic function:

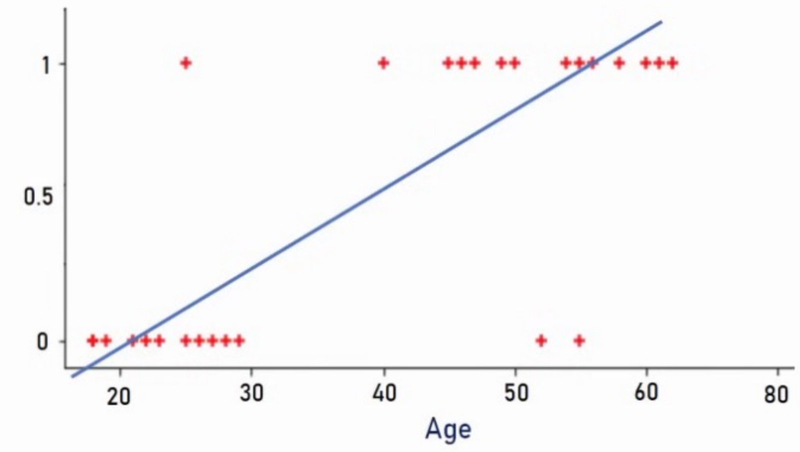
$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$



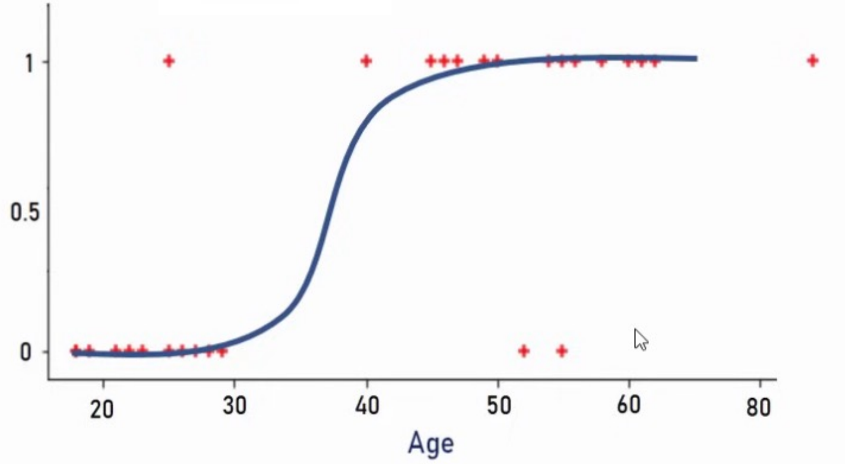
The Logistic Function: Converting to Probabilities

The key insight of logistic regression is to transform the linear predictor into a probability using the **logistic function** so that the predictions lie between 0 and 1:

$z = \beta_0 + \beta_1 X_1$

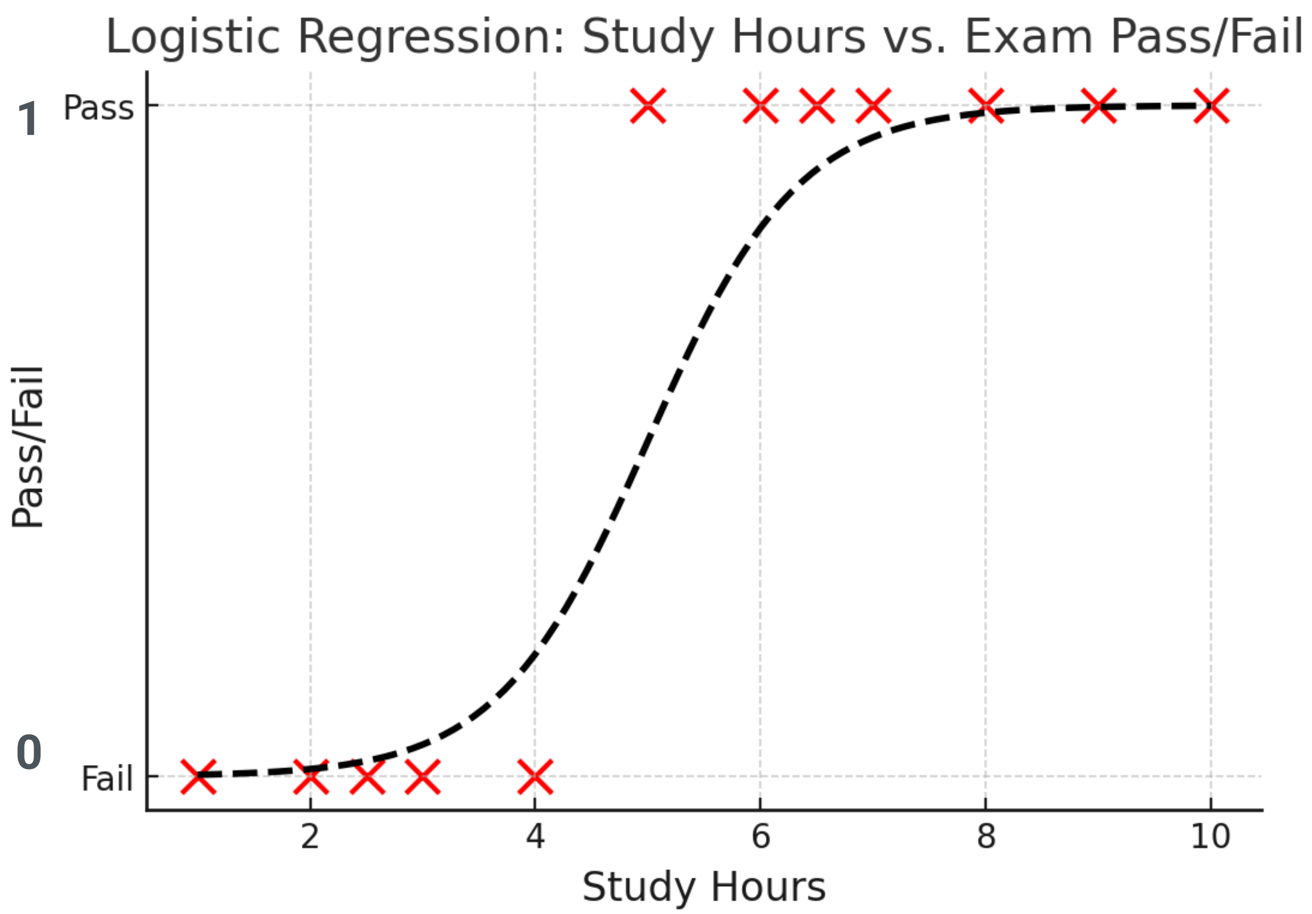


$$\frac{1}{1+e^{-z}}$$



How to fit the best fitted S shape line

Maximum Likelihood



Simple vs. Multiple Logistic Regression

Simple Logistic Regression

One predictor variable

Example: Smoking and cancer

Multiple Logistic Regression

Two or more predictor variables

Example: age, weight, ethnicity

Model Assumptions

1. Binary Outcome
Dependent variable must be binary (0/1, Yes/No)

2. Independence of Observations
Each observation should be independent

3. Linearity of Logit
Linear relationship between continuous predictors and log-odds

4. No Multicollinearity
Predictors should not be highly correlated

Evaluating Model Fit

Deviance

Measures model fit; lower is better

AIC

Model selection criterion

McFadden's R^2

Pseudo R-squared (0.2-0.4 = good)

```
> summary(model)

Call:
glm(formula = am ~ wt + hp, family = binomial, data = mtcars)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.86630    7.44356   2.535  0.01126 *
wt          -8.08348    3.06868  -2.634  0.00843 **
hp           0.03626    0.01773   2.044  0.04091 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 10.059  on 29  degrees of freedom
AIC: 16.059

Number of Fisher Scoring iterations: 8

> AIC(model)
[1] 16.05911
> BIC(model)
[1] 20.45632
> deviance(model)
[1] 10.05911
```

MLE is an iterative process (unlike OLS).

- AIC (Akaike’s Information Criterion)
  - -2lnL + 2k
- BIC (Bayesian Information Criterion)
  - -2lnL + k ln(N)
- k is the number of parameters

We will not interpret the numbers. Only used to compare between models and lower values represent better fit

Goodness-of-fit of the model

- **Deviance (D = -2ln[likelihood])**
  - lower D is associated with a better fit
  - D is conceptually equivalent to SSE in linear regression
  - It is an indicator of how much unexplained information there is after the model has been fitted

```
# Example logistic regression model
model <- glm(diabetes ~ bmi + age, data = patient_data, family = "binomial")
summary(model)

# R output
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -2.7243  -0.6780  -0.3793   0.6462   2.9048
#
# Null deviance: 1186.7  on 999  degrees of freedom
# Residual deviance: 917.8  on 997  degrees of freedom
# AIC: 923.8
```

Goodness-of-fit of the model

Measure if a more complex model fits the data better

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  patients$diabetes, fitted(model)
X-squared = 12.46, df = 8, p-value = 0.132
```

- Hosmer-Lemeshow statistic
  - The observed and expected values can be compared by calculation of a Pearson statistic.
  - H-L showed that if there are g groups, and the number of distinct covariate patterns equals the sample size, the statistic is approximately $\chi 2$ with g-2 d.f under H0, that the model is appropriate.

Discriminability

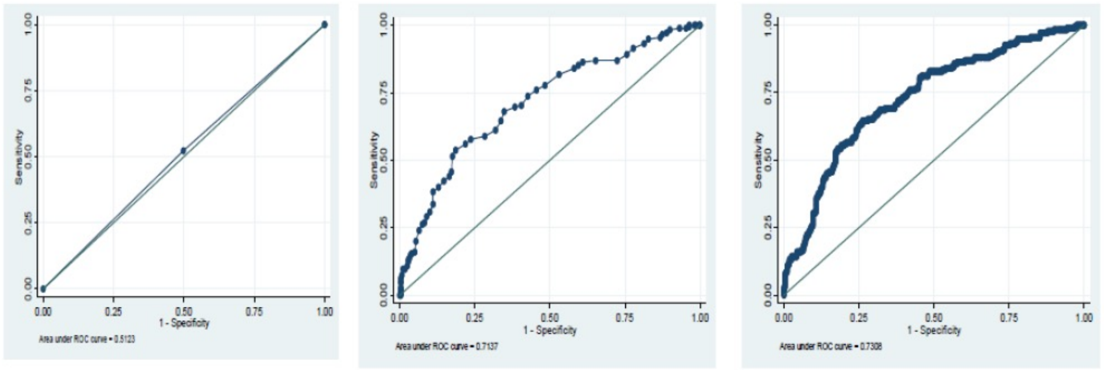
- How well does the model correctly distinguish those who have the outcome (from those who do not?
  - Sensitivity and Specificity
  - Classification Tables
  - Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC)
  - Somer’s D
  - Goodman Kruskal Gamma

Discriminability

- **ROC curve:**
- sensitivity (proportion of true positives) versus 1-specificity (proportion of true negatives) at carious cut points

- The area under the curve **(AUC)**, is a summary measure of the model's ability to discriminate between cases and controls (between 0 and 1)

Model (1), with treatment only | Model (2), adding APACHE to (1) | Model (3), adding Temp0 to (2)



AUC(1) = 0.5123 | AUC(2) = 0.7137 | AUC(3) = 0.7308

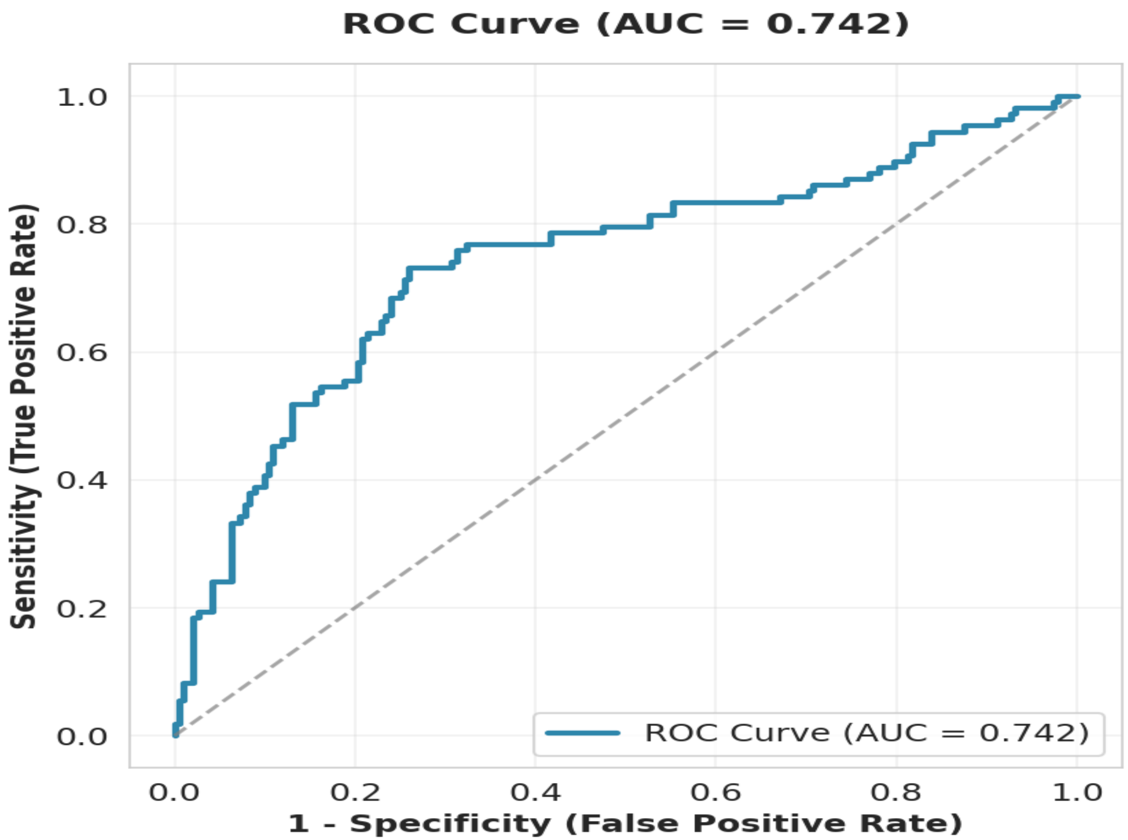
ROC Curve & AUC

ROC Curve

Plots True Positive Rate vs False Positive Rate

AUC :

- 0.9-1.0 = Excellent
- 0.8-0.9 = Good
- 0.7-0.8 = Fair
- 0.5 = No better than chance



Let’s practice!

Use the link in the chat:

Summary & Key Takeaways

What We Covered:

- Understanding logistic regression for binary outcomes
- Fitting models using glm() in R
- Interpreting coefficients and odds ratios
- Checking model assumptions
- Evaluating model performance

Best Practices

Always check assumptions and evaluate fit

Remember

Interpret in context of research question

Thank you!

- **Email:** Khades1@mcmaster.ca
- **Book an appointment** with DASH: https://library.mcmaster.ca/services/dash
- **Contact DASH:** Data Analysis Support Hub: libdash@mcmaster.ca

- **regression**