

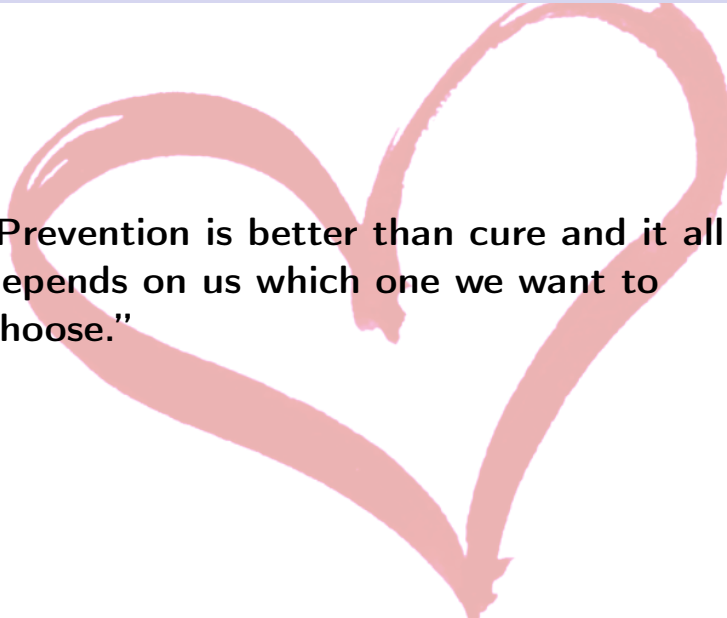
# HEART DISEASE PREDICTION USING LOGISTIC REGRESSION

ELIZABETH LIU

TWU

March 4, 2022

# MOTIVATION



**“Prevention is better than cure and it all depends on us which one we want to choose.”**

# PROJECT OUTLINE

1. BACKGROUND
2. PROJECT GOALS
3. MATH
4. DATASET DESCRIPTION
5. MODELLING
6. RESULTS
7. CONCLUSION
8. REFERENCES

# 1. BACKGROUND SECTION

1. INTRODUCTION

2. MATH BACKGROUND

## 1.1 BACKGROUND: INTRODUCTION (1/3)

### Cardiovascular Disease (CVD)

Heart disease or cardiovascular disease (CVD) is the leading cause of death for both men and women.

Because of the devastating effects of heart disease, prevention efforts must be made.

Prevention efforts are most effective when the person knows their risk.

## 1.1 BACKGROUND: INTRODUCTION (2/3)

### Data Mining

Data mining is the process of extracting and discovering patterns in large data sets.

Nowadays with the ability to store copious amounts of patient health data, data mining can be used in the healthcare sector to make meaningful decisions of the diagnosis and treatment of patients.

This includes the diagnosis of heart disease.

## 1.1 BACKGROUND: INTRODUCTION (3/3)

### Machine learning (ML)

Machine learning methods have been widely used in building a predictive models for heart disease using all kinds of patient data.

In machine learning, a model is built with training data and learns from the training data by making associations between different features and the outcome.

The model can then be evaluated by test data for improvement.

## 1.2 BACKGROUND: MATH (1/5)

### Logistic Regression

While there are many approaches to building a predictive model, Logistic Regression will be explored.

Logistic Regression is applicable when the outcome is binomial; the patient has heart disease and the patient does not have heart disease.

In the Logistic Regression model, the probability of a particular outcome can be predicted, which in this case is the risk of heart disease



## 1.2 BACKGROUND: MATH (2/5)

### Cost Function

The cost function is a function that calculates the distance between the predicted outcome and the actual outcome.

The cost function can be thought of as measuring the 'error' in the model.

## 1.2 BACKGROUND: MATH (3/5)

### Gradient Descent

Gradient descent is a method that can be used to solve the minimization of the cost function.

Gradient descent is a algorithm that searches for the lowest point in the gradient, which is the point of lowest 'error' or the most minimal cost.

This is where the most optimal combination of coefficients lie.

## 1.2 BACKGROUND: MATH (4/5)

### **Model Reliability**

The model needs to be able to provide a reliable diagnosis to influence health decisions in real life.

Therefore, the model will be evaluated and compared to the base R `glm()` model performance.

Both models will be evaluated using confusion matrix metrics.

## 1.2 BACKGROUND: MATH (5/5)

### **Confusion Matrix**

A confusion matrix is a table that is commonly used to visualize and evaluate the performance of a model.

From this table, useful metrics can be obtained that can be used to compare the two models.

## 2. PROJECT GOALS

- ▶ Logistic Regression will be used to build a model that can predict the risk of heart disease using a heart disease dataset.
- ▶ The model will be optimized by Gradient Descent.
- ▶ Afterward, the model will be evaluated using confusion matrix.
- ▶ Lastly, the model performance will be compared with the base R `glm()` model performance to see if the model built can compete with other model performances.

## 3. MATH SECTION

1. LOGISTIC REGRESSION
2. COST FUNCTION
3. GRADIENT DESCENT

## 3.1 MATH: LOGISTIC REGRESSION (1/4)

### Logistic Regression

Logistic regression can be thought of as a linear model with a binomial distribution and a logit link function.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Given the probability of  $p$ , the odds outcome is calculated as  $\frac{p}{(1-p)}$ . The logit function is the logarithm of the odds.

## 3.1 MATH: LOGISTIC REGRESSION (2/4)

The sigmoid function is the inverse of the logit function.

$$g(z) = \frac{1}{1 + e^{-z}}$$

The sigmoid function scales arbitrary real values to the range  $[0, 1]$ .

Using logit and sigmoid functions, variables and outcomes can be transform to fit a probabilistic scale.



## 3.1 MATH: LOGISTIC REGRESSION (3/4)

The linear relationship between  $x_i$  features and the outcome can be represented in this equation below.

$$\log\left(\frac{p^{(i)}}{1 - p^{(i)}}\right) = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots$$

This also can be written in matrix form.

## 3.1 MATH: LOGISTIC REGRESSION (4/4)

$$h(x) = \frac{1}{1 + e^{-\theta \cdot X}}$$

$X$  represents a matrix with vectors of all  $n$  data points.

$\theta$  is a vector with the parameters for each feature.

To create the model,  $\theta$  must be found.

## 3.2 MATH: COST FUNCTION (1/2)

### Cost Function

The cost function is defined as:

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

The cost function can be used to determine how fit the model is to the data.

For each data point, the cost function measures each distance the actual outcome or  $y$  is from the predicted  $h_{\theta}(x)$  value.

The "cost" or error increases the further the  $h_{\theta}(x)$  is from the actual outcome and vice versa

## 3.2 MATH: COST FUNCTION (2/2)

The combined version of the cost function is:

$$J = -\frac{1}{n} \sum_{i=1}^n (y^{(i)} \log(h(x^{(i)})) + (1 - y^{(i)}) \log(1 - h(x^{(i)})))$$

The cost function can be minimized to solve for the optimal  $\theta$  values. This optimization can be done with the gradient descent algorithm.

### 3.3 MATH: GRADIENT DESCENT (1/2)

Gradient descent is an iterative search algorithm that is used to search for  $\theta$  values that give minimum error in the cost function.

The general form of gradient descent is:

Repeat until convergence {  
 $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$   
}

### 3.3 MATH: GRADIENT DESCENT (2/2)

The derivative is what makes the search possible and delegates the direction the algorithm should search.

Positive derivative value means there is a positive slope up ahead in the gradient, which indicates a local maximum.

A negative derivative value means there is a negative slope and a local minimum.

In this case, the gradient descent algorithm would focus the search towards increasingly negative derivative values, and update the  $\theta$  values iteratively until the most optimal solution is reached.

## 4. DATASET DESCRIPTION (1/2)

The source of the data comes from the online UCI Machine Learning Repository.

The heart disease repository consists of 76 attributes.

The dataset used in this paper contains 270 observations and 14 attributes.

More details about the 14 attributes on the next slide.

# DATASET DESCRIPTION (2/2)

Name	Description	Value
Age	age in years	numeric
Sex	-	1 = male 0 = female
Chest	chest pain type	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic
BP	resting blood pressure in mm Hg	numeric
Cholesterol	serum cholestoral in mg/dl	numeric
FBS_over_120	fasting blood sugar >120 mg/dl	1 = true 0 = false
EKG	resting electrocardiographic results	0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular hypertrophy
Max_HR	maximum heart rate achieved	numeric
Exercise_angina	exercise induced angina	1 = yes 0 = no
ST_depression	ST depression induced by exercise relative to rest	numeric
Slope_ST	the slope of the peak exercise ST segment	1 = upsloping 2 = flat 3 = downsloping
vessels_fluro	number of major vessels colored by flourosopy	0 to 3
Thallium	test to visualize blood flow in heart	3 = normal 6 = fixed defect 7 = reversible defect
Heart_Disease, target variable	diagnosis of heart disease (angiographic disease status)	0 = <50% diameter narrowing 1 = >50% diameter narrowing



## 5. MODELLING SECTION

1. FEATURE SELECTION
2. BUILD model1
3. BUILD model2

## 5.1 MODELLING: FEATURE SELECTION (1/2)

A correlation matrix was plotted to determine which features had strong correlations with the target variable 'Heart\_Disease'.

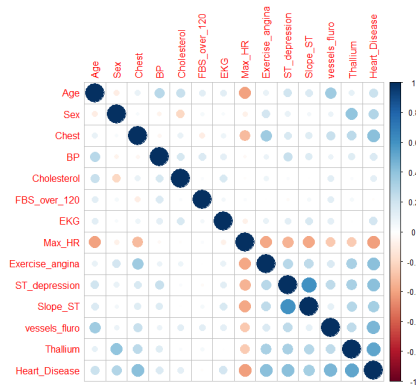


Figure 1: Feature Correlation plot

## 5.1 MODELLING: FEATURE SELECTION (2/2)

Coefficients:

	Estimate	Std. Error	z	value	Pr(> z )						
(Intercept)	-6.4996	1.0077	-6.450	1.12e-10	***						
Thallium	0.3991	0.1031	3.873	0.000107	***						
vessels_fluro	0.9397	0.2540	3.700	0.000216	***						
Chest	0.9370	0.2481	3.776	0.000159	***						
ST_depression	0.6242	0.1888	3.305	0.000948	***						
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

Figure 2: Feature Significance

Features 'Thallium', 'vessels\_fluro', 'Chest', and 'ST\_depression' had strong correlations with the target variable 'Heart\_Disease'. All four features, including the intercept, were significant enough features to build the Logistic Regression model.

## 5.2 MODELLING: BUILD model1 (1/6)

To build this model from scratch, the variables had to be defined.

$$\begin{matrix} & intercept & Thallium & vessels\_fluro & Chest & ST\_depression \\ \left( \begin{array}{ccccc} 1 & 3 & 3 & 4 & 2.4 \\ 1 & 7 & 0 & 3 & 1.6 \\ 1 & 7 & 0 & 4 & 0.3 \\ 1 & 7 & 1 & 3 & 0.2 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{array} \right) \end{matrix}$$

Figure 3: X matrix

The X variable is a  $n \times (i + 1)$  matrix with  $i$  number of features and  $n$  rows of datapoints.

## 5.2 MODELLING: BUILD model1 (2/6)

The  $y$  variable is a  $n \times 1$  matrix that holds the actual outcome for each row, with  $n$  rows of datapoints.

$$y = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

Figure 4:  $y$  matrix

## 5.2 MODELLING: BUILD model1 (3/6)

The  $\theta$  variable is a  $5 \times 1$  matrix that holds the each of the  $\theta$  values.  $\theta$  matrix will start with zeros. The zeros will eventually be that will be updated to include the optimal  $\theta$  values after gradient descent.

$$\theta = \begin{matrix} \textit{intercept} \\ \textit{Thallium} \\ \textit{vessels\_fluro} \\ \textit{Chest} \\ \textit{ST\_depression} \end{matrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Figure 5:  $\theta$  matrix

## 5.2 MODELLING: BUILD model1 (4/6)

After the variables were initialized, the functions in the Logistic Regression model were defined. The sigmoid function, cost function, and the gradient function were created.

```
**Create Sigmoid Function**
```

```
sigmoid <- function(z){1/(1+exp(-z))}
```

```
**Create Cost Function**
```

```
cost <- function(theta , X, y){  
  m <- length(y) # number of training examples  
  h <- sigmoid(X %*% theta)  
  J <- (t(-y)%*%log(h)-t(1-y)%*%log(1-h))/m  
  J  
}
```

## 5.2 MODELLING: BUILD model1 (5/6)

**\*\*Create Gradient Function\*\***

```
grad <- function(theta, X, y){  
  m <- length(y)  
  h <- sigmoid(X%*%theta)  
  grad <- (t(X)%*%(h - y))/m  
  grad  
}
```



## 5.2 MODELLING: BUILD model1 (6/6)

**\*\*Build Model & Return Coefficients\*\***

*#use the optim function to perform gradient descent*

```
model1 <- optim(theta, fn = cost, gr = grad,  
               X = X, y = y)
```

*#return coefficients*

```
model1$par
```

## 5.3 MODELLING: BUILD model2 (1/1)

Using base R `glm()`, I created a Logistic Regression model 'model2' to compare with model1. The base R `glm()` model uses Fisher scoring and maximum likelihood methods for optimization of  $\theta$ .

```
model2 <- glm(Heart_Disease ~ Thallium +  
              vessels_fluro + Chest + ST_depression ,  
              data=train , family = binomial("logit"))
```

## 6. RESULTS SECTION

1. model1 & model2  $\theta$  VALUES
2. CONFUSION MATRICES
3. CONFUSION MATRIX METRICS
4. AOC-AUC

## 6.1 RESULTS: model1 & model2 $\theta$ VALUES (1/2)

Table 1: model1  $\theta$  values

intercept	-6.4972526
Thallium	0.3988059
vessels_fluro	0.9395348
Chest	0.9369585
ST_depression	0.6239453

Table 2: model2  $\theta$  values

intercept	-6.4995820
Thallium	0.3991371
vessels_fluro	0.9397398
Chest	0.9370489
ST_depression	0.6242250

## 6.1 RESULTS: model1 & model2 $\theta$ VALUES (2/2)

From these two figures, it is obvious that the  $\theta$  values of model1 and model2 are very similar.

This demonstrates that both methods are very comparable.

Next, the predictive ability of the both models will be put to the test to see how much these slight  $\theta$  value differences effect the performance of the models.

## 6.2 RESULTS: CONFUSION MATRICES (1/2)

*Note: 0 represents the Absence of Heart Disease, and 1 represents the Presence of Heart Disease.*

Table 3: model1

	FALSE	TRUE
0	43	2
1	7	29

Table 4: model2

	FALSE	TRUE
0	43	2
1	7	29

## 6.2 RESULTS: CONFUSION MATRICES (2/2)

Looking at the confusion matrix, model1 and model2 perform the same.

This demonstrates that the slight differences in the  $\theta$  between model1 and model2 does not significantly make a difference in performance.

## 6.3 RESULTS: CONFUSION MATRIX METRICS (1/2)

Table 5: Accuracy, sensitivity, & specificity percentages

accuracy	0.89
sensitivity	0.81
specificity	0.96



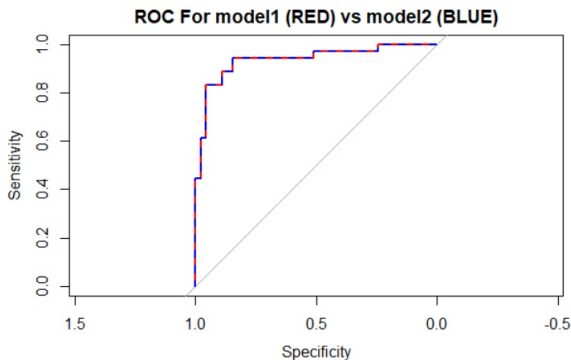
## 6.3 RESULTS: CONFUSION MATRIX METRICS (2/2)

The accuracy, sensitivity, and specificity determines how well both models can diagnose. Overall, accuracy, sensitivity, and specificity percentages are quite high, which means the models perform quite well and rarely give the incorrect diagnosis.

The sensitivity is 81%. This demonstrates the models were not too overgeneralized and still sensitive enough to detect of the presence of heart disease.

The specificity was at 95%, which is pretty high. A high specificity means that the models rarely give out false-positive results.

## 6.4 RESULTS: AOC-AUC (1/3)



"Area under curve of models: 0.937037037037037"

Figure 6: ROC graph for model1 & model 2

## 6.4 RESULTS: AOC (2/3)

The ROC graph can be plotted using the results from the confusion matrix.

ROC or Receiver Operator Characteristic graph summarizes the performance of both models by demonstrating the trade off between sensitivity and specificity.

The further the curve is from the gray line, the more predictive value there is in the model.

## 6.4 RESULTS: AUC (3/3)

The area under the curve or AUC score can also be calculated.

This score represents the probability that a randomly chosen positive instance will be achieved more often than negative instance.

The highest AUC score is 1.00.

The AUC score is 0.937, which is very close to 1.00. This demonstrates that both models performed quite well.

## 7. CONCLUSION (1/2)

### Summary of Project

Logistic regression with Gradient Descent is a useful way find the relationship between variables and a binomial outcome.

Logistic regression with Gradient Descent has equivalent results to the base R function `glm()` method in RStudio.

## 7. CONCLUSION (2/3)

### **Summary of Results**

Confusion matrix results indicated that the models perform quite well and rarely gives incorrect diagnoses.

High AUC score also demonstrated that both models have good predictive ability.

Overall, the results indicate that logistic regression with Gradient Descent is a good method to create a predictive model.

## 7. CONCLUSION (2/3)

### **Limitations of Study**

While the results of the project were seemingly successful, there are still ways the project can be improved.

## 8. REFERENCES (1/3)

- ▶ Berkson, J. (1944). Application to the logistic function to bio-assay. Journal of the American Statistical Association, 39(227), 357. <https://doi.org/10.2307/>
- ▶ Brixius, N. (2019). The logit and sigmoid functions, [<https://nathanbrixius.wordpress.com/2016/06/04/functions-i-have-known-logit-and-sigmoid/>] Nathan Brixius.
- ▶ Bhavsar, K. A., Abugabah, A., Singla, J., Alzubi, A. A., Bashir, A. K., & Nikita. (2021). comprehensive review on medical diagnosis using machine learning. Computers, Materials & Continua, 67(2), 1997–2014. <https://doi.org/10.32604/cmc.2021.014943>
- ▶ Institute of Medicine (US) Committee on a National Surveillance System for Cardiovascular and Select Chronic Diseases. (2011). Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung (2, Cardiovascular Disease). Washington (DC): National Academies Press (US); Retrieved 2021, from: <https://www.ncbi.nlm.nih.gov/books/NBK83160>



## 8. REFERENCES (2/3)

- ▶ Katz, M. H. (2003). Multivariable Analysis: A Primer for readers of Medical Research. Annals of Internal Medicine, 138(8), 644.  
<https://doi.org/10.7326/0003-4819-138-8-200304150-00012>
- ▶ M., J. (2019). Logistic regression from scratch in R, [<https://towardsdatascience.com/logistic-regression-from-scratch-in-r-b5b122fd8e83>], Medium.
- ▶ M., J. (2019). Logistic regression from scratch in R, [<https://towardsdatascience.com/logistic-regression-from-scratch-in-r-b5b122fd8e83>], Medium.
- ▶ Ogundele, I. O., Popoola, O. L., Oyesola, O. O., & Orija, K. T. (2018). A Review on Data Mining in Healthcare. International Journal of Advanced Research in Computer Engineering & Technology, 7(9).

## 8. REFERENCES (3/3)

- ▶ PM. Murphy, KW. Aha, UCI Repository of machine learning databases: Heart Disease Data Set, [<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>], Irvine, CA: University of California, Department of Information and Computer Science, 1994.
- ▶ Sperandei S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12–18.  
<https://doi.org/10.11613/BM.2014.003>
- ▶ Upadhyay, R. (2018). Gradient descent for logistic regression simplified - step by Step Visual Guide, [<http://ucanalytics.com/blogs/gradient-descent-logistic-regression-simplified-step-step-visual-guide>]. YOU CANalytics.