Logistic Regression

Cost Function & Gradient Descent

Rahul Singh
rsingh@arrsingh.com

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression is used to predict a binary value

A model that returns a binary value (true / false)

x is the independent variable

$$y = f(x)$$

y (dependent variable) that returns 0 (false) or 1 (true)

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Function

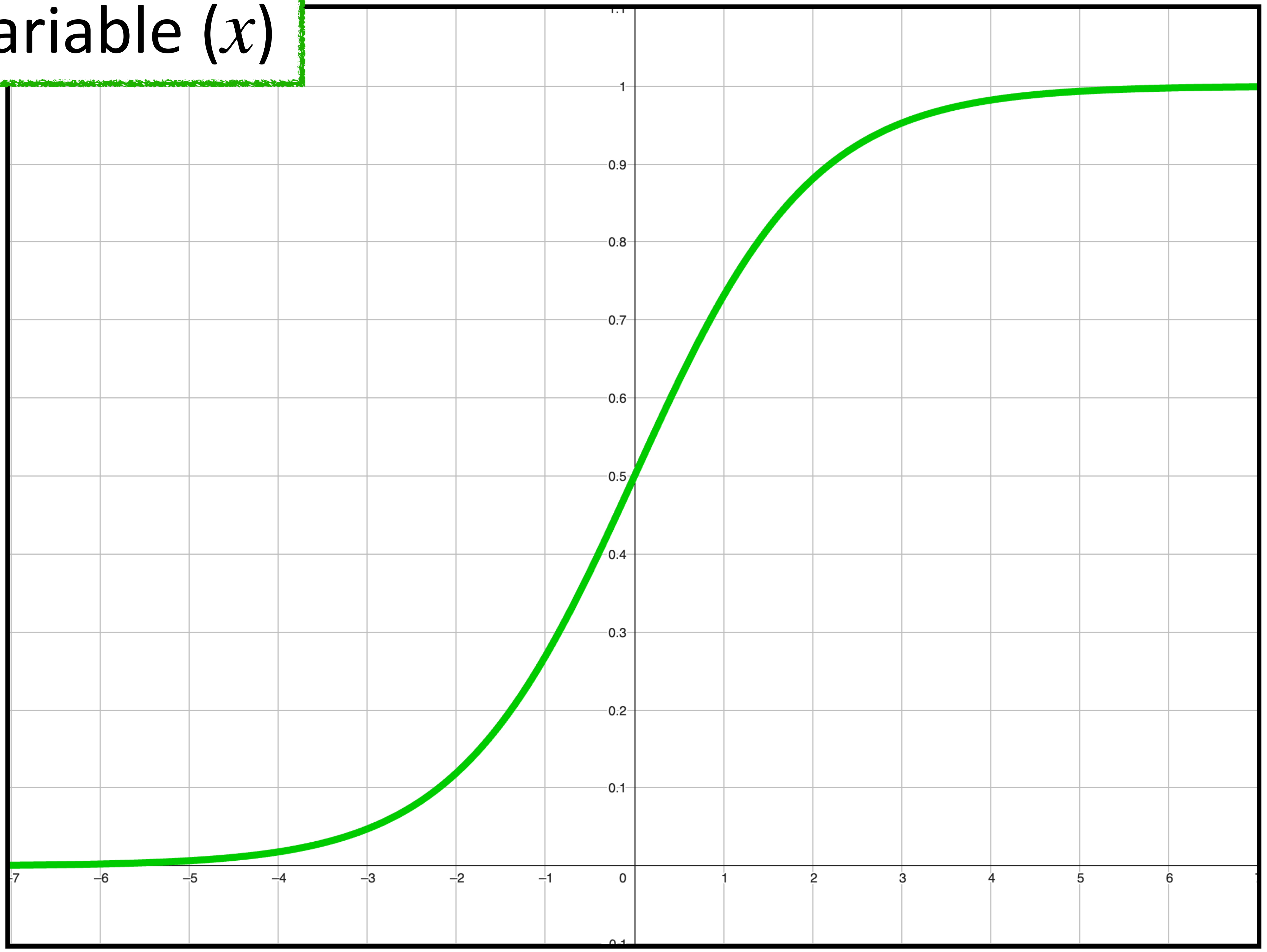
Logistic regression uses a sigmoid curve (Logistic Function) that converts a linear combination of input features (x values) into probabilities (y values)

$$y = \beta + wx$$

1 independent variable (x)

$$\sigma(y) = \frac{1}{1 + e^{-(\beta + wx)}}$$

1 dependent variable (y)



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Generalizing this to k independent variables…

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Function

The model is generalized to k independent variables and $k+1$ parameters

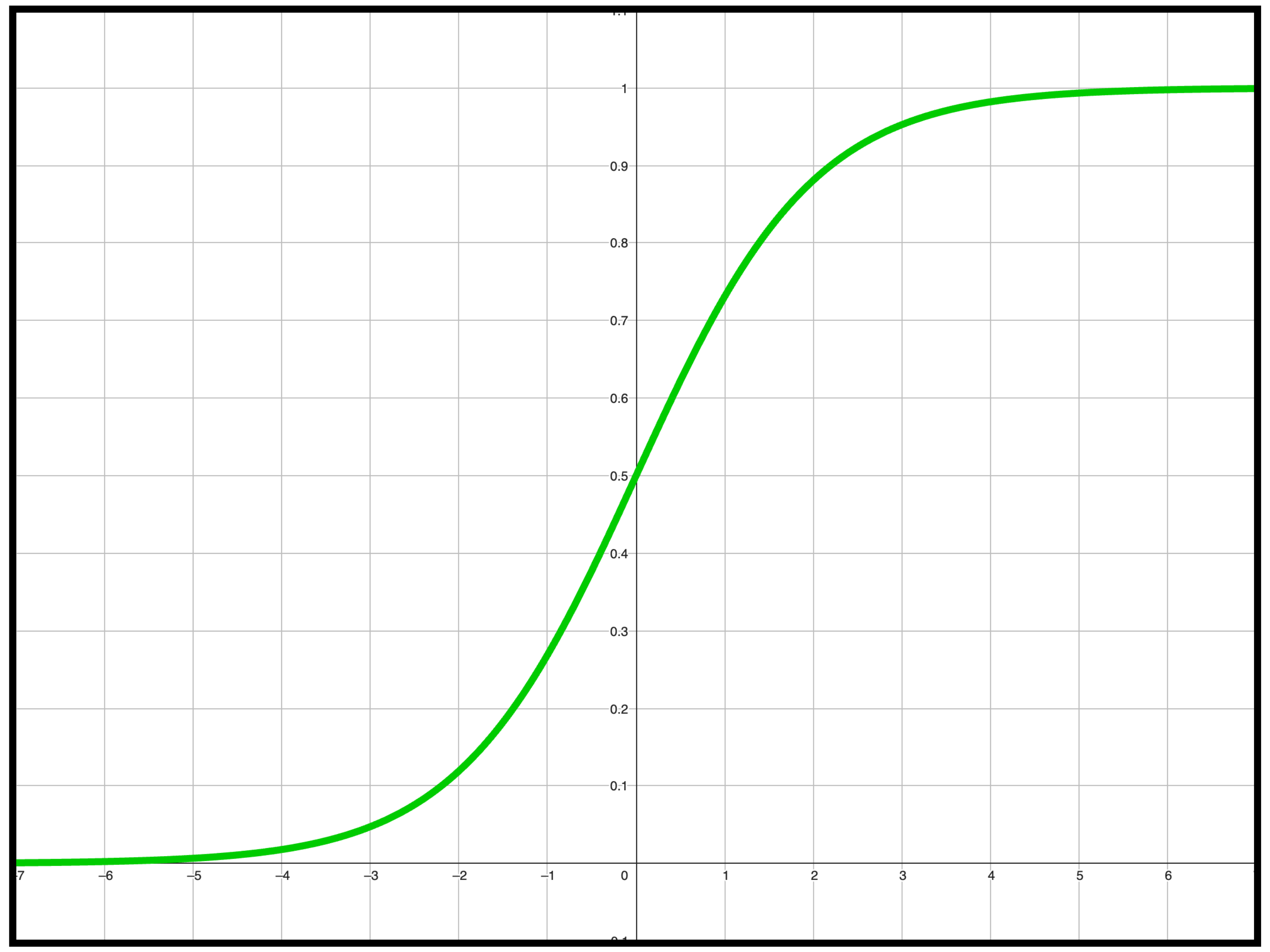
$$y = \beta + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_kx_k$$

1 dependent variable y

k independent variables $x_1 \,..\, x_k$

$$\sigma(y) = \frac{1}{1 + e^{-(\beta + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_kx_k)}}$$

This model has
$k+1$ parameters
$\beta, w_1, w_2 \ldots w_k$



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Function

The model is generalized to k independent variables and $k+1$ parameters

$$y = \beta + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_kx_k$$

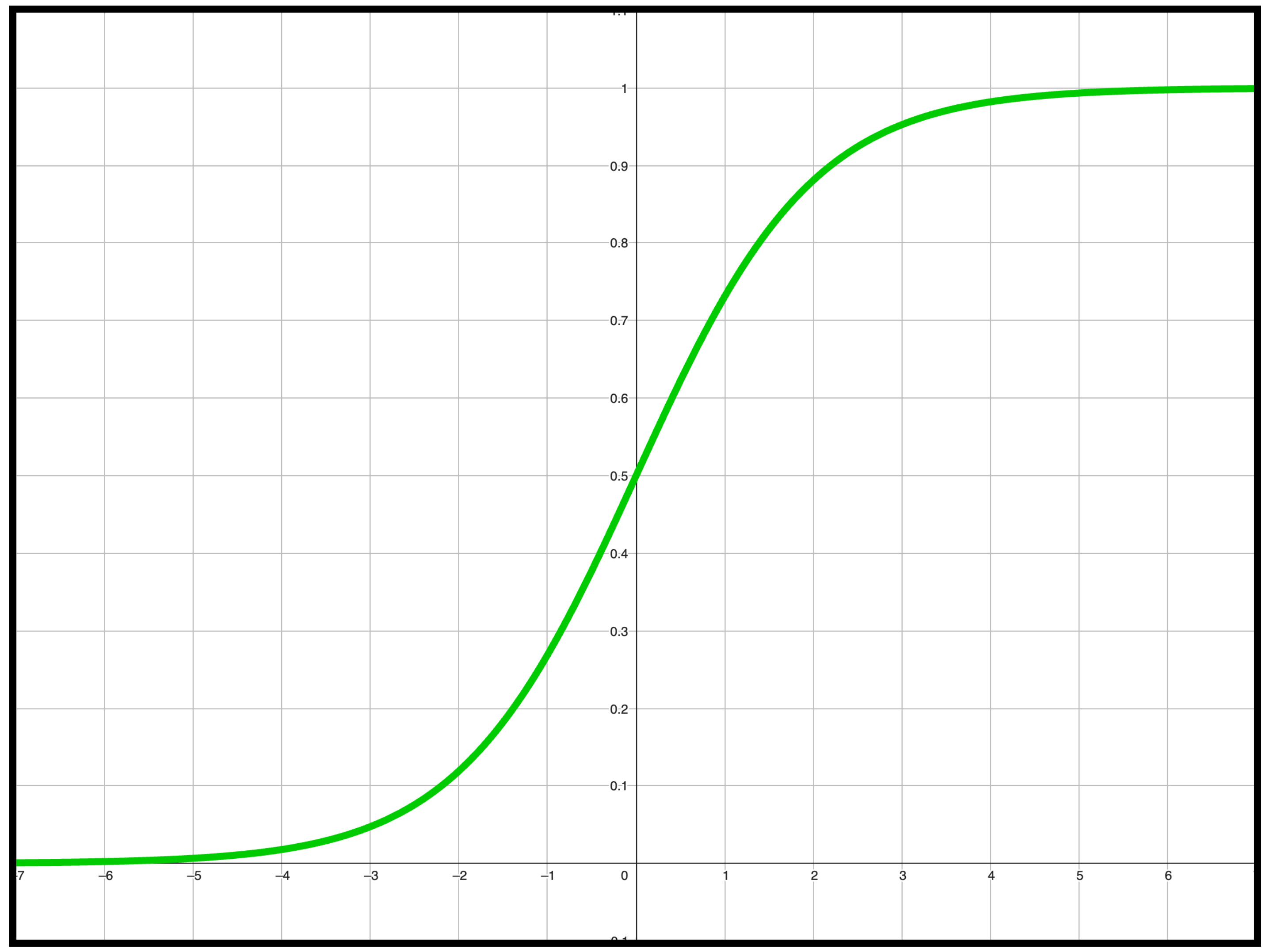
General Matrix form:

$$\hat{Y} = \sigma(W^TX + \beta)$$

W is a $k \times 1$ vector of weights $w_1, w_2, w_3 \ldots w_k$

X is a $k \times n$ matrix of n observations $x_1, x_2, x_3 \ldots x_k$

β is a scalar



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Function

The model is generalized to k independent variables and $k+1$ parameters

$$y = \beta + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_kx_k$$

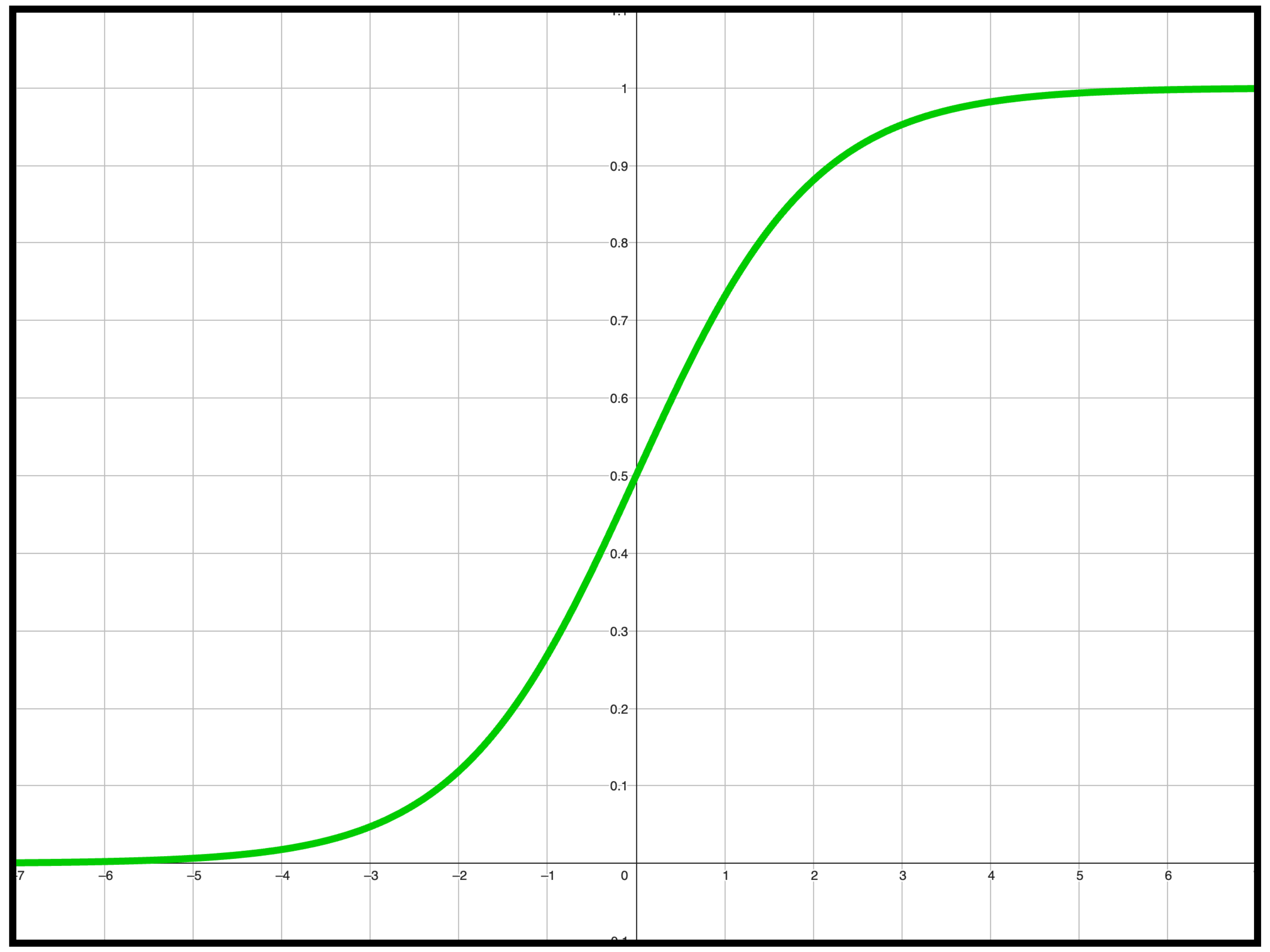
General Matrix form:

$$\hat{Y} = \sigma(W^TX + \beta)$$

$\hat{Y}$ is the vector of predicted values from the model.
$\hat{Y}$ is a vector of probabilities each between 0 and 1

A Threshold converts a given probability to a binary value

if $\hat{y}_i \geq 0.5$ then 1
if $\hat{y}_i < 0.5$ then 0



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression

Logistic Function

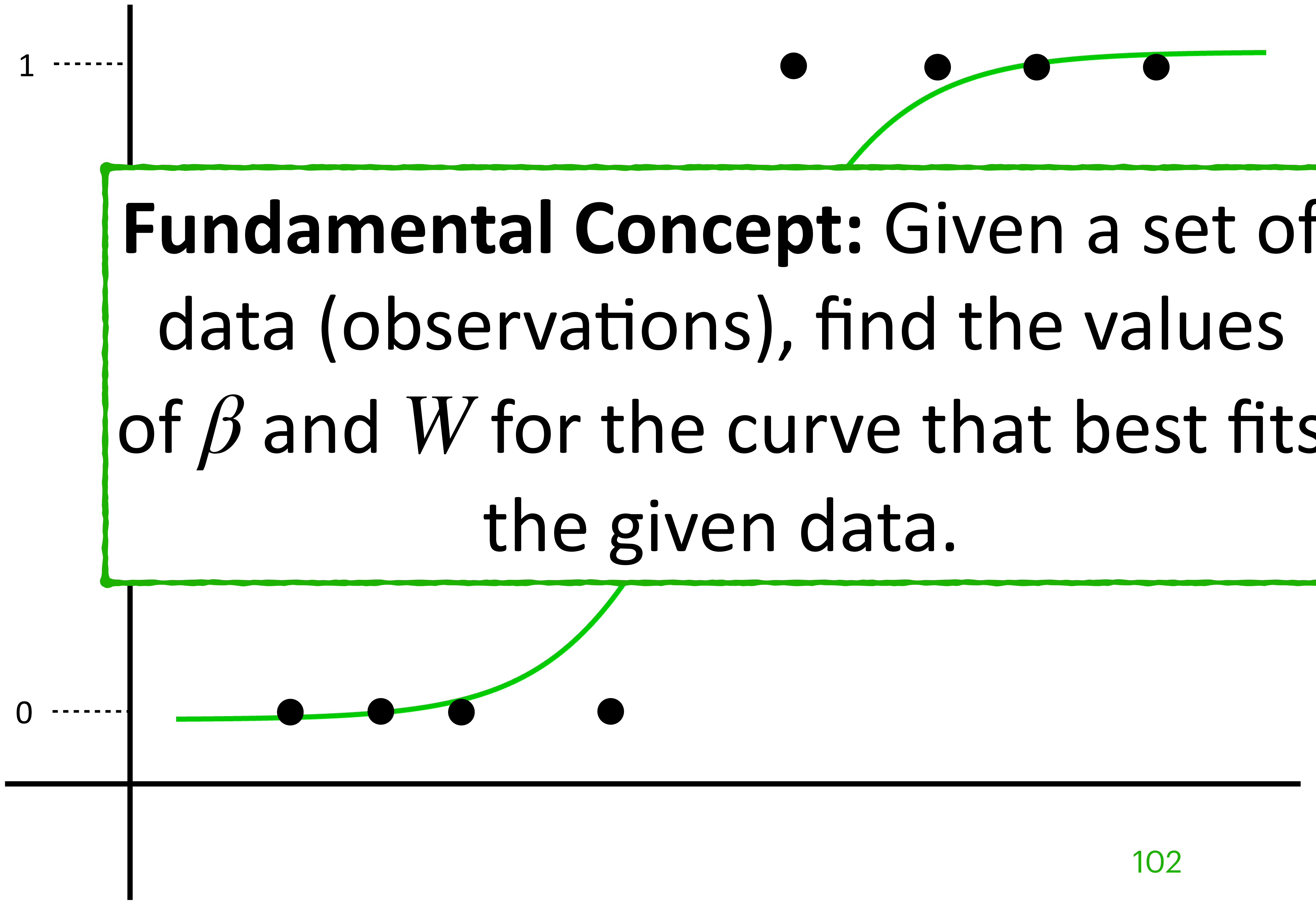
Logistic regression uses a sigmoid curve (Logistic Function) that converts a linear combination of input features (x values) into probabilities (y values)

Curve of best fit is…

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$\Rightarrow \hat{y}_i = \frac{1}{1 + e^{-(w_i x_i + \beta)}}$$

We can use **Gradient Descent** to find the optimal values of β and W



Fundamental Concept: Given a set of data (observations), find the values of β and W for the curve that best fits the given data.

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Recap: Gradient Descent and Linear Regression

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Multiple Regression

Linear Model in k Dimensions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_{k-1} x_{k-1}$$

$$\hat{Y} = X\beta$$

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & . & . & . & x_{(k-1)1} \\ 1 & x_{12} & x_{22} & & & & . & x_{(k-1)2} \\ 1 & x_{13} & x_{23} & & & & . & x_{(k-1)3} \\ \vdots & \vdots & \vdots & & & & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} & . & . & . & x_{(k-1)n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

$$\underset{n \times 1}{\hat{Y}} = \underset{n \times k}{X} \; \underset{k \times 1}{\beta}$$

The **Mean Squared Error (MSE):**

$$\frac{1}{2n} \| Y - X\beta \|^2$$

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Data is synthetic and not plotted to scale

Multiple Regression

$$\hat{Y} = X\beta$$

Linear Model in k Dimensions

The **Mean Squared Error (MSE):**

$$\frac{1}{2n} \| Y - X\beta \|^2$$

Cost Function

Gradient descent can be used to minimize the cost function.

We compute the partial derivative of the cost function w.r.t the parameters β

Partial Derivative w.r.t β:

$$\frac{\partial}{\partial \beta} \frac{1}{2n} \| Y - X\beta \|^2 = \frac{1}{n} X^T (X\beta - Y)$$

Partial Derivative w.r.t β

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Data is synthetic and not plotted to scale

Gradient Descent

A linear model in $k + 1$ dimensions…

$$\hat{Y} = X\beta$$

Cost Function (**Mean Squared Error (MSE)**):

$$\frac{1}{2n} \| Y - X\beta \|^2$$

Gradient Vector (Partial Derivative w.r.t β):

$$\frac{\partial}{\partial \beta} \frac{1}{2n} \| Y - X\beta \|^2 = \frac{1}{n} X^T (X\beta - Y)$$

Gradient Descent: Basic Concept

Step 1: Start with random values for β

Step 2: Compute the partial derivative of the cost function w.r.t β

Step 3: Calculate a step size that is proportional to the slope

Step 4: Calculate new values for β by subtracting the step size

Step 5: Go to step 2 and repeat

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

We need a cost function for Logistic Regression

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

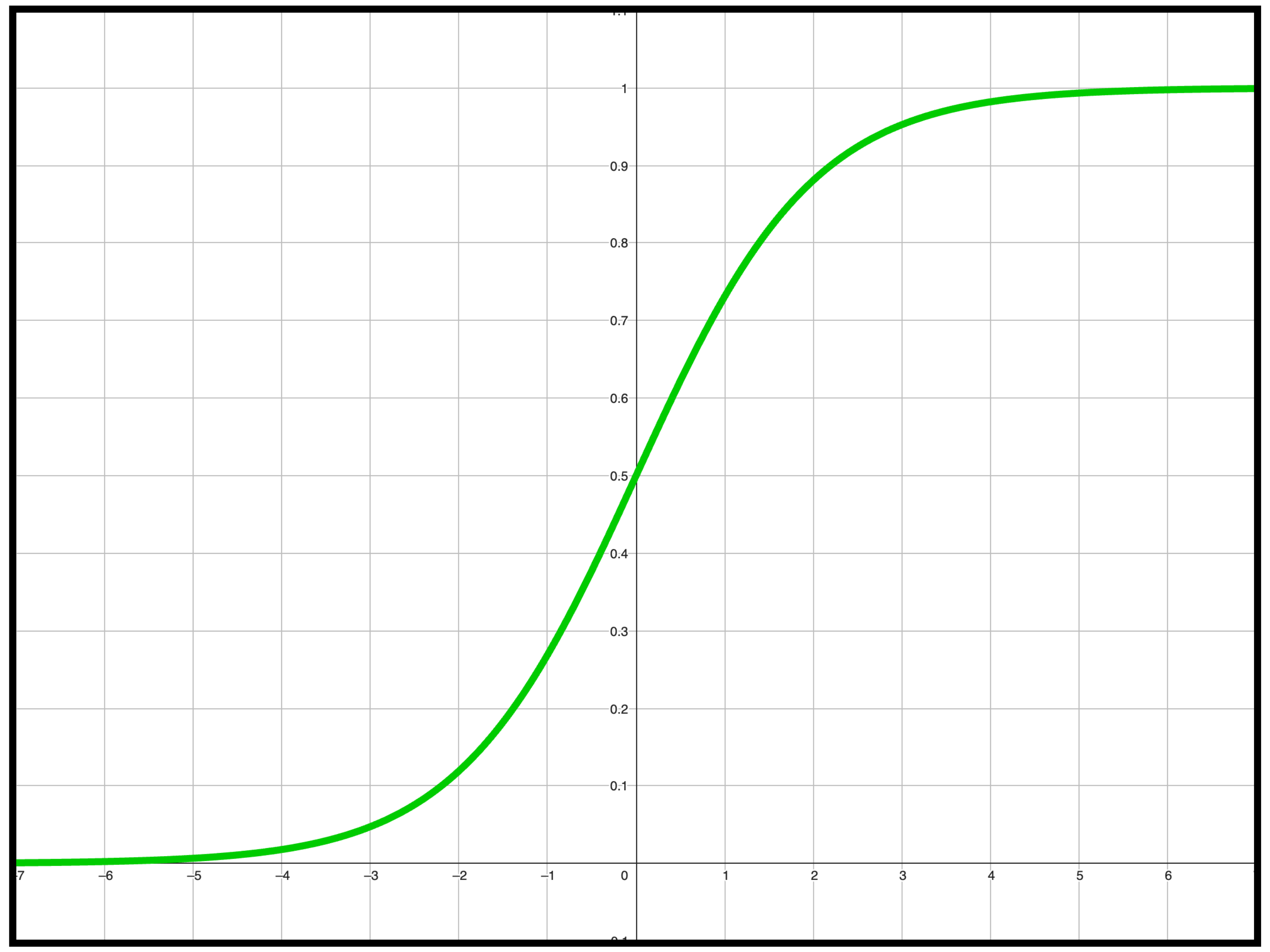
Logistic Regression

$$\hat{Y} = \sigma(W^T X + \beta)$$

W is a $k \times 1$ vector of weights $\omega_1, \omega_2, \omega_3 \ldots \omega_k$
X is a $k \times n$ matrix of observations $x_1, x_2, x_3 \ldots x_k$
β is a scalar

$\hat{Y}$ is the predicted values for the model with parameters W and β

Logistic Regression uses Maximum Likelihood Estimation to find the parameters W and β



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression

Logistic Regression uses Maximum Likelihood Estimation to find the parameters W and β

$$p(y\,|\,x; W, \beta)$$

Likelihood of a value y given a value x for a model with parameters W and β

Logistic Regression Cost Function

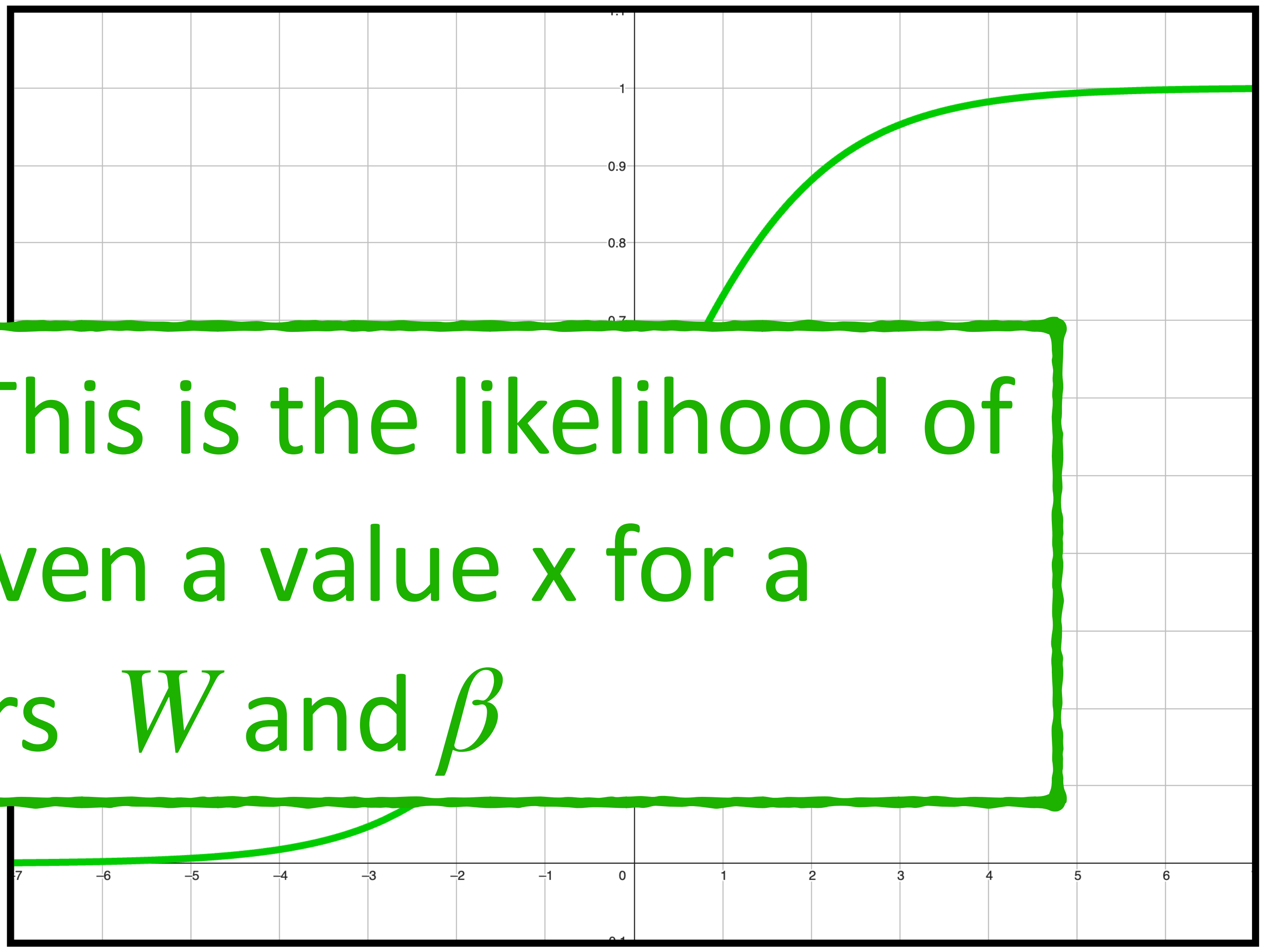
if $y = 1$ then $p(y\,|\,x; W, \beta) = \hat{y}$

if $y = 0$ then $p(y\,|\,x; W, \beta) = 1 - \hat{y}$

Logistic Regression Cost Function

A compact representation of the cost function

$$p(y\,|\,x; W, \beta) = \hat{y}^{y}\,(1 - \hat{y})^{(1-y)}$$



Likelihood Function: This is the likelihood of predicting a value y given a value x for a model with parameters W and β

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression

Logistic Regression uses Maximum Likelihood Estimation to find the parameters W and β

Logistic Regression Cost Function

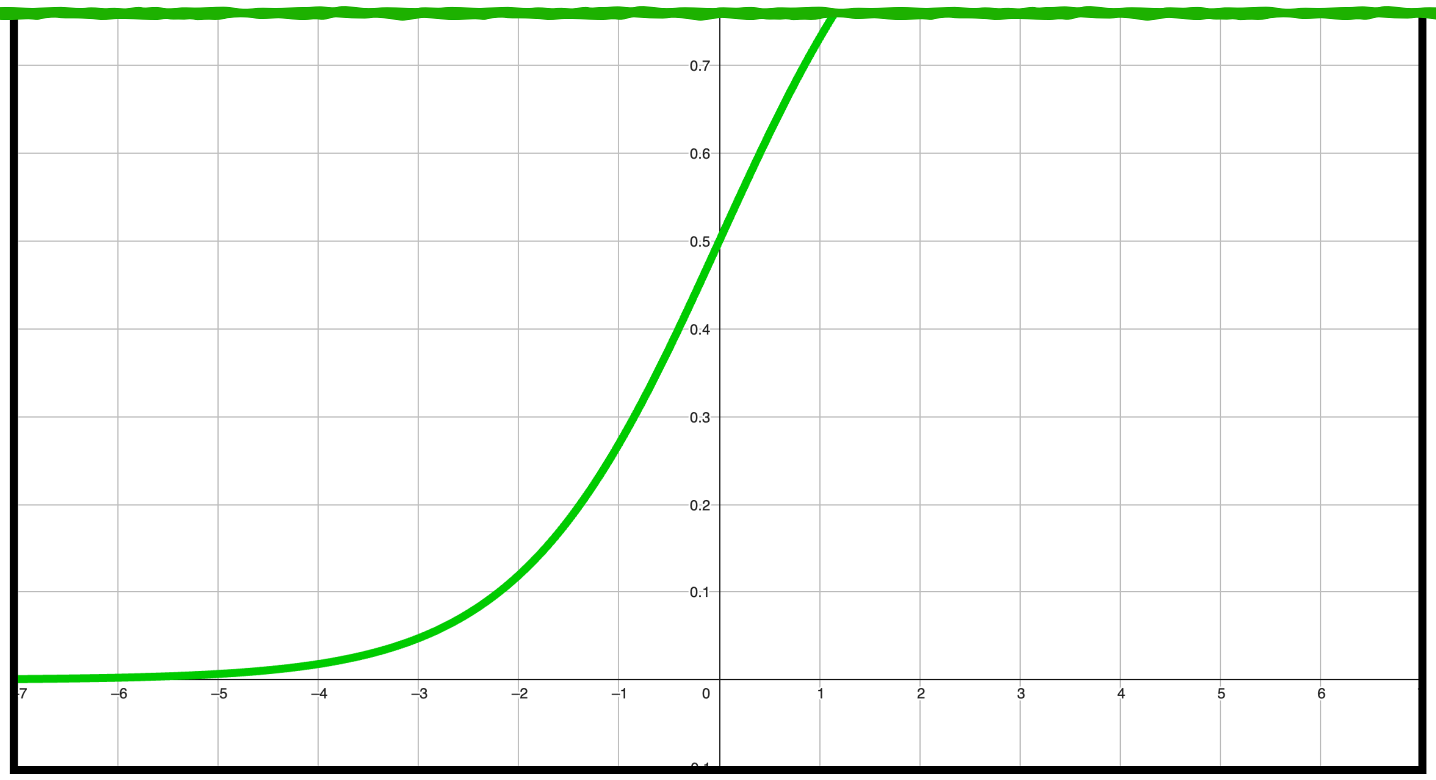
A compact representation of the cost function

$$p(y \mid x; W, \beta) = \hat{y}^{y}\,(1 - \hat{y})^{(1-y)}$$

Maximizing the Likelihood is the same as maximizing the log likelihood

Likelihood Function: This is the likelihood of predicting a value y given a value x for a model with parameters W and β

We want to find the values of W and β that maximize this function. Hence Maximum Likelihood Estimation



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression

Logistic Regression uses Maximum Likelihood Estimation to find the parameters W and β

Logistic Regression Cost Function

A compact representation of the cost function

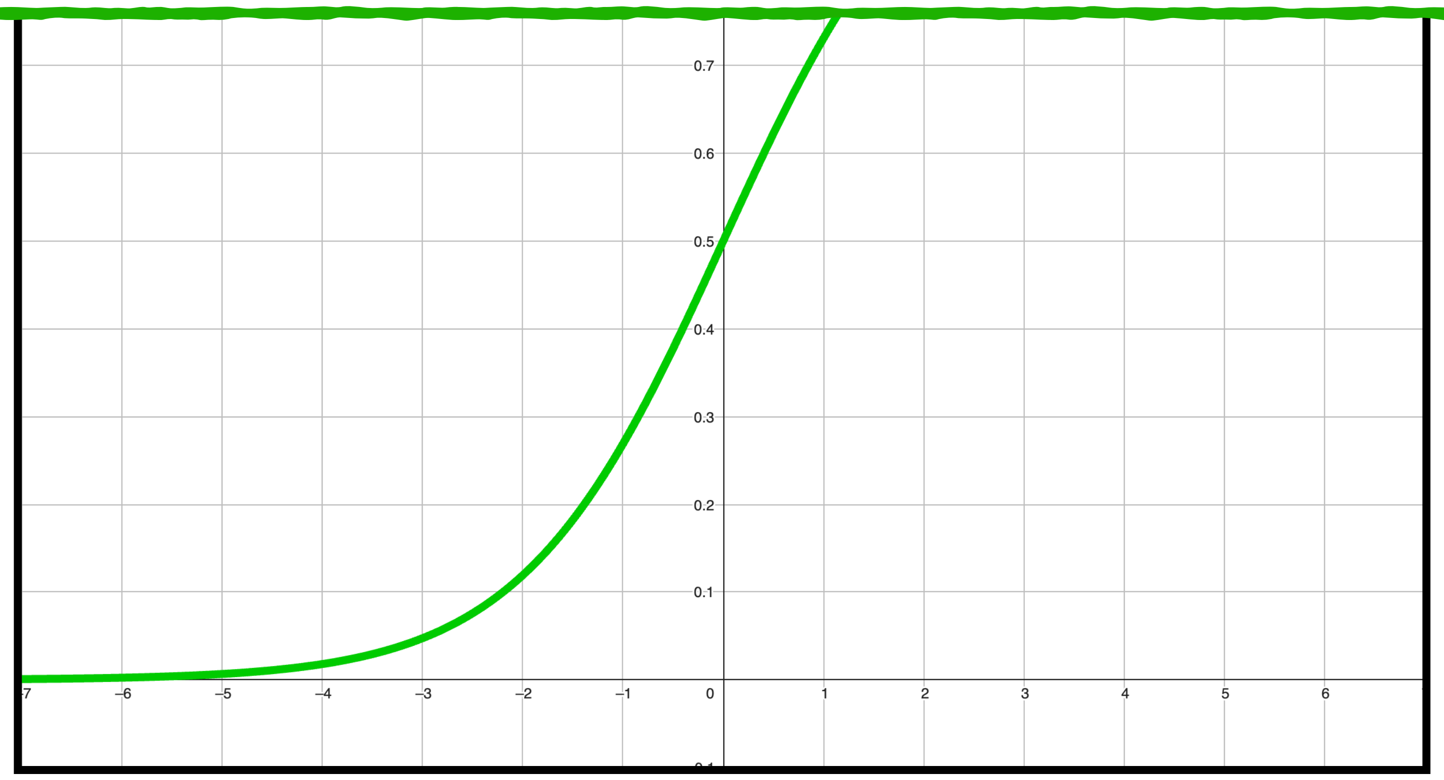
$$p(y \mid x; W, \beta) = \hat{y}^{y} (1 - \hat{y})^{(1-y)}$$

Likelihood Function: This is the likelyhood of predicting a value y given a value x for a model with parameters W and β

Maximizing the Likelihood is the same as maximizing the log likelihood

We want to find the values of W and β that maximize this function. Hence Maximum Likelihood Estimation

$$\begin{aligned} log_e\, p(y \mid x; W, \beta)) &= log_e\, \hat{y}^{y} (1 - \hat{y})^{(1-y)} \\ &= y\, log_e \hat{y} + (1 - y)\, log_e (1 - \hat{y}) \end{aligned}$$



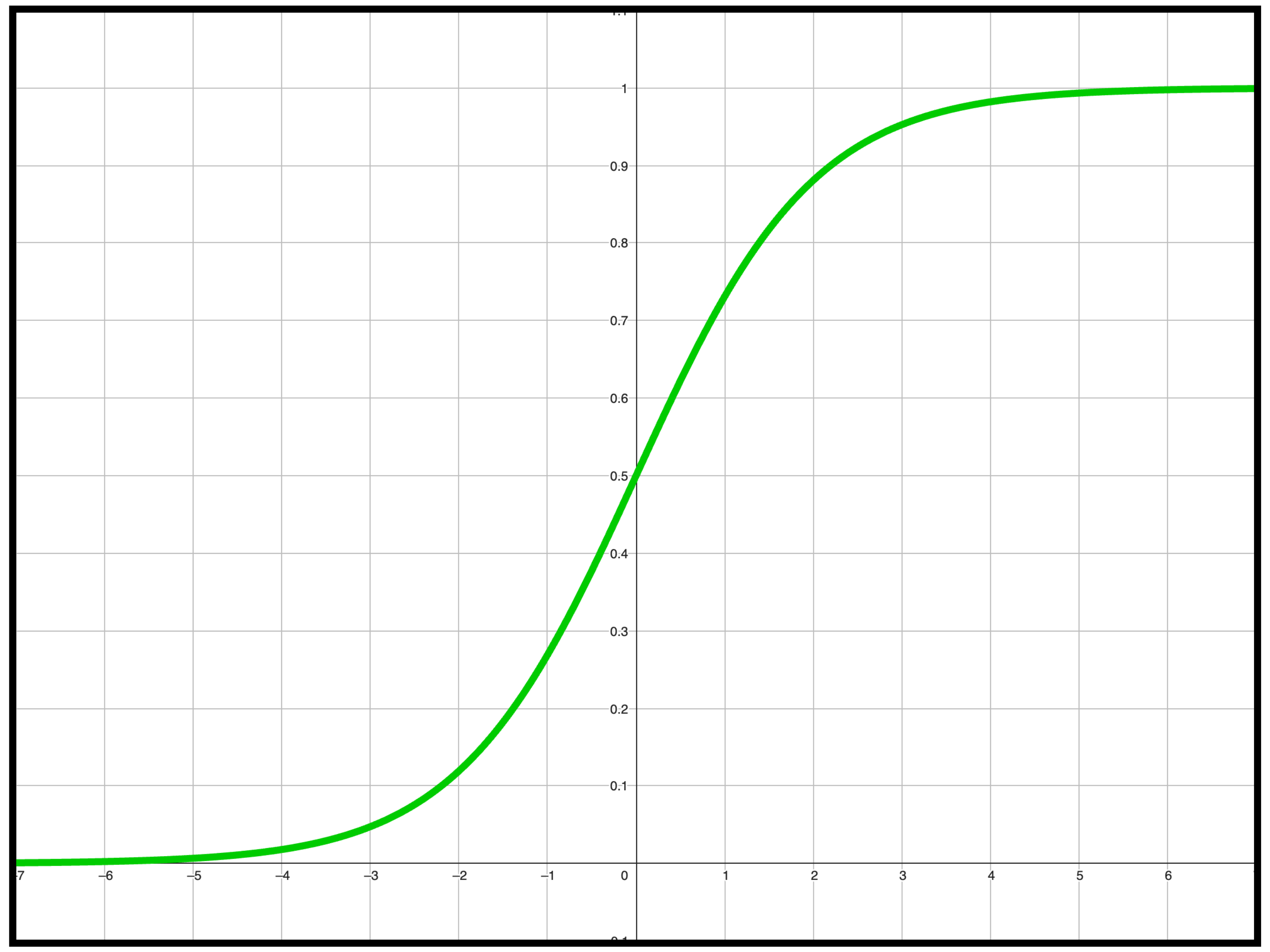
Next we average the cost across all observations…

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n} \sum_{i=1}^{n} y\, log_e \hat{y} + (1 - y)\, log_e (1 - \hat{y})$$



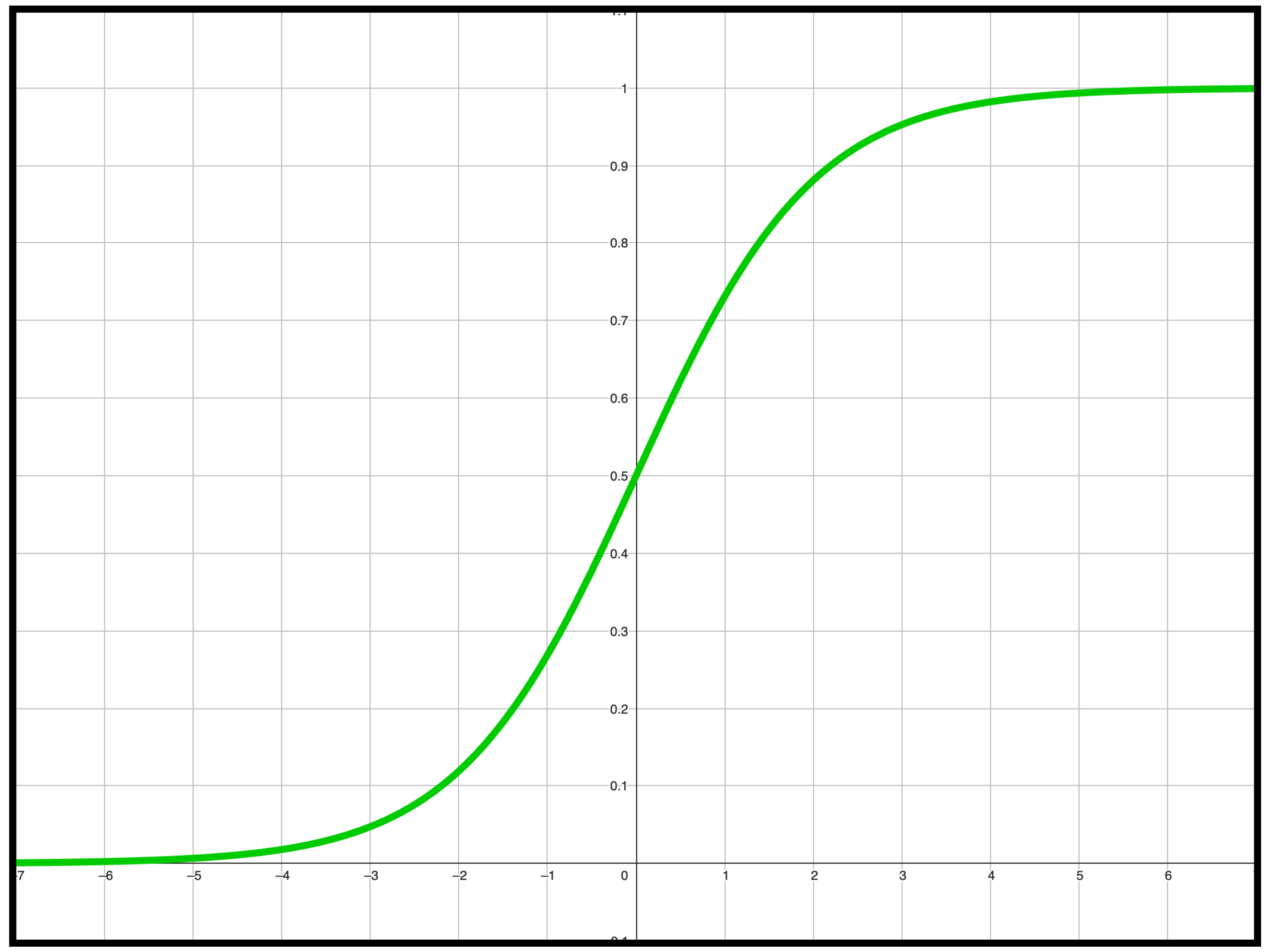
Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n}\sum_{i=1}^{n} y\, log_e \hat{y} + (1 - y)\, log_e (1 - \hat{y})$$

The negative sign indicates that we want to minimize the cost (aka maximize the likelihood)



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

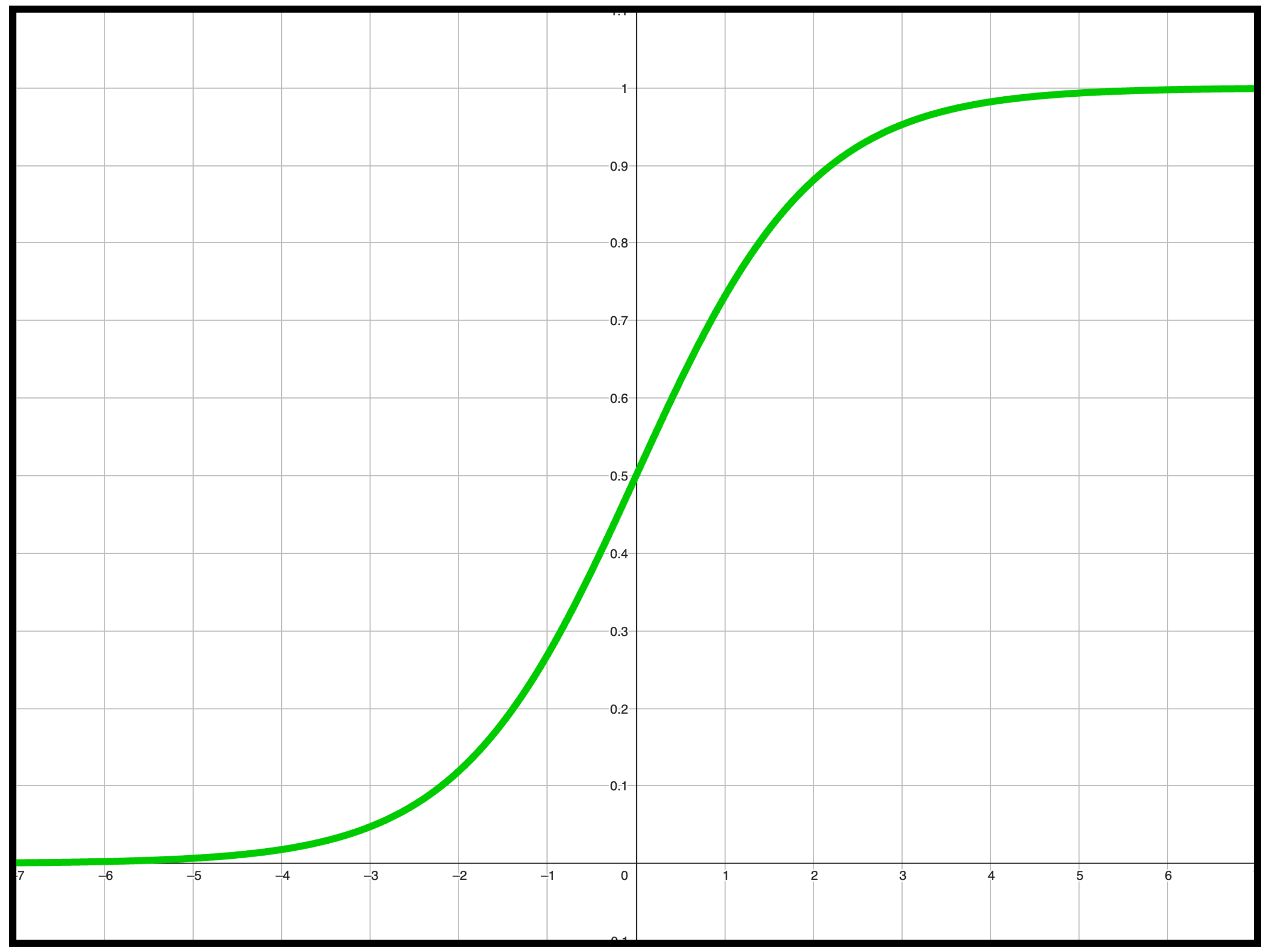
$$L(W, \beta) = -\frac{1}{n} \sum_{i=1}^{n} y \, log_e \hat{y} + (1 - y) \, log_e (1 - \hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y) \, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

With the cost function and the first derivatives above we can use gradient descent to estimate the values of W and β that minimize the cost



Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n} \sum_{i=1}^{n} y\, log_e \hat{y} + (1 - y)\, log_e (1 - \hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y)\, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Gradient Descent for Logistic Regression

Step 1: Start with random values for W and β

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n} \sum_{i=1}^{n} y \, log_e \hat{y} + (1 - y) \, log_e (1 - \hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y) \, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Gradient Descent for Logistic Regression

Step 1: Start with random values for W and β

Step 2: Compute the partial derivative of the cost function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y) \, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n}\sum_{i=1}^{n} y\, log_e \hat{y} + (1-y)\, log_e(1-\hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y)\, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Gradient Descent for Logistic Regression

Step 1: Start with random values for W and β

Step 2: Compute the partial derivative of the cost function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y)\, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Step 3: Calculate a step size that is proportional to the slope

$$step\_size_W = \frac{\partial}{\partial W} L(W, \beta) \times learning\_rate$$

$$step\_size_\beta = \frac{\partial}{\partial \beta} L(W, \beta) \times learning\_rate$$

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n}\sum_{i=1}^{n} y\, log_e \hat{y} + (1-y)\, log_e(1-\hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y)\, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Gradient Descent for Logistic Regression

Step 1: Start with random values for W and β

Step 2: Compute the partial derivative of the cost function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y)\, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Step 3: Calculate a step size that is proportional to the slope

$$step\_size_W = \frac{\partial}{\partial W} L(W, \beta) \times learning\_rate$$

$$step\_size_\beta = \frac{\partial}{\partial \beta} L(W, \beta) \times learning\_rate$$

Step 4: Calculate new values for W and β by subtracting the step size

$$W = W - step\_size_W$$

$$\beta = \beta - step\_size_\beta$$

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n}\sum_{i=1}^{n} y\, log_e \hat{y} + (1-y)\, log_e (1-\hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y) X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Gradient Descent for Logistic Regression

Step 1: Start with random values for W and β

Step 2: Compute the partial derivative of the cost function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y) X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Step 3: Calculate a step size that is proportional to the slope

$$step\_size_W = \frac{\partial}{\partial W} L(W, \beta) \times learning\_rate$$

$$step\_size_\beta = \frac{\partial}{\partial \beta} L(W, \beta) \times learning\_rate$$

Step 4: Calculate new values for W and β by subtracting the step size

$$W = W - step\_size$$

$$\beta = \beta - step\_size$$

Step 5: Go to step 2 and repeat

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Logistic Regression Cost Function

$$\hat{Y} = \sigma(W^T X + \beta)$$

$$L(W, \beta) = -\frac{1}{n}\sum_{i=1}^{n} y\, log_e \hat{y} + (1-y)\, log_e(1-\hat{y})$$

Partial Derivatives of the Cost Function w.r.t W and β

$$\frac{\partial}{\partial W} L(W, \beta) = (\hat{Y} - Y)\, X$$

$$\frac{\partial}{\partial \beta} L(W, \beta) = (\hat{Y} - Y)$$

Gradient Descent for Logistic Regression

Step 1: Start with random values for W and β

Step 2: Compute the partial derivative of the cost function w.r.t W and β

Gradient Descent continues in this manner until the step size is close to zero or a fixed number of iterations

$$step\_size_\beta = \frac{\partial}{\partial \beta} L(W, \beta) \times learning\_rate$$

Step 4: Calculate new values for W and β by subtracting the step size

$$W = W - step\_size$$
$$\beta = \beta - step\_size$$

Step 5: Go to step 2 and repeat

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)

Related Tutorials & Textbooks

Logistic Regression

An introduction to Logistic Regression. A Logistic Regression model use used to predict a binary value (the dependent variable) for one or more independent variables using a threshold to classify a probability.

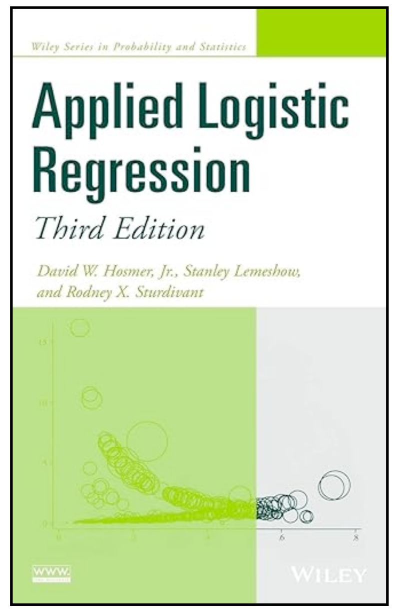
Multiple Regression

Multiple regression extends the two dimensional linear model introduced in Simple Linear Regression to k + 1 dimensions with one dependent variable, k independent variables and k+1 parameters.

Gradient Descent for Multiple Regression

Gradient Descent algorithm for multiple regression and how it can be used to optimize k + 1 parameters for a Linear model in multiple dimensions.

Recommended Textbooks



Applied Logistic Regression

by David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant

For a complete list of tutorials see:
https://arrsingh.com/ai-tutorials

Copyright (c) 2025, Rahul Singh, licensed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/)