Gradient Descent

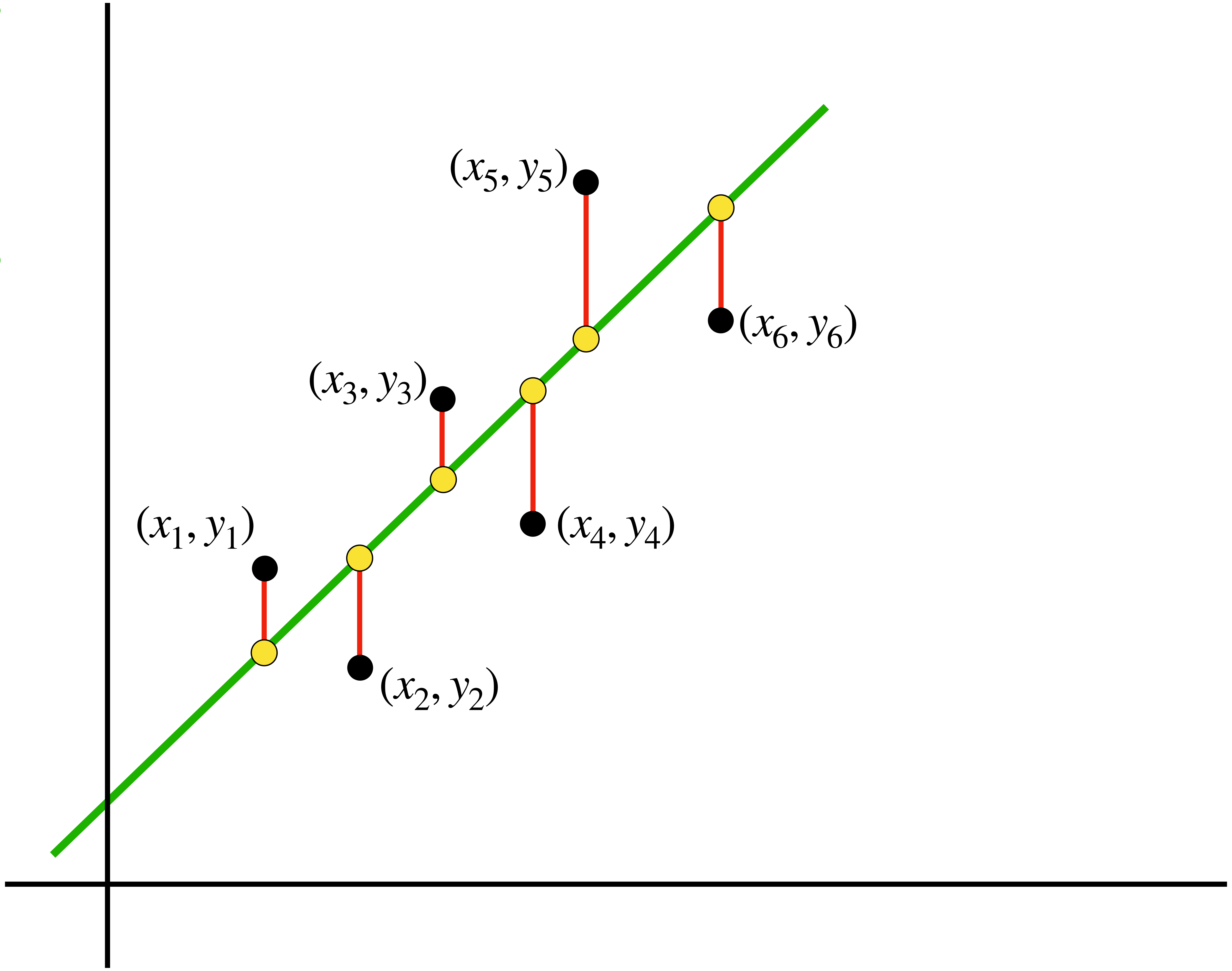
Multiple Regression using Gradient Descent

Rahul Singh
rsingh@arrsingh.com

Simple Linear Regression

The Problem Statement:

Simple Linear Regression: Find the values of β_0 and β_1 such that the **Mean Squared Error (MSE)** is minimized.



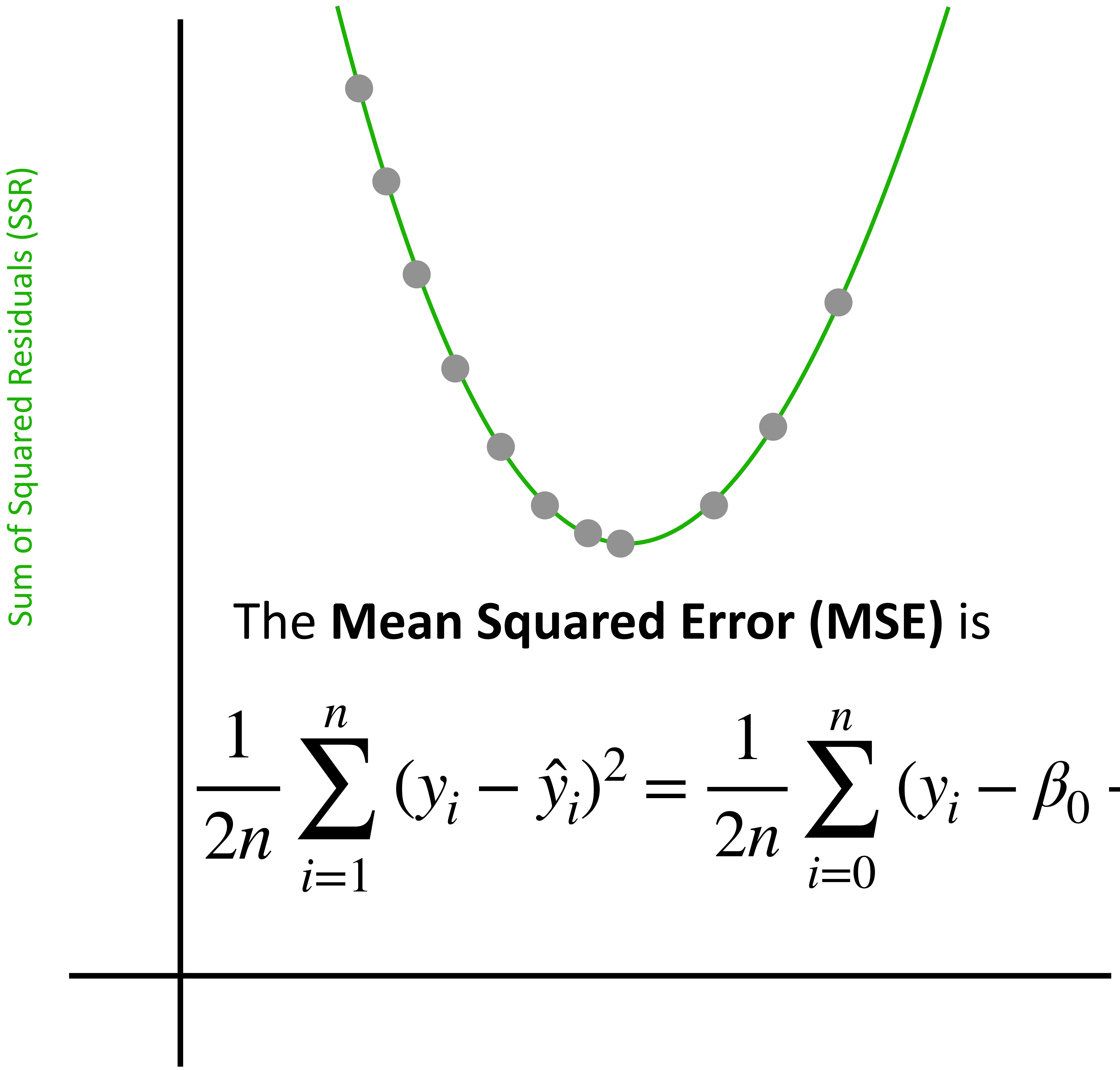
The line of best fit is $\hat{y} = \beta_0 + \beta_1 \hat{x}$

Mean Squared Error (MSE)

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

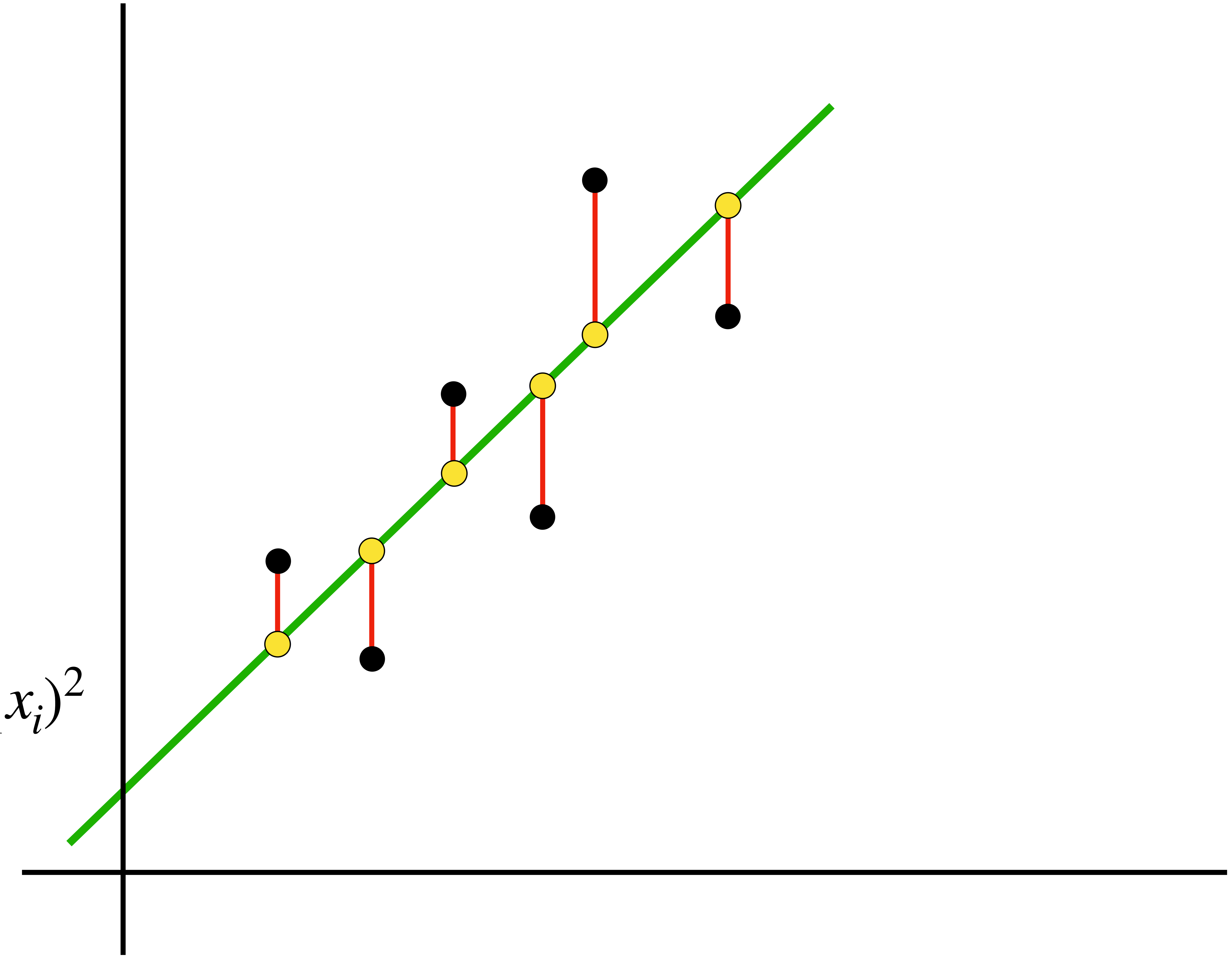
Data is synthetic and not plotted to scale

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



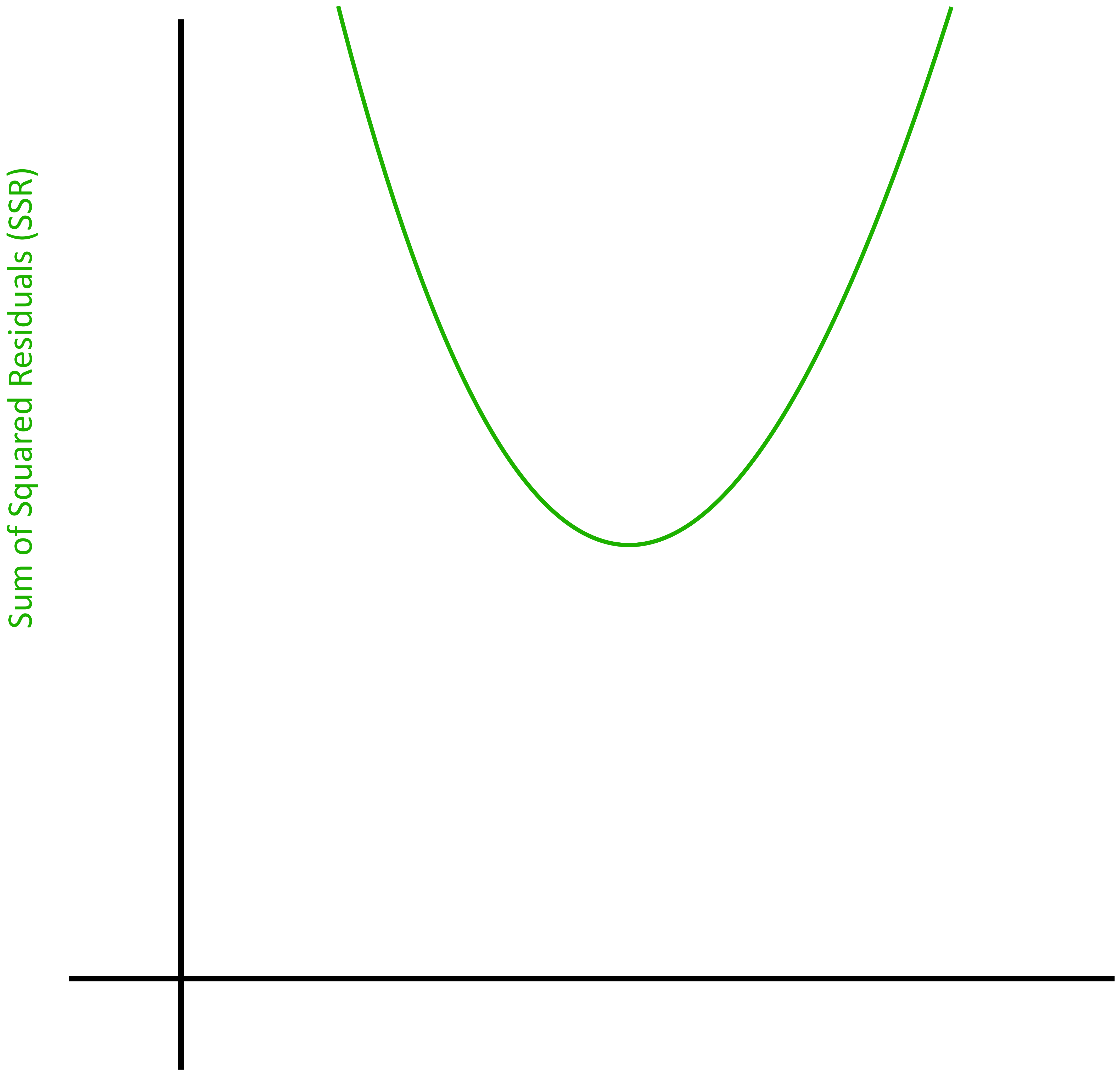
The **Mean Squared Error (MSE)** is

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$



Data is synthetic and not plotted to scale

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

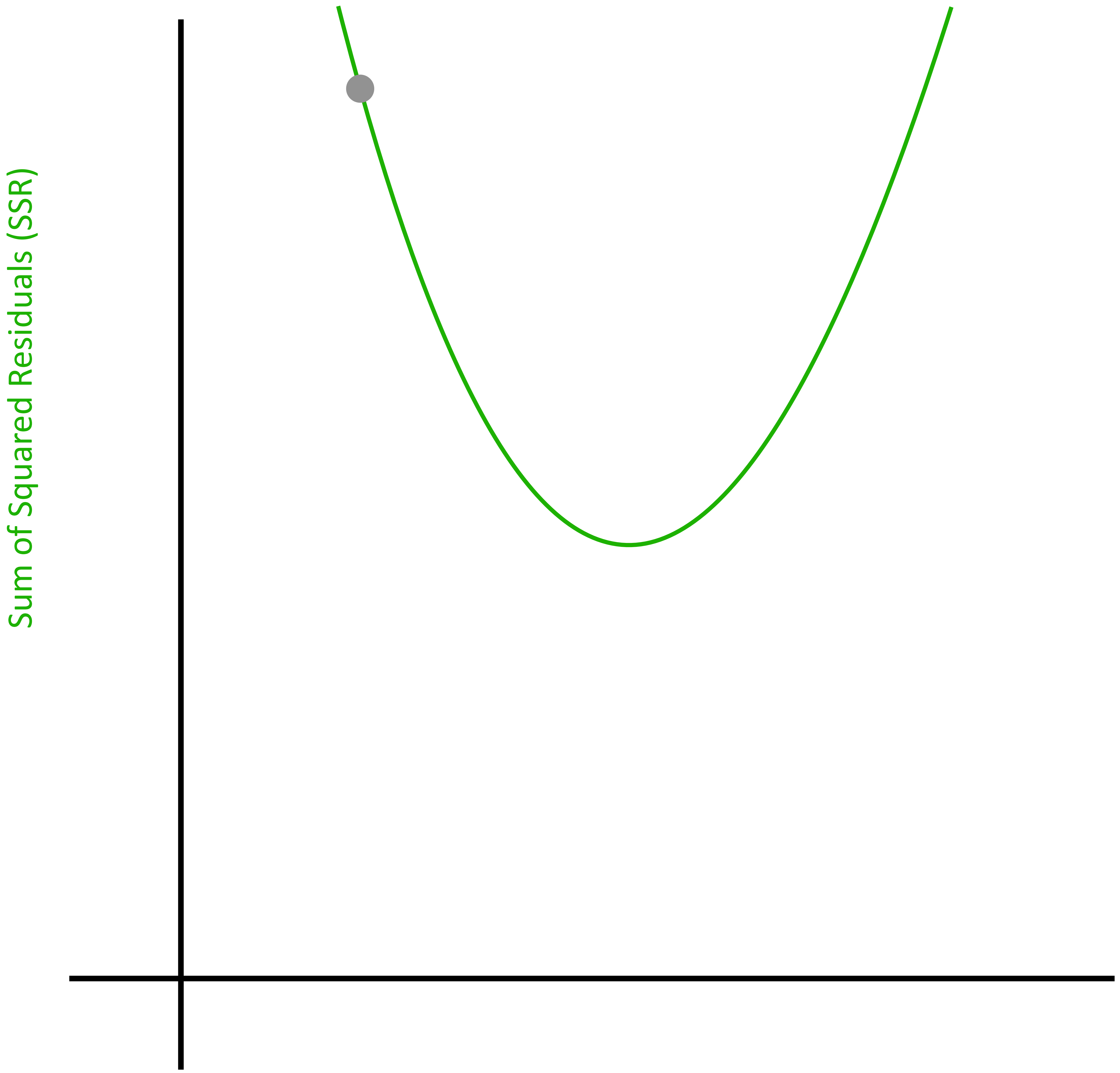
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent

Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

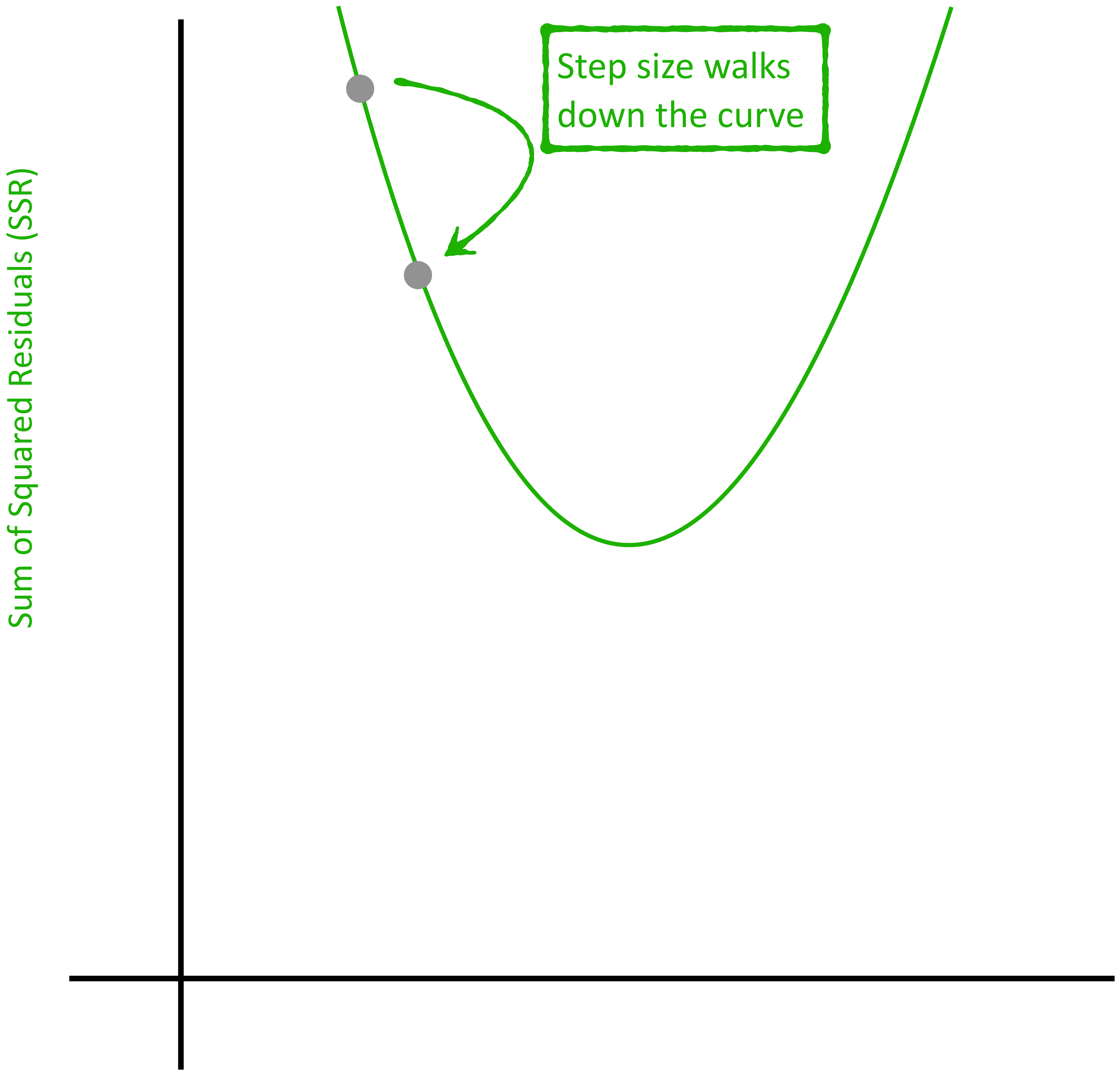
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent

Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

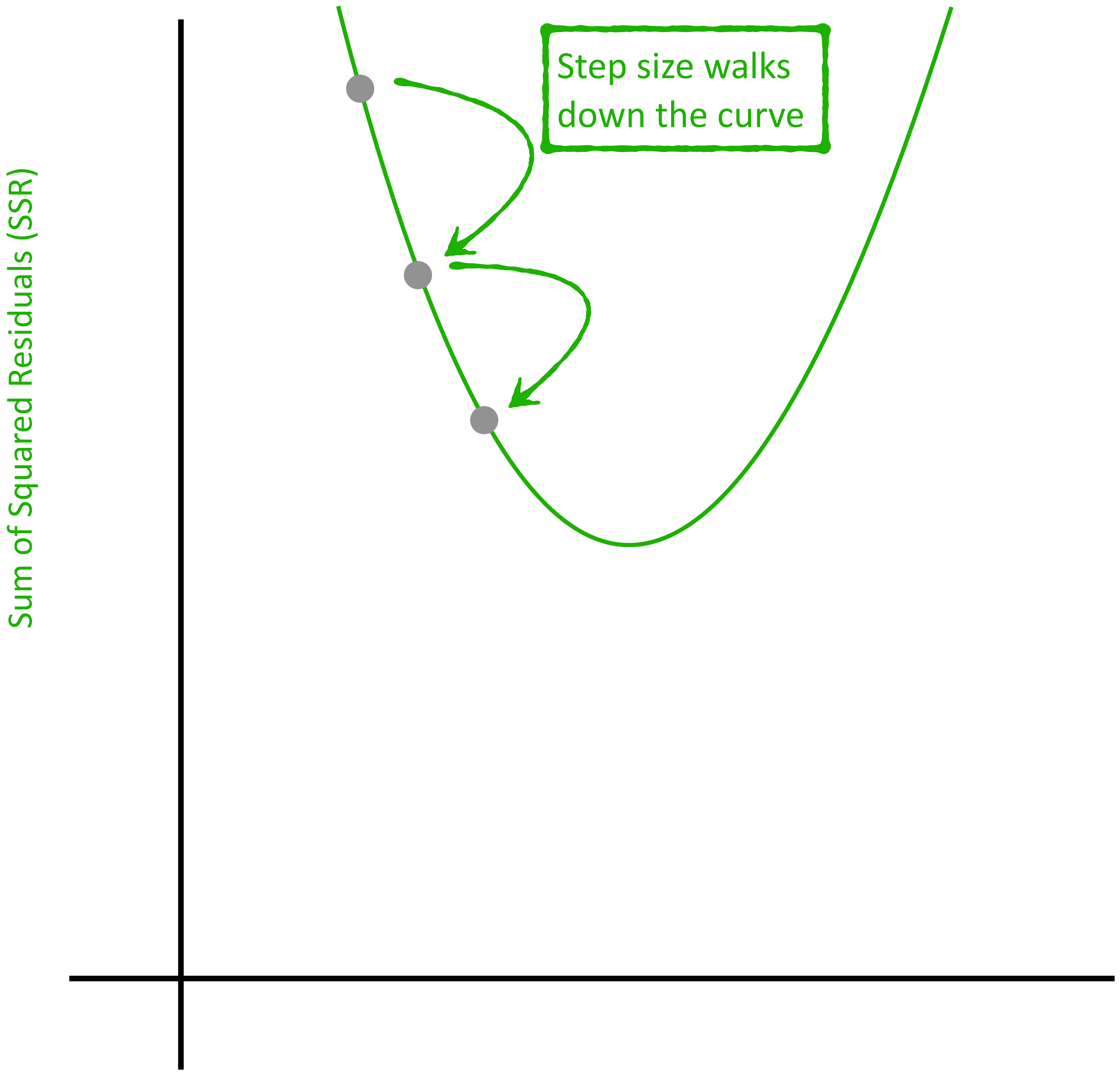
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

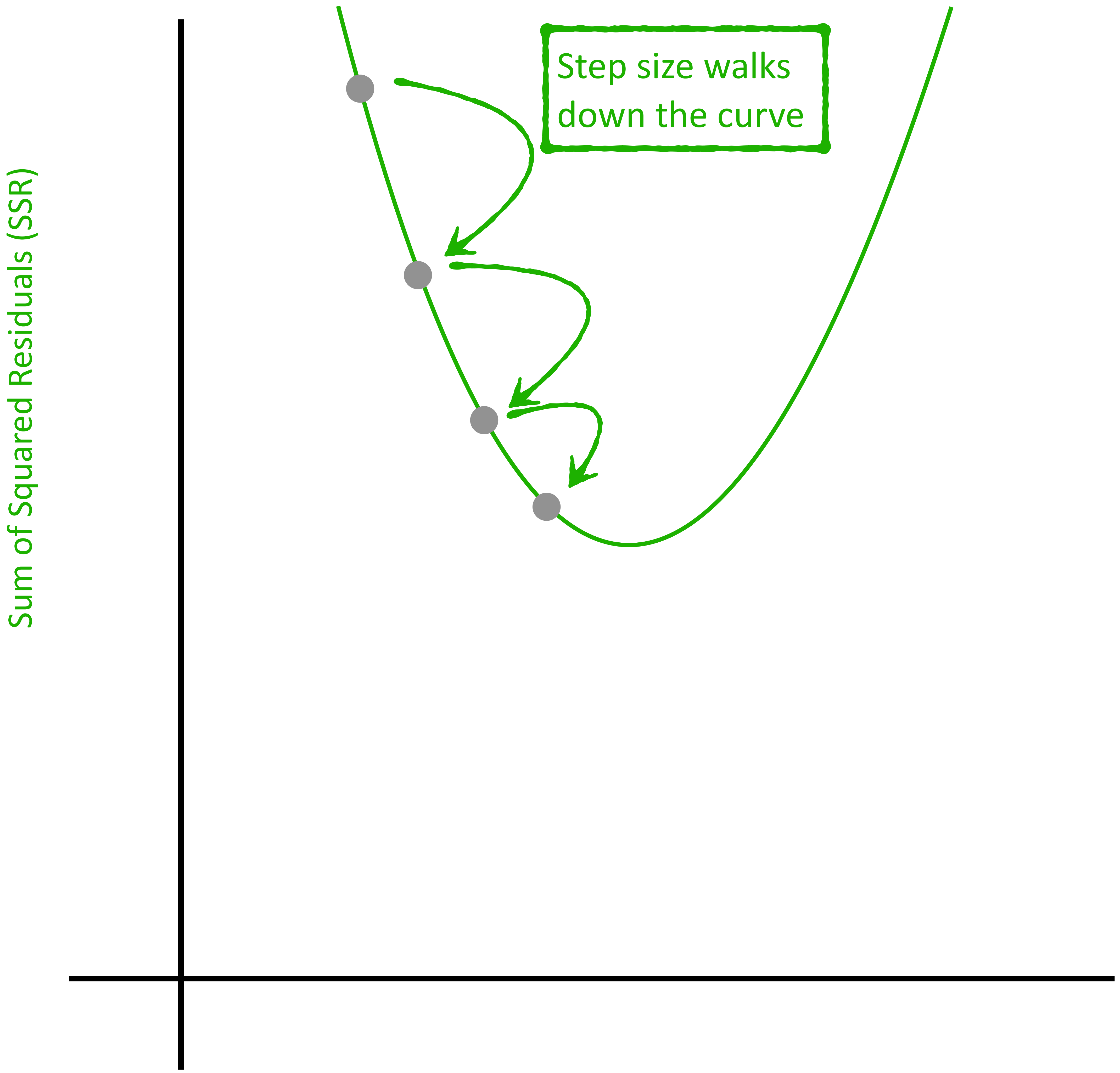
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

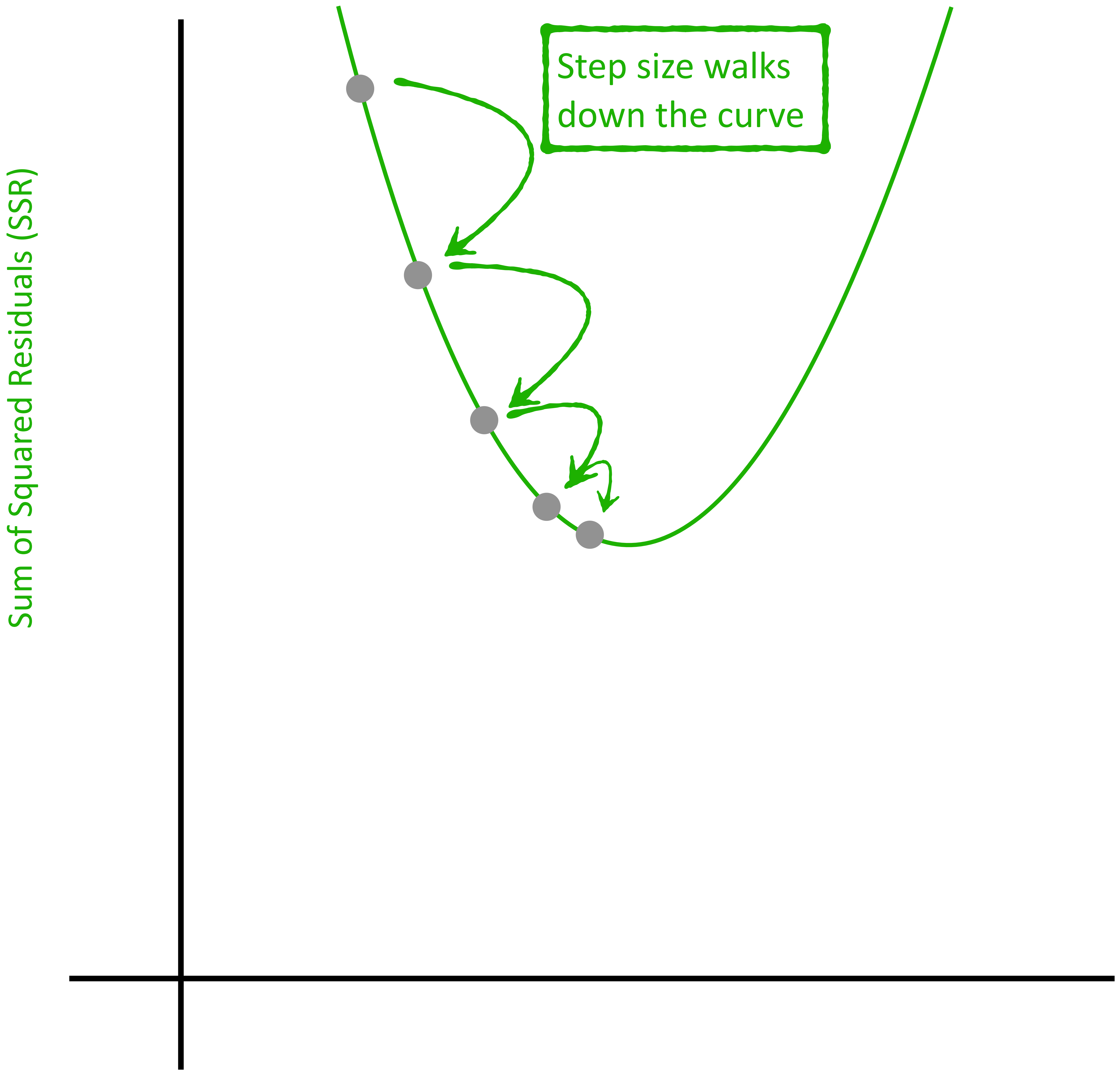
Step 3: Calculate a step size that is proportional to the slope

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Gradient Descent

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

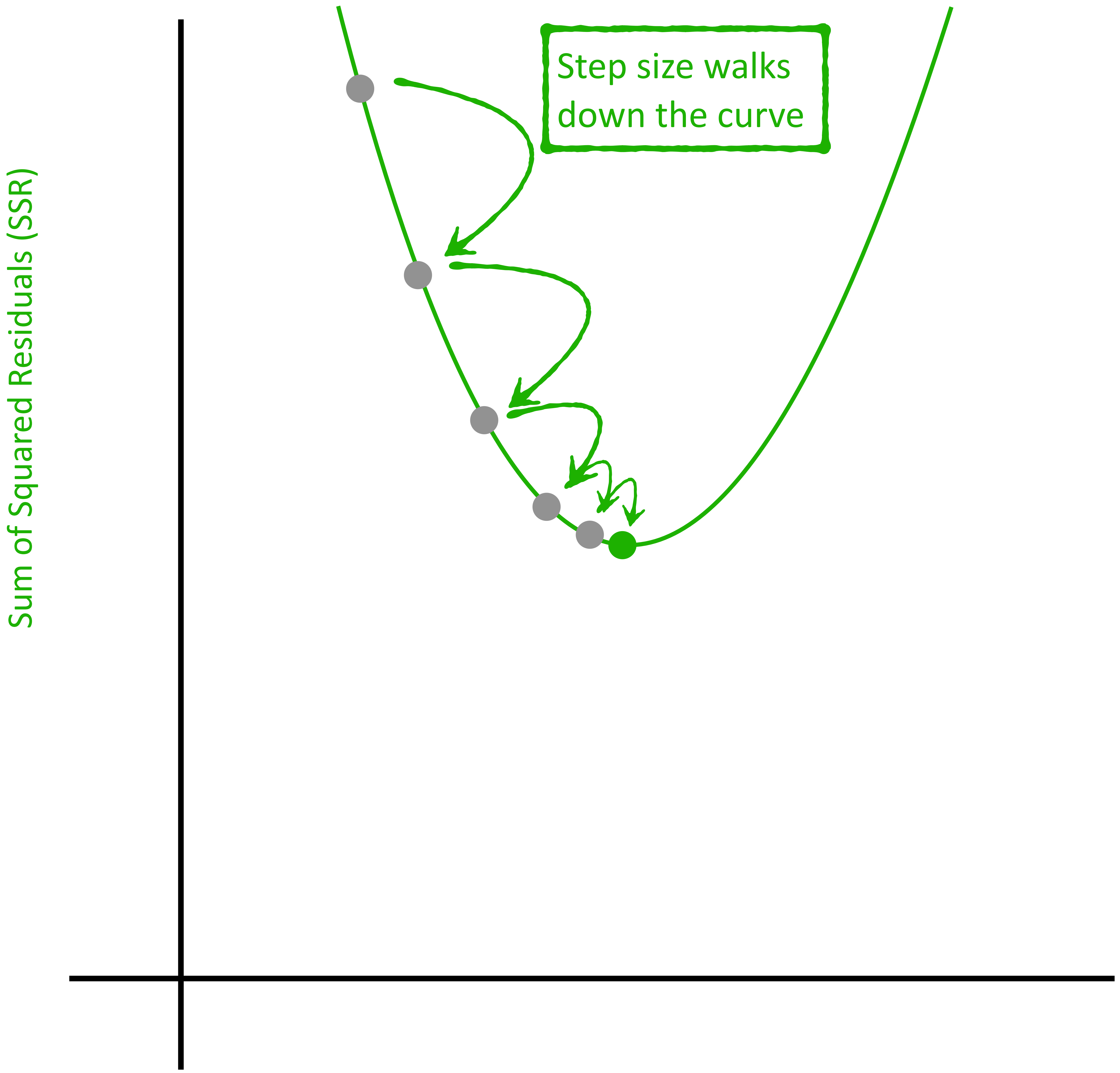
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent

Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

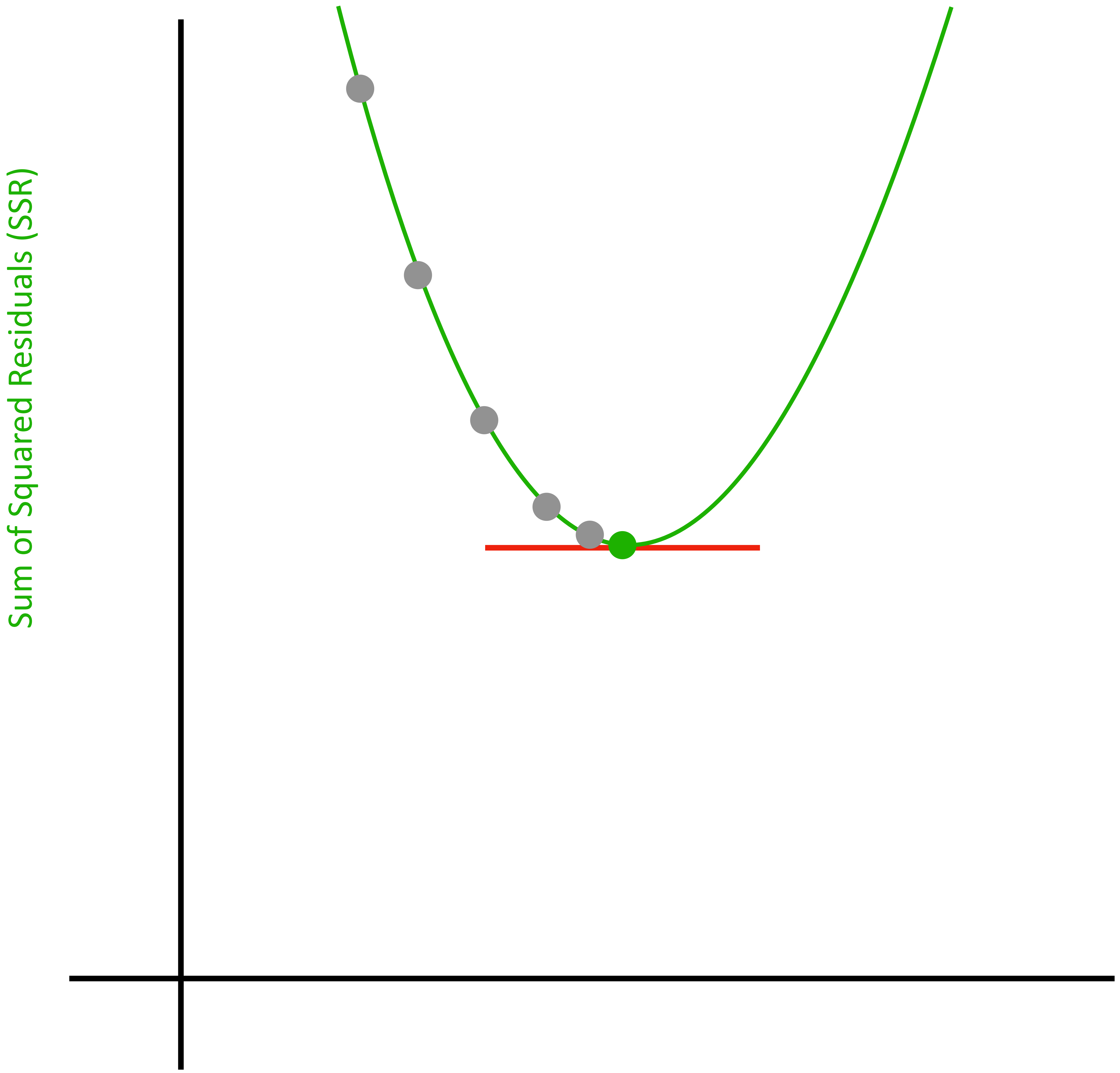
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent: Basic Concept

Step 1: Start with random values for β_0 and β_1

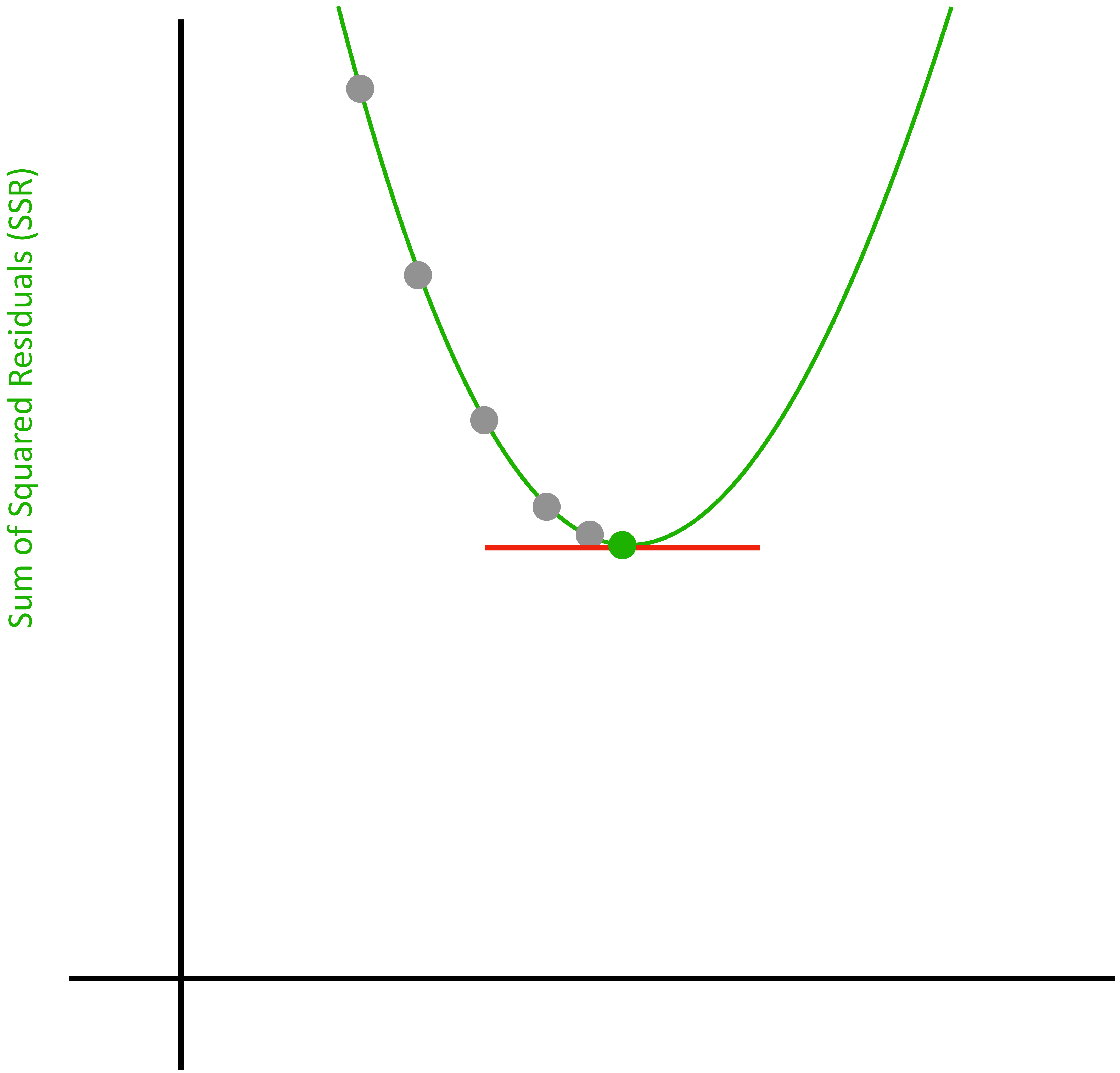
Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

Mean Squared Error (MSE) for various values of β_0 and β_1 follows this curve



Gradient Descent: Basic Concept

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

A linear model in 2 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x$$

Has 2 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Gradient Descent Algorithm

Step 1: Start with random values for β_0 and β_1

Step 2: **Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point**

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

A linear model in 2 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x$$

Has 2 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Compute 2 partial derivatives

Gradient Descent Algorithm

Step 1: Start with random values for β_0 and β_1

Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

$$\frac{\partial}{\partial \beta_0}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial}{\partial \beta_1}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

A linear model in 2 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x$$

Has 2 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Compute 2 step sizes

Gradient Descent Algorithm

Step 1: Start with random values for β_0 and β_1

Step 2: Compute the partial derivative of the MSE w.r.t β_0 and β_1 - this is the slope at that point

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

$$step\_size_{\beta_0} = \frac{\partial}{\beta_0} MSE \times learning\_rate$$

$$step\_size_{\beta_1} = \frac{\partial}{\beta_1} MSE \times learning\_rate$$

A linear model in 3 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Has 3 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1 and β_2

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1 and β_2 - this is the slope at that point

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

Step 4: Calculate new values for β_0, β_1 and β_2 by subtracting the step size

Step 5: Go to step 2 and repeat

A linear model in 3 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Has 3 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

Compute 3 partial derivatives

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1 and β_2

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1 and β_2 - this is the slope at that point

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

$$\frac{\partial}{\partial \beta_0}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i})^2$$

$$\frac{\partial}{\partial \beta_1}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i})^2$$

$$\frac{\partial}{\partial \beta_1}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i})^2$$

A linear model in 3 dimensions...

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Has 3 parameters

And a cost function...

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

Compute 3 step sizes

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1 and β_2

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1 and β_2 - this is the slope at that point

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

$$step\_size_{\beta_0} = \frac{\partial}{\beta_0} MSE \times learning\_rate$$

$$step\_size_{\beta_1} = \frac{\partial}{\beta_1} MSE \times learning\_rate$$

$$step\_size_{\beta_2} = \frac{\partial}{\beta_2} MSE \times learning\_rate$$

A linear model in 4 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Has 4 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i})^2$$

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1, β_2 and β_3

Step 2: **Compute the partial derivative of the MSE w.r.t β_0, β_1, β_2 and β_3 - this is the slope at that point**

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

Step 4: Calculate new values for β_0, β_1, β_2 and β_3 by subtracting the step size

Step 5: Go to step 2 and repeat

A linear model in 4 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Has 4 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i})^2$$

Compute 4 partial derivatives

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1, β_2 and β_3

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1, β_2 and β_3 - this is the slope at that point

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

$$\frac{\partial}{\partial \beta_0}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i})^2$$

$$\frac{\partial}{\partial \beta_1}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i})^2$$

$$\frac{\partial}{\partial \beta_2}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i})^2$$

$$\frac{\partial}{\partial \beta_3}\frac{1}{2n}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i})^2$$

A linear model in 4 dimensions…

Gradient Descent

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Has 4 parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i})^2$$

Compute 4 step sizes

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1, β_2 and β_3

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1, β_2 and β_3 - this is the slope at that point

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

$$step\_size_{\beta_0} = \frac{\partial}{\beta_0} MSE \times learning\_rate$$

$$step\_size_{\beta_1} = \frac{\partial}{\beta_1} MSE \times learning\_rate$$

$$step\_size_{\beta_2} = \frac{\partial}{\beta_2} MSE \times learning\_rate$$

$$step\_size_{\beta_3} = \frac{\partial}{\beta_3} MSE \times learning\_rate$$

A linear model in k dimensions...

Gradient Descent

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

Has k parameters

And a cost function...

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1, β_2 and β_3

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1, β_2 and β_3 - this is the slope at that point

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

Step 4: Calculate new values for β_0, β_1, β_2 and β_3 by subtracting the step size

Step 5: Go to step 2 and repeat

A linear model in k dimensions…

Gradient Descent

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

Has k parameters

And a cost function…

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

Compute k partial derivatives

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1, β_2 and β_3

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1, β_2 and β_3 - this is the slope at that point

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

$$\frac{\partial}{\partial \beta_0}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

$$\frac{\partial}{\partial \beta_1}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

$$\frac{\partial}{\partial \beta_2}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

$$\vdots$$

$$\frac{\partial}{\partial \beta_{k-1}}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \hat{x}_{1i} - \beta_2 \hat{x}_{2i} - \beta_3 \hat{x}_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

A linear model in k dimensions...

Gradient Descent

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

Has k parameters

And a cost function...

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \ldots - \beta_{k-1} x_{(k-1)i})^2$$

Compute k step sizes

Gradient Descent Algorithm

Step 1: Start with random values for β_0, β_1, β_2 and β_3

Step 2: Compute the partial derivative of the MSE w.r.t β_0, β_1, β_2 and β_3 - this is the slope at that point

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

$$step\_size_{\beta_0} = \frac{\partial}{\beta_0} MSE \times learning\_rate$$

$$step\_size_{\beta_1} = \frac{\partial}{\beta_1} MSE \times learning\_rate$$

$$step\_size_{\beta_2} = \frac{\partial}{\beta_2} MSE \times learning\_rate$$

$$\vdots$$

$$step\_size_{\beta_k} = \frac{\partial}{\beta_{k-1}} MSE \times learning\_rate$$

$$\hat{y} = \beta_0 + \beta_1\hat{x}_1 + \beta_2\hat{x}_2 + \beta_3\hat{x}_3 + \ldots + \beta_k\hat{x}_k$$

Gradient Descent: Basic Concept

Step 1: Start with random values for $\beta_0, \beta_1, \beta_2$ and β_3

Step 2: Compute the partial derivative of the SSR

t that point

ortional

Has k +

And a c

Computing k partial derivatives isn't practical

$$\sum_{i=0}^{n} (y_i - \beta_0 - \beta_1\hat{x}_{1i} - \beta_2\hat{x}_{2i} - \beta_3\hat{x}_{3i} - \ldots - \beta_k\hat{x}_{ki})^2$$

$$step\_size_{\beta_2} = \frac{\partial}{\beta_2}SSR \times learning\_rate$$

$$\vdots$$

$$step\_size_{\beta_k} = \frac{\partial}{\beta_k}SSR \times learning\_rate$$

Compute k+ 1 step sizes

Lets use a Matrix

Multiple Regression

Linear Model in k Dimensions

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

$$\hat{y}_n = 1 \times \beta_0 + x_{1n} \times \beta_1 + x_{2n} \times \beta_2 + x_{3n} \times \beta_3 + \ldots + x_{k-1n} \times \beta_{k-1}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & . & . & . & x_{(k-1)1} \\ 1 & x_{12} & x_{22} & x_{32} & . & . & . & x_{(k-1)2} \\ 1 & x_{13} & x_{23} & x_{33} & . & . & . & x_{(k-1)3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} & . & . & . & x_{(k-1)n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

- 1 dependent variable $\hat{y}$
- $k-1$ independent variables $x_1, x_2, x_3 \ldots x_{k-1}$
- k parameters - $\beta_0, \beta_1, \beta_2, \beta_3 \ldots \beta_{k-1}$

Data is synthetic and not plotted to scale

Multiple Regression

Linear Model in k Dimensions

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

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

$\hat{Y}$ and β are column vectors and X is a matrix

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

- 1 dependent variable $\hat{y}$
- $k-1$ independent variables $x_1, x_2, x_3 \ldots x_{k-1}$
- k parameters - $\beta_0, \beta_1, \beta_2, \beta_3 \ldots \beta_{k-1}$

Data is synthetic and not plotted to scale

Linear Model in k Dimensions

Multiple Regression

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

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

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

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

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

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

Partial Derivative w.r.t β:

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

Lets compute this matrix derivative

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

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

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

$$let\, A = (Y - X\beta)$$

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

$$= \frac{1}{2n}\frac{\partial}{\partial\beta}(A)^T(A)\ \frac{\partial}{\partial\beta}A$$

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

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

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

Euclidean norm of a matrix:

$$\| A \| = \sqrt{A^T A}$$

$$let\, A = (Y - X\beta)$$

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

$$= \frac{1}{2n}\frac{\partial}{\partial\beta}(A)^T(A)\ \frac{\partial}{\partial\beta}A$$

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

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

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

Euclidean norm of a matrix:

$\| A \| = \sqrt{A^T A}$

$$let\, A = (Y - X\beta)$$

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

$$= \frac{1}{2n}\frac{\partial}{\partial\beta}(A)^T(A)\ \frac{\partial}{\partial\beta}A$$

Chain Rule for Derivative

See Tutorial on Differential Calculus

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

Linear Model in k Dimensions

Partial Derivative w.r.t β:

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

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

$$= \frac{1}{n}A^T \; (-X)$$

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

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

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

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

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

$$= \frac{1}{n}A^T\ (-X)$$

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

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

Chain Rule for Derivative

See Tutorial on Differential Calculus

Multiple Regression

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

Linear Model in k Dimensions

Partial Derivative w.r.t β:

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

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

$$= \frac{1}{n}A^T\ (-X)$$

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

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

Chain Rule for Derivative

See Tutorial on Differential Calculus

$$\frac{\partial}{\partial A}A^T A = 2A^T$$

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

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

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

$$= \frac{1}{n}A^T\ (-X)$$

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

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

Chain Rule for Derivative

See Tutorial on Differential Calculus

$$\frac{\partial}{\partial A}A^T A = 2A^T$$

$$\frac{\partial}{\partial\beta}(Y - X\beta) = -X$$

Multiple Regression

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

Linear Model in k Dimensions

Partial Derivative w.r.t β:

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

Gradient Vector: $1 \times (k+1)$ row vector

because…

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

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

Gradient Vector: $1 \times (k+1)$ row vector

because...

Y is a $(n+1) \times 1$ column vector

$X\beta$ is a $(n+1) \times 1$ column vector

$(Y - X\beta)$ is a $(n+1) \times 1$ column vector

$(Y - X\beta)^T$ is a $1 \times (n+1)$ row vector

X is a $(n+1) \times (k+1)$ matrix

$(Y - X\beta)^T(X)$ is a $1 \times (k+1)$ row vector

Multiple Regression

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

Linear Model in k Dimensions

Partial Derivative w.r.t β:

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

Gradient Vector: $1 \times (k+1)$ row vector

Transpose it to get the column vector

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

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

Gradient Vector: $(k+1) \times 1$ column vector

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

$\hat{Y} = X\beta$

Linear Model in k Dimensions

Multiple Regression

Partial Derivative w.r.t β:

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

Gradient Vector: $1 \times (k+1)$ row vector

Transpose it to get the column vector

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

$(AB)^T = B^T A^T$

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

Gradient Vector: $(k+1) \times 1$ column vector

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

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

Linear Model in k Dimensions

Multiple Regression

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

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

Cost Function

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

Lets walk through gradient descent using this matrix representation of the Cost Function and its partial derivative (the gradient vector)

A linear model in k 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

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

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

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

$$d\_cost = \frac{1}{n}X^T(X\beta - Y)$$

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

Step 4: Calculate new values for β_0 and β_1 by subtracting the step size

Step 5: Go to step 2 and repeat

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

$$d\_cost = \frac{1}{n} X^T (X\beta - Y)$$

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

$$step\_size = d\_cost \times learning\_rate$$

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

Step 5: Go to step 2 and repeat

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

$$d\_cost = \frac{1}{n} X^T (X\beta - Y)$$

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

$$step\_size = d\_cost \times learning\_rate$$

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

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

Step 5: Go to step 2 and repeat

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

$$d\_cost = \frac{1}{n} X^T (X\beta - Y)$$

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

$$step\_size = d\_cost \times learning\_rate$$

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

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

Step 5: Go to step 2 and repeat

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

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

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

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

Step 5: Go to step 2 and repeat

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):

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

Matrix algebra allows us to compute gradients and step sizes in a single computation

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

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

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

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

Step 5: Go to step 2 and repeat

Related Tutorials & Textbooks

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.

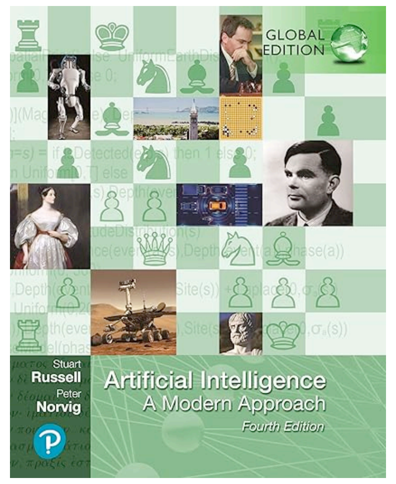
Gradient Descent for Simple Linear Regression

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

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.

Recommended Textbooks



Artificial Intelligence: A Modern Approach

by Peter Norvig, Stuart Russell

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