DATA MINING 2
Gradient Descent

Riccardo Guidotti

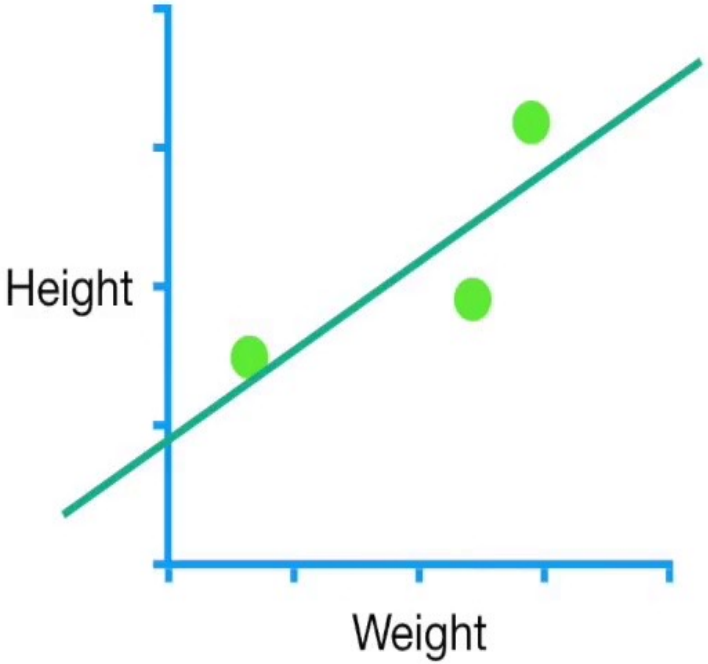
a.a. 2021/2022

Contains edited slides from StatQuest

Gradient Descent

- GD is a very effective and widely usable mathematical technique to find the best parameters in many and various tasks such as
- Linear Regression
- Logistic Regression
- Neural Networks
- …

GD for Linear Regression

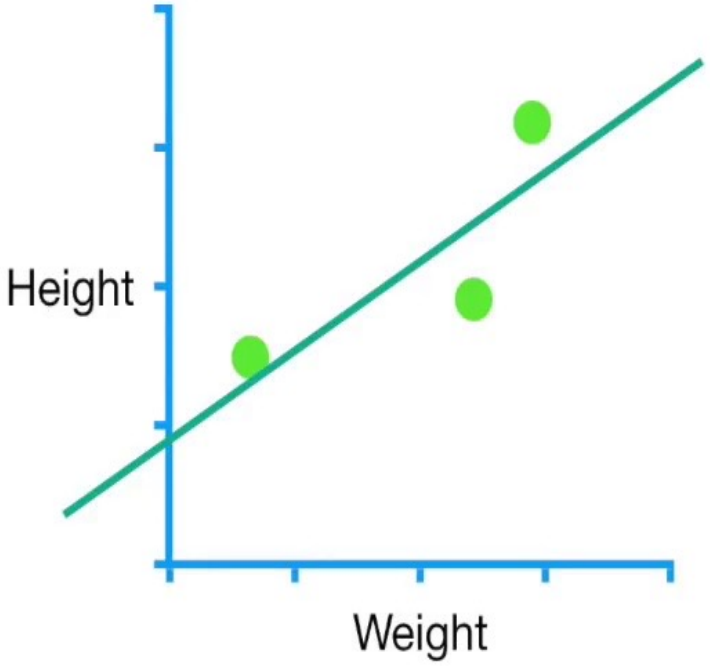


Predicted Height = intercept + slope × **Weight**

So let's learn how **Gradient Descent** can fit a line to data by finding the optimal values for the **Intercept** and the **Slope**.

GD to find *b*

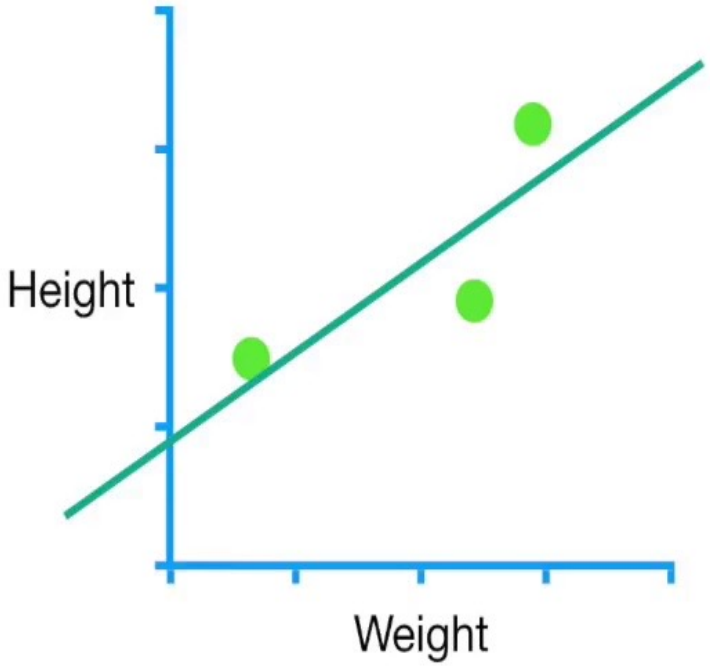
Predicted Height = intercept + slope × **Weight**



Actually, we'll start by using **Gradient Descent** to find the **Intercept**.

GD to find *b*

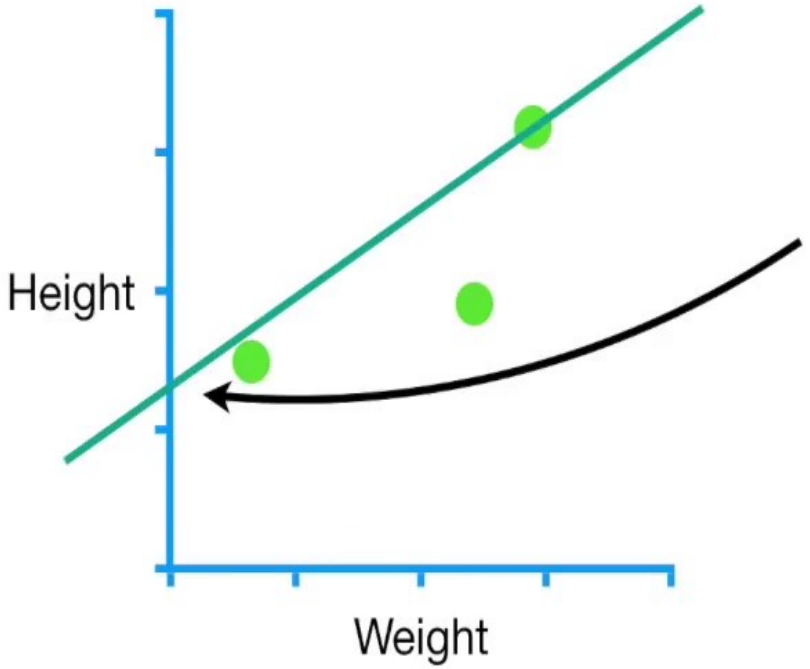
Predicted Height = intercept + slope × **Weight**



So for now, let's just plug in the **Least Squares** estimate for the **Slope**, **0.64**.

GD to find *b*

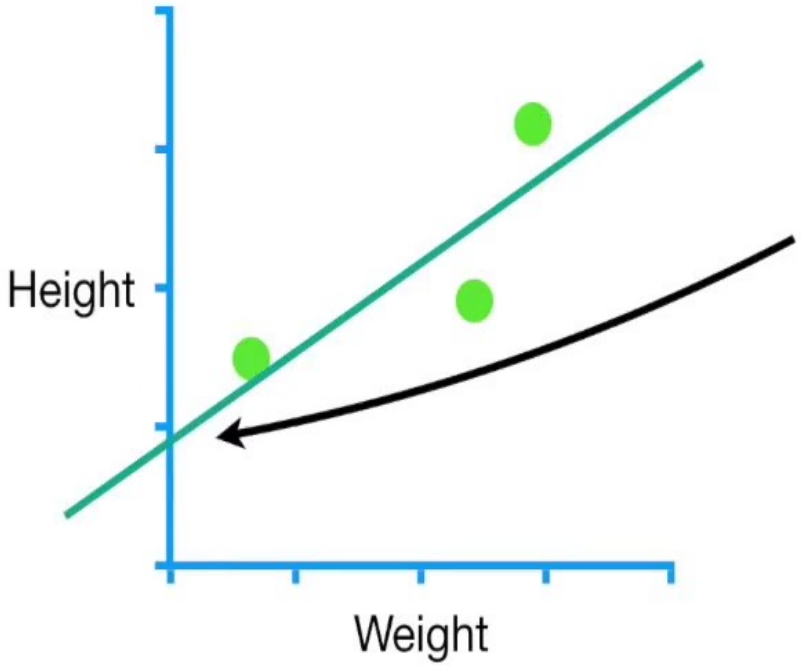
Predicted Height = intercept + 0.64 x **Weight**



…and we'll use **Gradient Descent** to find the the optimal value for the Intercept.

GD to find *b*

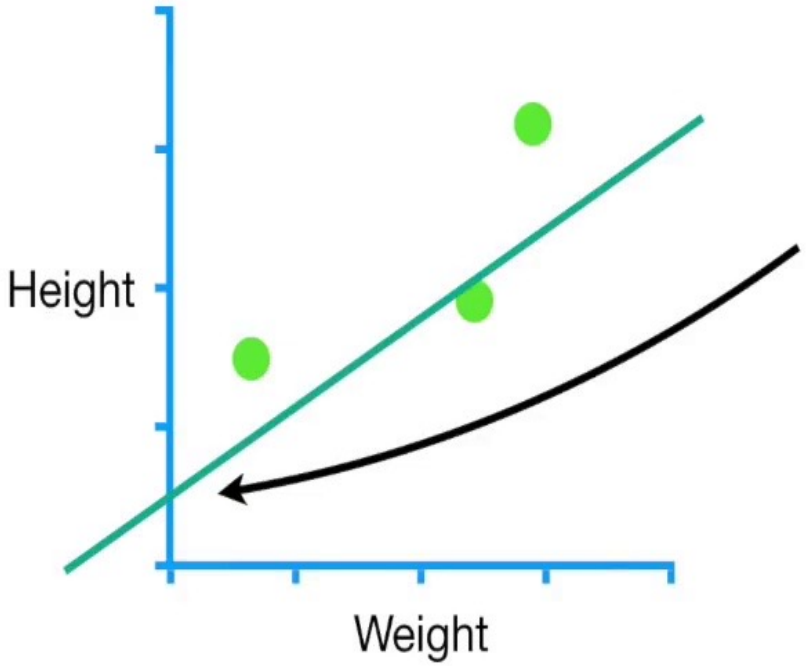
Predicted Height = intercept + 0.64 x **Weight**



…and we'll use **Gradient Descent** to find the the optimal value for the Intercept.

GD to find *b*

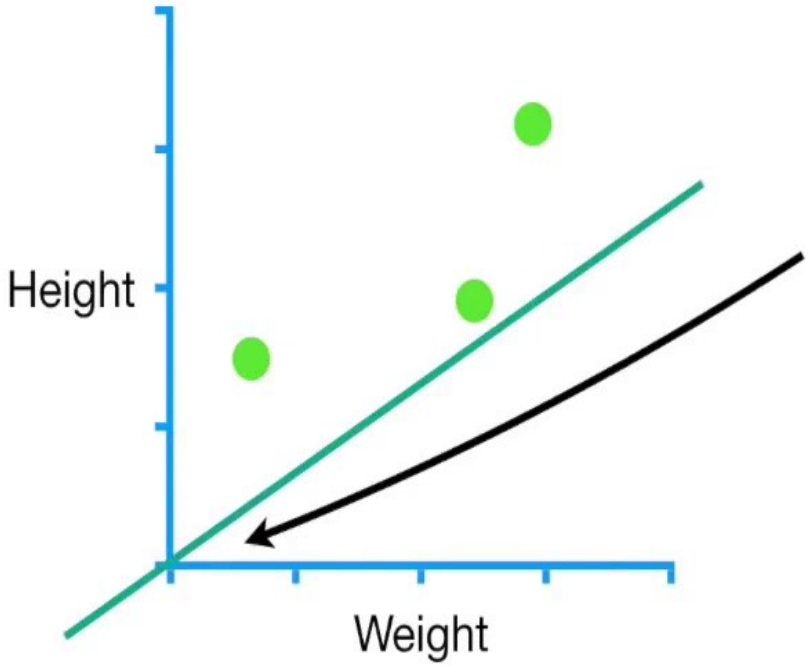
Predicted Height = intercept + 0.64 x **Weight**



…and we'll use **Gradient Descent** to find the the optimal value for the Intercept.

GD to find *b*

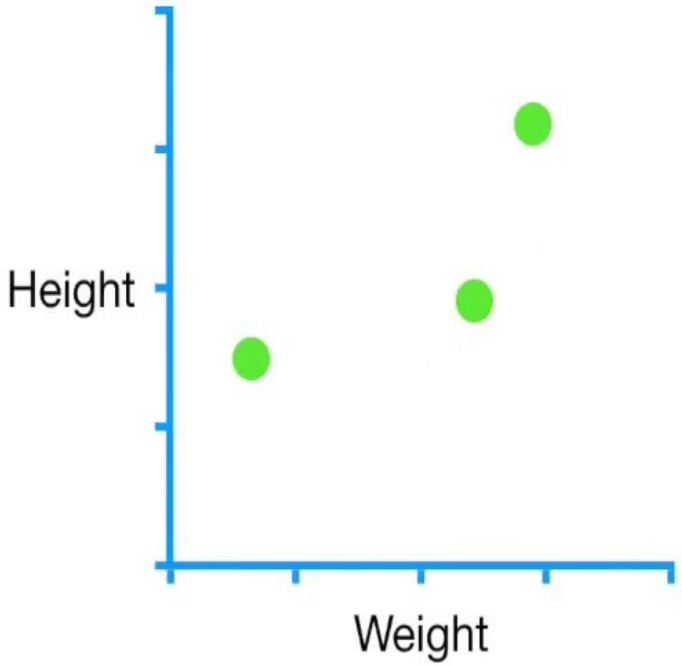
Predicted Height = intercept + 0.64 x **Weight**



…and we'll use **Gradient Descent** to find the the optimal value for the Intercept.

GD to find *b*

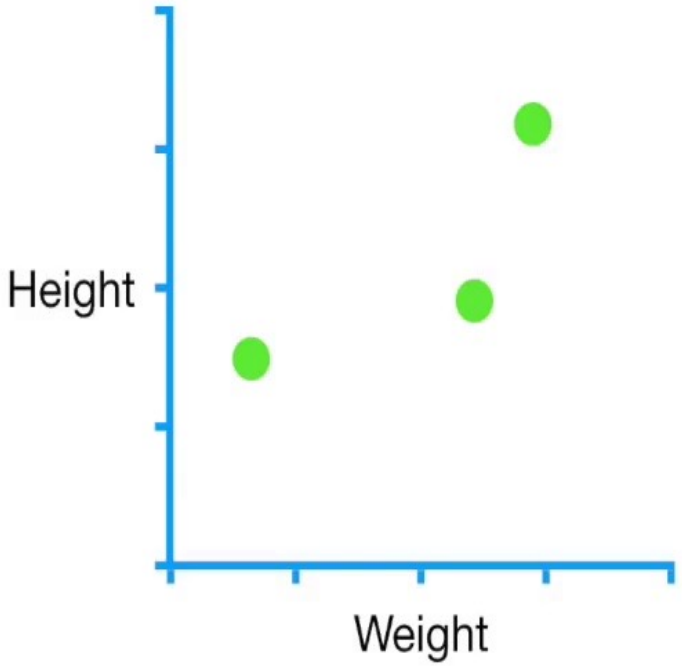
Predicted Height = intercept + 0.64 x **Weight**



The first thing we do is pick a random value for the **Intercept**.

GD to find *b*

Predicted Height = intercept + 0.64 x **Weight**



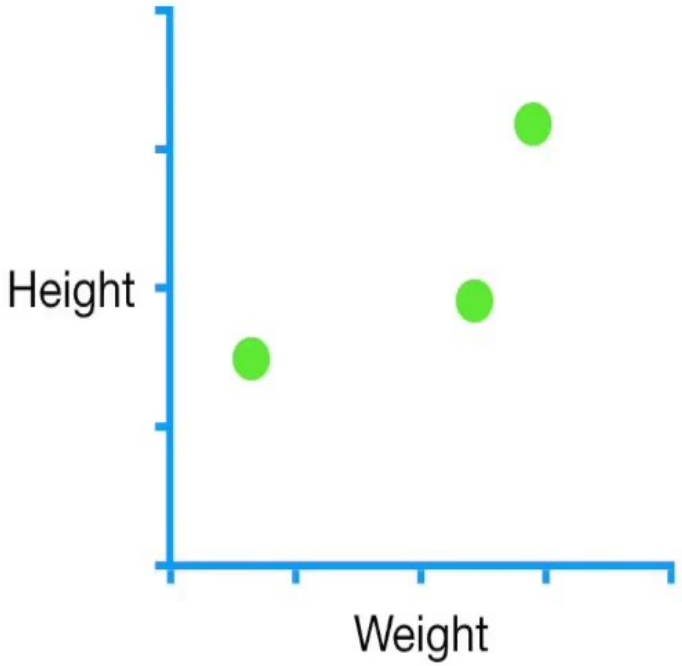
The first thing we do is pick a random value for the **Intercept**.

This is just an initial guess that gives **Gradient Descent** something to improve upon.

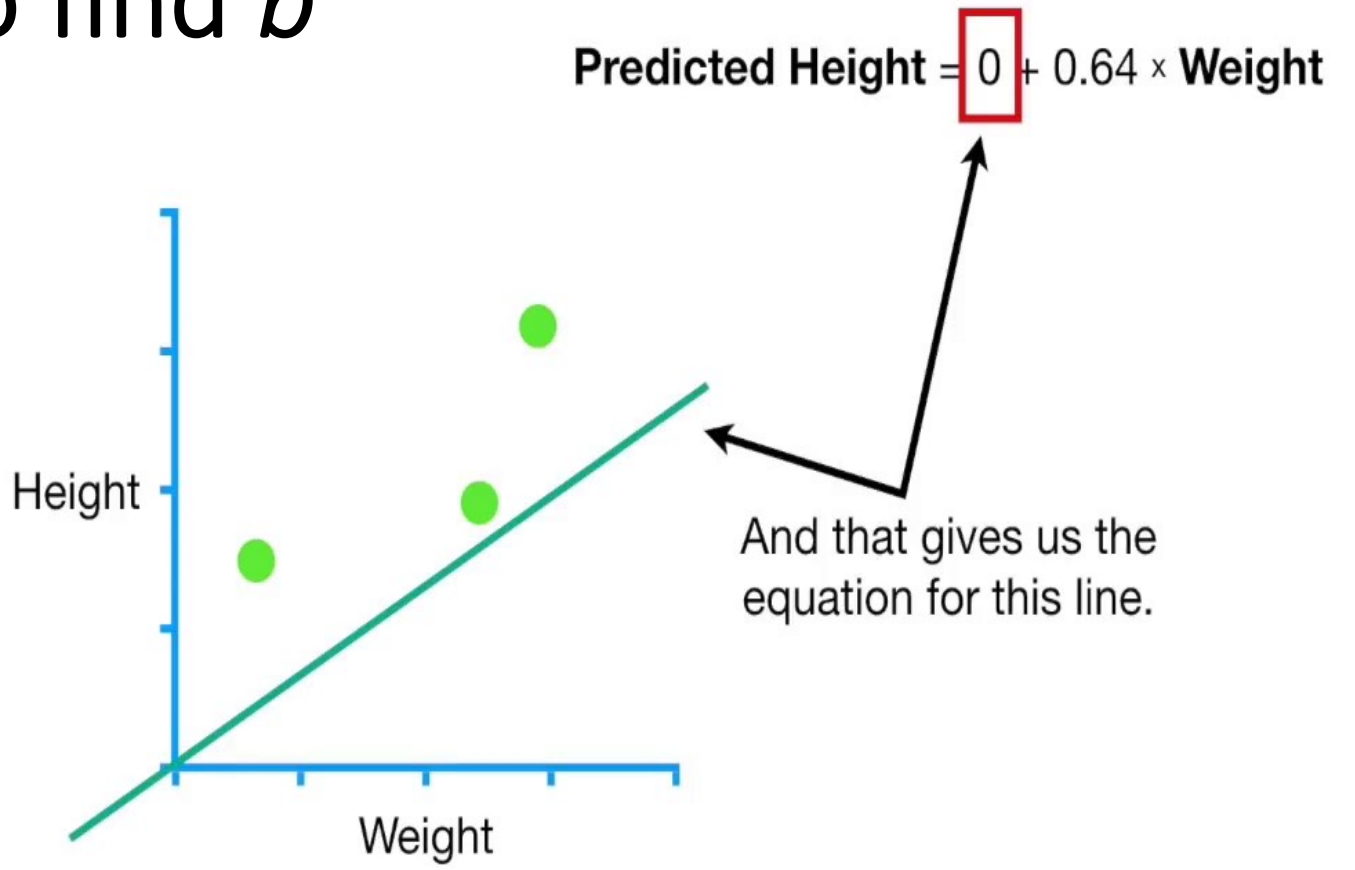
GD to find *b*

Predicted Height = 0 + 0.64 × **Weight**

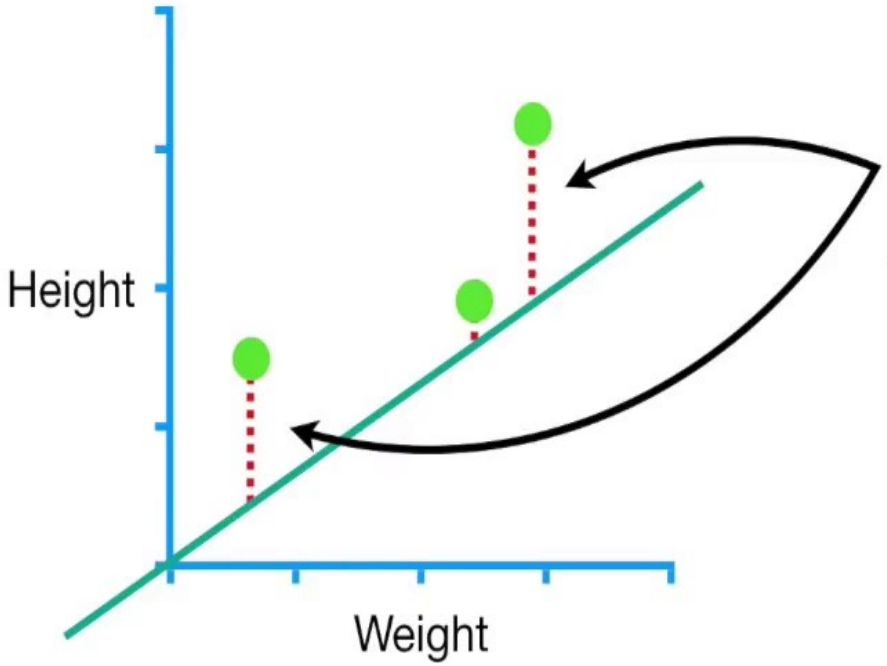
In this case, we'll use **0**, but any number will do.



GD to find *b*

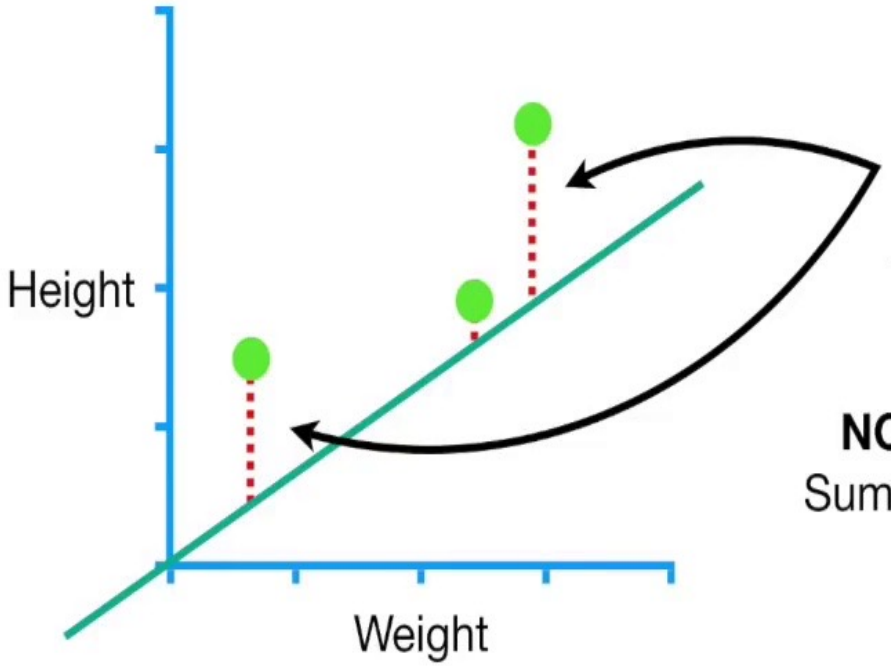

Predicted Height = 0 + 0.64 x Weight
Height
Weight
And that gives us the equation for this line.

GD to find *b*



In this example, we will evaluate how well this line fits the data with the **Sum of the Squared Residuals**.

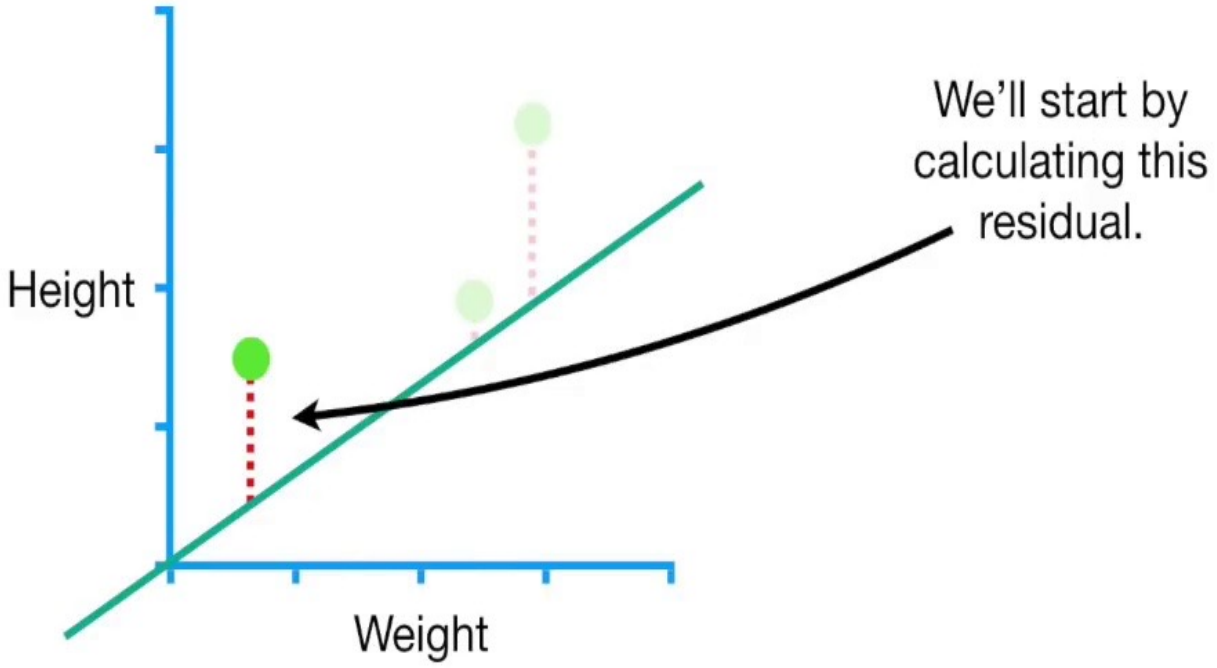
GD to find *b*



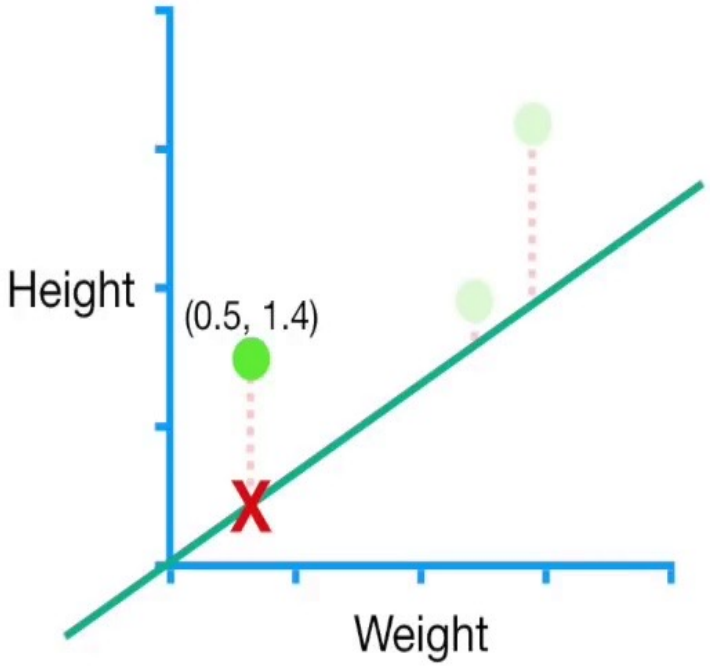
In this example, we will evaluate how well this line fits the data with the **Sum of the Squared Residuals**.

NOTE: In Machine Learning lingo, The Sum of the Squared Residuals is a type of **Loss Function**.

GD to find *b*



GD to find *b*

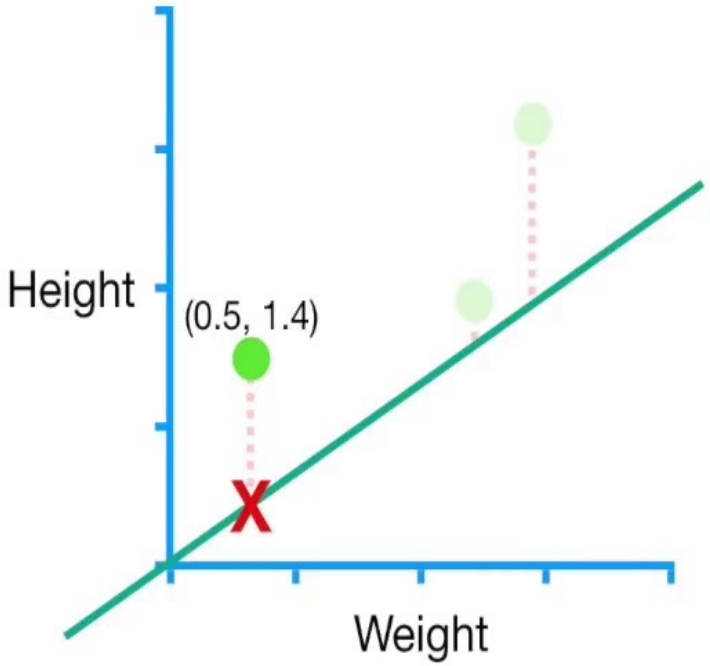


We get the **Predicted Height**, the point on the line…

…by plugging **Weight = 0.5** into the equation for the line…

Predicted Height = 0 + 0.64 × **Weight**

GD to find *b*

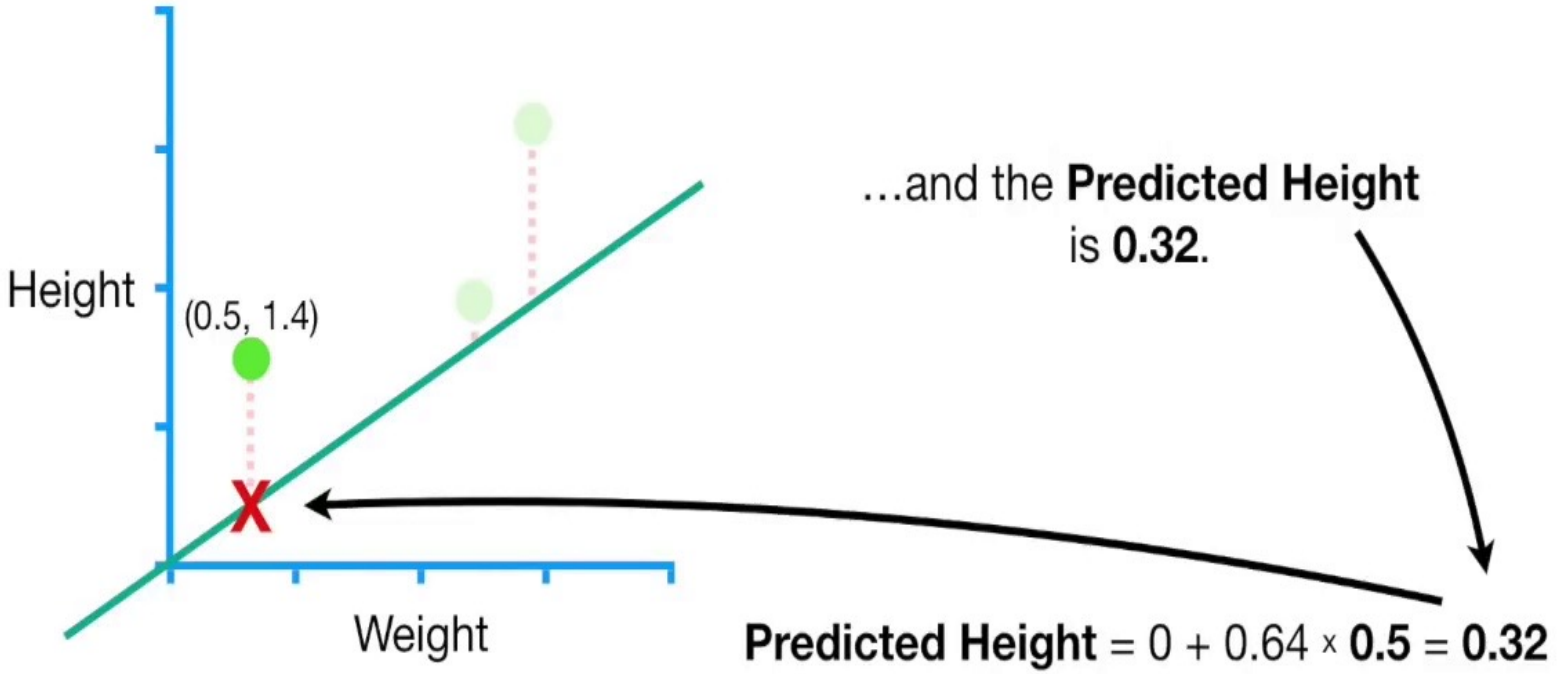


We get the **Predicted Height**, the point on the line…

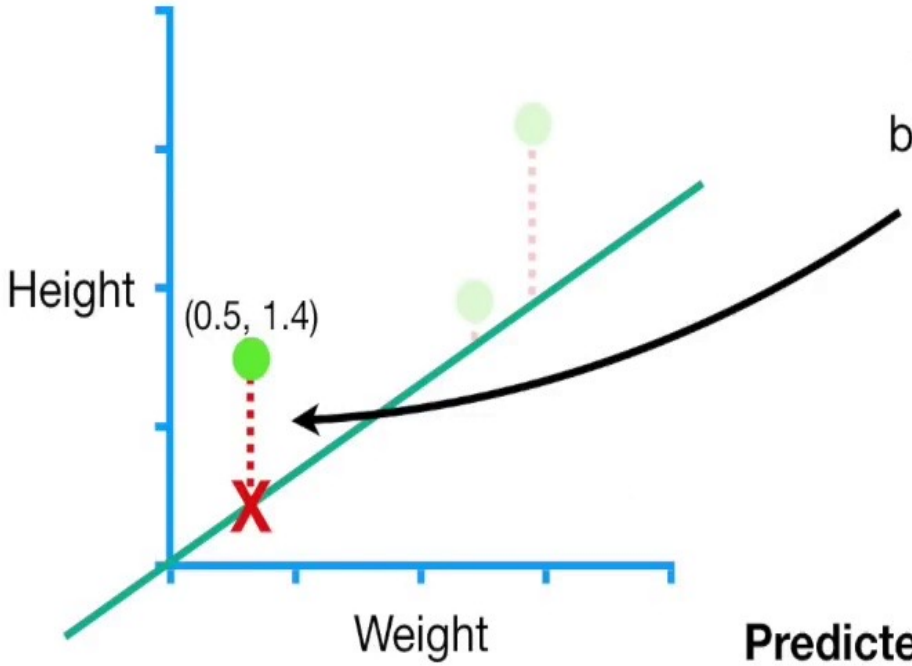
…by plugging **Weight = 0.5** into the equation for the line…

Predicted Height = 0 + 0.64 × **0.5**

GD to find *b*


Height
(0.5, 1.4)
Weight
…and the **Predicted Height** is **0.32**.
Predicted Height = 0 + 0.64 × **0.5** = **0.32**

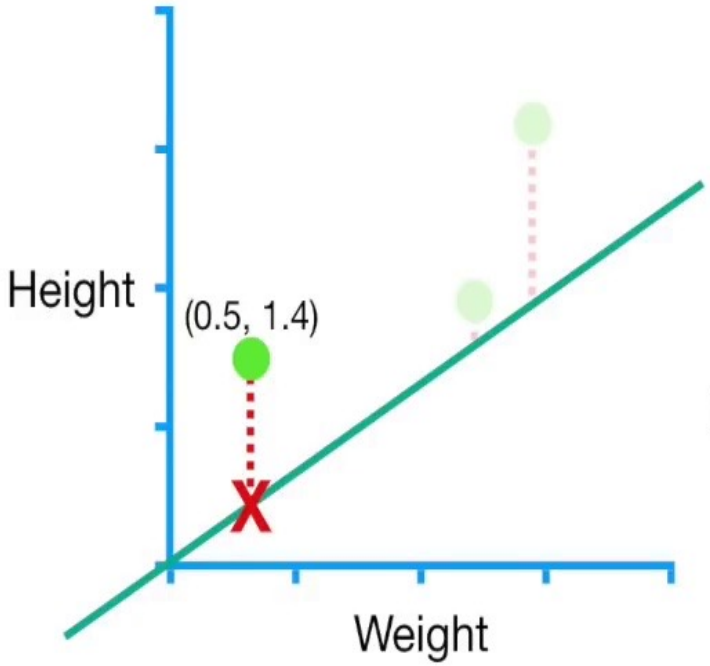
GD to find *b*



The residual is the difference between the **Observed Height**, and the **Predicted Height**…

Predicted Height = 0 + 0.64 × **0.5** = **0.32**

GD to find *b*



The residual is the difference between the **Observed Height**, and the **Predicted Height**…

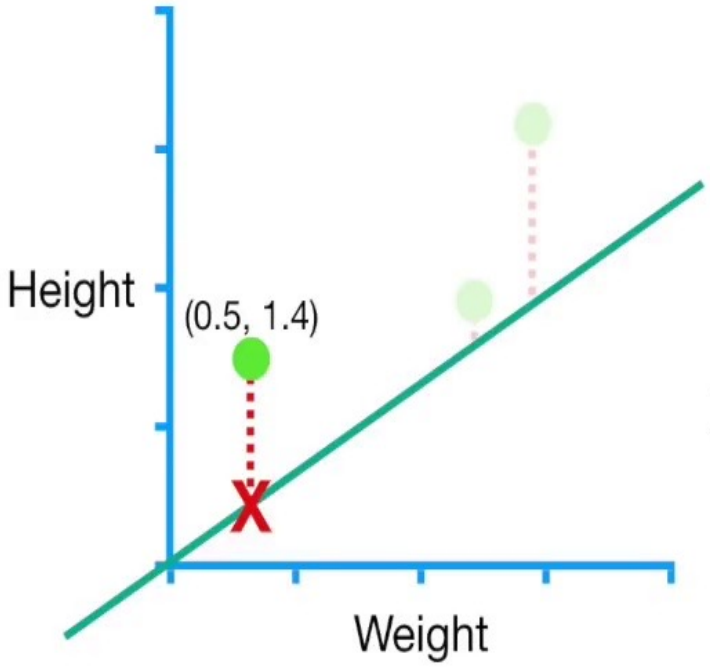
Residual = **Observed Height** - **Predicted Height**

Predicted Height = 0 + 0.64 × **0.5** = **0.32**

GD to find *b*

Sum of squared residuals =

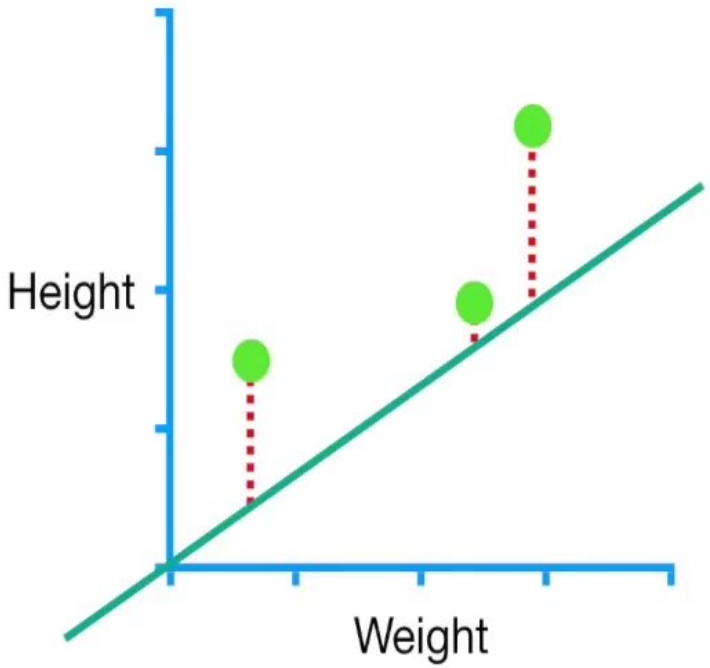
We'll keep track of the Sum of the Squared Residuals up here.



Residual = 1.4 - 0.32 = **1.1**

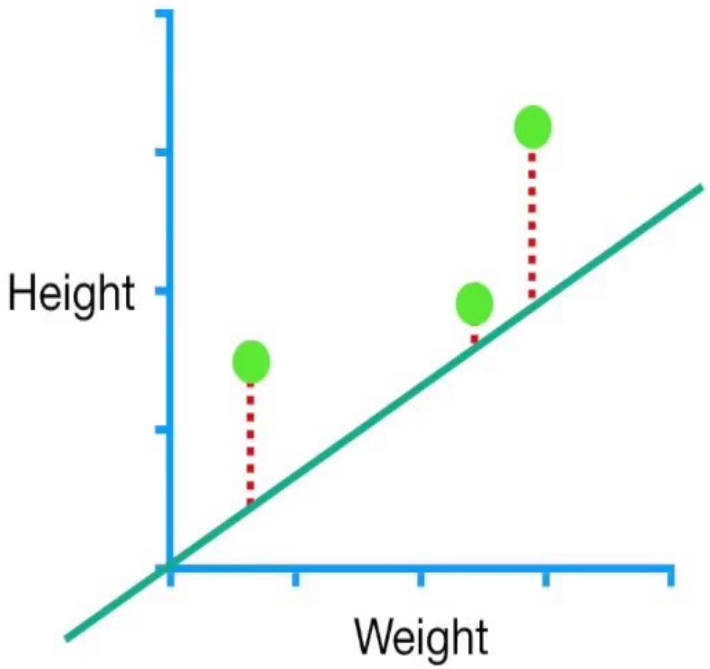
Predicted Height = 0 + 0.64 × **0.5** = **0.32**

Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

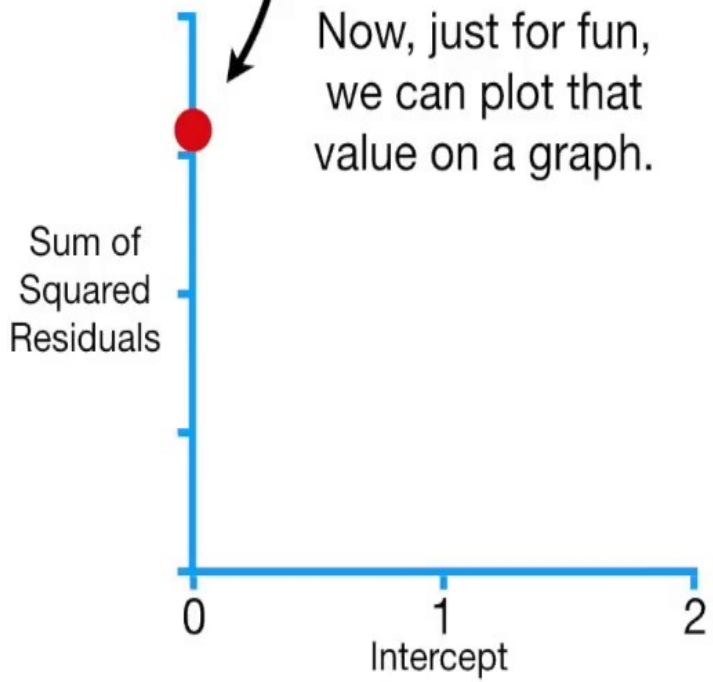


In the end, **3.1** is the Sum of the Squared Residuals.

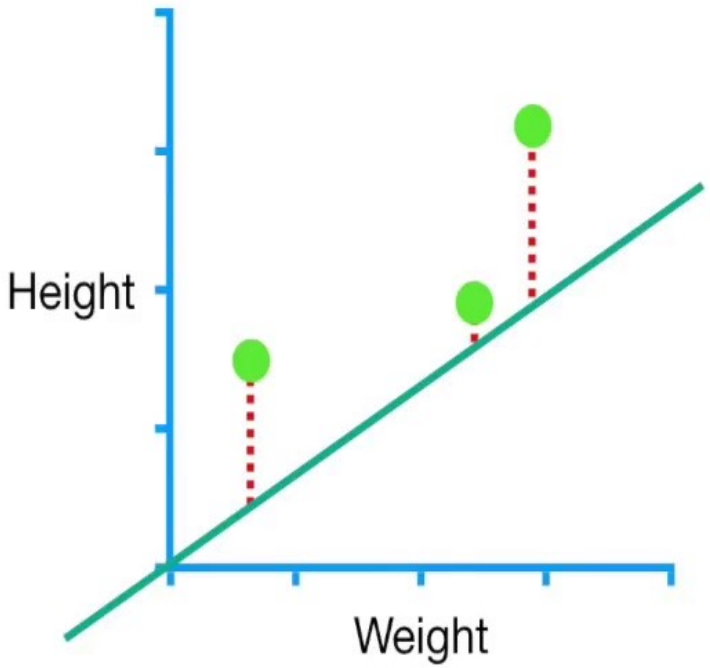
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

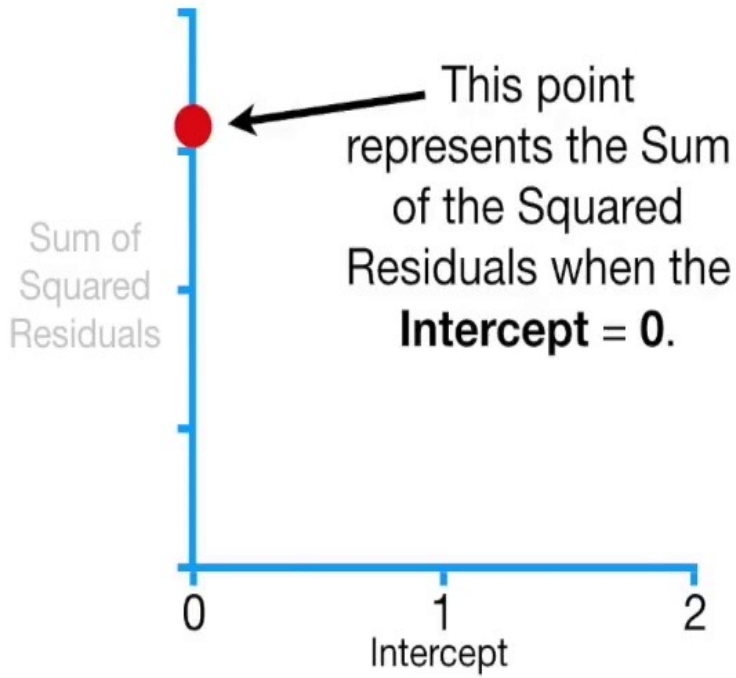


Now, just for fun, we can plot that value on a graph.

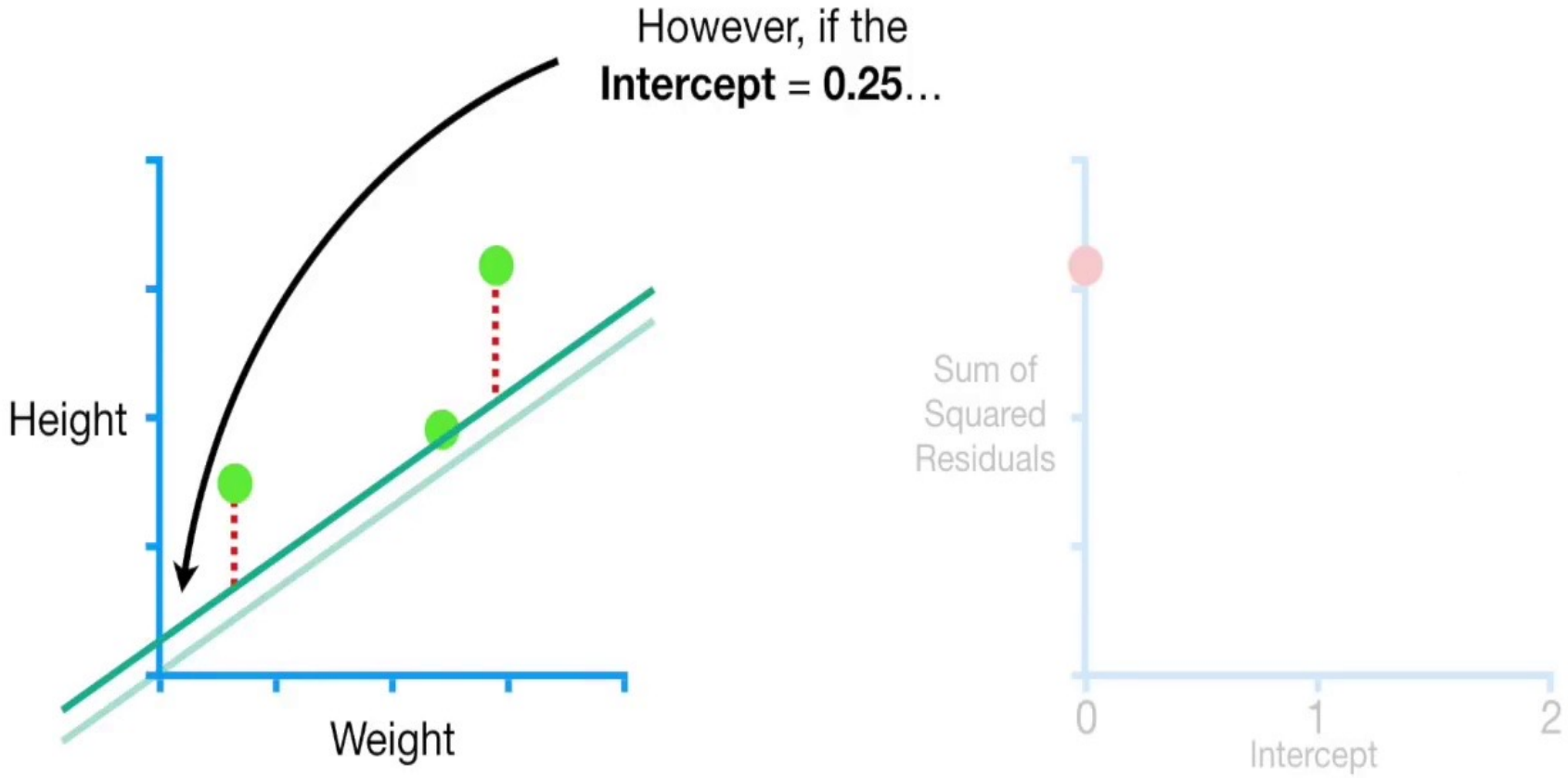


Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

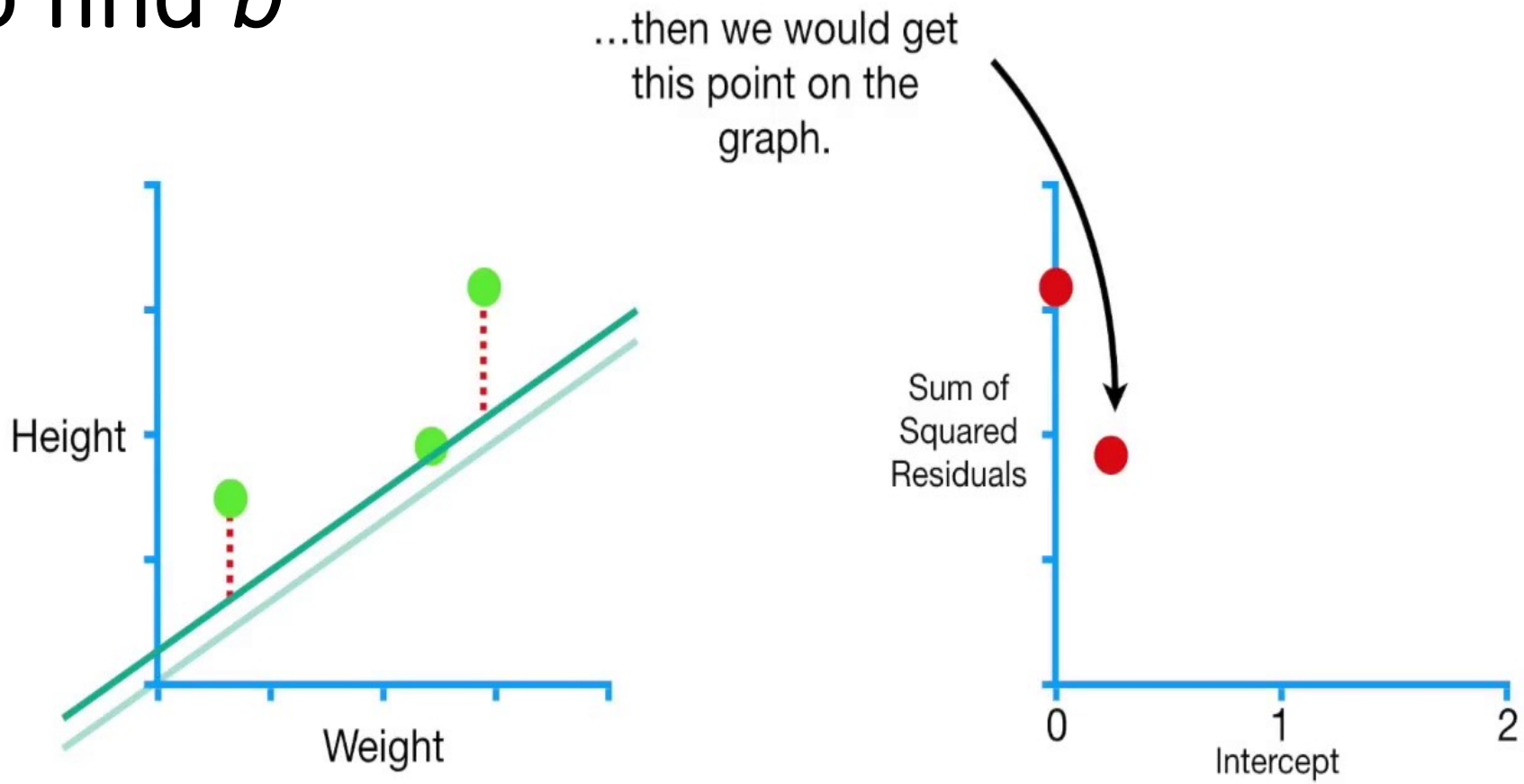




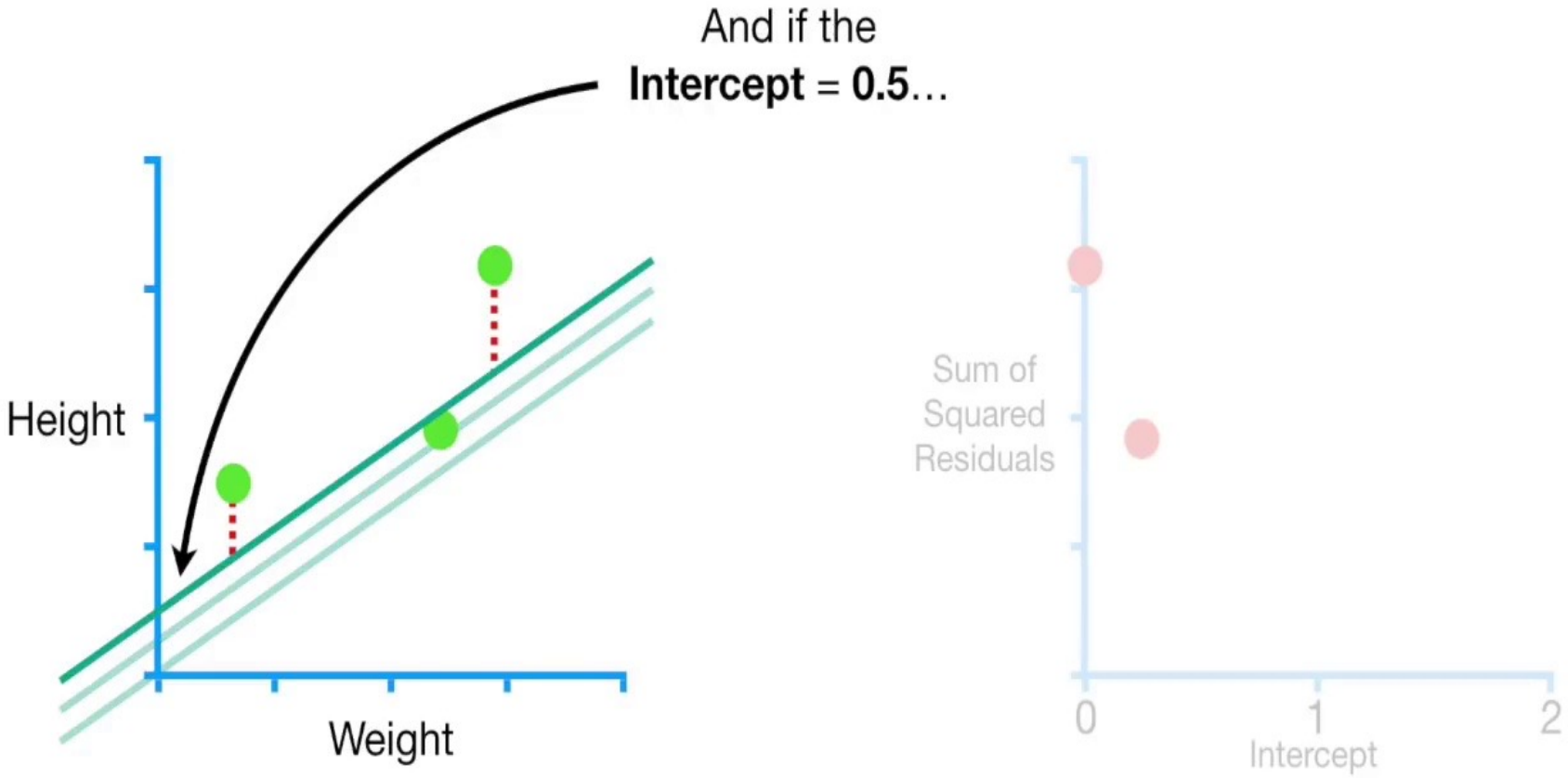
GD to find *b*


However, if the
Intercept = 0.25…
Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept

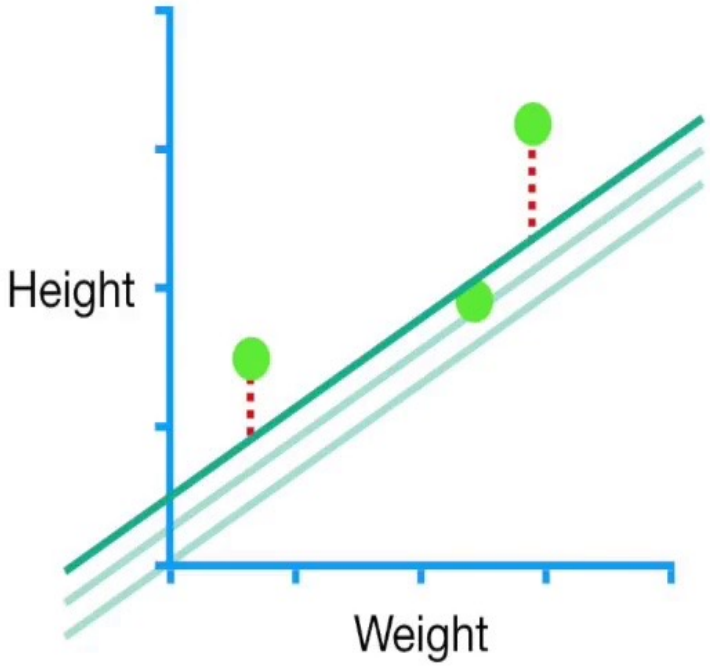
GD to find *b*

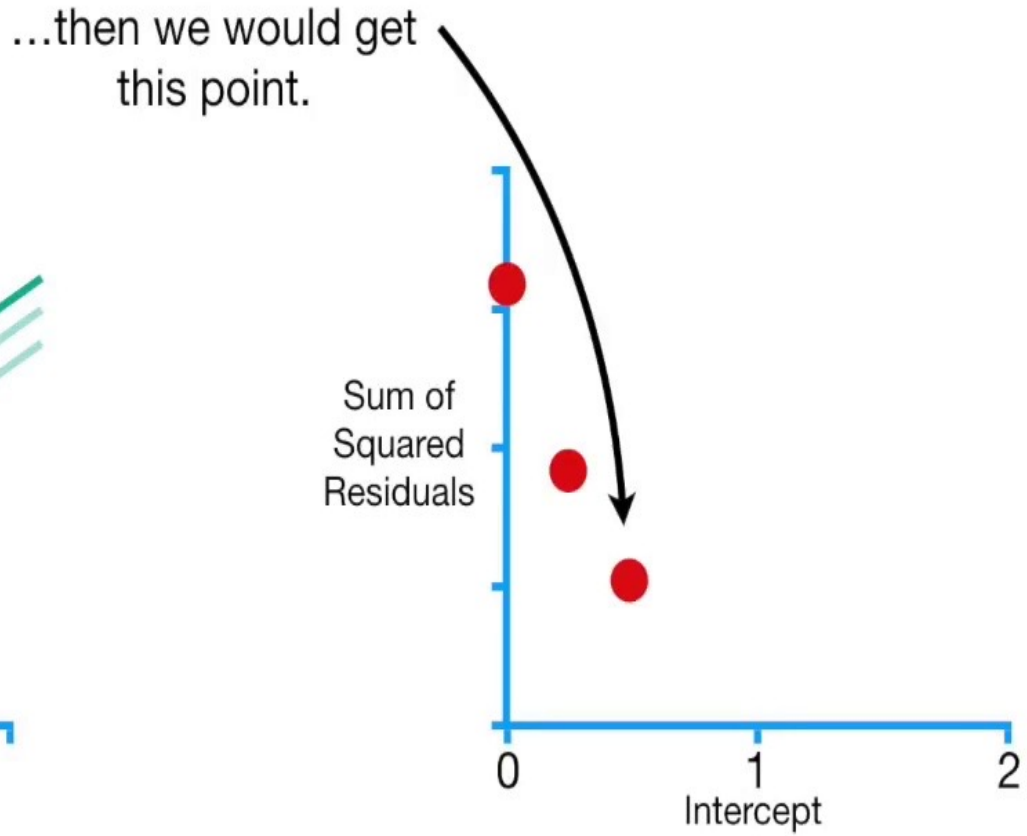

...then we would get this point on the graph.
Height
Weight
Sum of Squared Residuals
0
1
2
Intercept

GD to find *b*

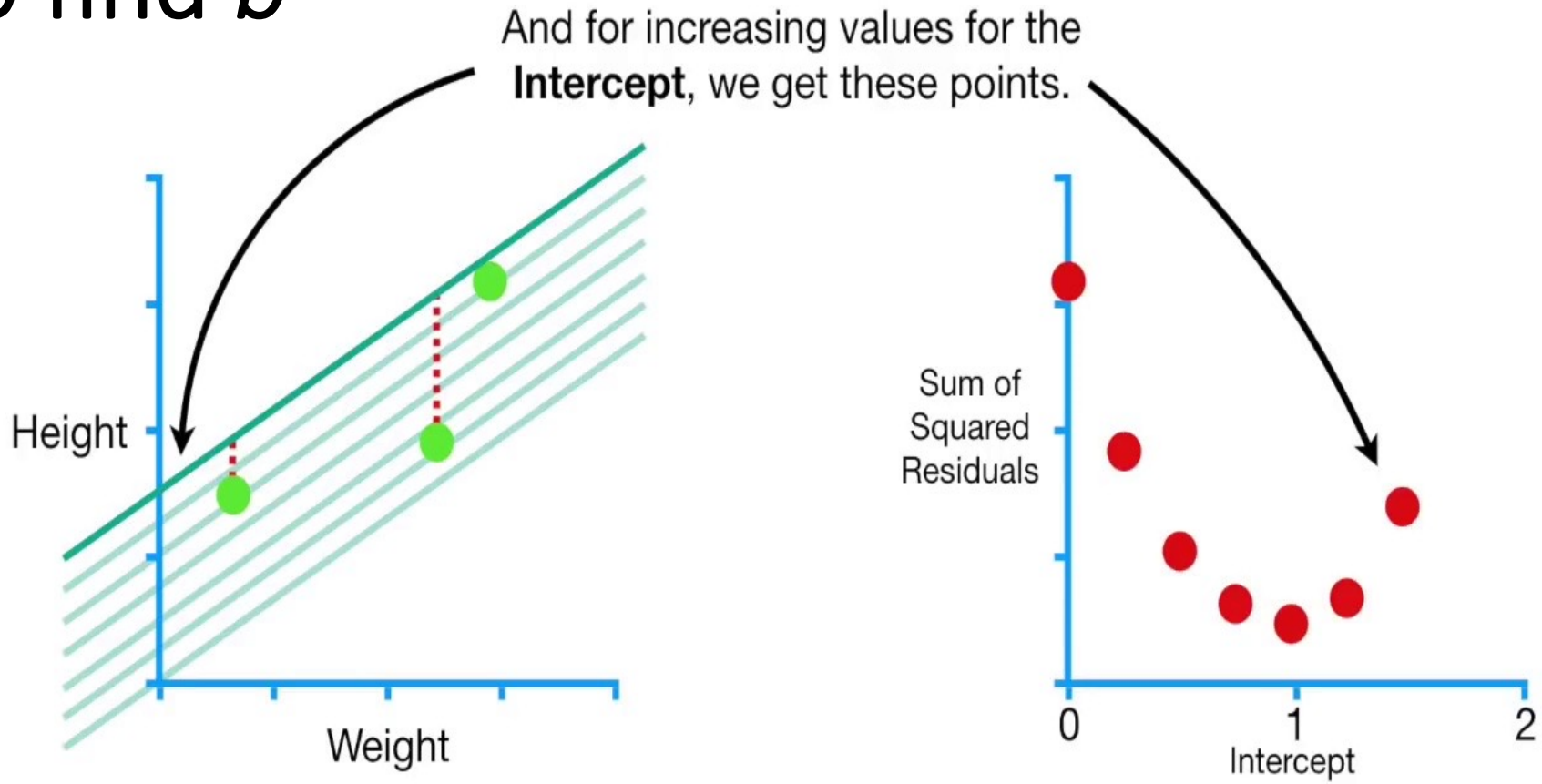

And if the
Intercept = 0.5...
Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept

GD to find *b*

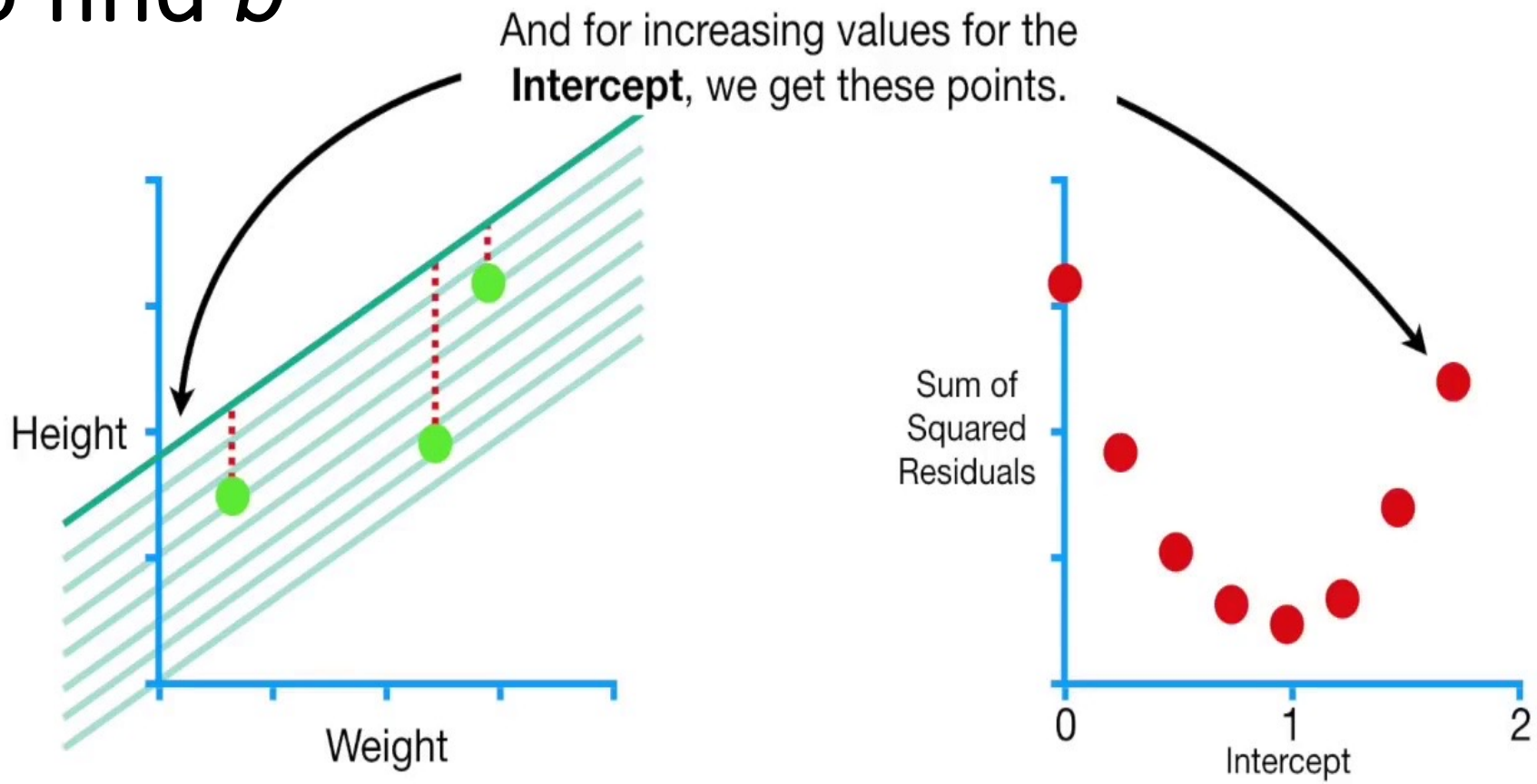

Height
Weight


...then we would get this point.
Sum of Squared Residuals
0
1
2
Intercept

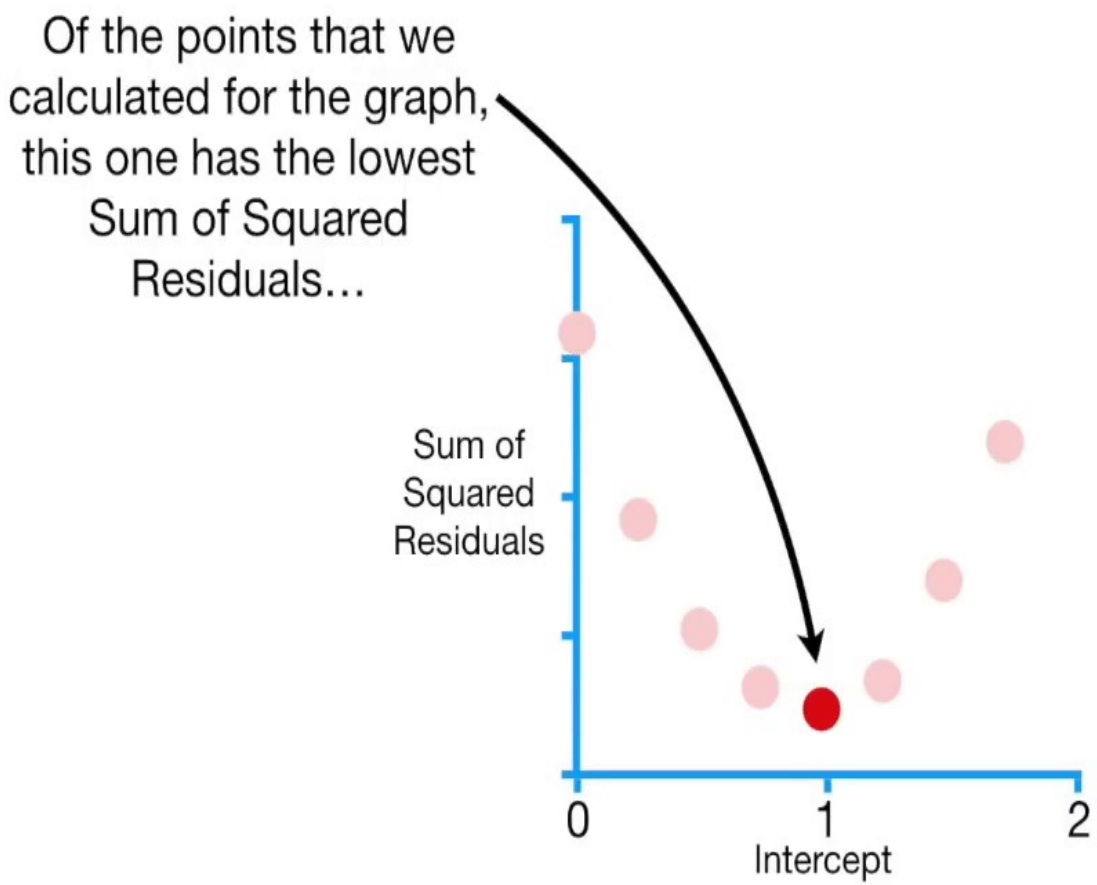
GD to find *b*


And for increasing values for the
Intercept, we get these points.
Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept

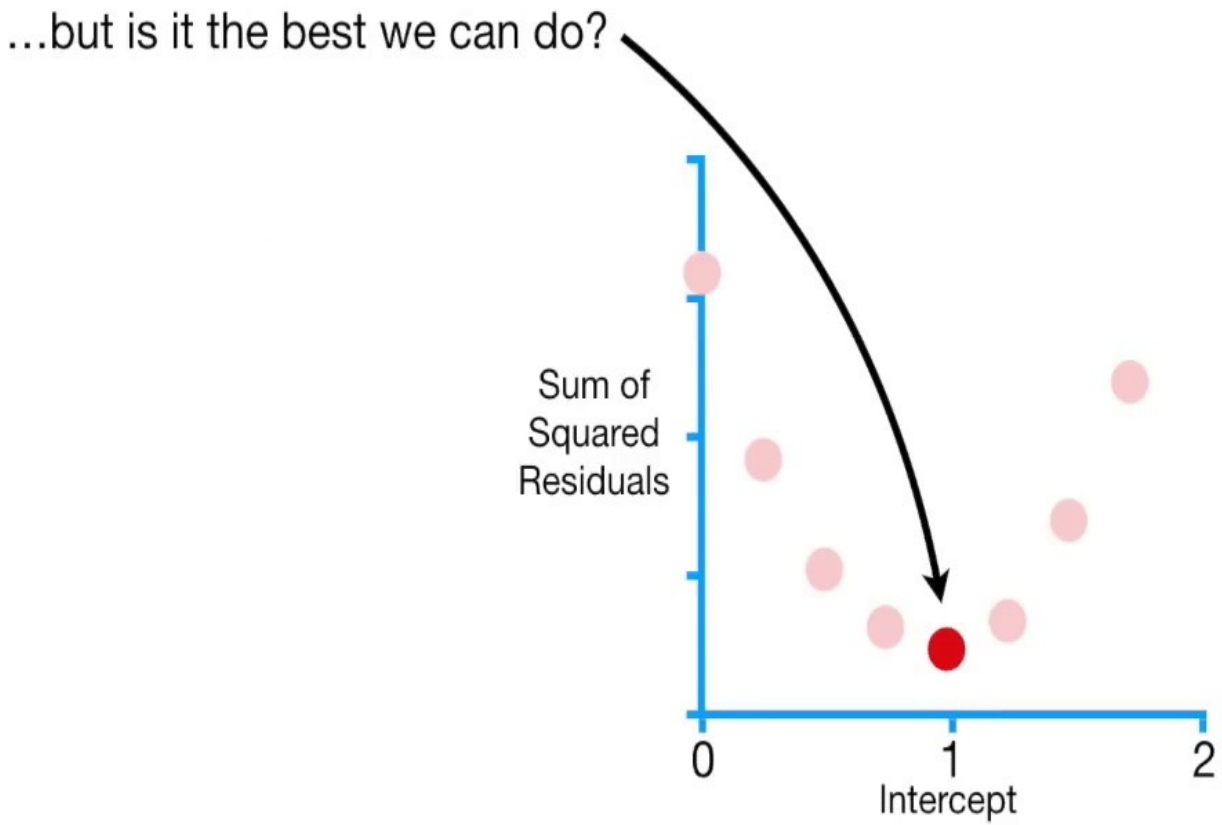
GD to find *b*


And for increasing values for the
Intercept, we get these points.
Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept

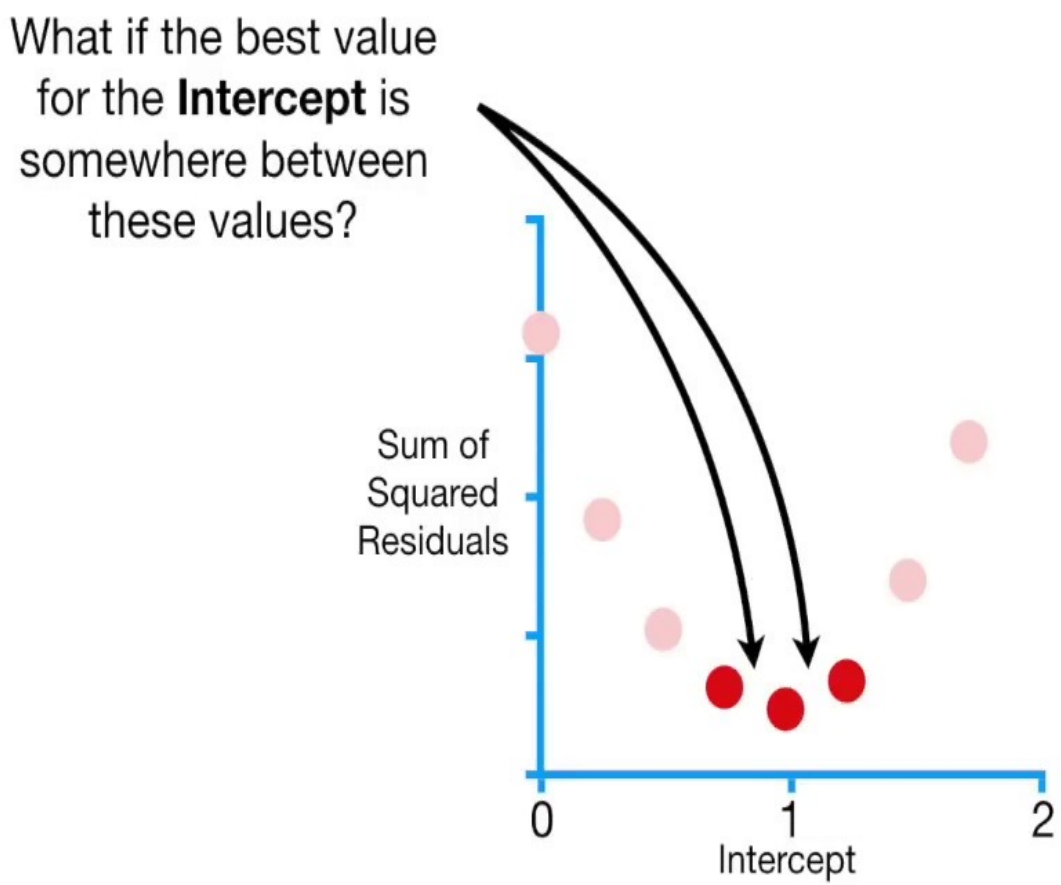
GD to find *b*


Of the points that we calculated for the graph, this one has the lowest Sum of Squared Residuals...
Sum of Squared Residuals
0
1
2
Intercept

GD to find *b*

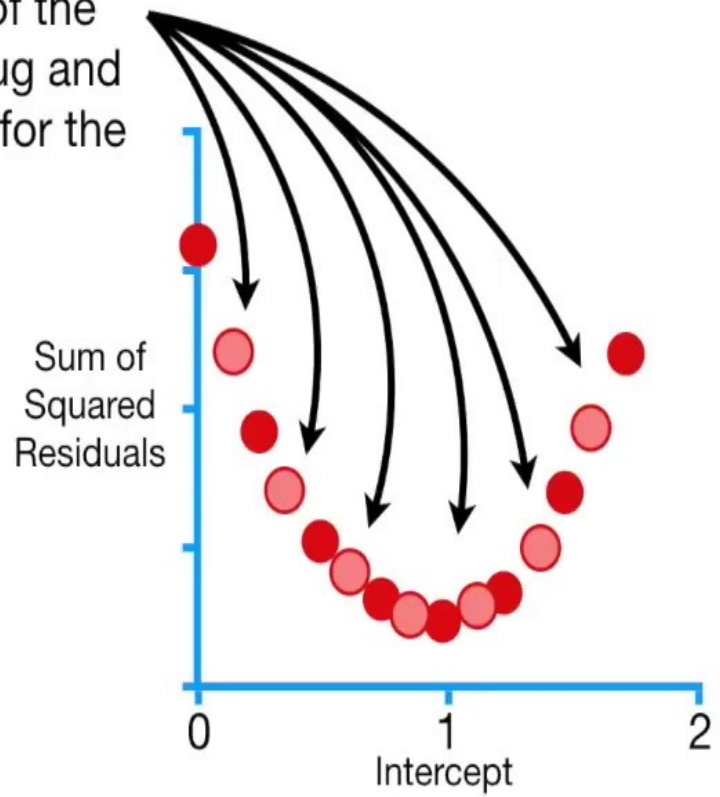

...but is it the best we can do?
Sum of Squared Residuals
0
1
2
Intercept

GD to find *b*

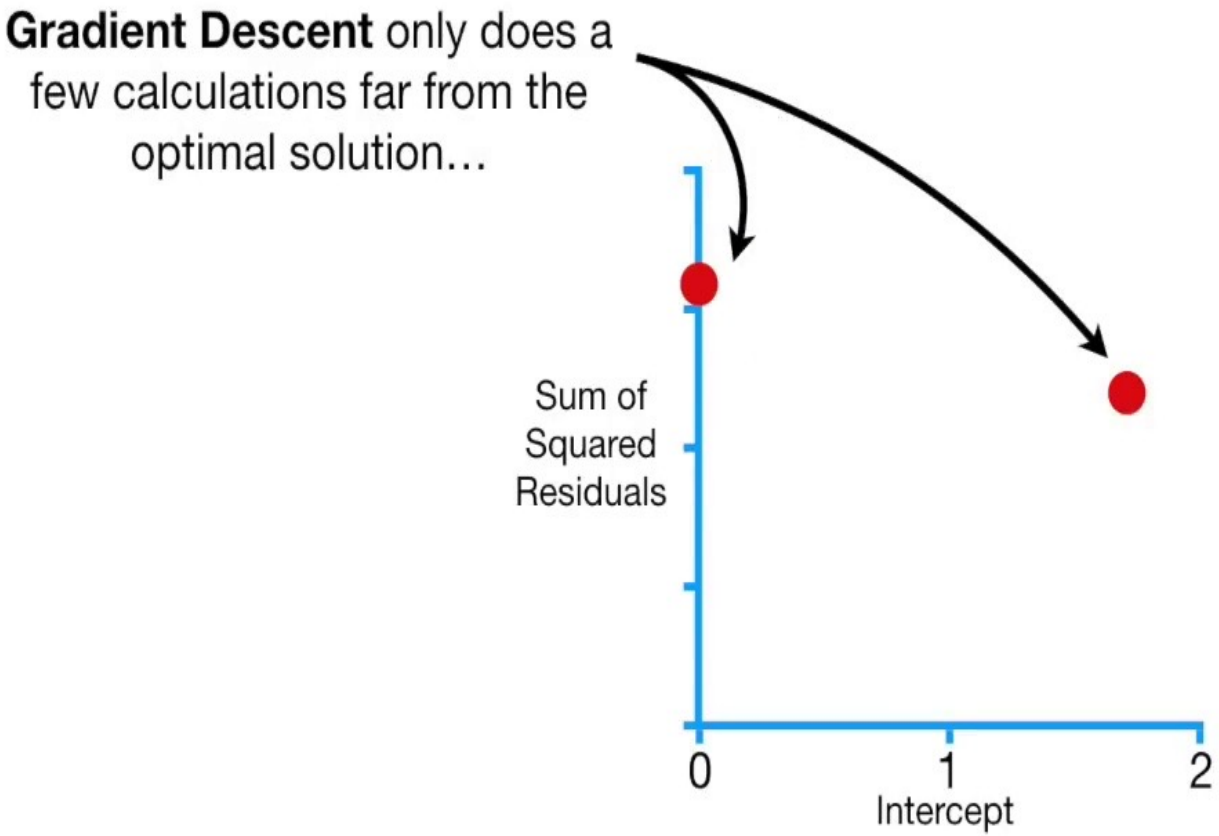

What if the best value for the **Intercept** is somewhere between these values?
Sum of Squared Residuals
0
1
2
Intercept

GD to find *b*

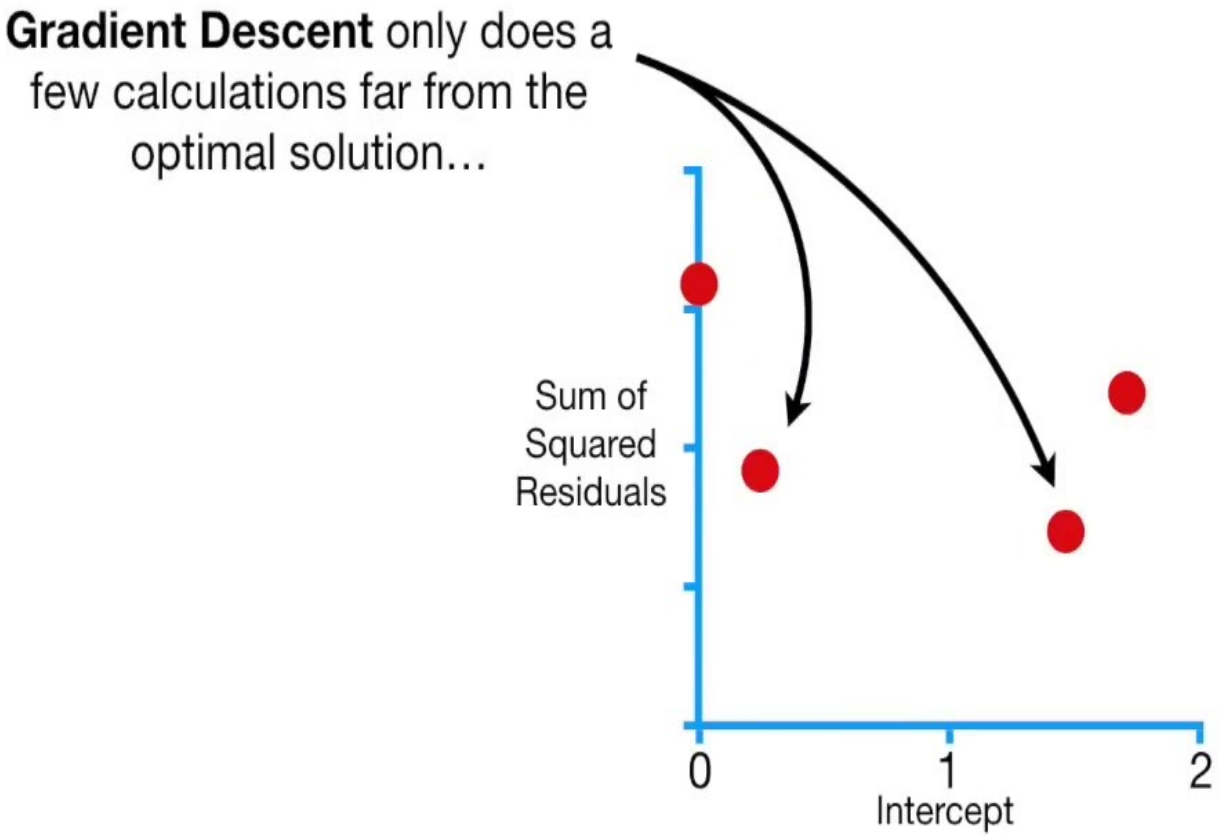
A slow and painful method for finding the minimal Sum of the Squared Residuals is to plug and chug a bunch more values for the **Intercept**.



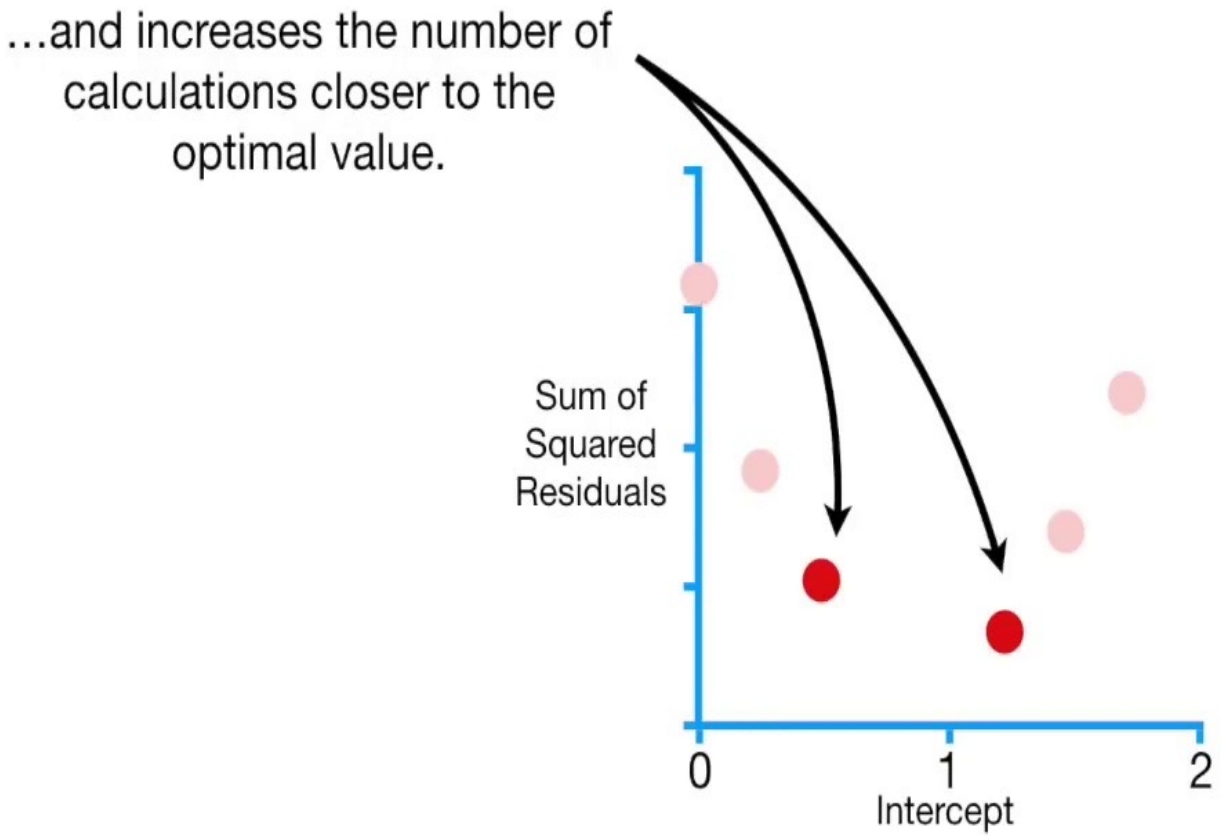
GD to find *b*


Gradient Descent only does a few calculations far from the optimal solution…
Sum of Squared Residuals
0
1
2
Intercept

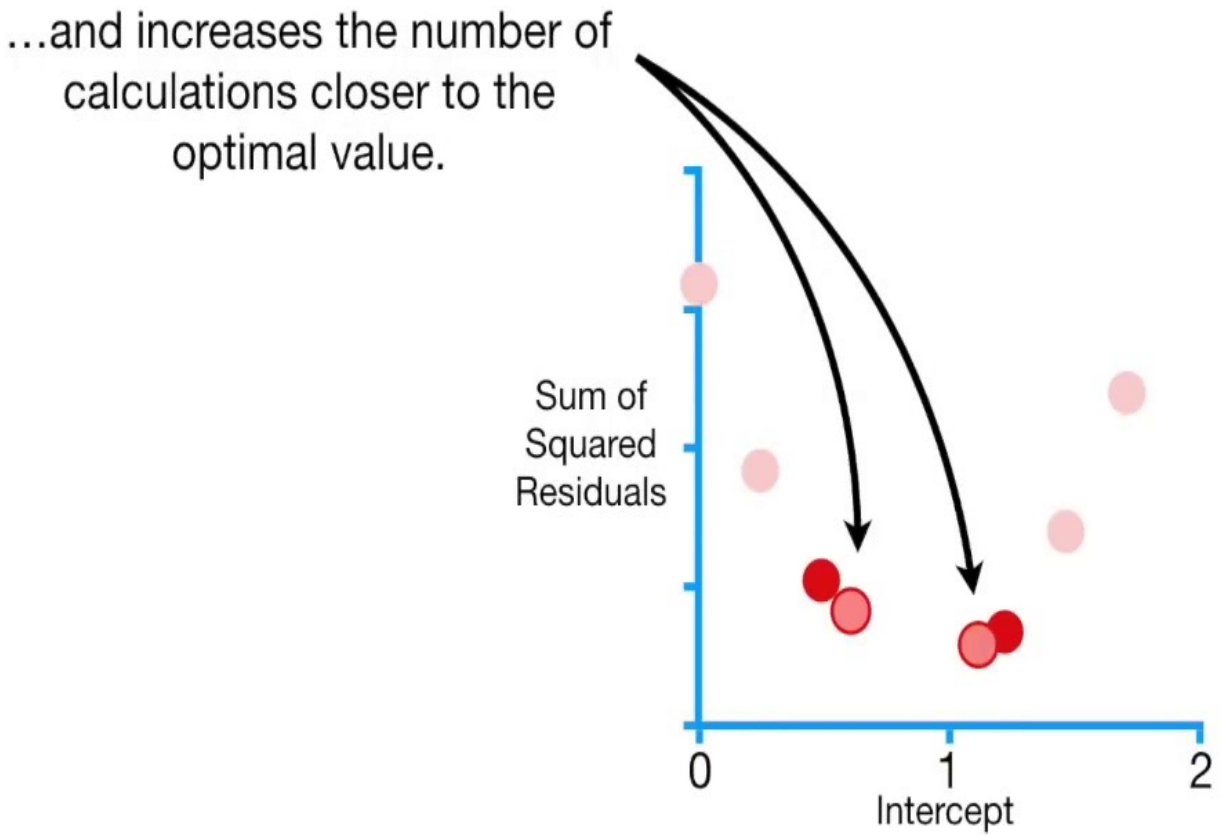
GD to find *b*


Gradient Descent only does a few calculations far from the optimal solution…
Sum of Squared Residuals
0
1
2
Intercept

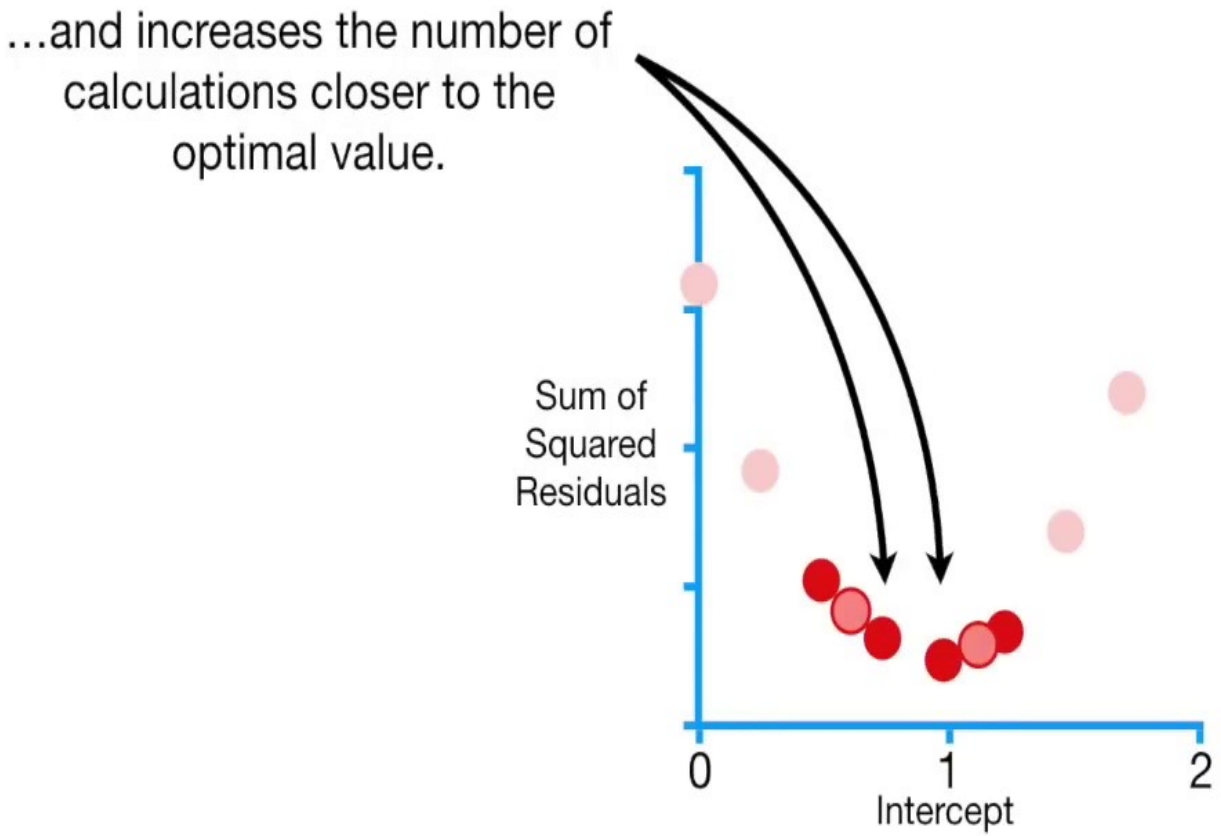
GD to find *b*


...and increases the number of calculations closer to the optimal value.
Sum of Squared Residuals
0
1
2
Intercept

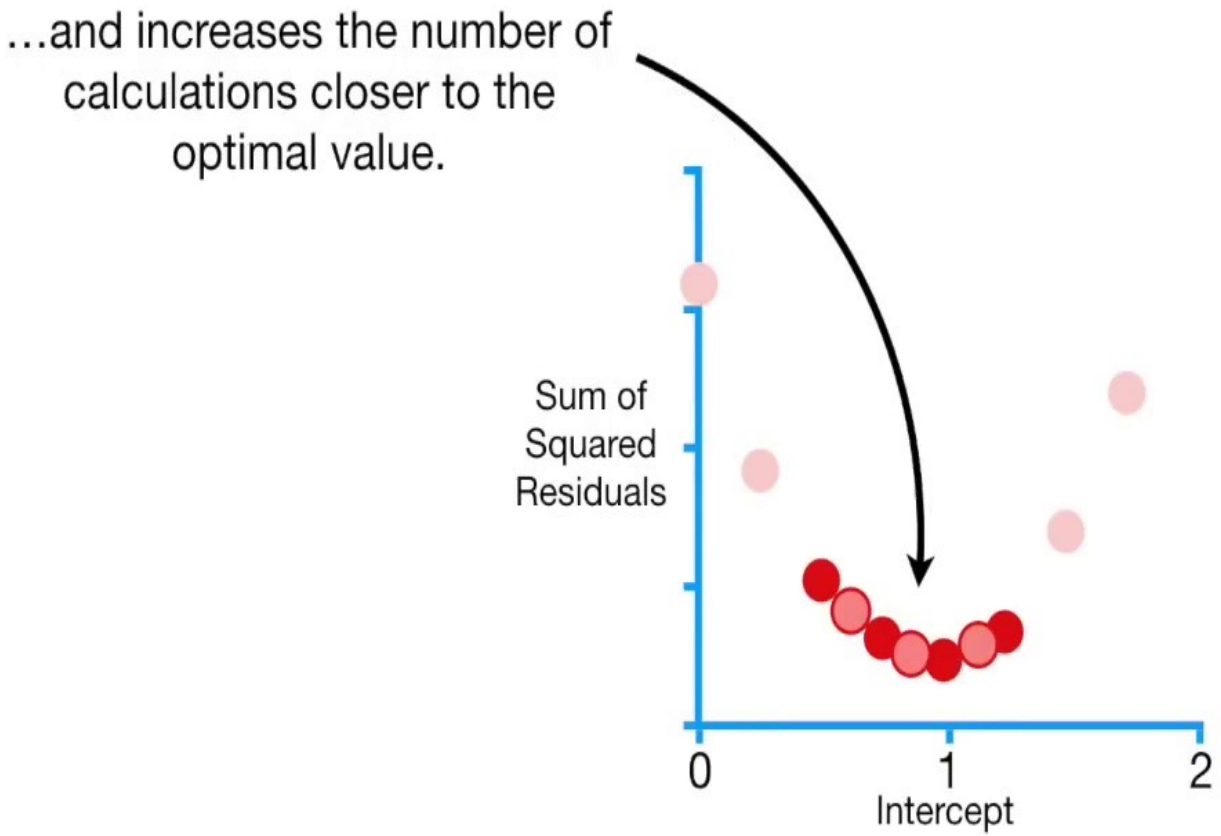
GD to find *b*


...and increases the number of calculations closer to the optimal value.
Sum of Squared Residuals
0
1
2
Intercept

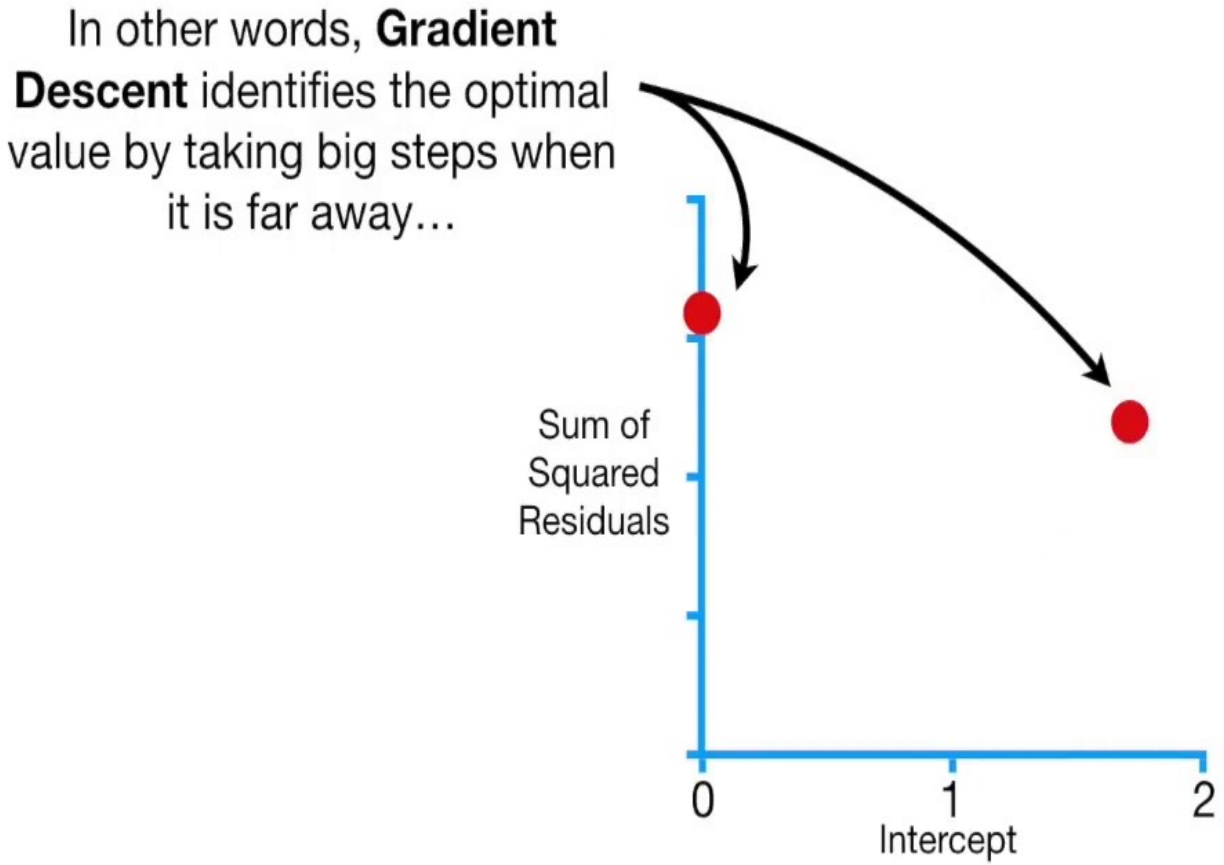
GD to find *b*


...and increases the number of calculations closer to the optimal value.
Sum of Squared Residuals
0
1
2
Intercept

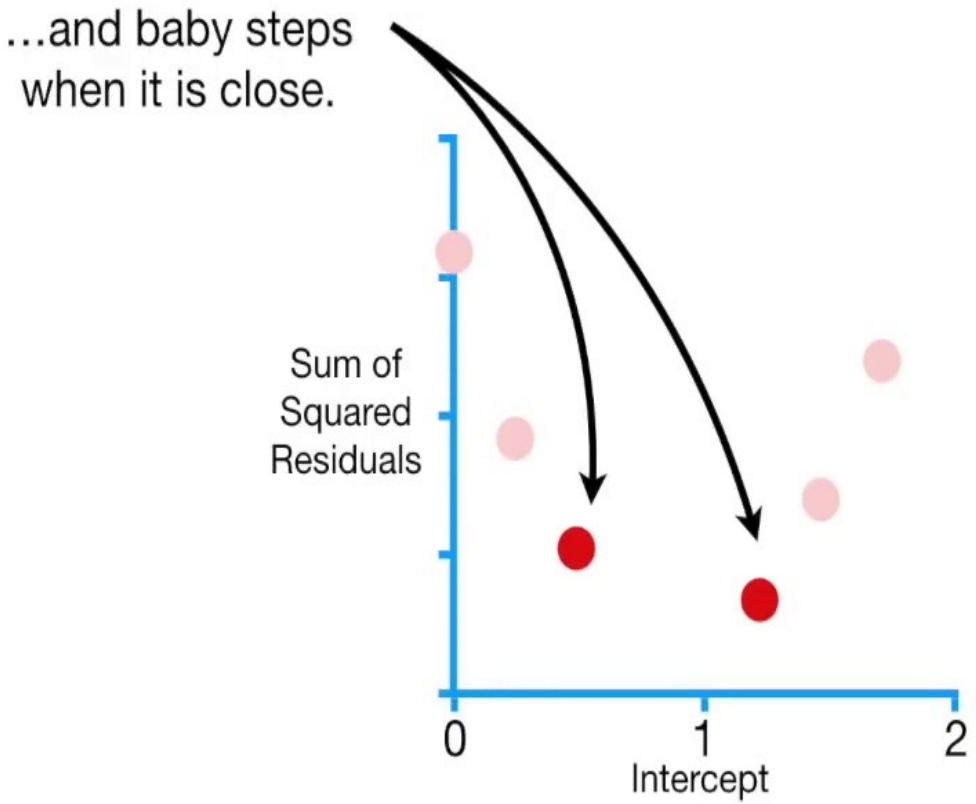
GD to find *b*


...and increases the number of calculations closer to the optimal value.
Sum of Squared Residuals
0
1
2
Intercept

GD to find *b*

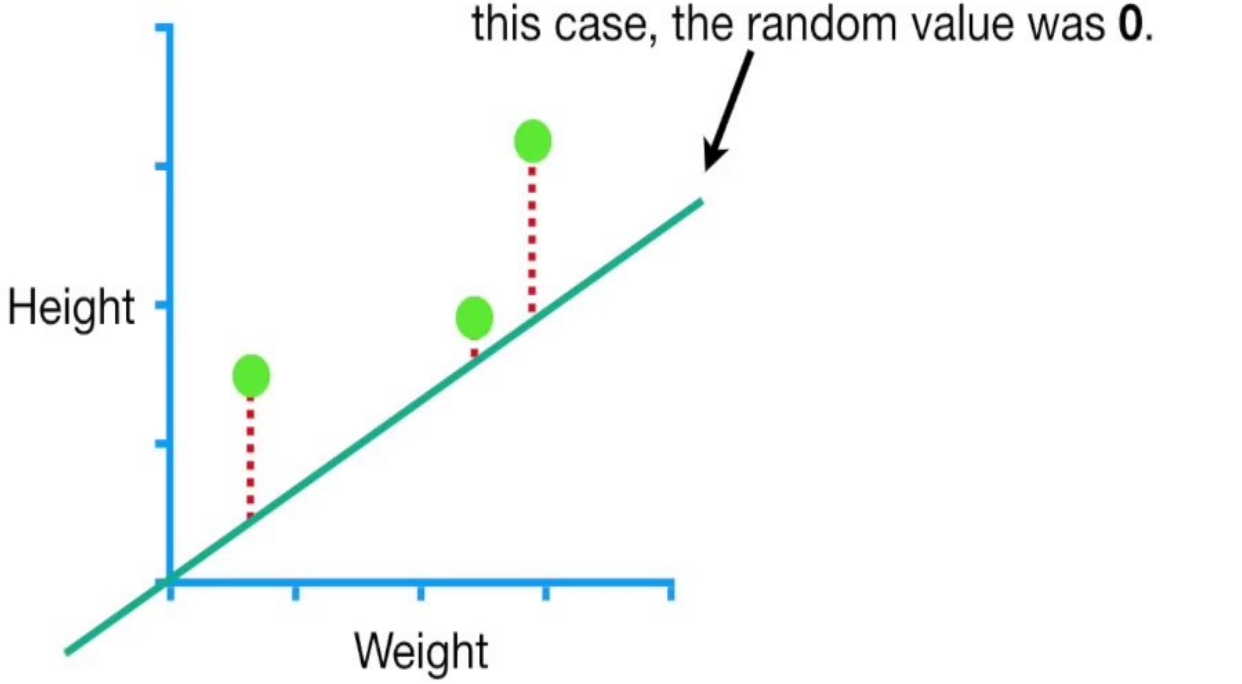

In other words, **Gradient Descent** identifies the optimal value by taking big steps when it is far away…
Sum of Squared Residuals
0
1
2
Intercept

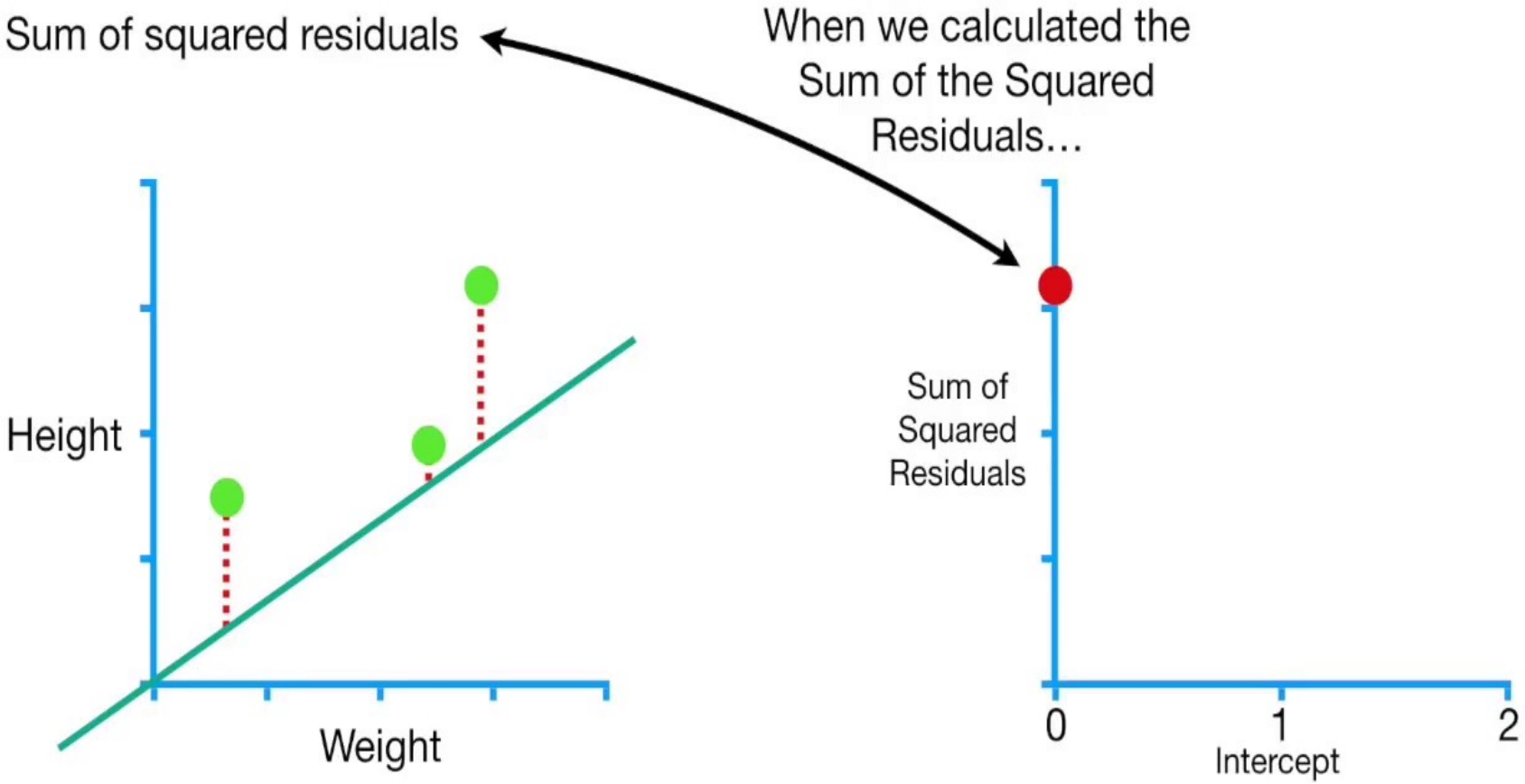
GD to find *b*


...and baby steps
when it is close.
Sum of
Squared
Residuals
0
1
2
Intercept

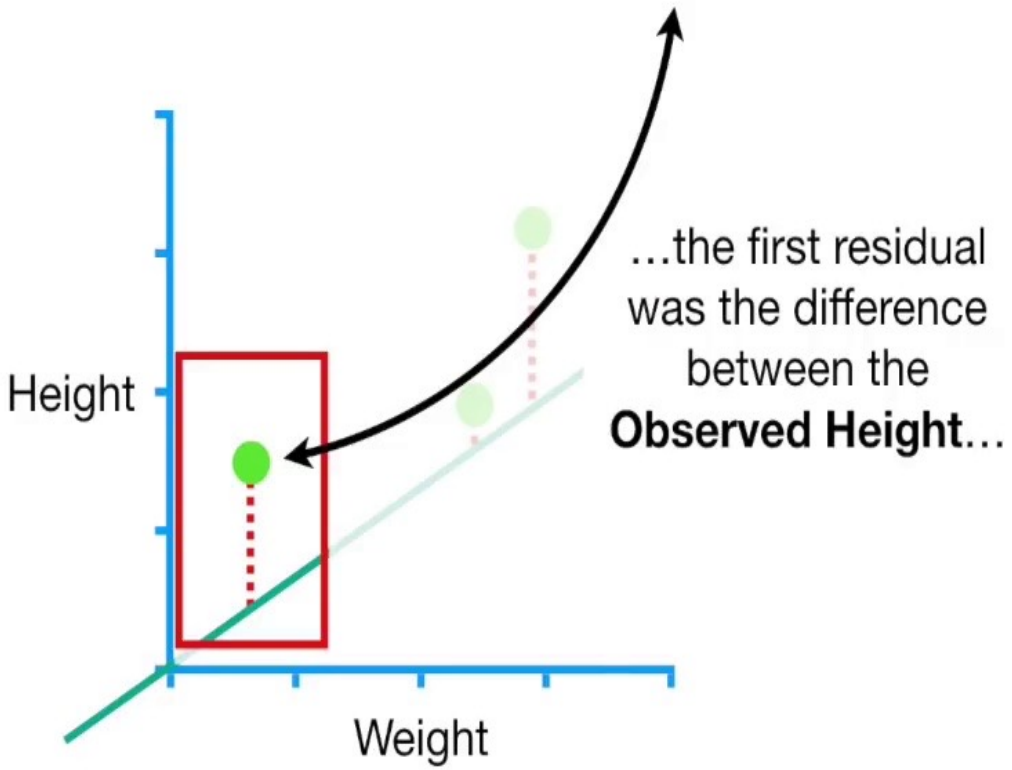
GD to find *b*

So let's get back to using **Gradient Descent** to find the optimal value for the **Intercept**, starting from a random value. In this case, the random value was **0**.

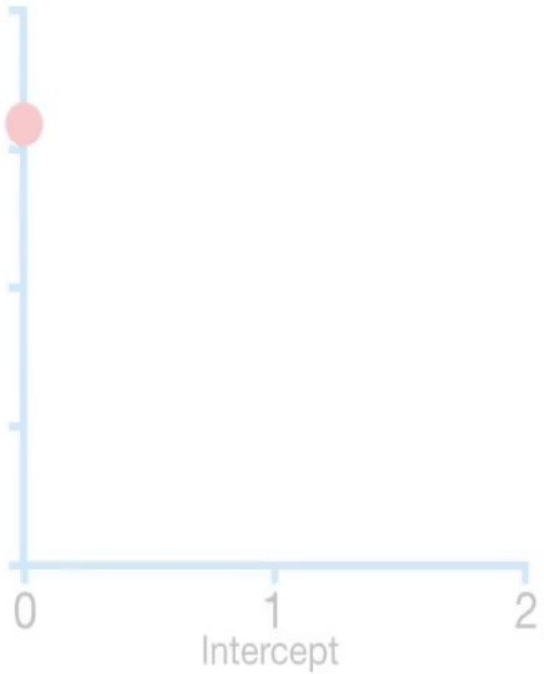


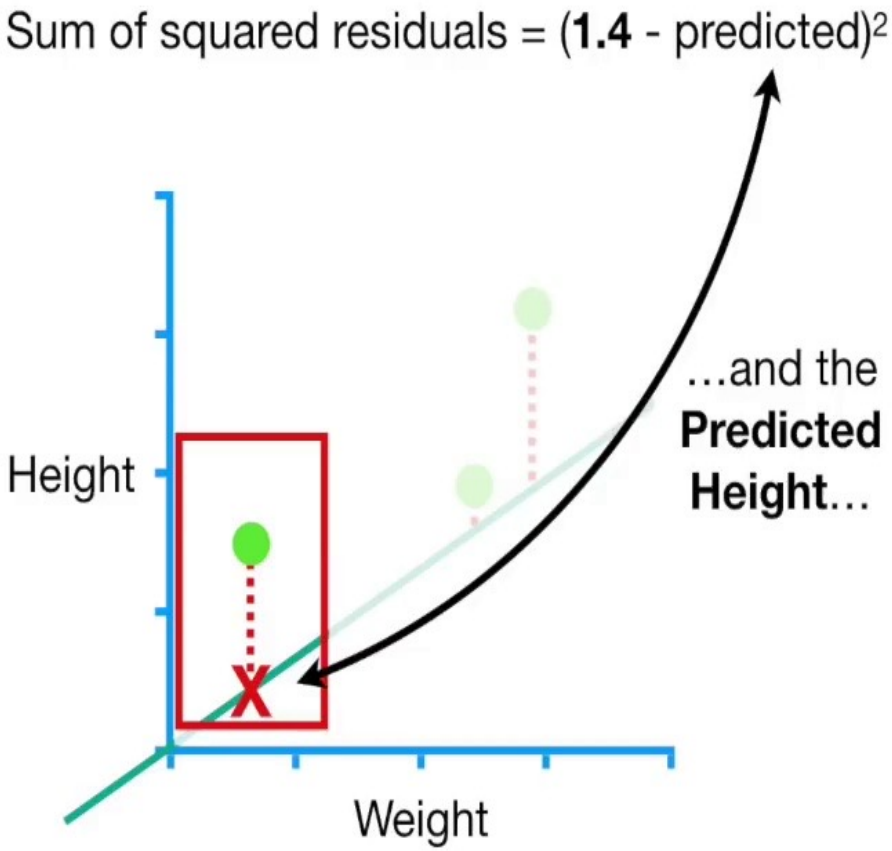

Sum of squared residuals
When we calculated the Sum of the Squared Residuals…
Height
Weight
Sum of Squared Residuals
0
1
2
Intercept

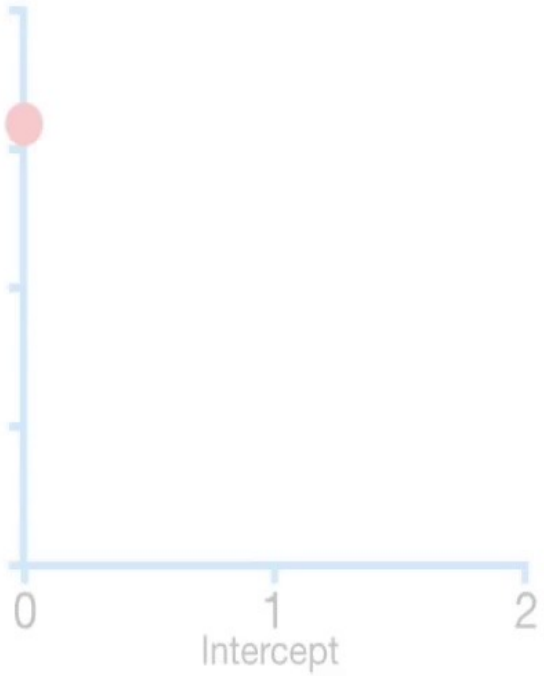
Sum of squared residuals = (observed - predicted)^2



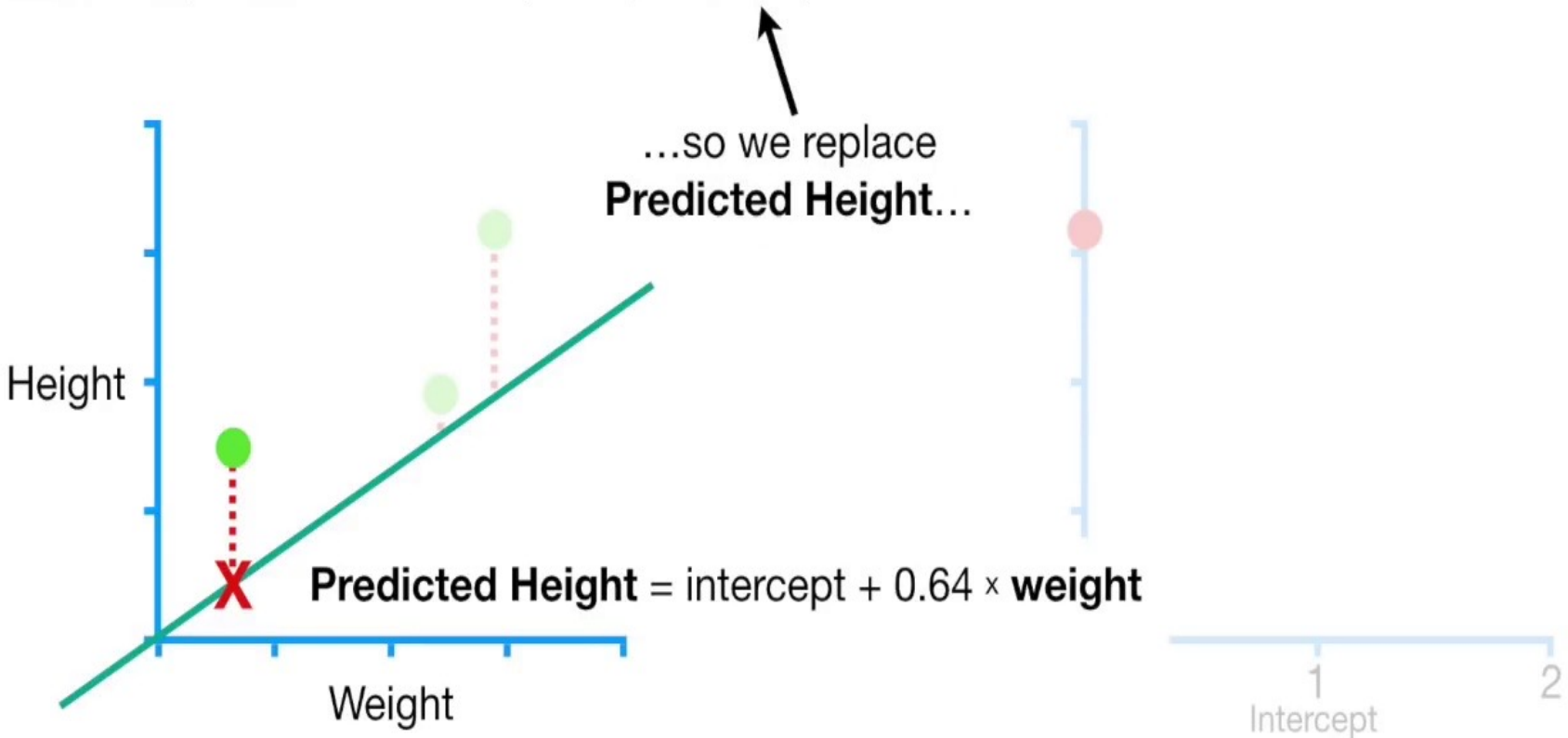
...the first residual was the difference between the **Observed Height**...



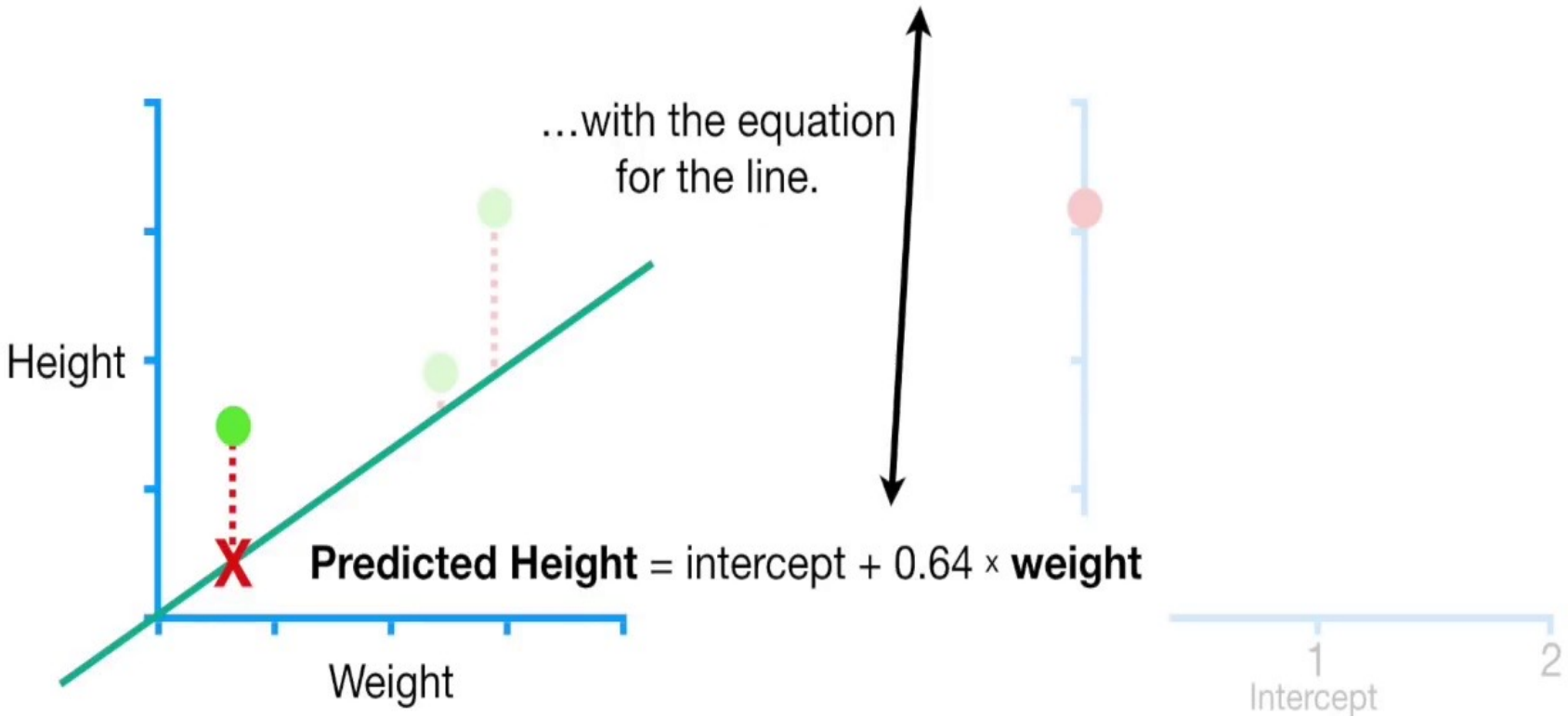

Sum of squared residuals = (**1.4** - predicted)^2
…and the **Predicted Height**…
Height
Weight


0
1
2
Intercept

Sum of squared residuals = (**1.4** - predicted)^2

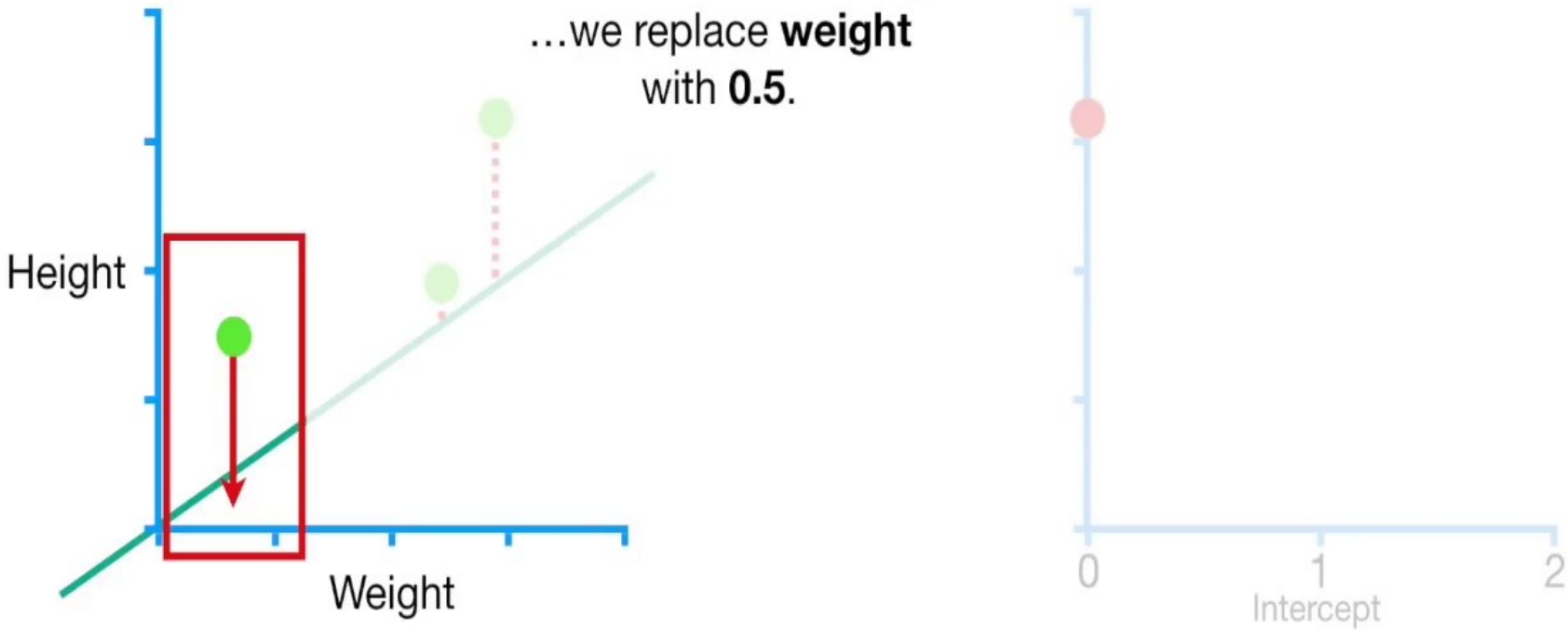


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **weight**))^2



Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

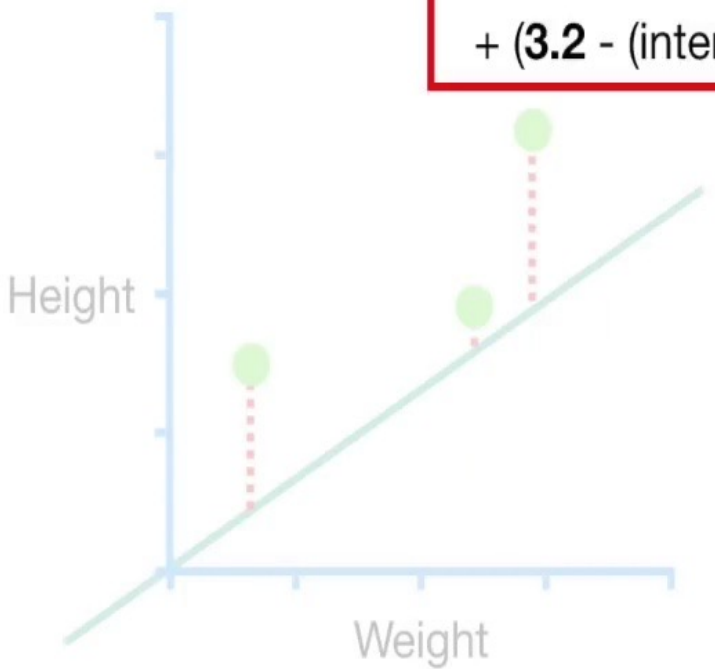
…we replace **weight** with **0.5**.



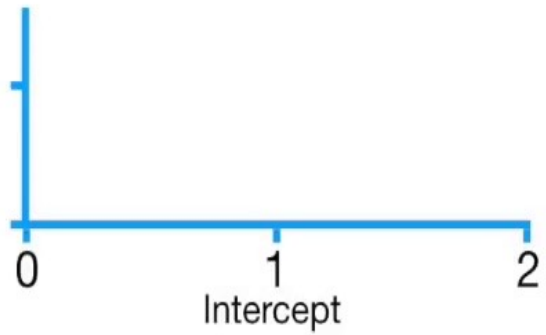
Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

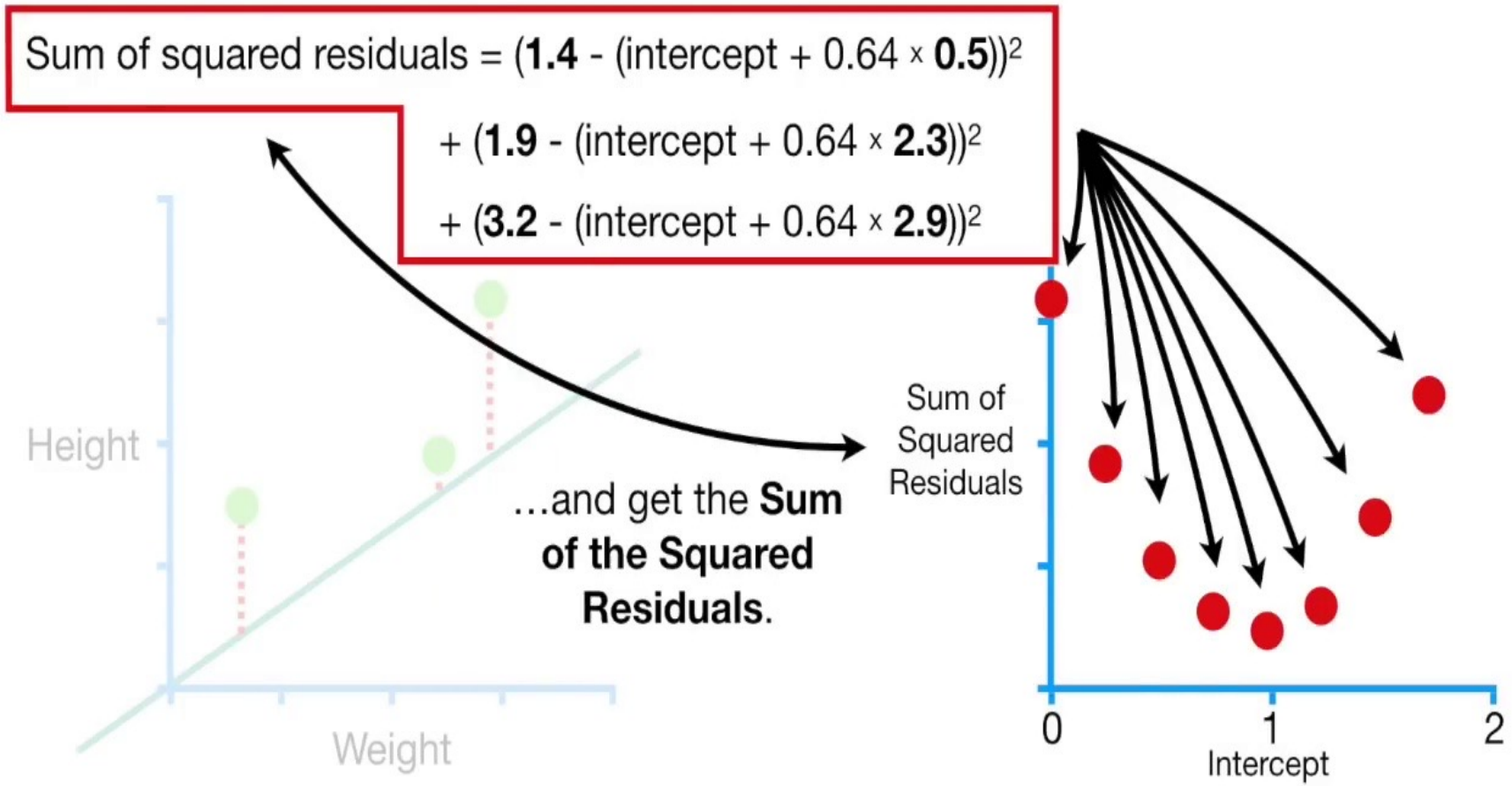
+ (**1.9** - (intercept + 0.64 x **2.3**))^2

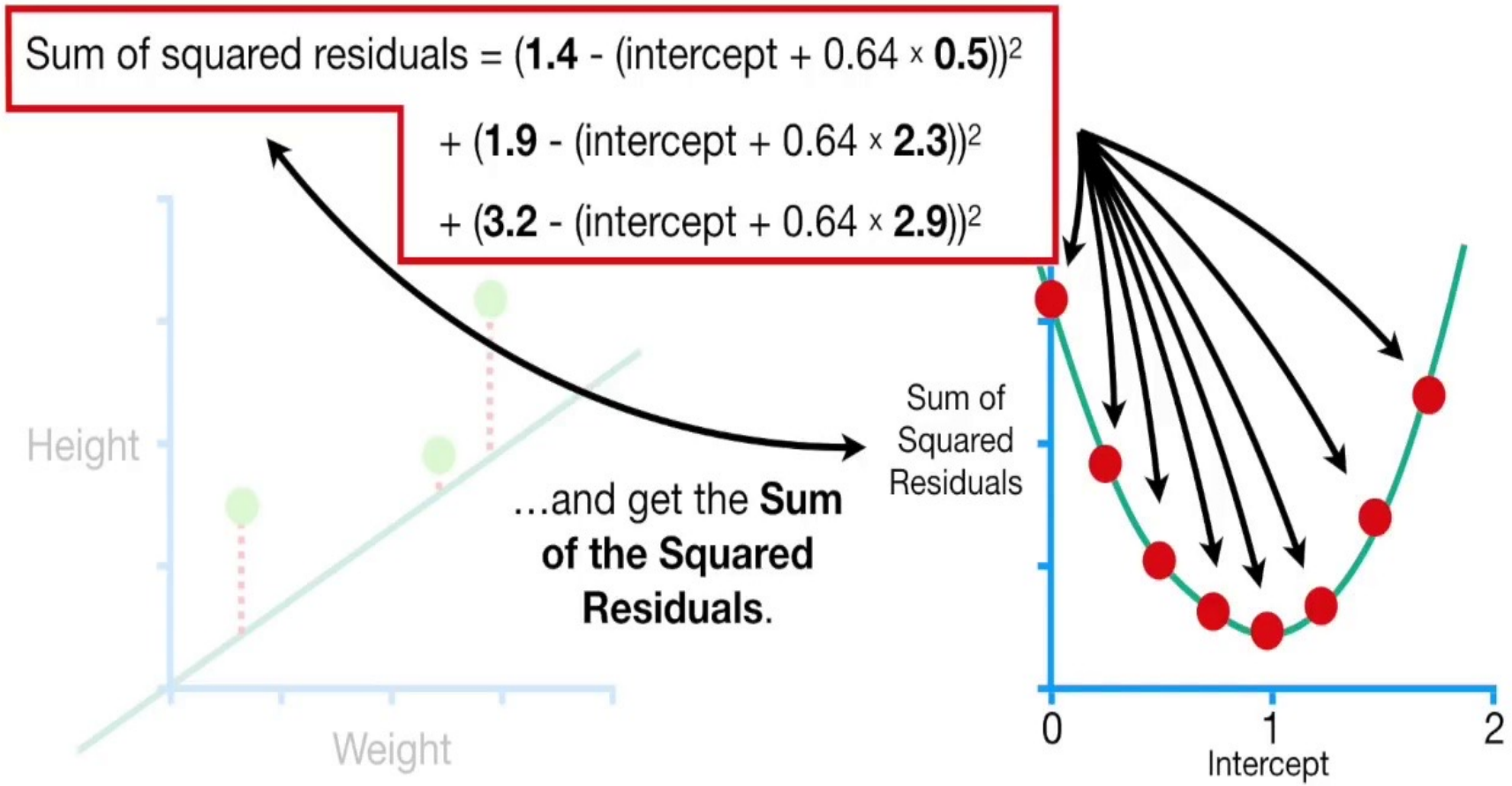
+ (**3.2** - (intercept + 0.64 x **2.9**))^2



Now we can easily plug in any value for the **intercept**…




Sum of squared residuals = (1.4 - (intercept + 0.64 x 0.5))²
+ (1.9 - (intercept + 0.64 x 2.3))²
+ (3.2 - (intercept + 0.64 x 2.9))²
Height
Weight
…and get the Sum of the Squared Residuals.
Sum of Squared Residuals
0
1
2
Intercept

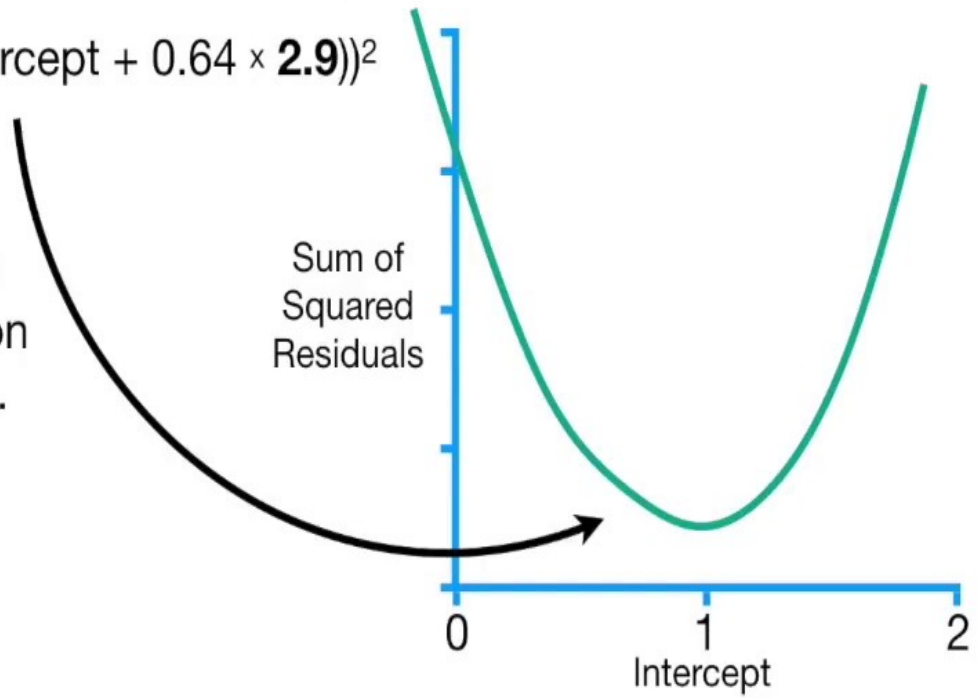

Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2
+ (**1.9** - (intercept + 0.64 x **2.3**))^2
+ (**3.2** - (intercept + 0.64 x **2.9**))^2
Height
Weight
…and get the **Sum of the Squared Residuals**.
Sum of Squared Residuals
0
1
2
Intercept

Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

+ (**1.9** - (intercept + 0.64 x **2.3**))^2

+ (**3.2** - (intercept + 0.64 x **2.9**))^2

Thus, we now have an equation for this curve…

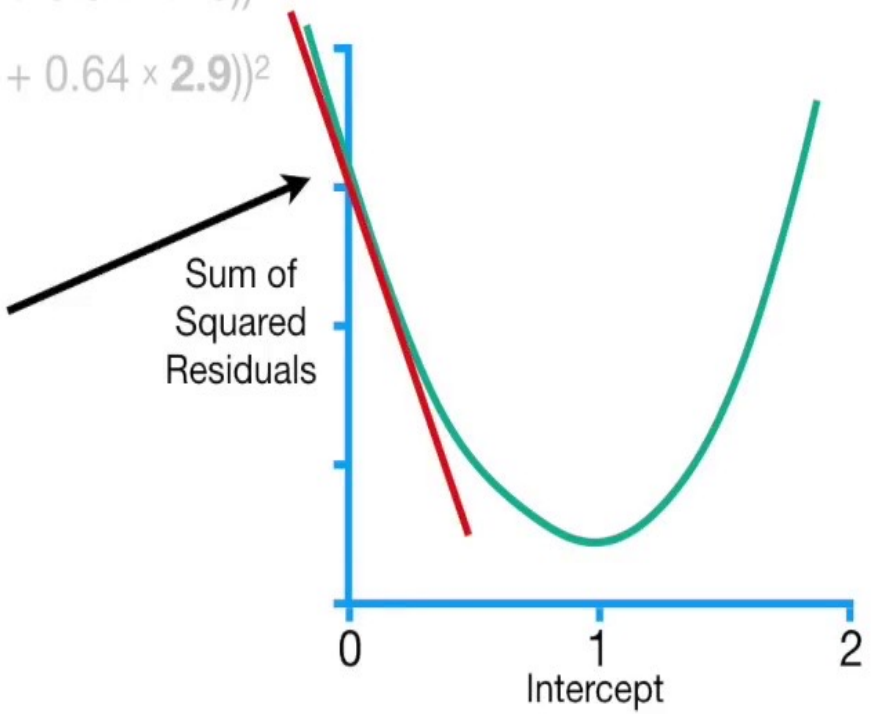


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

+ (**1.9** - (intercept + 0.64 x **2.3**))^2

+ (**3.2** - (intercept + 0.64 x **2.9**))^2

…and we can take the derivative of this function and determine the slope at any value for the **Intercept**.

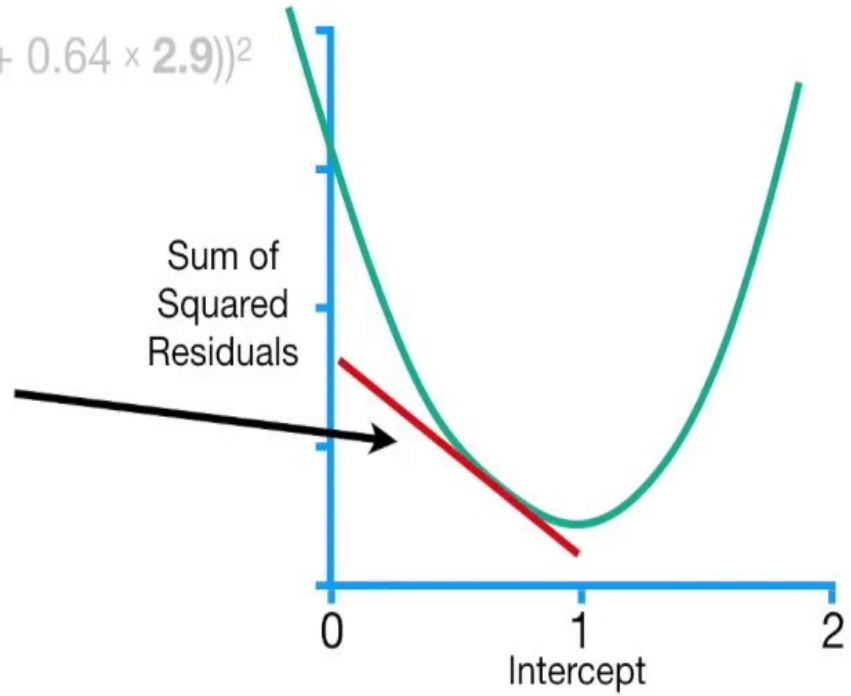


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

+ (**1.9** - (intercept + 0.64 x **2.3**))^2

+ (**3.2** - (intercept + 0.64 x **2.9**))^2

…and we can take the derivative of this function and determine the slope at any value for the **Intercept**.

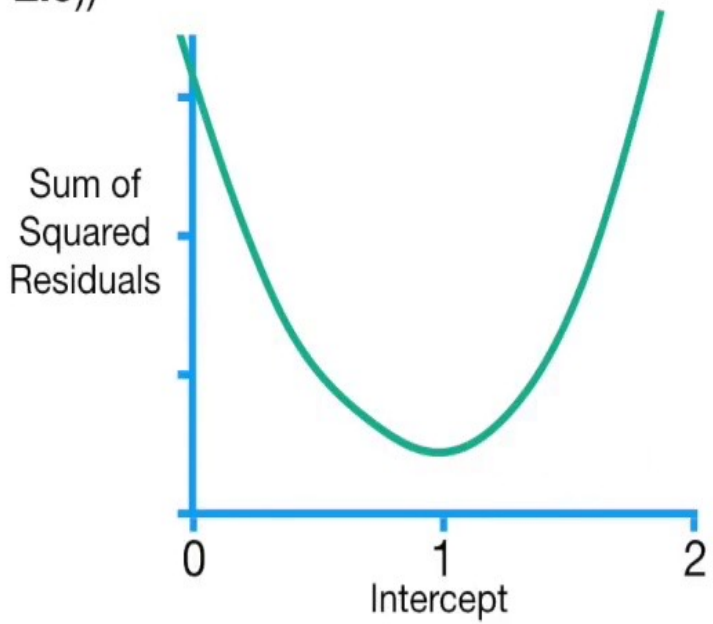


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

+ (**1.9** - (intercept + 0.64 x **2.3**))^2

+ (**3.2** - (intercept + 0.64 x **2.9**))^2

So let's take the derivative of the Sum of the Squared Residuals with respect to the **Intercept**.



$$\text{Sum of squared residuals} = (\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2 + (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2 + (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$$

$$\frac{d}{d\ \text{intercept}} \text{Sum of squared residuals} =$$

$$\text{Sum of squared residuals} = (\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$$
$$+ (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$$
$$+ (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$$

…the derivative of the first part…

$$\frac{d}{d\ \text{intercept}} \text{Sum of squared residuals} = \frac{d}{d\ \text{intercept}} (\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$$

Sum of squared residuals = $(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$

$+ (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$

$+ (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$

…plus the derivative of the second part…

$$\frac{d}{d\ intercept} \text{Sum of squared residuals} = \frac{d}{d\ intercept} (\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\ intercept} (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$$

Sum of squared residuals = $(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$

$+ (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$

$+ (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$

…plus the derivative of the third part.

$$\frac{d}{d\,\text{intercept}}\,\text{Sum of squared residuals} = \frac{d}{d\,\text{intercept}}(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\,\text{intercept}}(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\,\text{intercept}}(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$$

$$\frac{d}{d\ intercept}(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})) \times -1$$

$$= -2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))$$

…and this…

…is the derivative of the first part…

$$\frac{d}{d\ intercept} \text{Sum of squared residuals} = \frac{d}{d\ intercept}(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\ intercept}(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\ intercept}(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))^2$$

$$\frac{d}{d\,intercept}(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})) \times -1$$

$$= -2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))$$

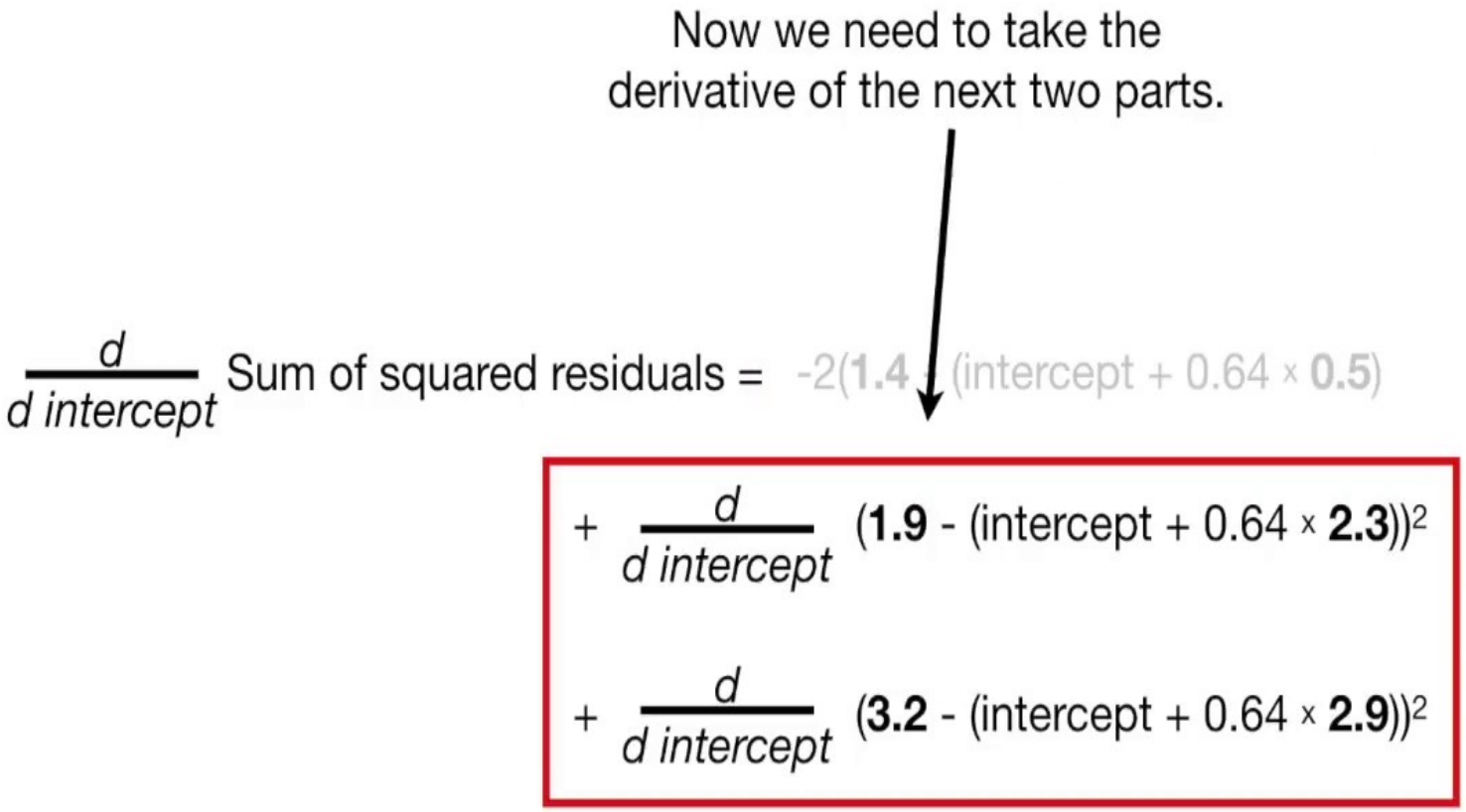
…so we plug it in.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))$$

$$+ \frac{d}{d\,intercept}(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\,intercept}(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))^2$$

GD to find *b*


Now we need to take the derivative of the next two parts.
$\frac{d}{d\ intercept}$ Sum of squared residuals = -2(1.4 - (intercept + 0.64 × 0.5)
+ $\frac{d}{d\ intercept}$ (1.9 - (intercept + 0.64 × 2.3))²
+ $\frac{d}{d\ intercept}$ (3.2 - (intercept + 0.64 × 2.9))²

GD to find *b*

$$\frac{d}{d\ intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$

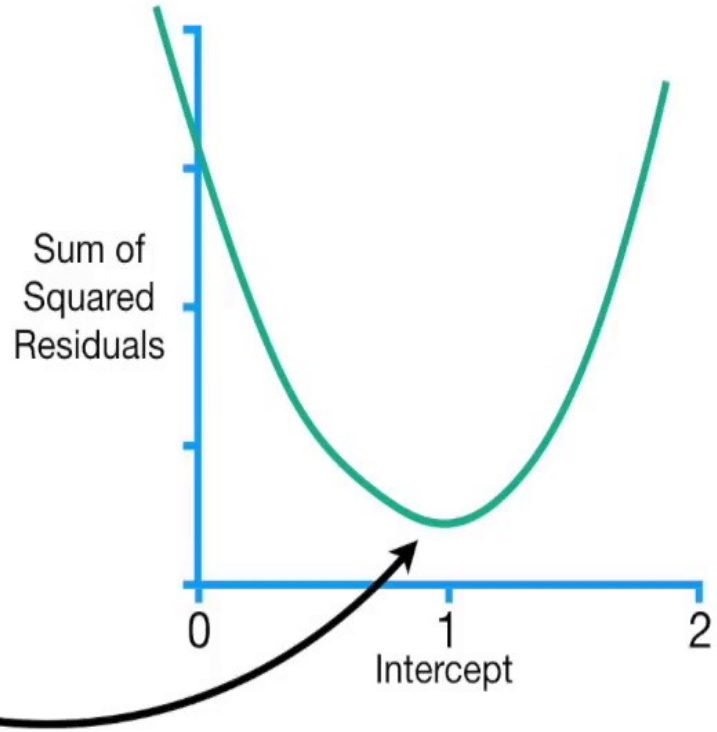
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$

Now that we have the derivative, **Gradient Descent** will use it to find where the Sum of Squared Residuals is lowest.



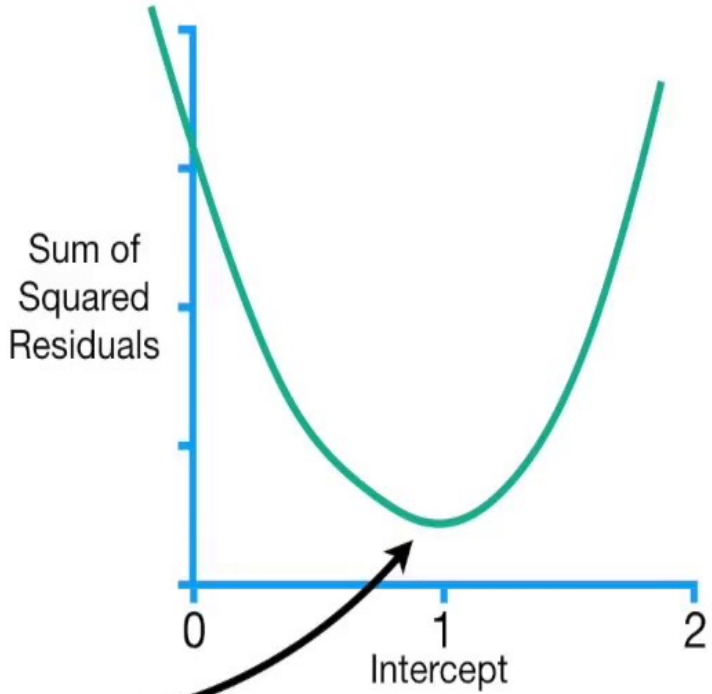
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))$$

NOTE: If we were using **Least Squares** to solve for the optimal value for the **Intercept**, we would simply find where the the slope of the curve = **0**.



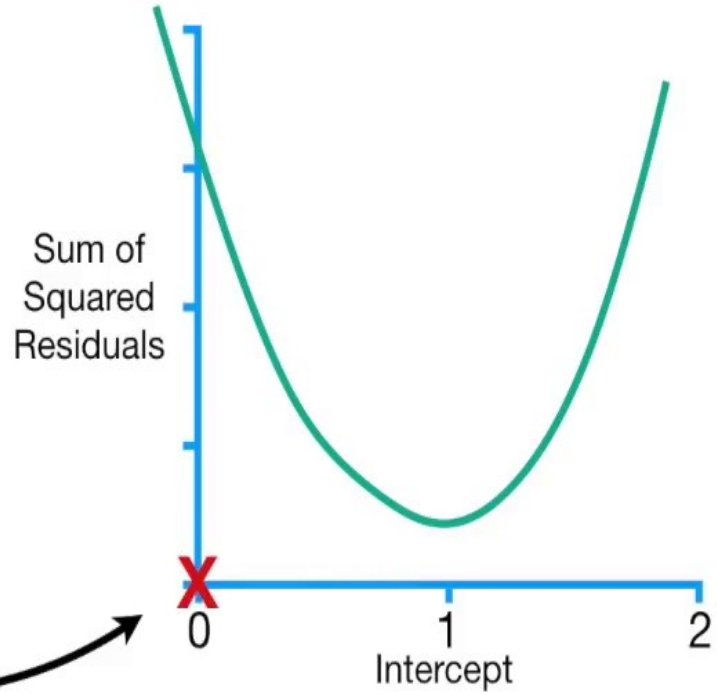
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))$$

In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.



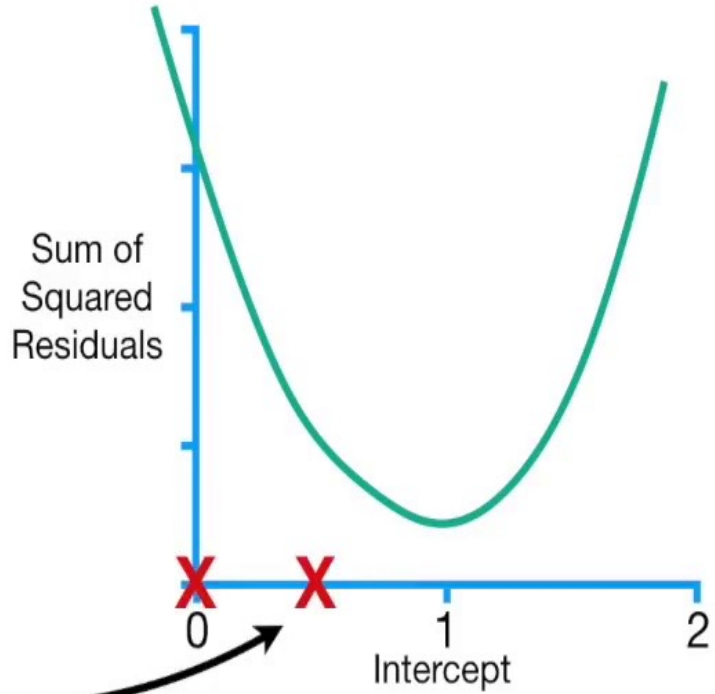
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))$$

In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.

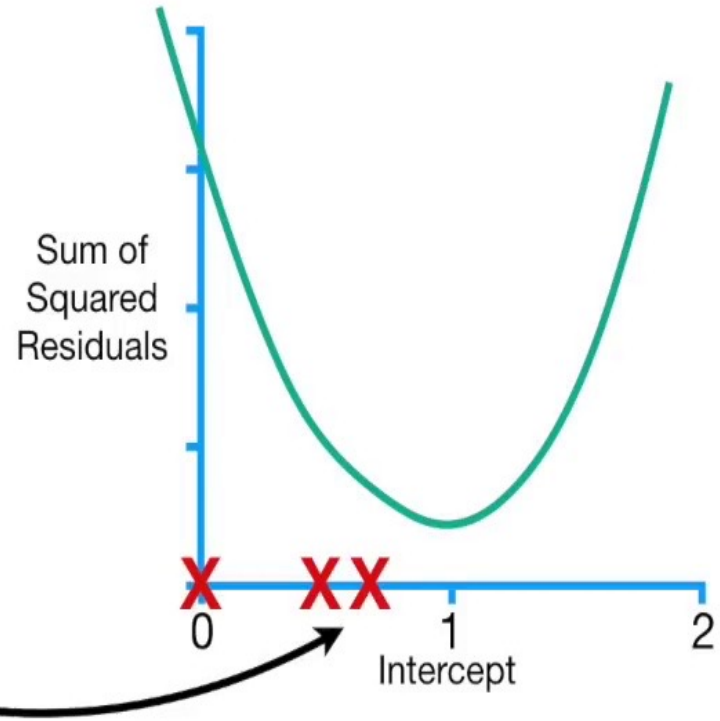


$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$



In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.

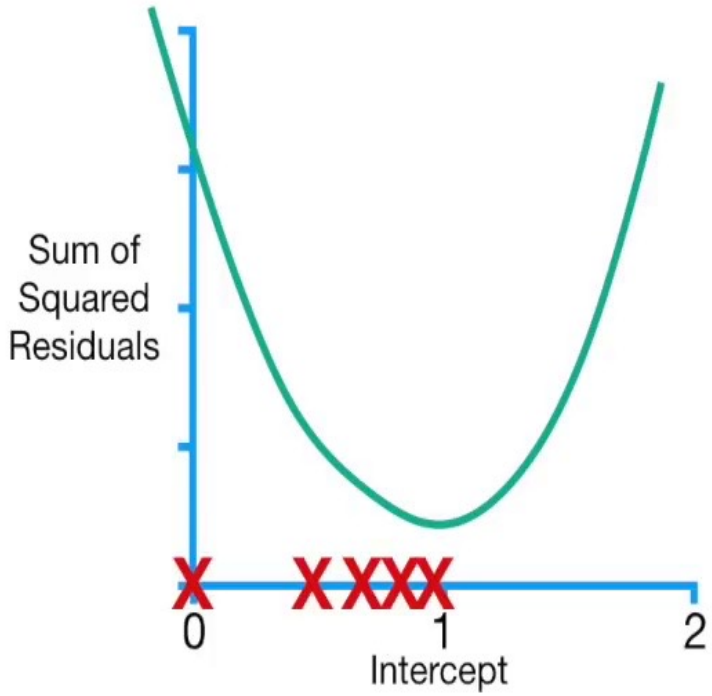
$$\frac{d}{d\,intercept}\text{ Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$

This makes **Gradient Descent** very useful when it is not possible to solve for where the derivative = **0**, and this is why **Gradient Descent** can be used in so many different situations.



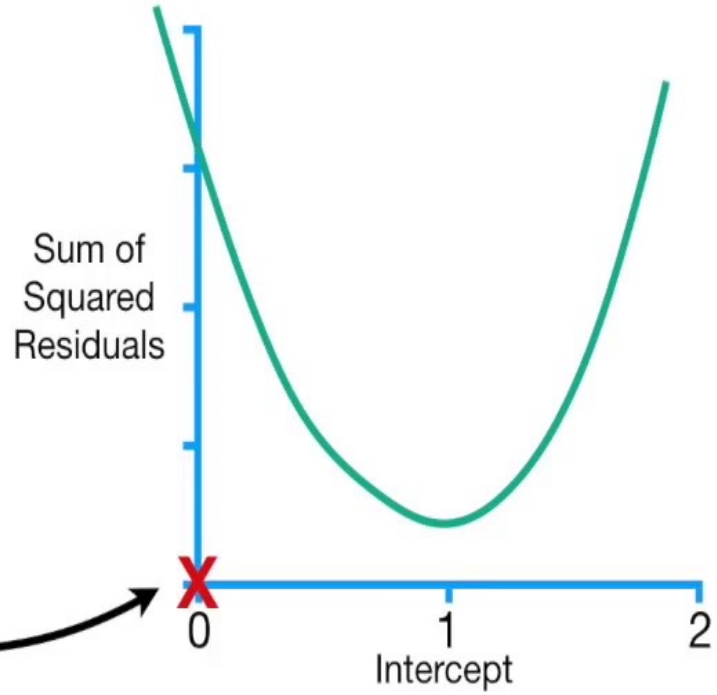
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))$$

Remember, we started by setting the **Intercept** to a random number. In this case, that was **0**.



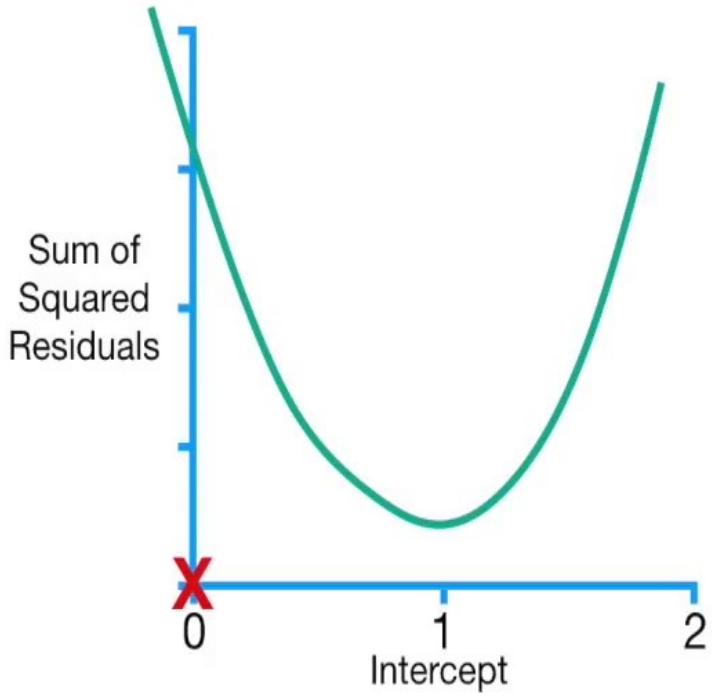
$$\frac{d}{d\,intercept}\text{ Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

So we plug **0** into the derivative…



$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

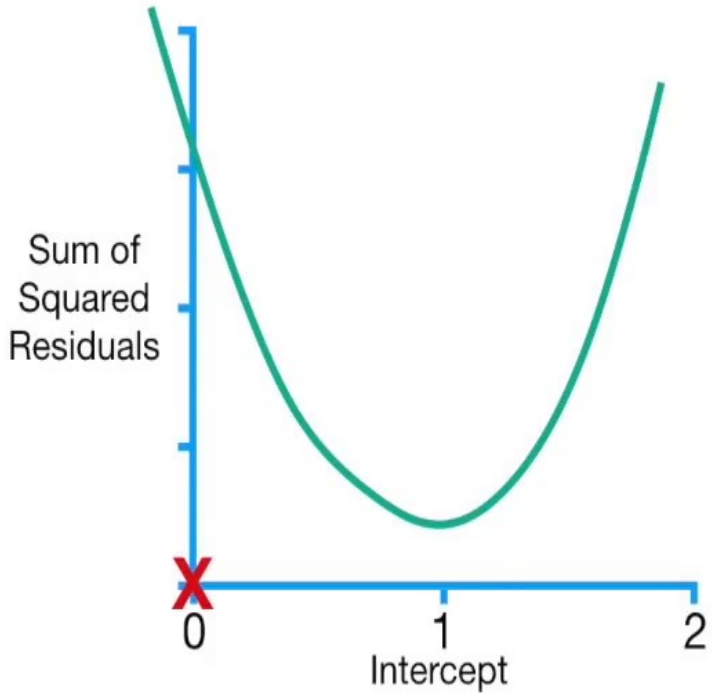
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

…and we get **-5.7**.



$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

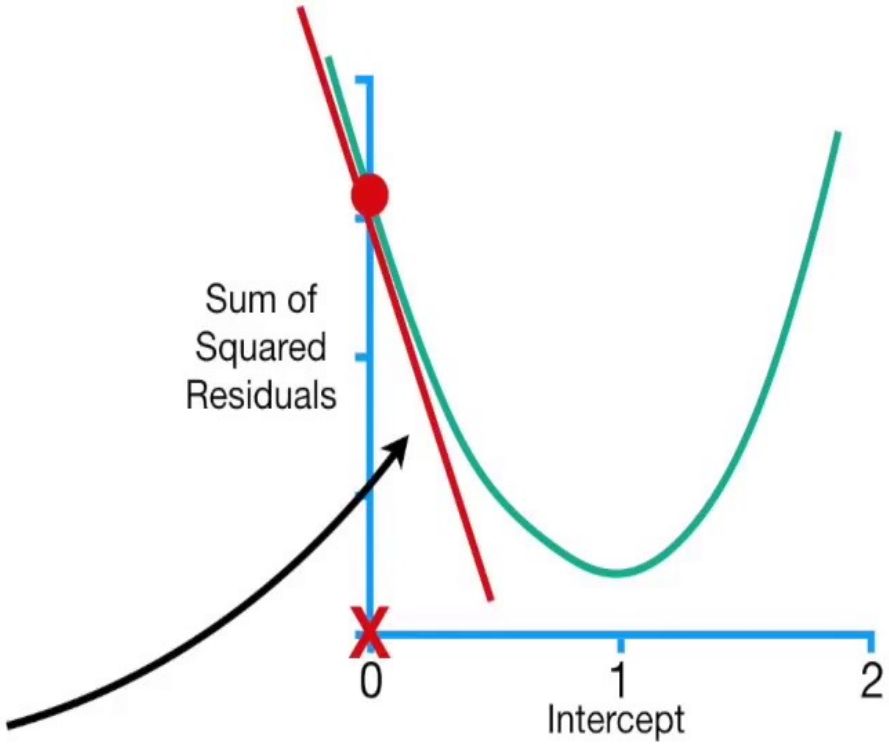
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

So when the **Intercept** = **0**, the slope of the curve = **-5.7**.



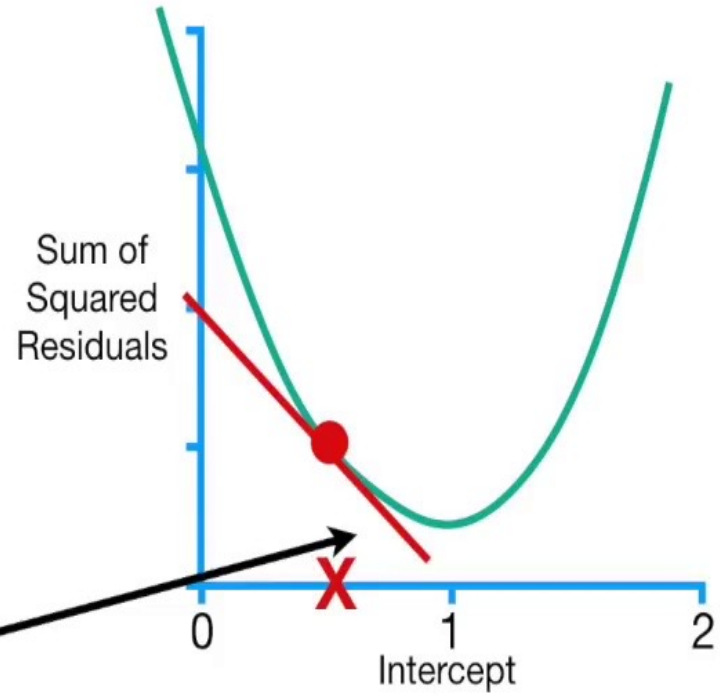
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



NOTE: The closer we get to the optimal value for the **Intercept**, the closer the slope of the curve gets to **0**.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

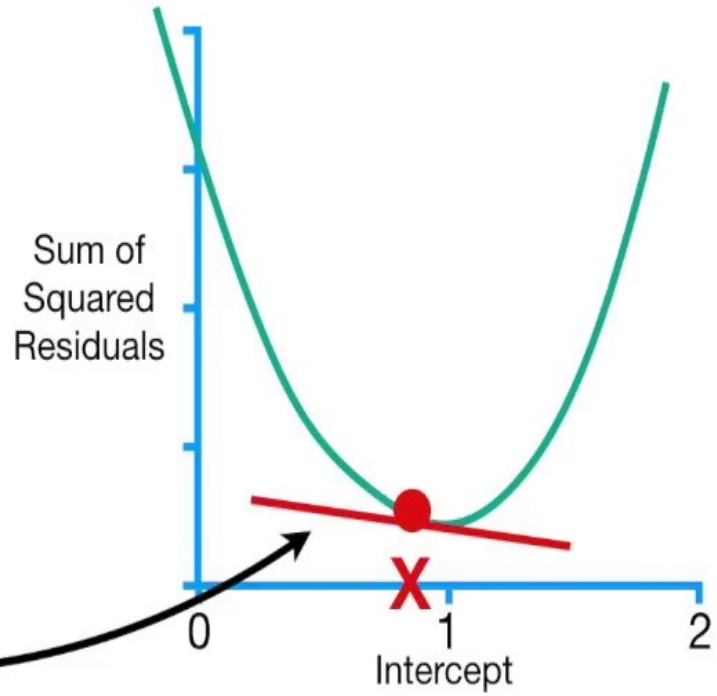
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

This means that when the slope of the curve is close to **0**...



$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

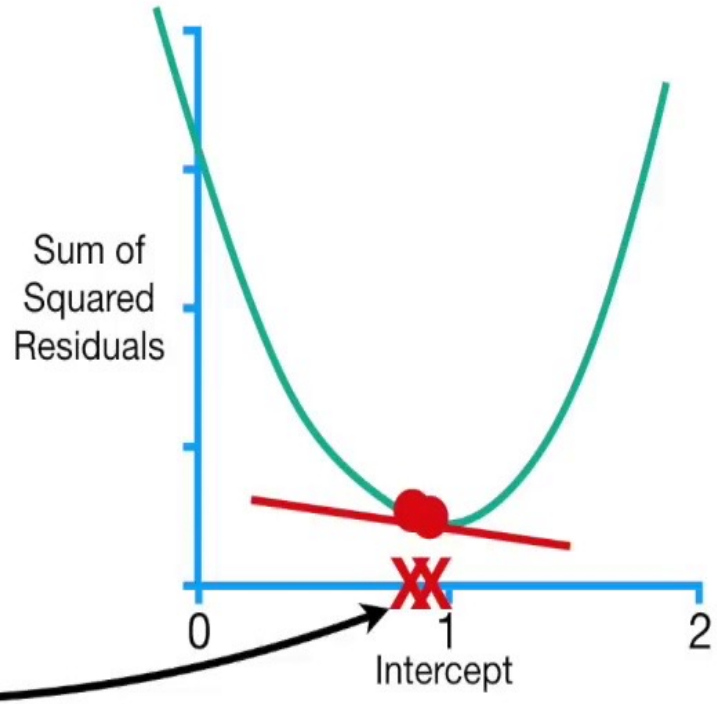
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

…then we should take baby steps, because we are close to the optimal value…



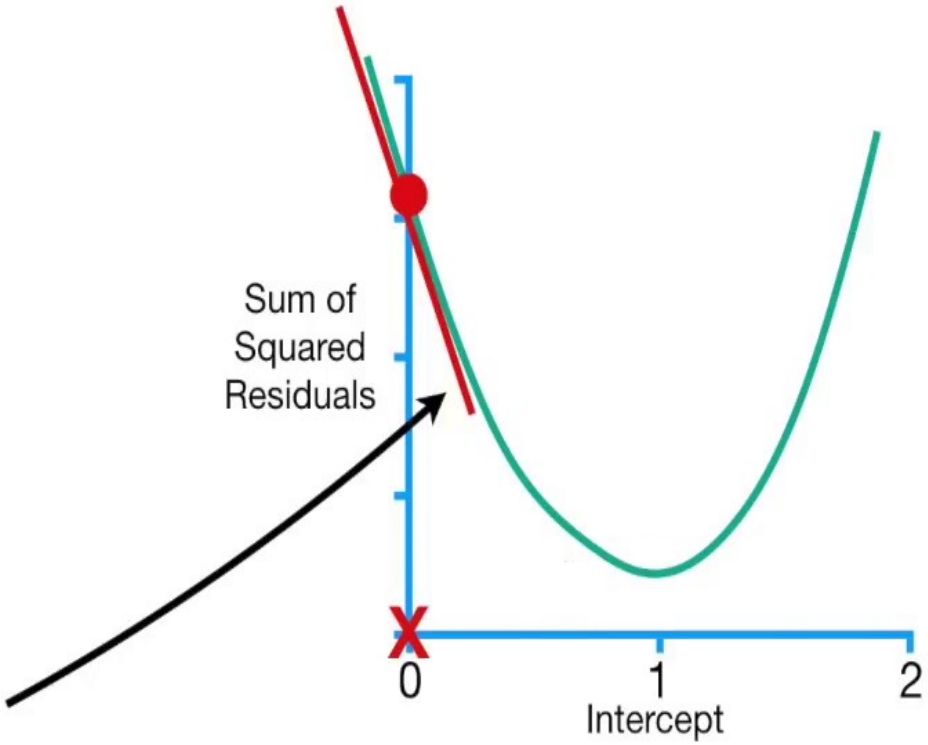
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



...and when the slope is far from **0**...

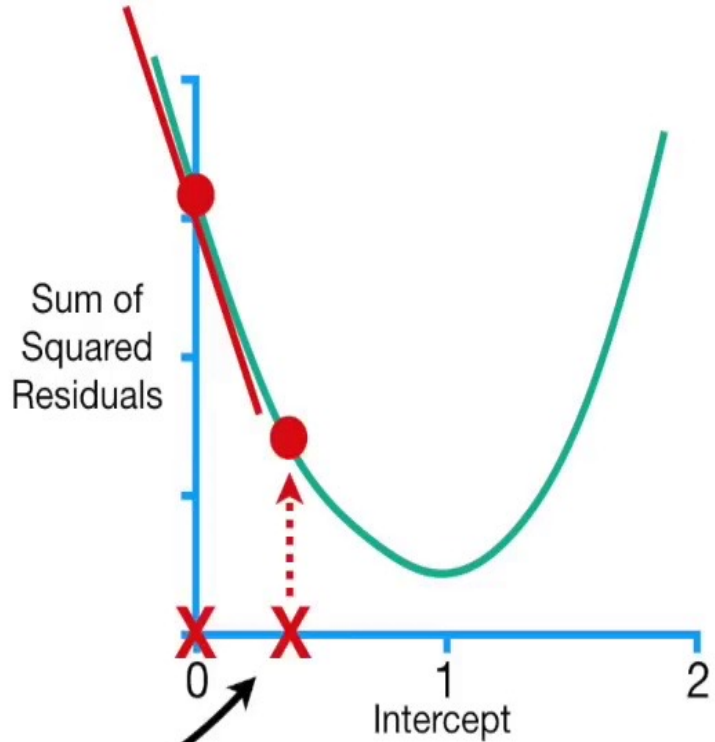
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



…then we should take big steps, because we are far from the optimal value.

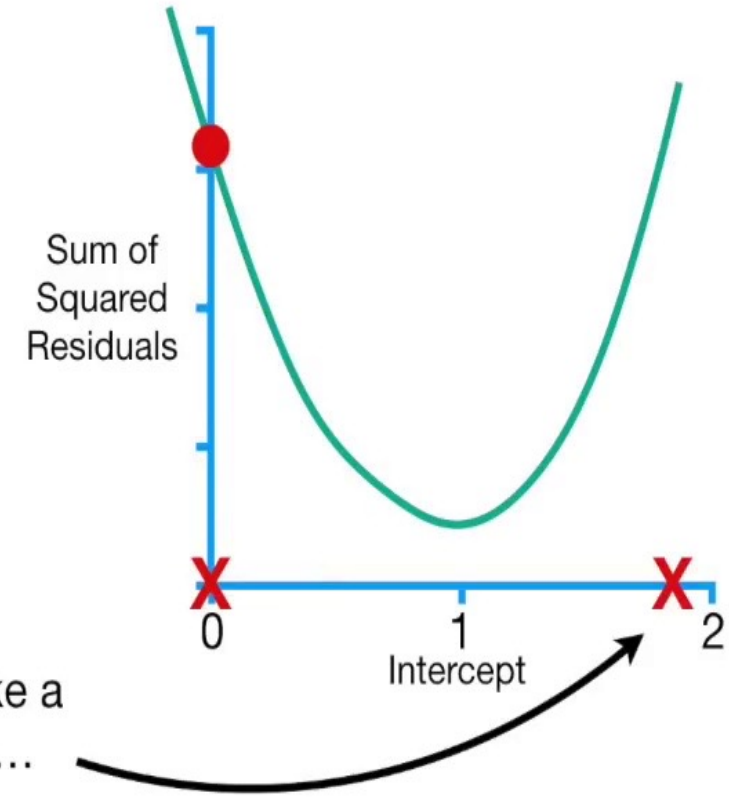
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



However, if we take a super huge step…

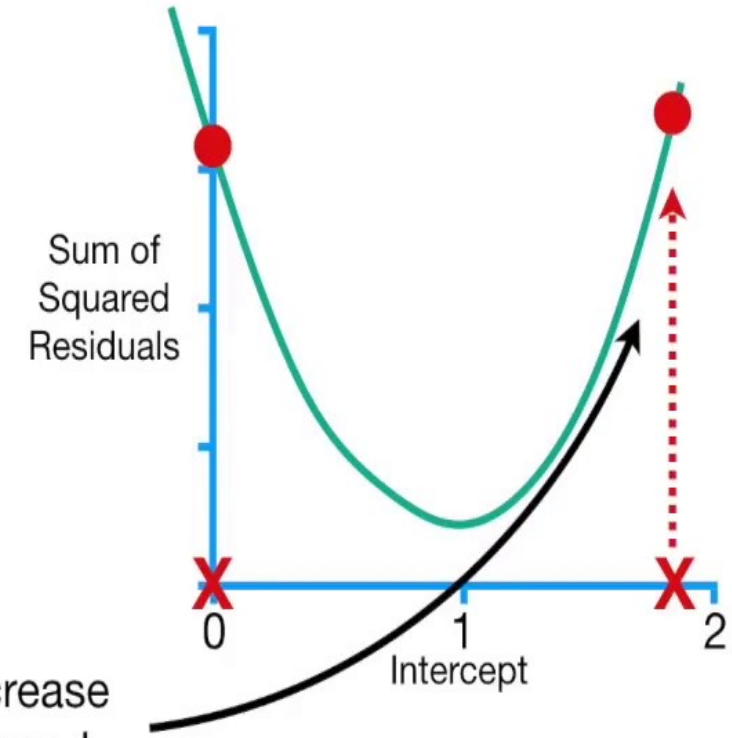
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



...then we would increase the Sum of the Squared Residuals!

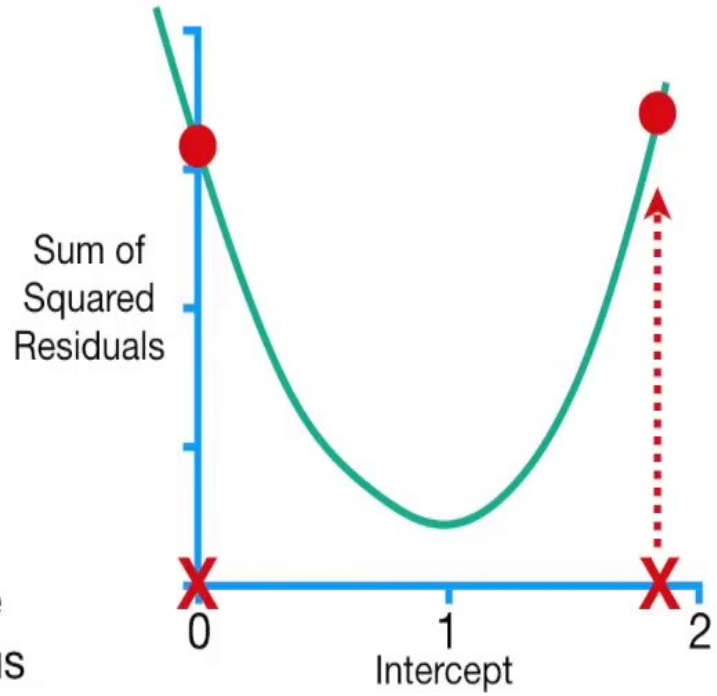
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



So the size of the step should be related to the slope, since it tells us if we should take a baby step or a big step, but we need to make sure the big step is not too big.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

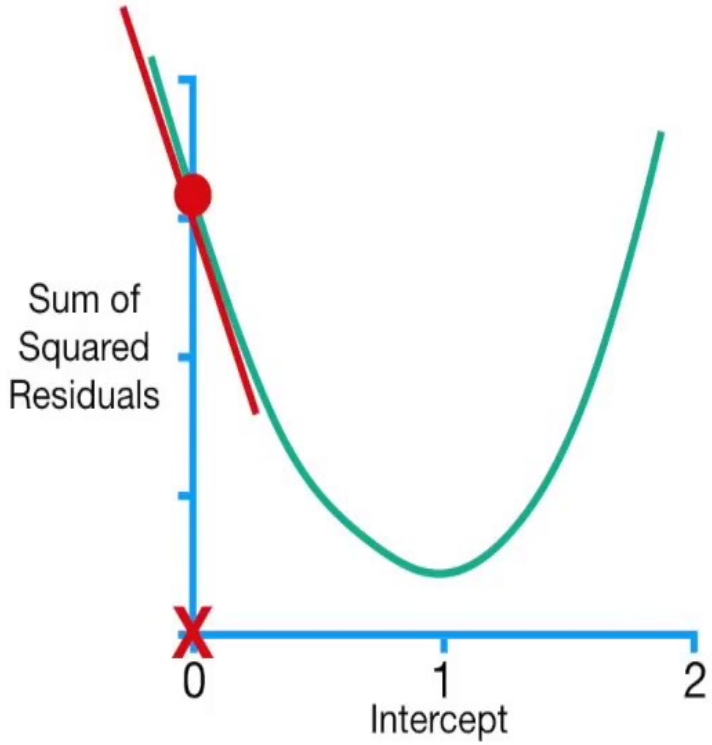
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

Step Size = -5.7

Gradient Descent determines the **Step Size** by multiplying the **slope**…



$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

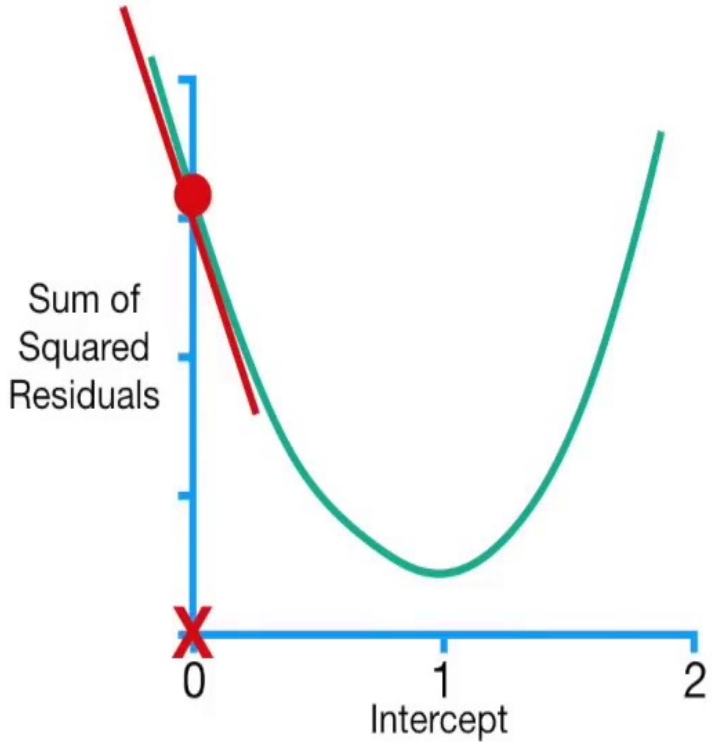
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

Step Size = -5.7 x 0.1

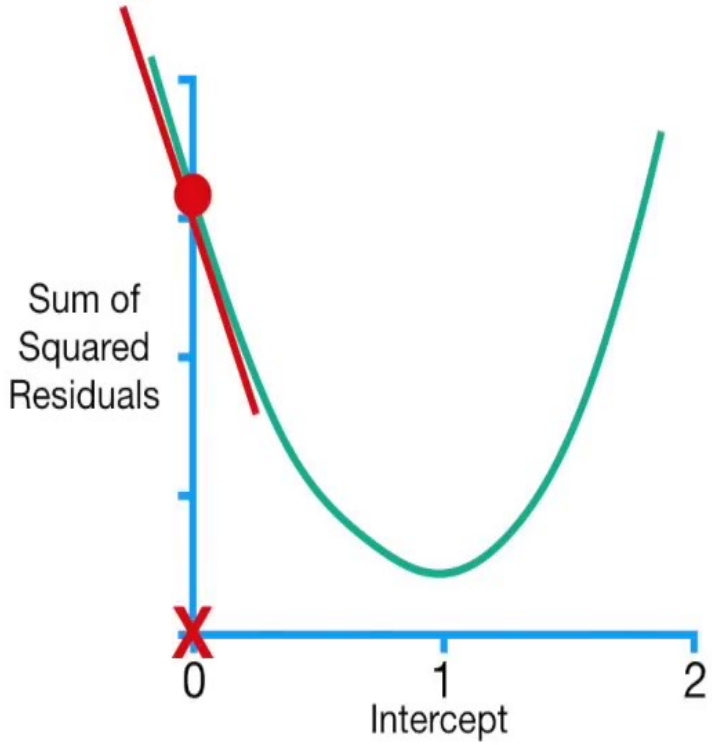
…by a small number called
The Learning Rate.



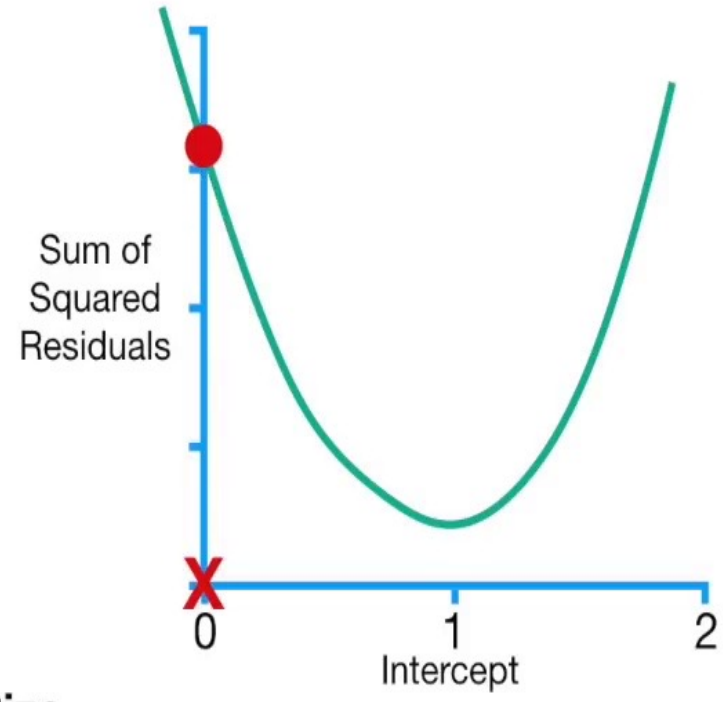
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9})) = -5.7$$

Step Size = -5.7 x 0.1 = -0.57

When the **Intercept** = **0**, the **Step Size** = **-0.57**.



$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9})) = -5.7$$

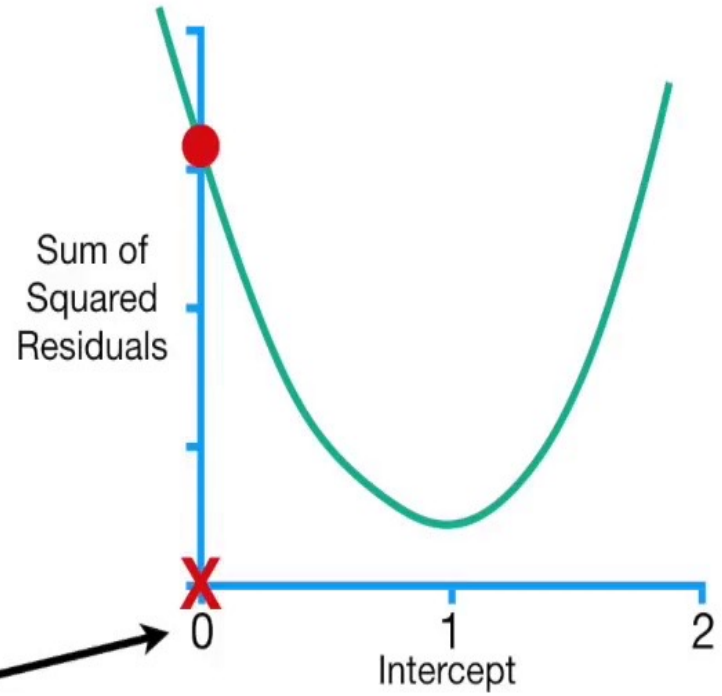


Step Size = -5.7 x 0.1 = -0.57

New Intercept = ← With the **Step Size**, we can calculate a **New Intercept**.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9})) = -5.7$$

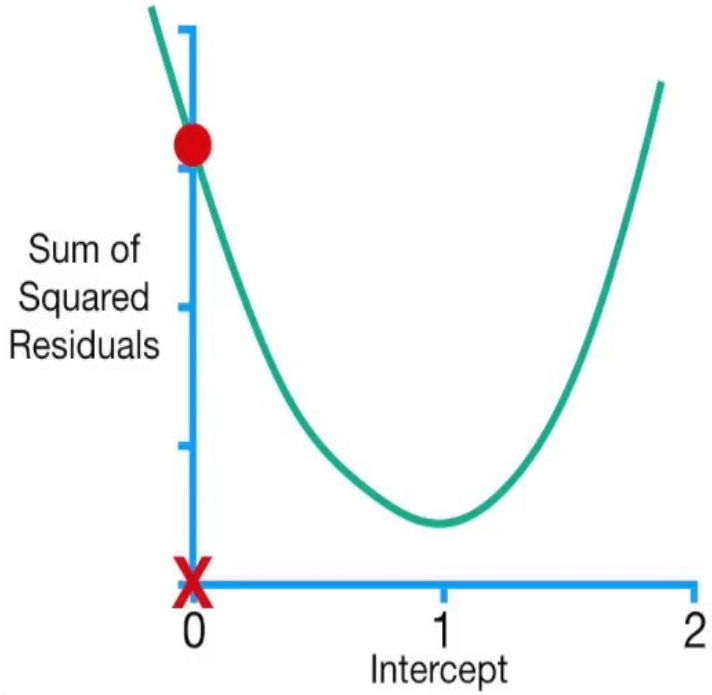
Step Size = -5.7 x 0.1 = -0.57



New Intercept = **Old Intercept** - Step Size

The **New Intercept** is
the **Old Intercept**...

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9})) = -5.7$$



Step Size = -5.7 x 0.1 = -0.57

New Intercept = **Old Intercept** - **Step Size**

…minus the **Step Size**.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

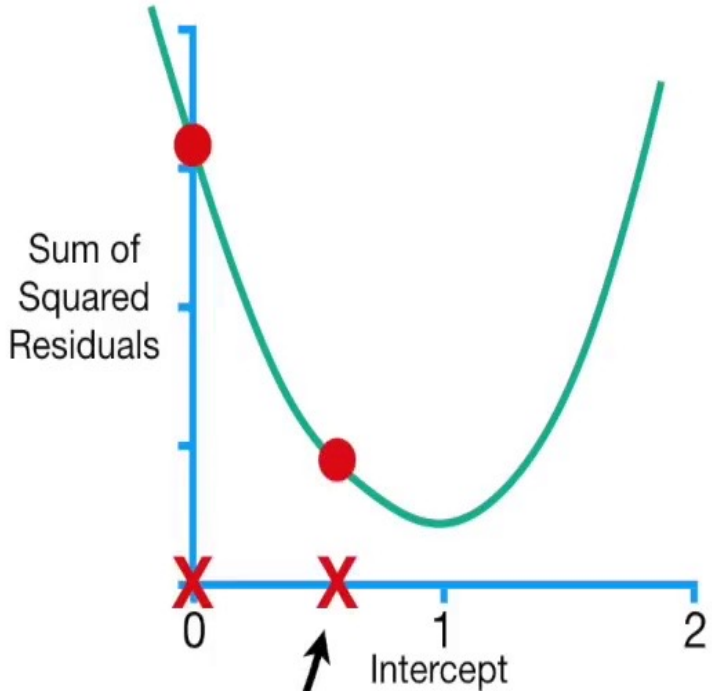
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

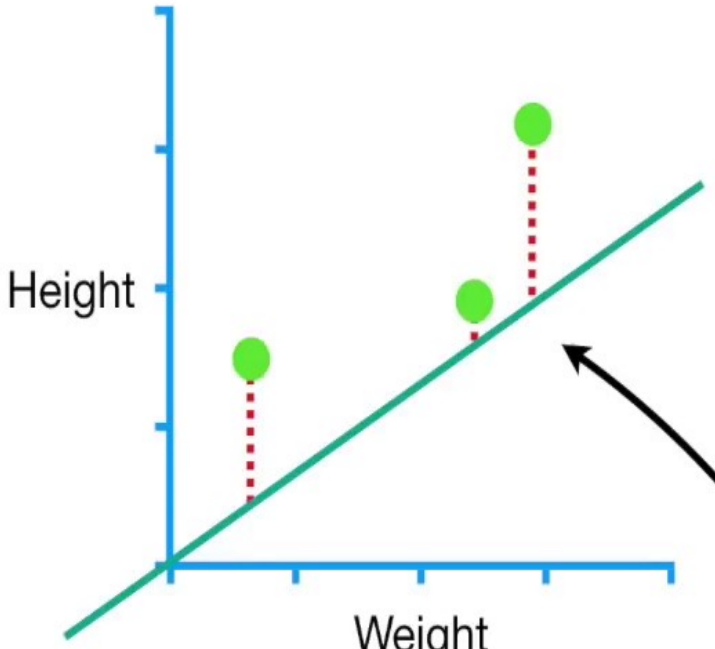
$$= -5.7$$

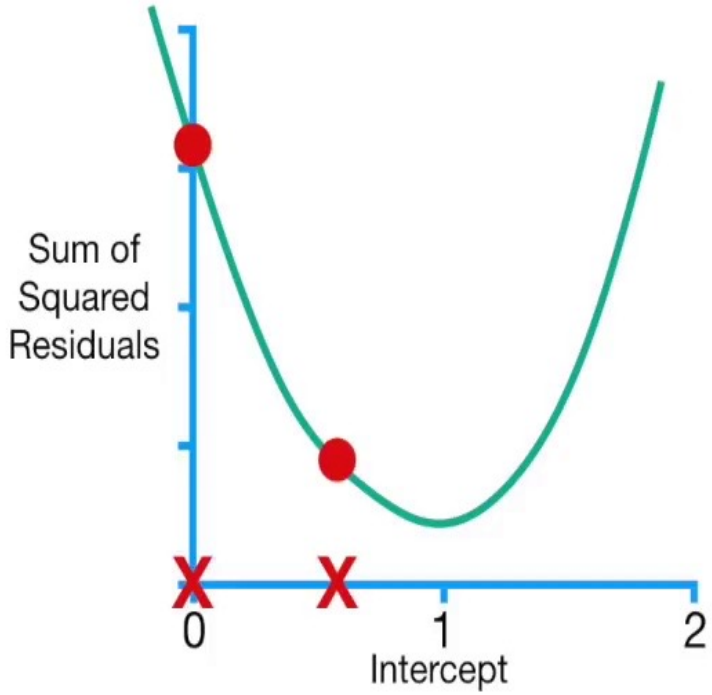
Step Size = -5.7 x 0.1 = -0.57



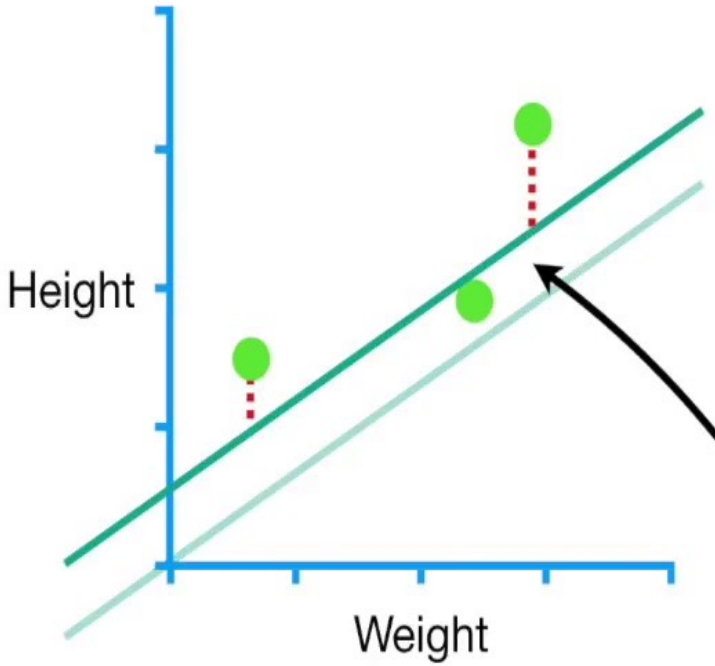
New Intercept = 0 - (-0.57) = **0.57**

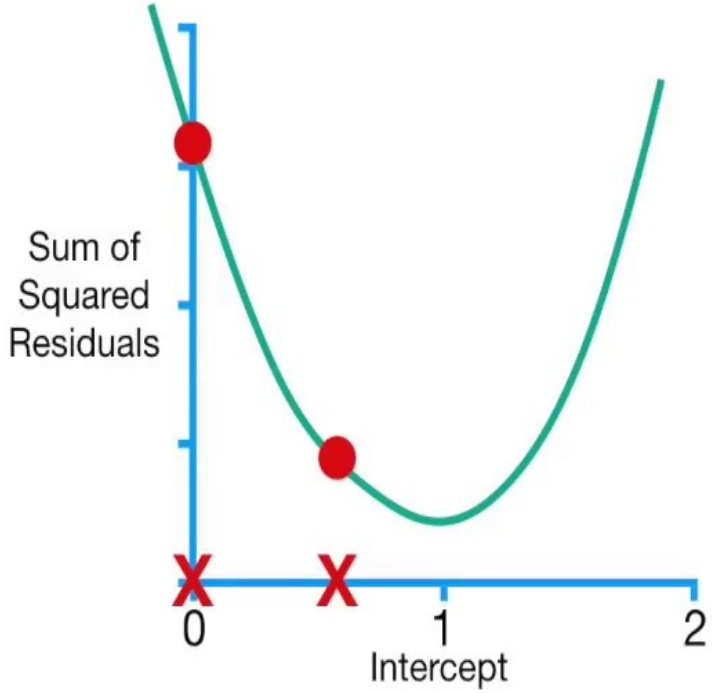
…and the the **New Intercept** = **0.57**.





Going back to the original data and the original line, with the **Intercept** = **0**...





…we can see how much the residuals shrink when the **Intercept** = **0.57**.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

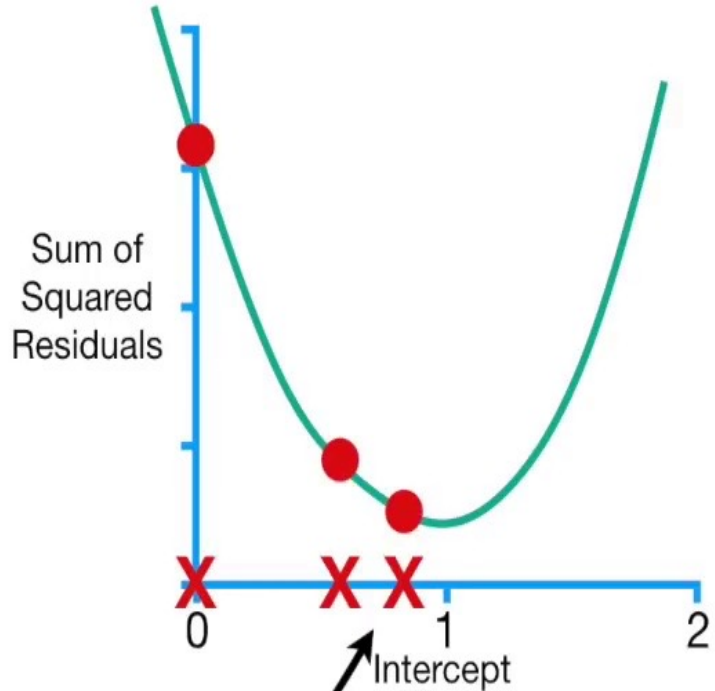
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0.57 + 0.64 \times \mathbf{2.9}))$$

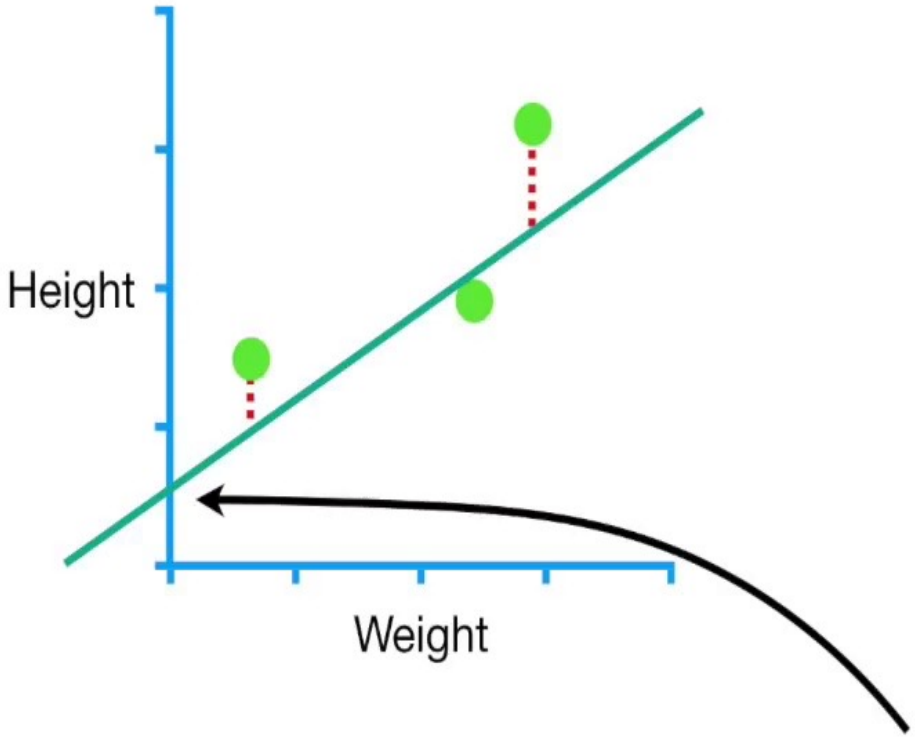
$$= -2.3$$

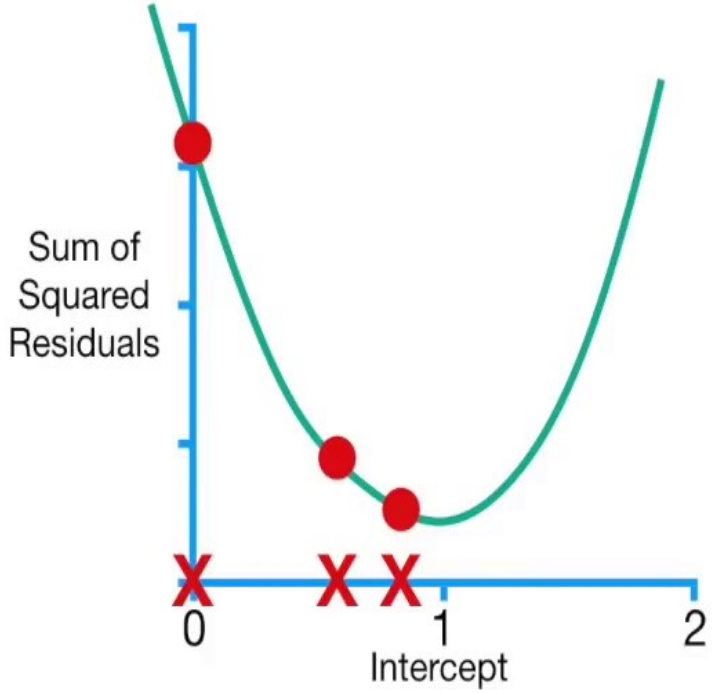
Step Size = $-2.3 \times 0.1 = \mathbf{-0.23}$



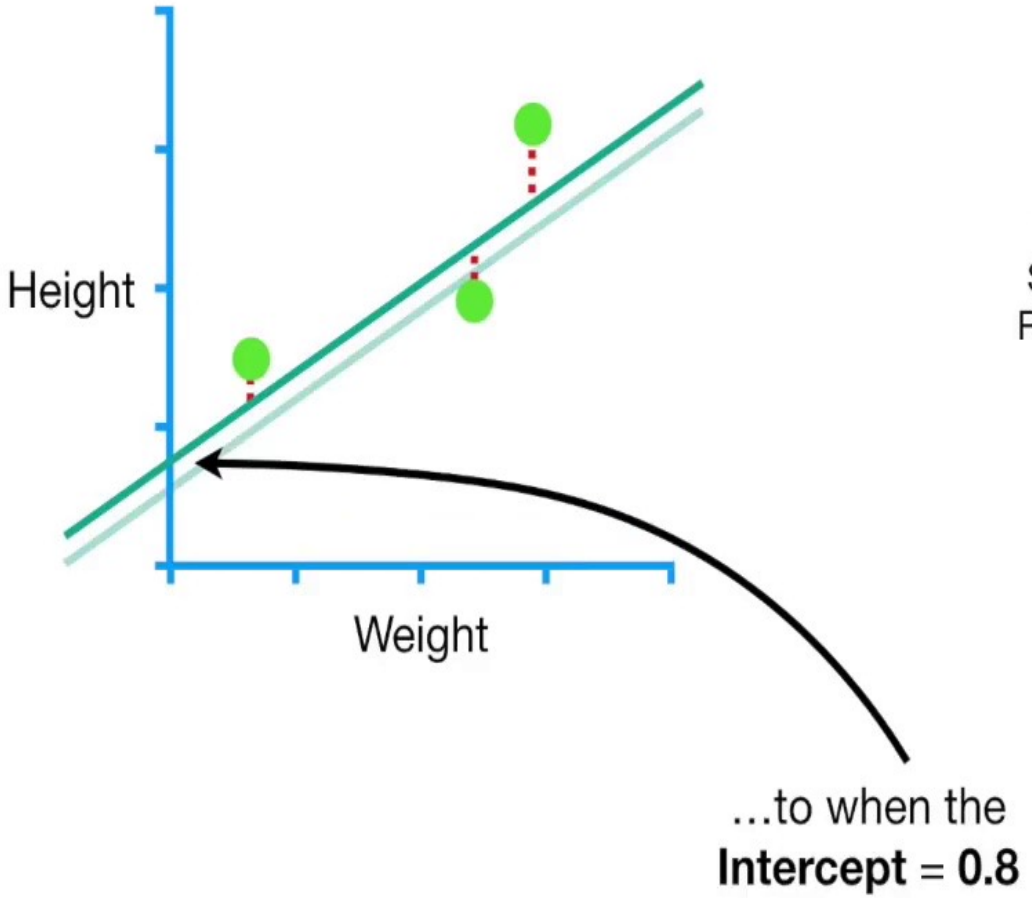
New Intercept = 0.57 - (-0.23) = **0.8**

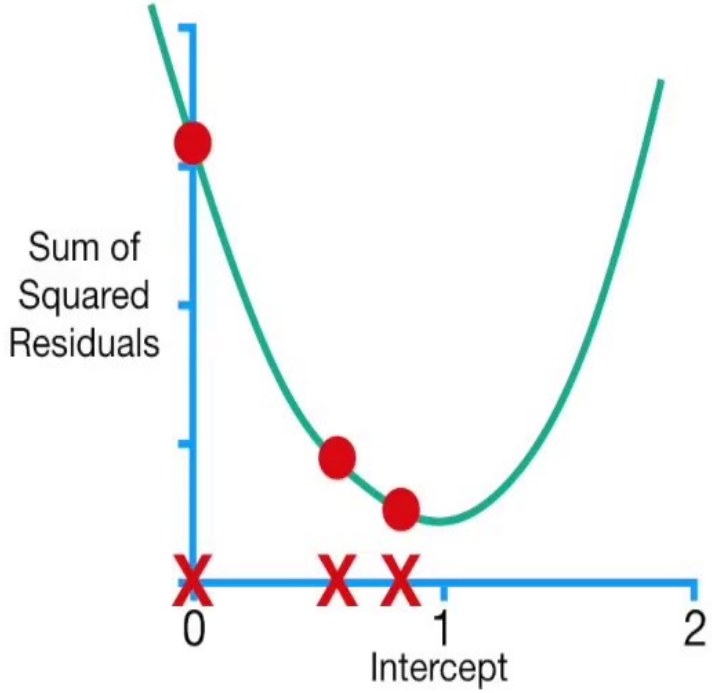
…and the **New Intercept** = **0.8**



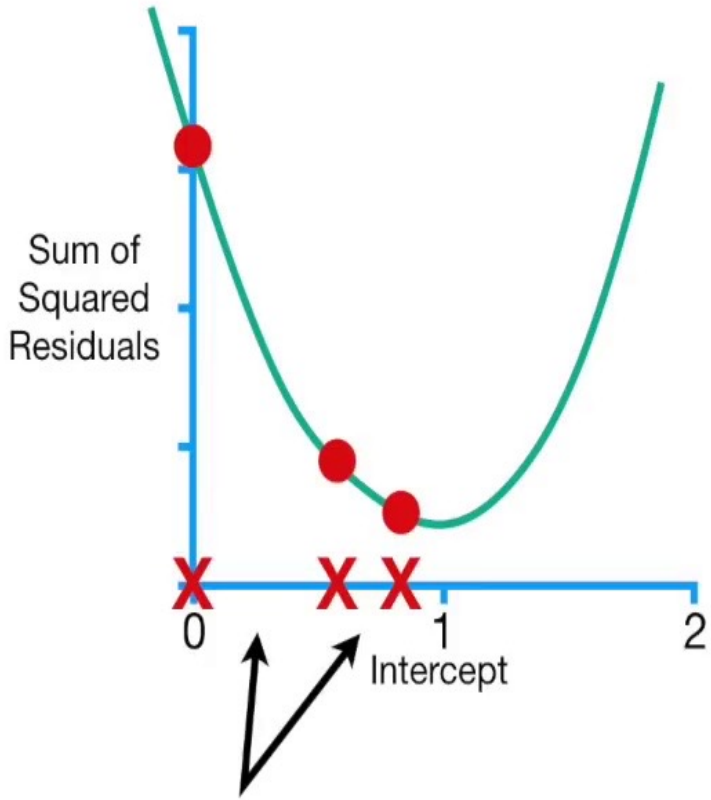


Now we can compare the residuals when the **Intercept** = **0.57**…

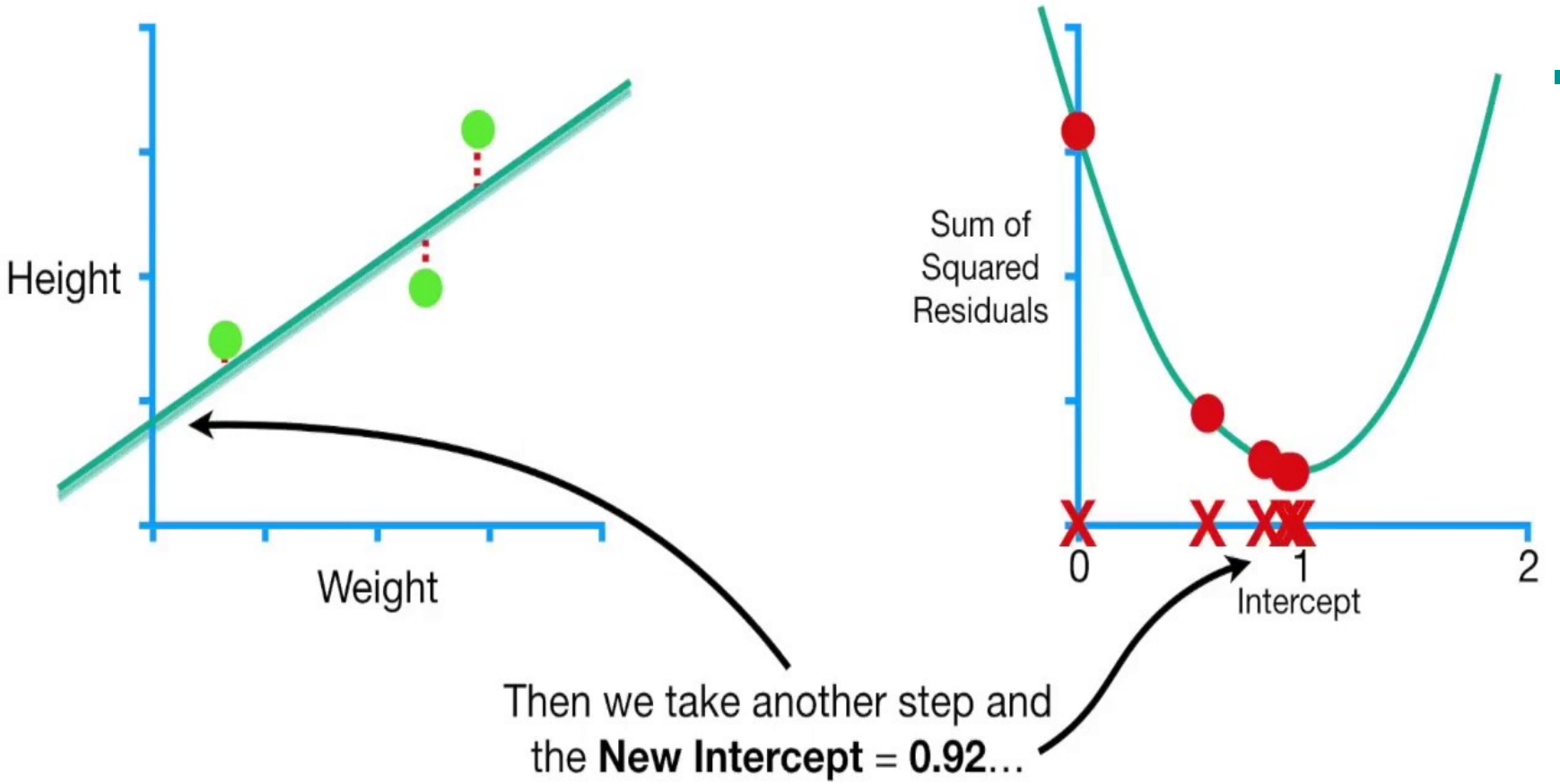


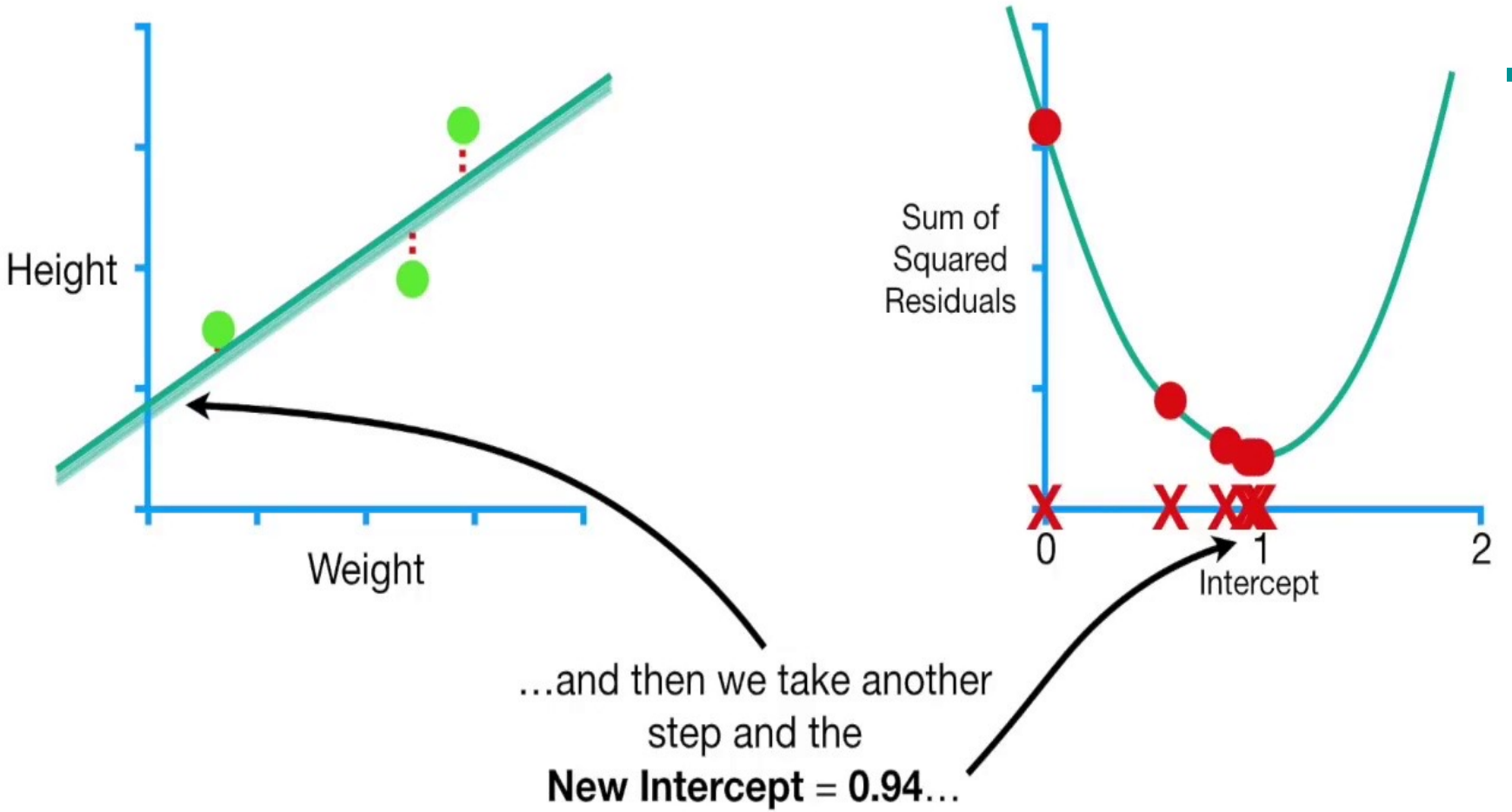


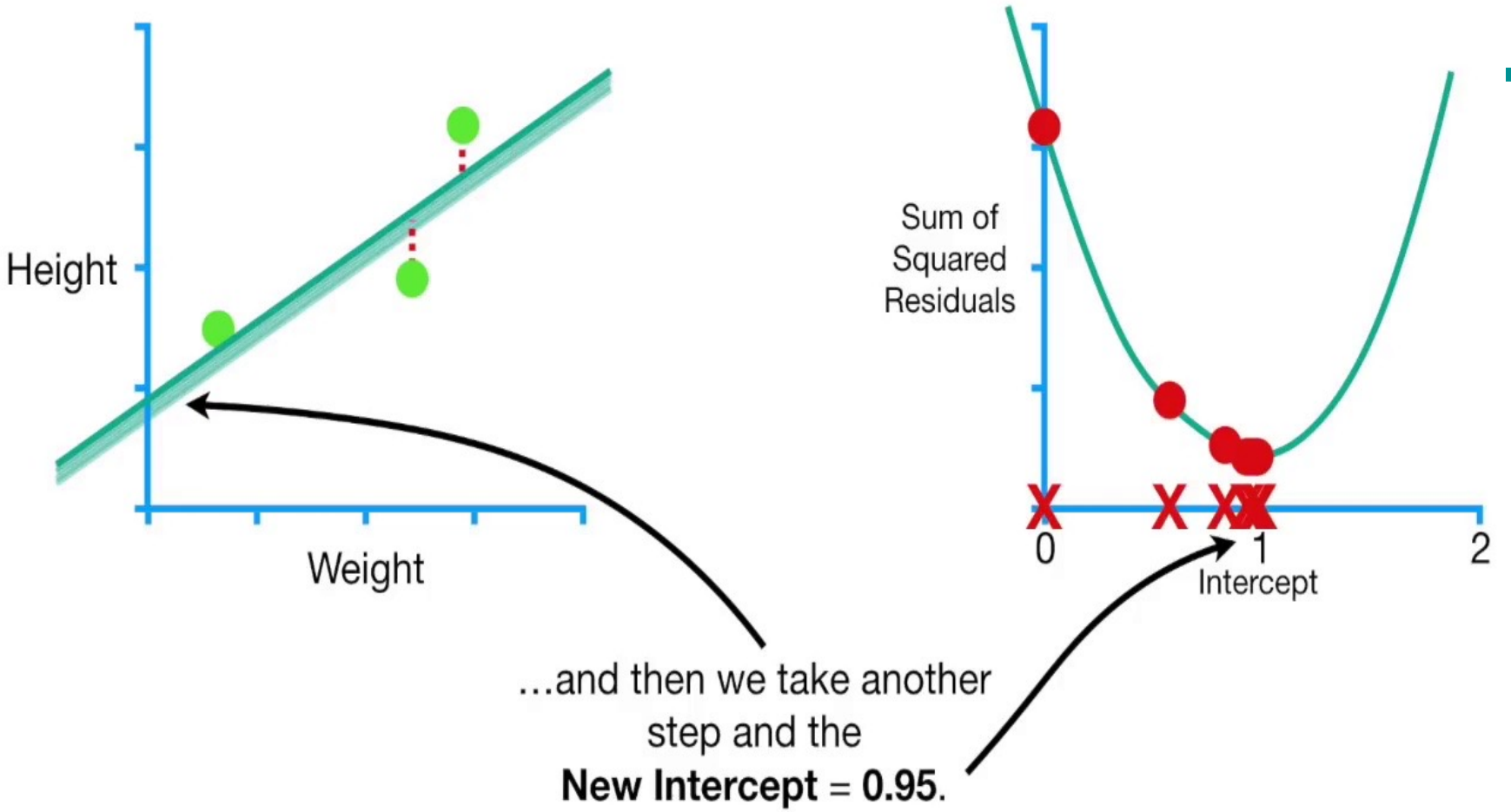
…to when the
Intercept = 0.8



Notice that the first step was relatively large compared to the second step.

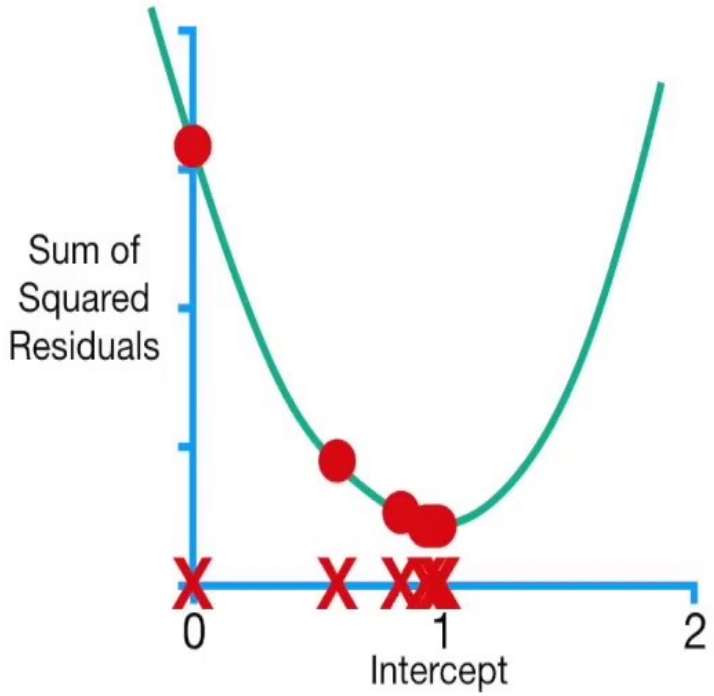

Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept
Then we take another step and
the **New Intercept** = **0.92**...


Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept
...and then we take another
step and the
New Intercept = 0.94...


Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept
...and then we take another
step and the
New Intercept = 0.95.

GD to find *b*

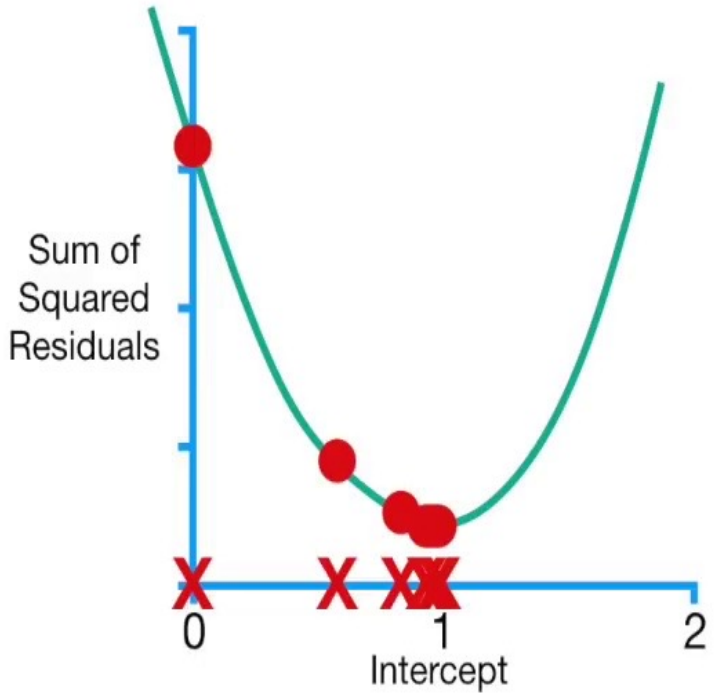
After 6 steps, the **Gradient Descent** estimate for the **Intercept** is **0.95**.



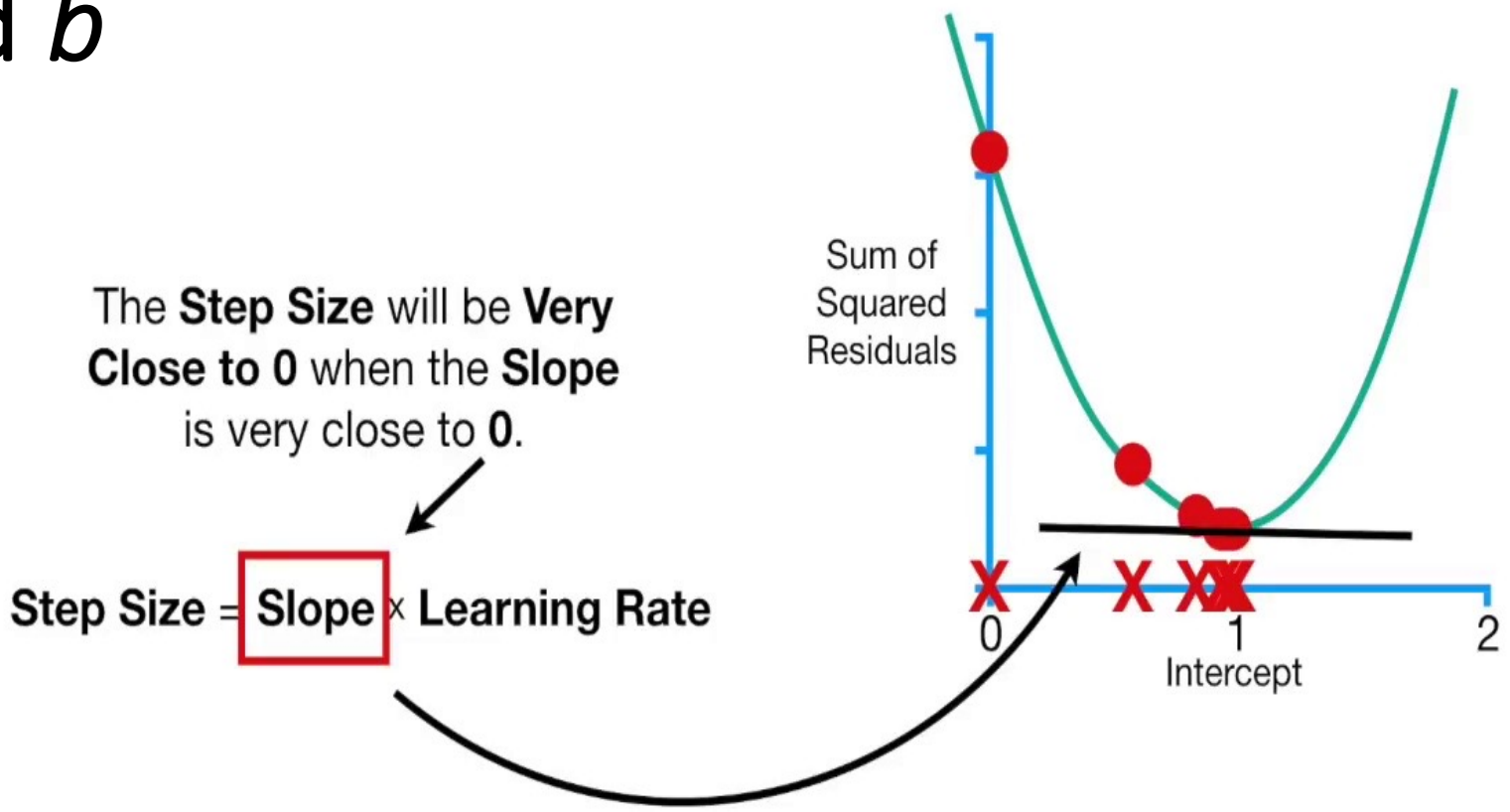
GD to find *b*

Gradient Descent stops when the **Step Size** is **Very Close To 0.**

Step Size = **Slope** x **Learning Rate**



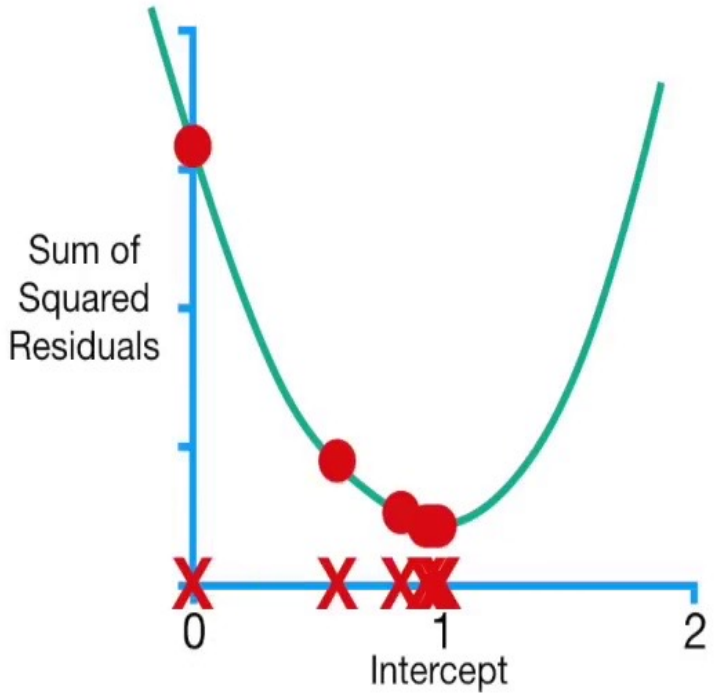
GD to find *b*


The **Step Size** will be **Very Close to 0** when the **Slope** is very close to **0**.
Step Size = **Slope** x **Learning Rate**
Sum of Squared Residuals
0
1
2
Intercept

GD to find *b*

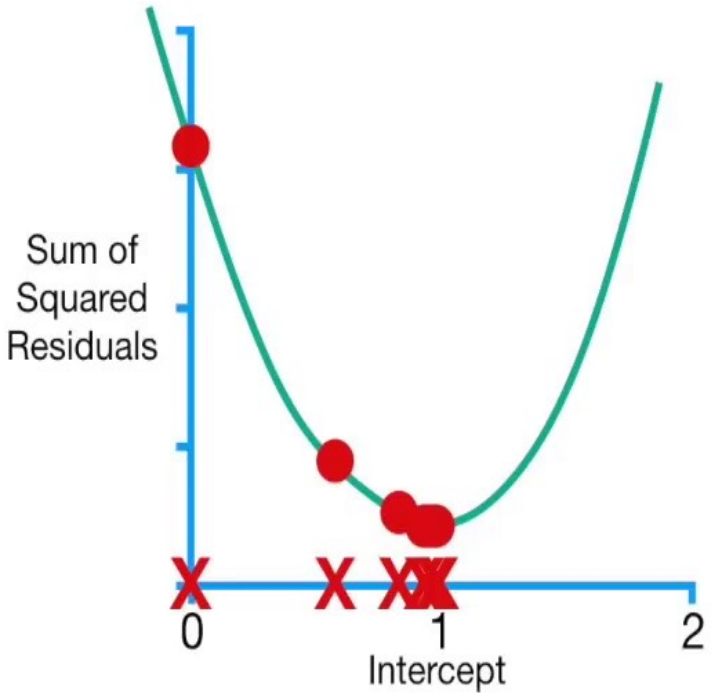
In practice, the
Minimum Step Size = **0.001**
or smaller.

Step Size = **Slope** x **Learning Rate**



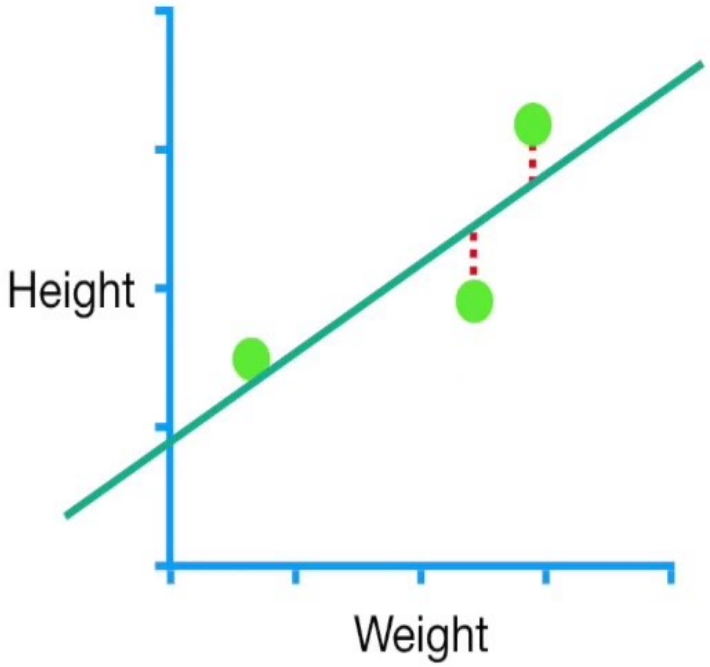
GD to find *b*

That said, **Gradient Descent** also includes a limit on the number of steps it will take before giving up.



GD for *m, b*

Predicted Height = intercept + 0.64 x **Weight**

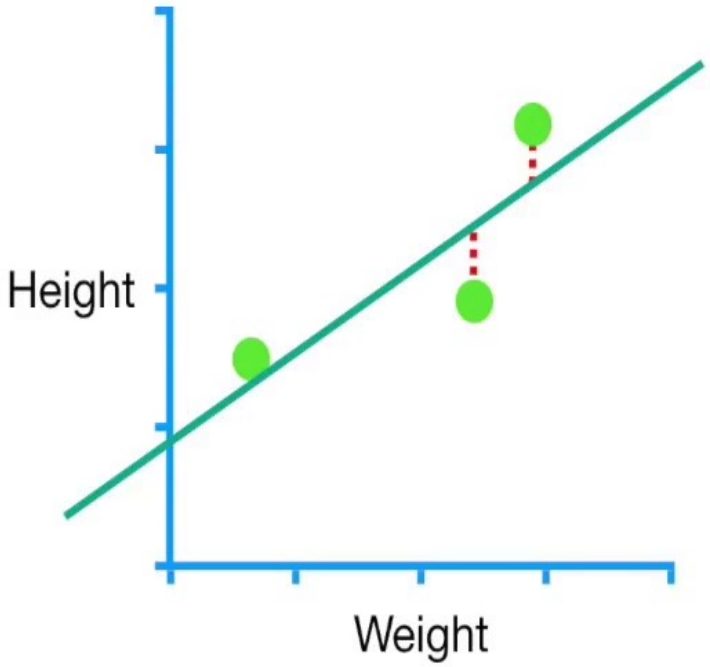


Now that we understand how **Gradient Descent** can estimate the **Intercept**...

GD for *m, b*

Predicted Height = intercept + slope x **Weight**

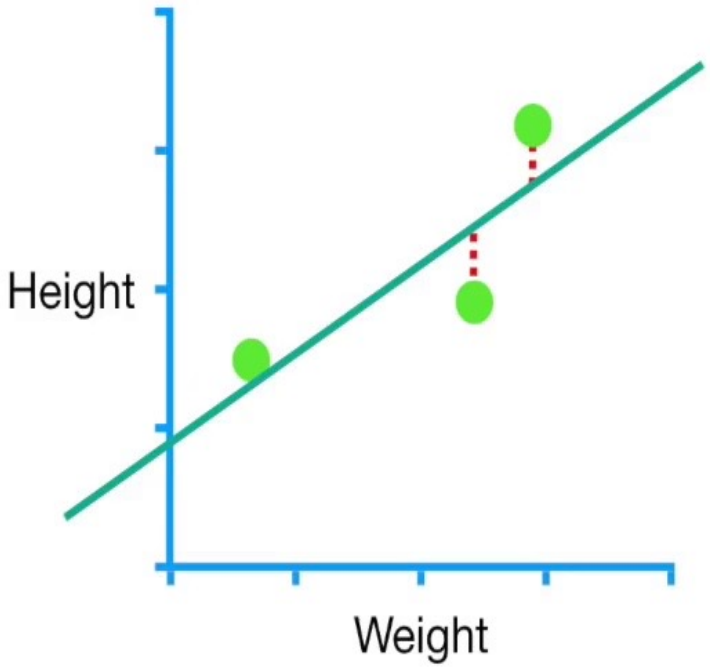
…let's talk abut how to estimate the **Intercept** *and* the **Slope**.



Sum of squared residuals = (**1.4** - (intercept + slope x **0.5**))^2

+ (**1.9** - (intercept + slope x **2.3**))^2

+ (**3.2** - (intercept + slope x **2.9**))^2



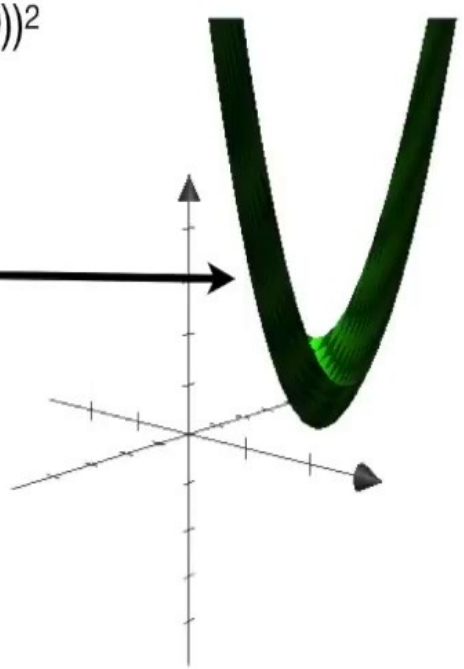
Just like before, we will use the Sum of the Squared Residuals as the **Loss Function**

Sum of squared residuals = (**1.4** - (intercept + slope x **0.5**))^2

+ (**1.9** - (intercept + slope x **2.3**))^2

+ (**3.2** - (intercept + slope x **2.9**))^2

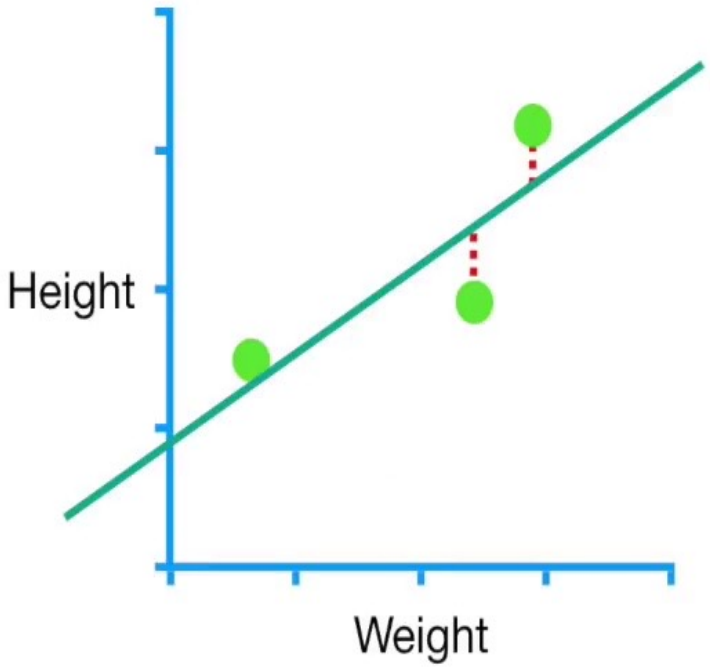
This is a 3-D graph of the **Loss Function** for different values for the **Intercept** and the **Slope**



Sum of squared residuals = (**1.4** - (intercept + slope x **0.5**))^2

+ (**1.9** - (intercept + slope x **2.3**))^2

+ (**3.2** - (intercept + slope x **2.9**))^2

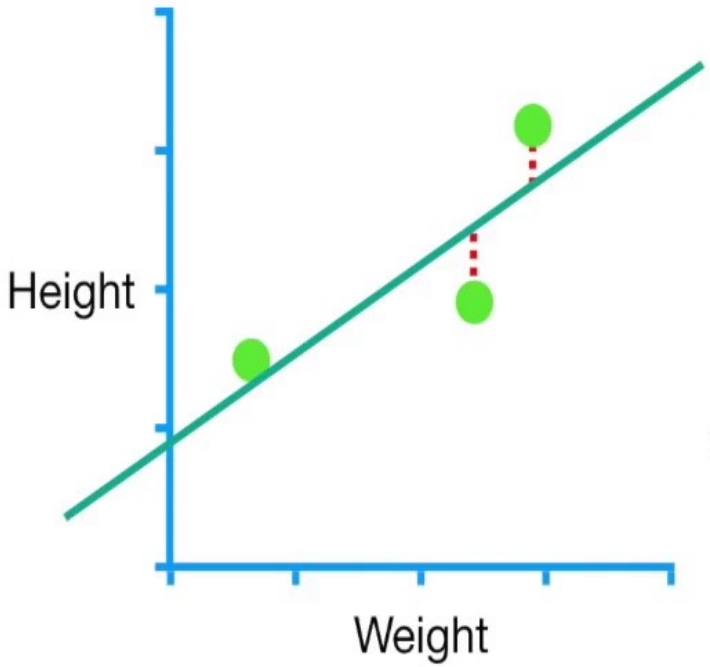


So, just like before, we need to take the derivative of this function…

Sum of squared residuals = (**1.4** - (intercept + slope x **0.5**))^2

+ (**1.9** - (intercept + slope x **2.3**))^2

+ (**3.2** - (intercept + slope x **2.9**))^2



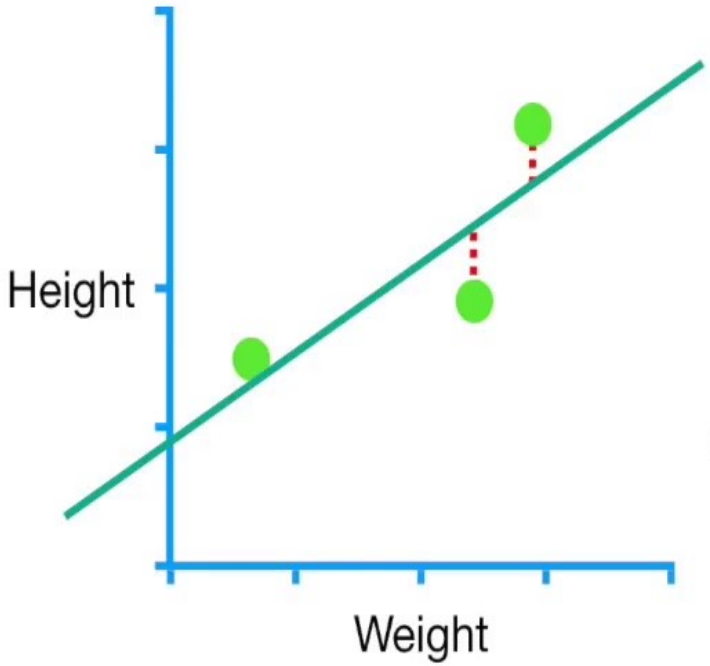
…and just like before, we'll take the derivative with respect to the **Intercept**…

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals}$$

Sum of squared residuals = (**1.4** - (intercept + slope × **0.5**))^2

+ (**1.9** - (intercept + slope × **2.3**))^2

+ (**3.2** - (intercept + slope × **2.9**))^2



…but unlike before, we'll also take the derivative with respect to the **Slope**!

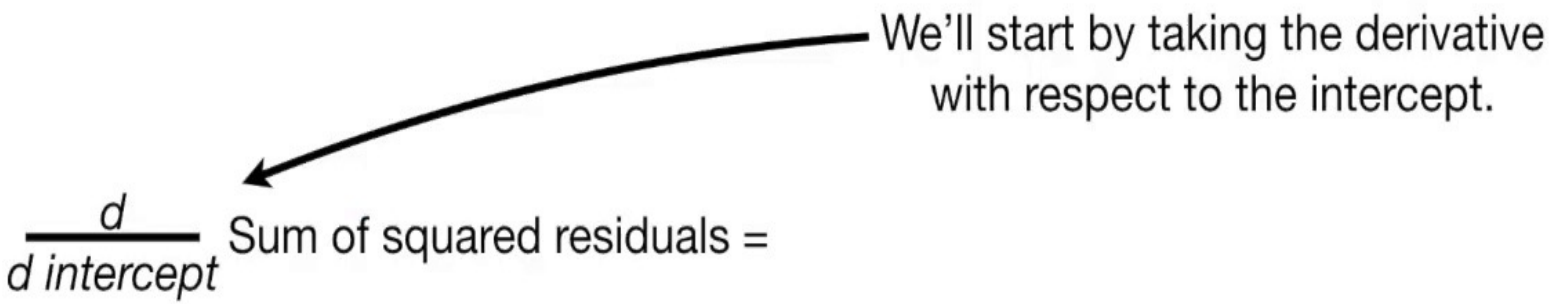
$$\frac{d}{d\,intercept}\text{Sum of squared residuals}$$

$$\frac{d}{d\,slope}\text{Sum of squared residuals}$$

Sum of squared residuals = (**1.4** - (intercept + slope x **0.5**))^2

+ (**1.9** - (intercept + slope x **2.3**))^2

+ (**3.2** - (intercept + slope x **2.9**))^2



$$\text{Sum of squared residuals} = (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2 + (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2 + (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

Just like before, we take the derivative of each part…

$$\frac{d}{d\ \text{intercept}} \text{Sum of squared residuals} = \frac{d}{d\ \text{intercept}} (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2$$

Sum of squared residuals = $(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2$

$+ (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$

$+ (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$

Just like before, we take the derivative of each part…

$$\frac{d}{d\,\text{intercept}} \text{Sum of squared residuals} = \frac{d}{d\,\text{intercept}} (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\,\text{intercept}} (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\,\text{intercept}} (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$\frac{d}{d\ intercept}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5})) \times -1$$

$$= -2(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))$$

…and this…

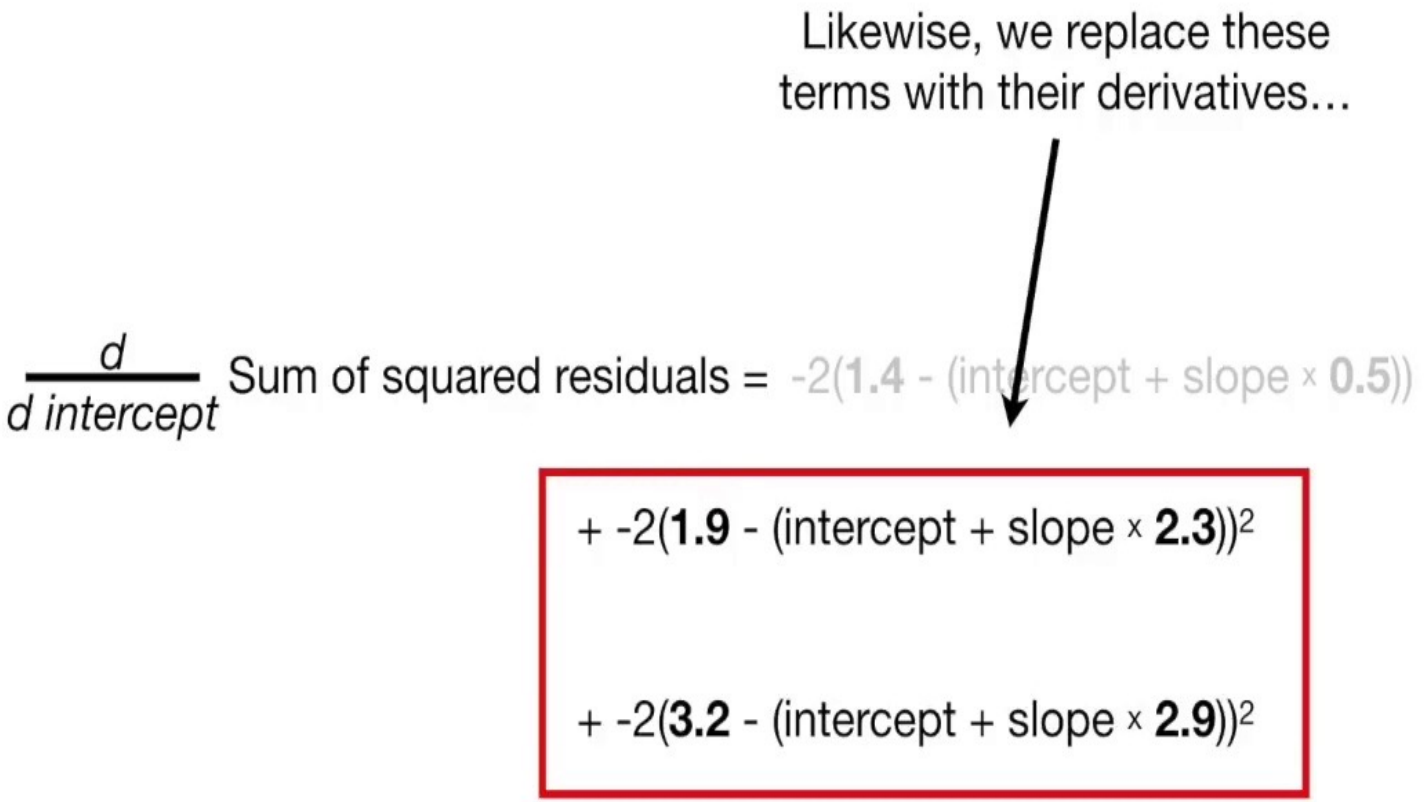
…is the derivative of the first part…

$$\frac{d}{d\ intercept} Sum\ of\ squared\ residuals = \frac{d}{d\ intercept}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\ intercept}(\mathbf{1.9} - (intercept + slope \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\ intercept}(\mathbf{3.2} - (intercept + slope \times \mathbf{2.9}))^2$$

GD for *m, b*


Likewise, we replace these terms with their derivatives…
$\frac{d}{d\ intercept}$ Sum of squared residuals = -2(**1.4** - (intercept + slope × **0.5**))
+ -2(**1.9** - (intercept + slope × **2.3**))²
+ -2(**3.2** - (intercept + slope × **2.9**))²

Sum of squared residuals = (**1.4** - (intercept + slope x **0.5**))^2

+ (**1.9** - (intercept + slope x **2.3**))^2

+ (**3.2** - (intercept + slope x **2.9**))^2

Now let's take the derivative of the Sum of the Squared Residuals with respect to the **Slope**.

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

$$\text{Sum of squared residuals} = (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2 + (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2 + (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

Just like before, we take the derivative of each part…

$$\frac{d}{d\,\text{slope}} \text{Sum of squared residuals} = \frac{d}{d\,\text{slope}} (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2$$

Sum of squared residuals = $(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2$

$+ (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$

$+ (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$

Just like before, we take the derivative of each part…

$$\frac{d}{d\,slope}\text{ Sum of squared residuals} = \frac{d}{d\,slope}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\,slope}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\,slope}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$\frac{d}{d\,slope}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5})) \times \mathbf{-0.5}$$

$$= -2 \times \mathbf{0.5}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))$$

…and this…

…is the derivative of the first part…

$$\frac{d}{d\,slope} Sum\ of\ squared\ residuals = \frac{d}{d\,slope}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))^2$$

$$+ \frac{d}{d\,slope}(\mathbf{1.9} - (intercept + slope \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\,slope}(\mathbf{3.2} - (intercept + slope \times \mathbf{2.9}))^2$$

$$\frac{d}{d\,slope}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5})) \times \mathbf{-0.5}$$

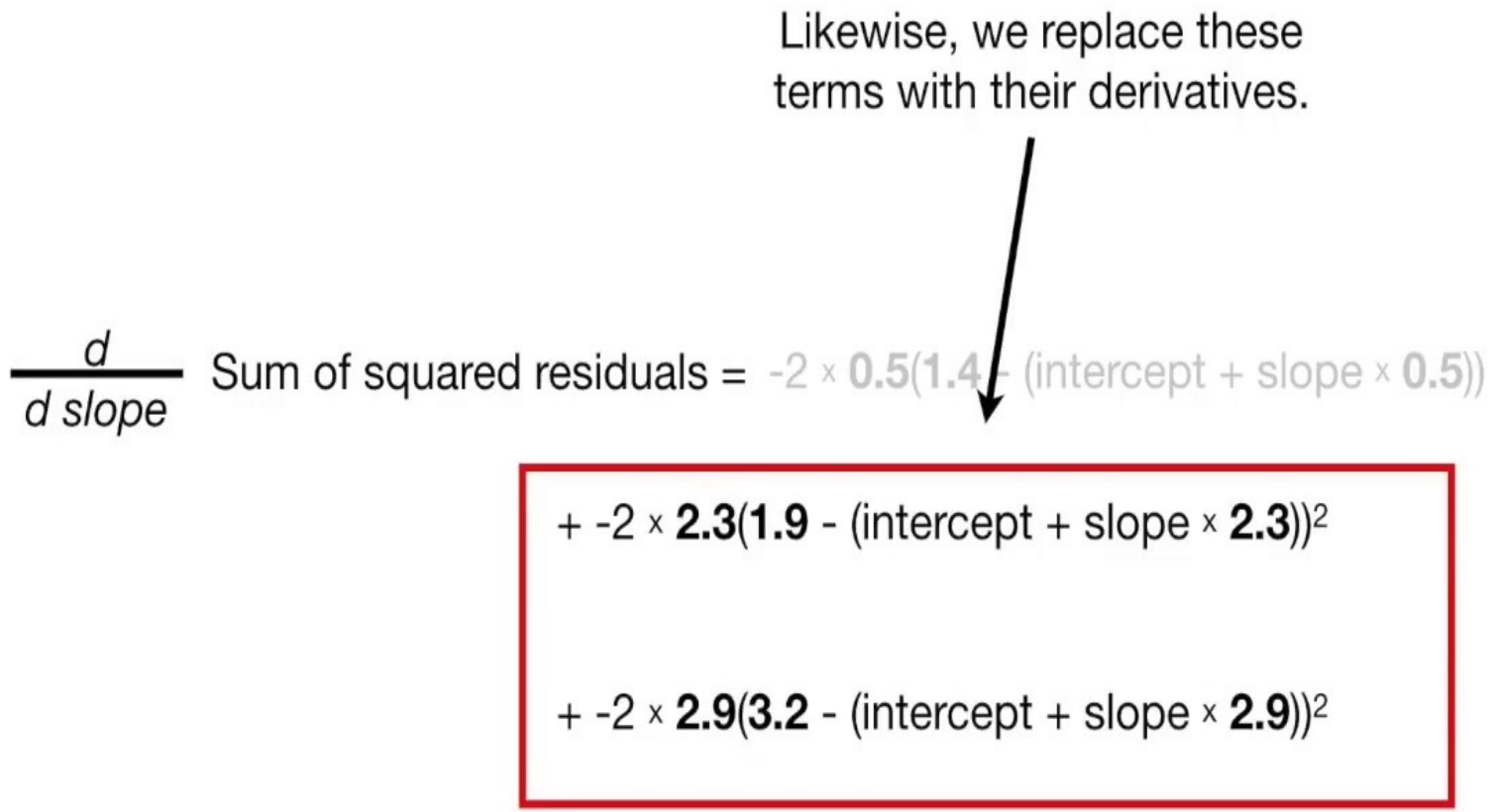
$$= -2 \times \mathbf{0.5}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))$$

…so we plug it in.

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))$$

$$+ \frac{d}{d\,slope}(\mathbf{1.9} - (intercept + slope \times \mathbf{2.3}))^2$$

$$+ \frac{d}{d\,slope}(\mathbf{3.2} - (intercept + slope \times \mathbf{2.9}))^2$$


Likewise, we replace these terms with their derivatives.
d / d slope
Sum of squared residuals =
-2 x 0.5(1.4 - (intercept + slope x 0.5))
+ -2 x 2.3(1.9 - (intercept + slope x 2.3))²
+ -2 x 2.9(3.2 - (intercept + slope x 2.9))²

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

Here's the derivative of the Sum of the Squared Residuals with respect to the **Intercept**...

...and here's the derivative with respect to the **Slope**.

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

NOTE: When you have two or more derivatives of the same function, they are called a **Gradient**.

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

We will use this **Gradient** to **descend** to lowest point in the **Loss Function**, which, in this case, is the Sum of the Squared Residuals…

…thus, this is why this algorithm is called **Gradient Descent**!

$$\frac{d}{d\,slope}\text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

Just like before, we will start by picking a random number for the **Intercept**. In this case we'll set the **Intercept** = **0**...

...and we'll pick a random number for the **Slope**. In this case we'll set the **Slope** = **1**.

$$\frac{d}{d\,slope}\text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

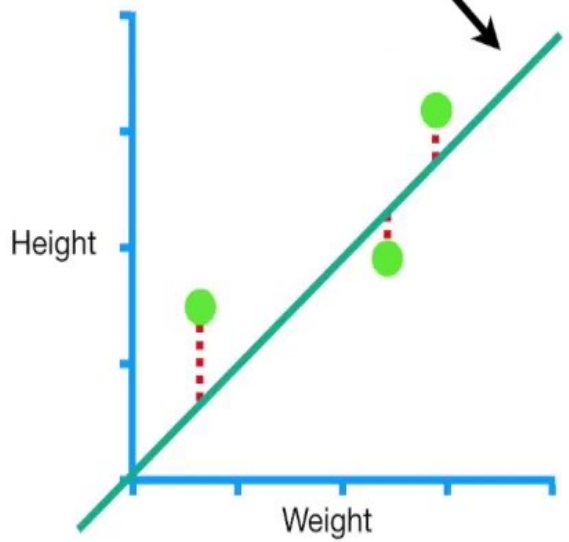
$$\frac{d}{d\,slope}\text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

Thus, this line, with **Intercept** = **0** and **Slope** = **1**, is where we will start.



$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + slope \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + slope \times \mathbf{2.9}))$$

Now let's plug in **0** for the **Intercept** and **1** for the **Slope**...

$$\frac{d}{d\ slope}$$ Sum of squared residuals =

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (intercept + slope \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (intercept + slope \times \mathbf{2.3}))^2$$

$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

…and that gives us two **Slopes**…

$$\frac{d}{d\ slope}$$ Sum of squared residuals =

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

Step Size$_{Intercept}$ = -1.6 x **Learning Rate**

…now we plug the **Slopes** into the **Step Size** formulas…

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})) + -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2 + -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

Step Size$_{Slope}$ = -0.8 x **Learning Rate**

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})) + -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

Step Size$_{Intercept}$ = -1.6 x **Learning Rate**

…and multiply by the **Learning Rate**, which this time we set to **0.01**…

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})) + -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2 + -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

Step Size$_{Slope}$ = -0.8 x **Learning Rate**

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})) + -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

$$\mathbf{Step\ Size}_{Intercept} = -1.6 \times 0.01 = \mathbf{-0.016}$$

$$\mathbf{New\ Intercept} = \mathbf{Old\ Intercept} - \mathbf{Step\ Size}$$

Now we calculate the **New Intercept** and **New Slope** by plugging in the **Old Intercept** and the **Old Slope**…

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})) + -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2 + -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

$$\mathbf{Step\ Size}_{Slope} = -0.8 \times 0.01 = \mathbf{-0.008}$$

$$\mathbf{New\ Slope} = \mathbf{Old\ Slope} - \mathbf{Step\ Size}$$

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

$\mathbf{Step\ Size}_{Intercept} = -1.6 \times 0.01 = \mathbf{-0.016}$

New Intercept = 0 - (-0.016) = 0.016

…and we end up with a **New Intercept** and a **New Slope**.

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

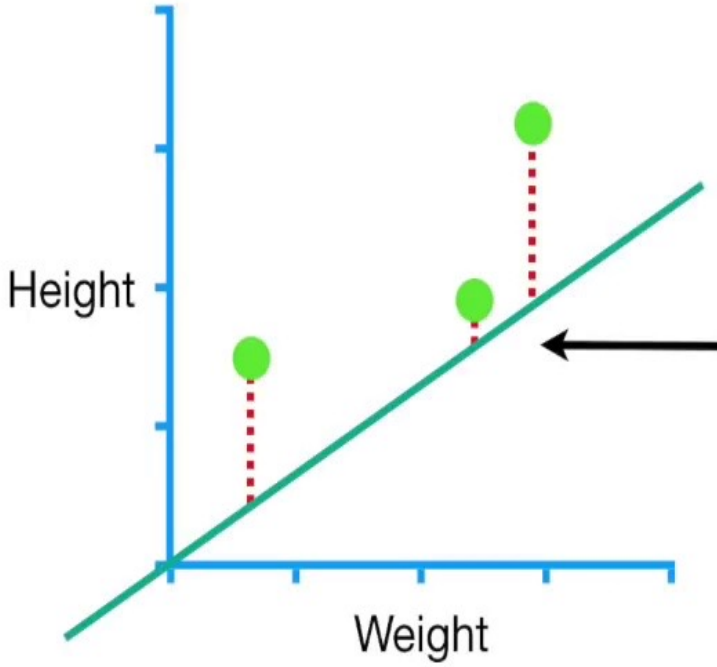
$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

$\mathbf{Step\ Size}_{Slope} = -0.8 \times 0.01 = \mathbf{-0.008}$

New Slope = 1 - (-0.008) = 1.008

GD for *m, b*

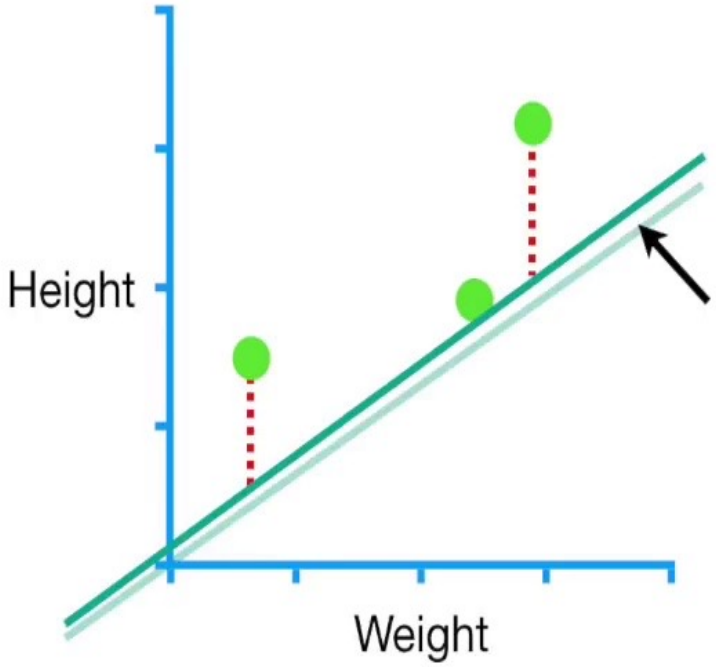


New Intercept = 0 - (-0.016) = 0.016

This is the line we started with… (**Slope = 1** and **Intercept = 0**)

New Slope = 1 - (-0.008) = 1.008

GD for *m, b*

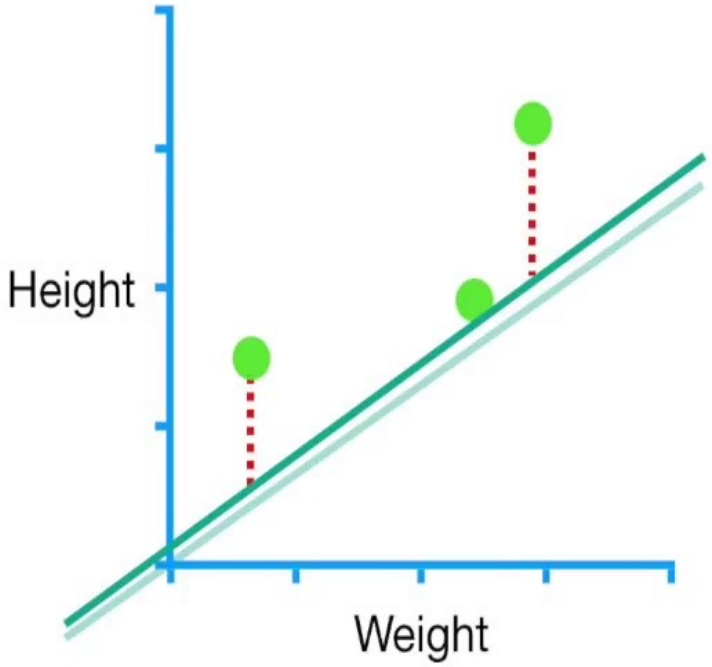


New Intercept = 0 - (-0.016) = 0.016

...and this is the new line (with **Slope = 1.008** and **Intercept = 0.016**) after the first step.

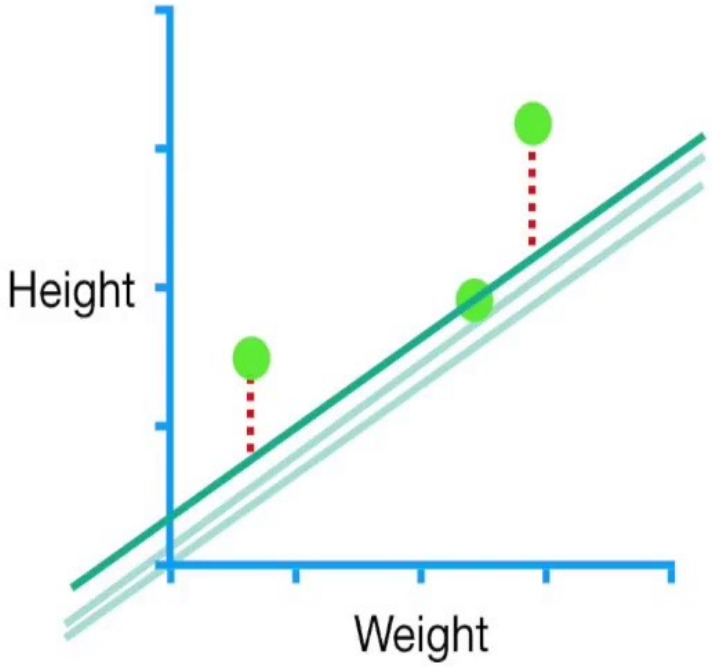
New Slope = 1 - (-0.008) = 1.008

GD for *m*, *b*



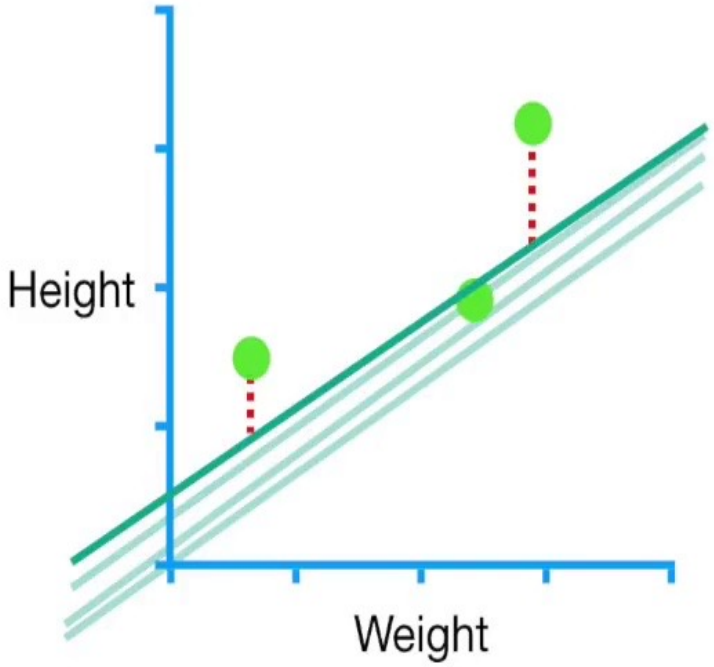
Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

GD for *m*, *b*



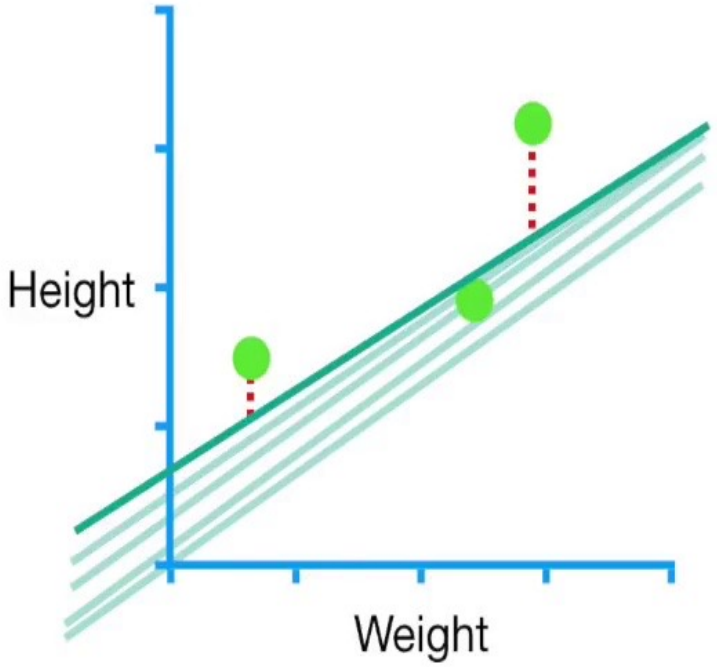
Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

GD for *m*, *b*



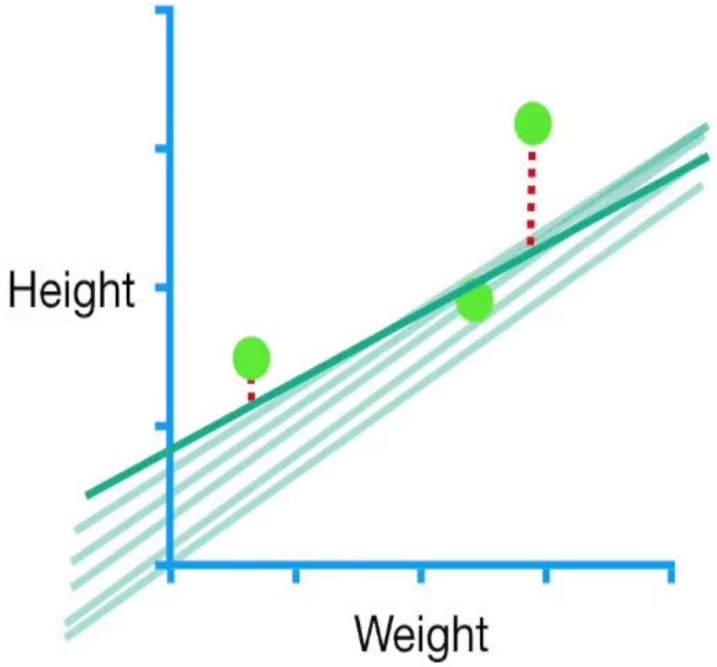
Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

GD for *m*, *b*



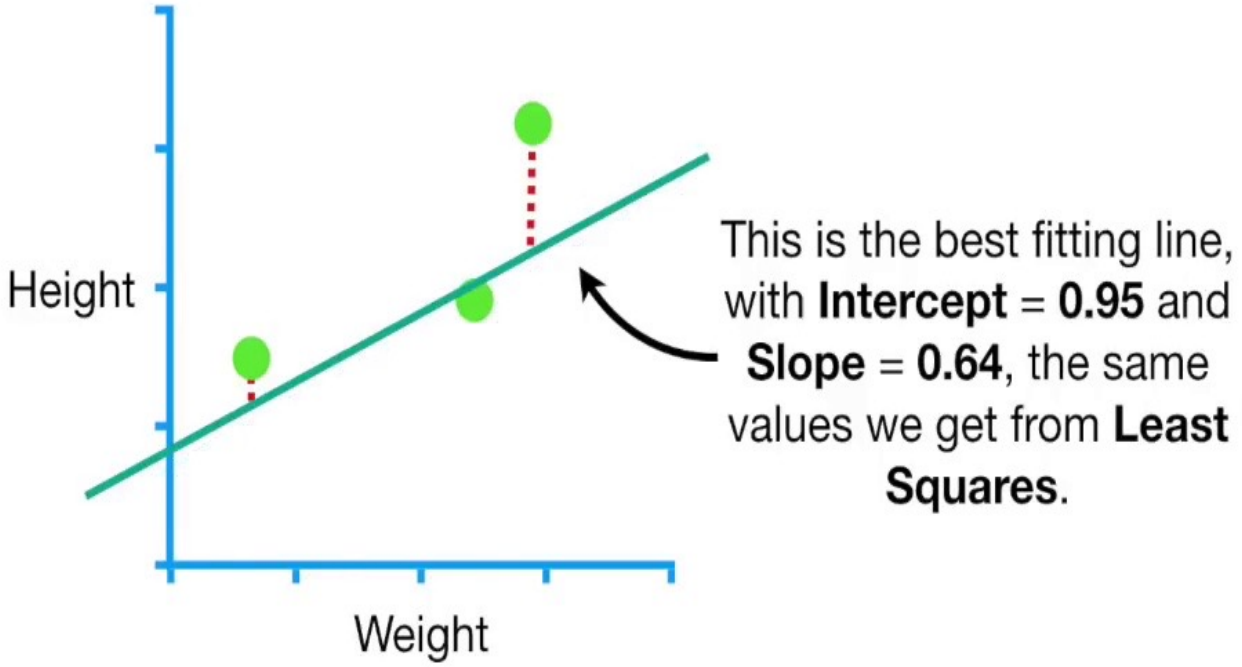
Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

GD for *m*, *b*

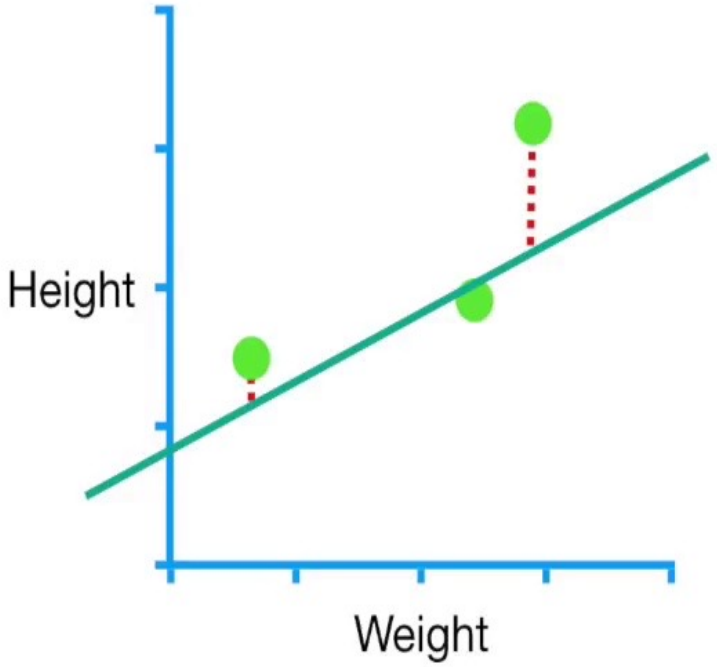


Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

GD for *m, b*

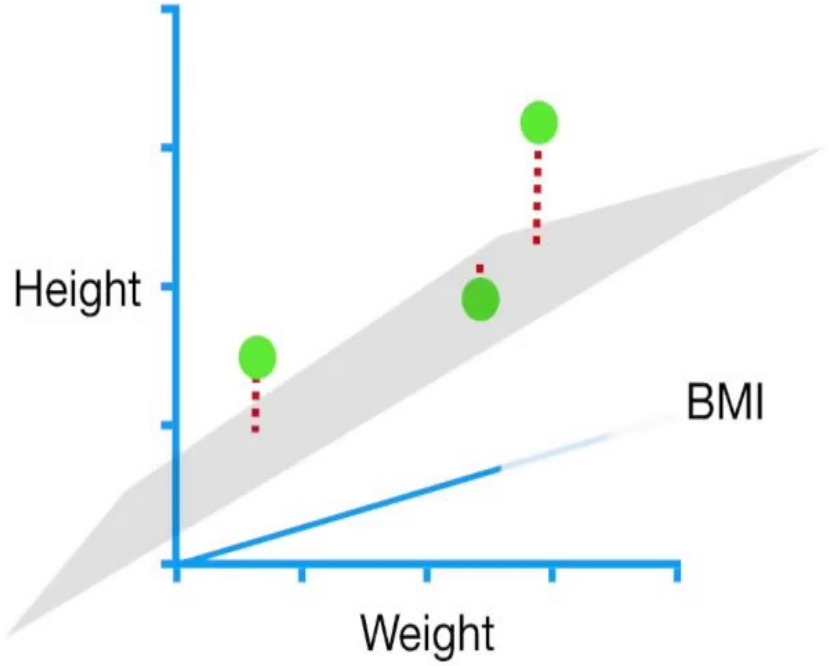

Height
Weight
This is the best fitting line, with **Intercept** = **0.95** and **Slope** = **0.64**, the same values we get from **Least Squares**.

GD for *m, b*



We now know how **Gradient Descent** optimizes two parameters, the **Slope** and **Intercept**.

GD for more parameters and variables



If we had more parameters, then we'd just take more derivatives and everything else stays the same.

Gradient Descent Recap

Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Gradient Descent Recap

Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Step 2: Pick random values for the parameters.

Gradient Descent Recap

Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Step 2: Pick random values for the parameters.

Step 3: Plug the parameter values into the derivatives (ahem, the **Gradient**).

Gradient Descent Recap

Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Step 2: Pick random values for the parameters.

Step 3: Plug the parameter values into the derivatives (ahem, the **Gradient**).

Step 4: Calculate the Step Sizes: **Step Size** = **Slope** x **Learning Rate**

Gradient Descent Recap

Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Step 2: Pick random values for the parameters.

Step 3: Plug the parameter values into the derivatives (ahem, the **Gradient**).

Step 4: Calculate the Step Sizes: **Step Size** = **Slope** x **Learning Rate**

Step 5: Calculate the New Parameters:

New Parameter = **Old Parameter** - **Step Size**

Gradient Descent Recap

Now go back to **Step 3** and repeat until **Step Size** is very small, or you reach the **Maximum Number of Steps**.

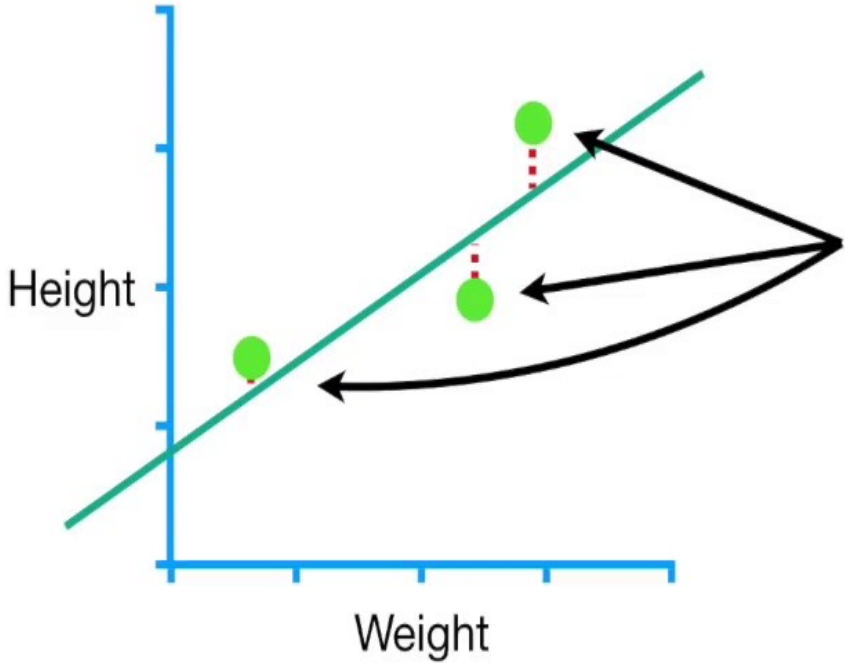
Step 3: Plug the parameter values into the derivatives (ahem, the **Gradient**).

Step 4: Calculate the Step Sizes: **Step Size** = **Slope** x **Learning Rate**

Step 5: Calculate the New Parameters:

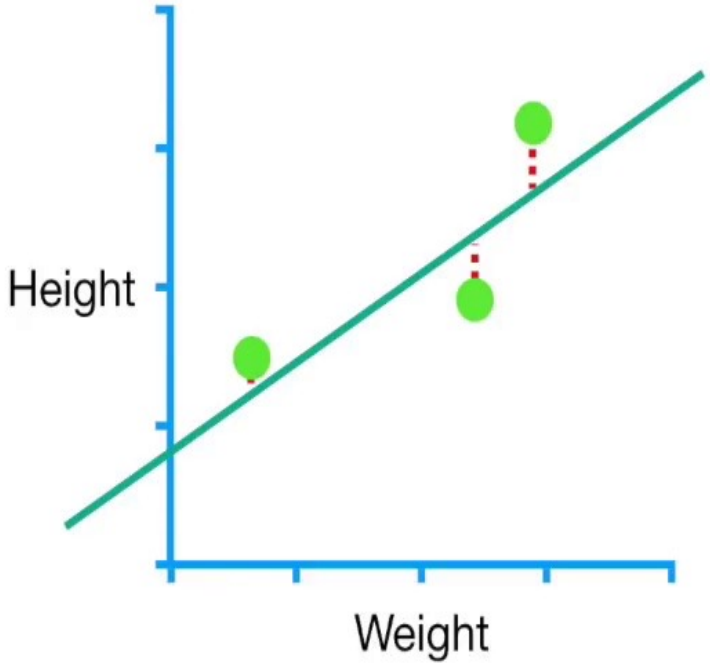
New Parameter = **Old Parameter** - **Step Size**

Stochastic Gradient Descent



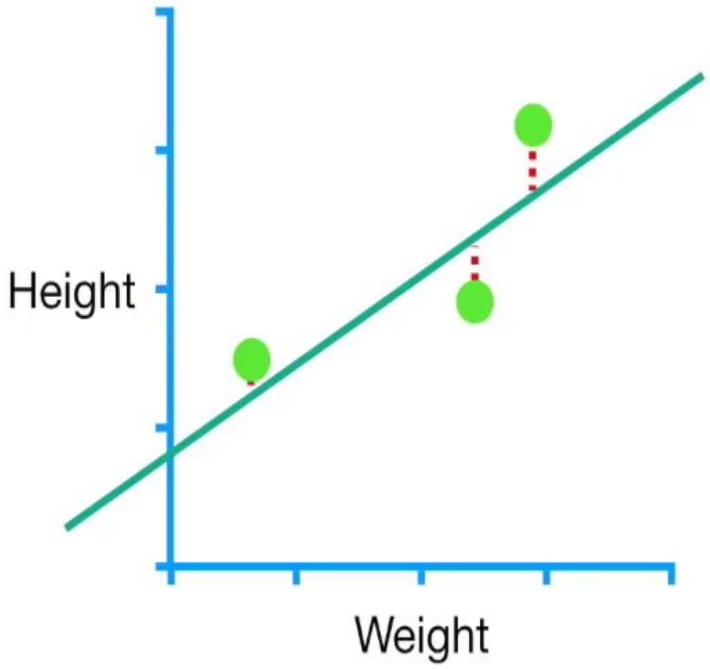
In our example, we only had three data points, so the math didn't take very long…

Stochastic Gradient Descent



...but when you have millions of data points, it can take a long time.

Stochastic Gradient Descent



So there is a thing called **Stochastic Gradient Descent** that uses a randomly selected subset of the data at every step rather than the full dataset.

This reduces the time spent calculating the derivatives of the **Loss Function**.

References

- Lemaréchal, C. (2012). "Cauchy and the Gradient Method" (PDF). Doc Math Extra: 251–254.
- An overview of gradient descent optimization algorithms [pdf]