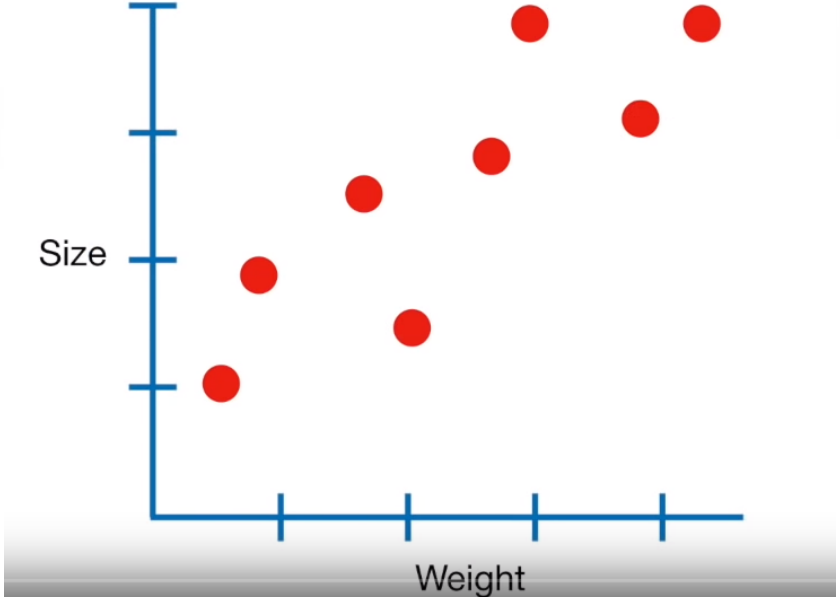
Ridge Regression and Lasso Regression

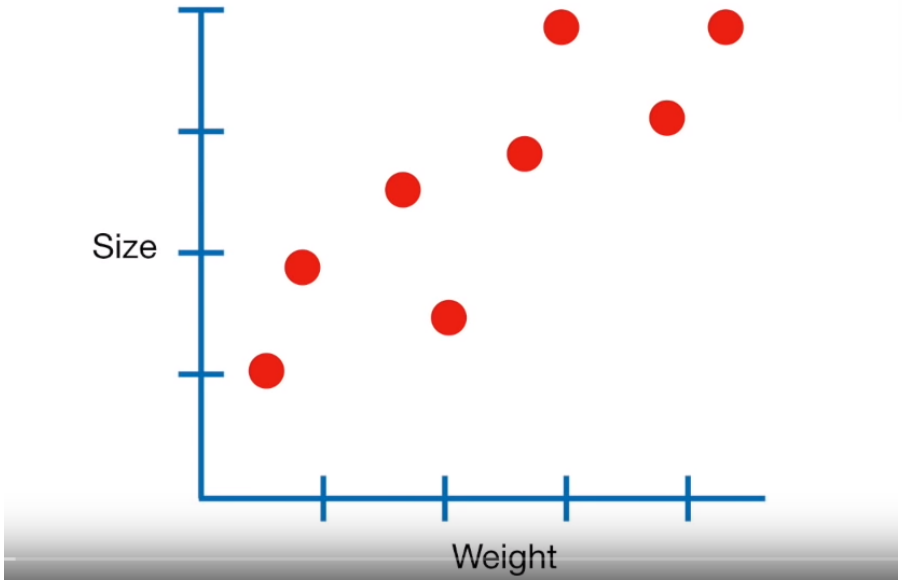
By,
Divesh R. Kubal
Data Scientist, eClerx Services
Center of Excellence – Machine Learning

Ridge Regression

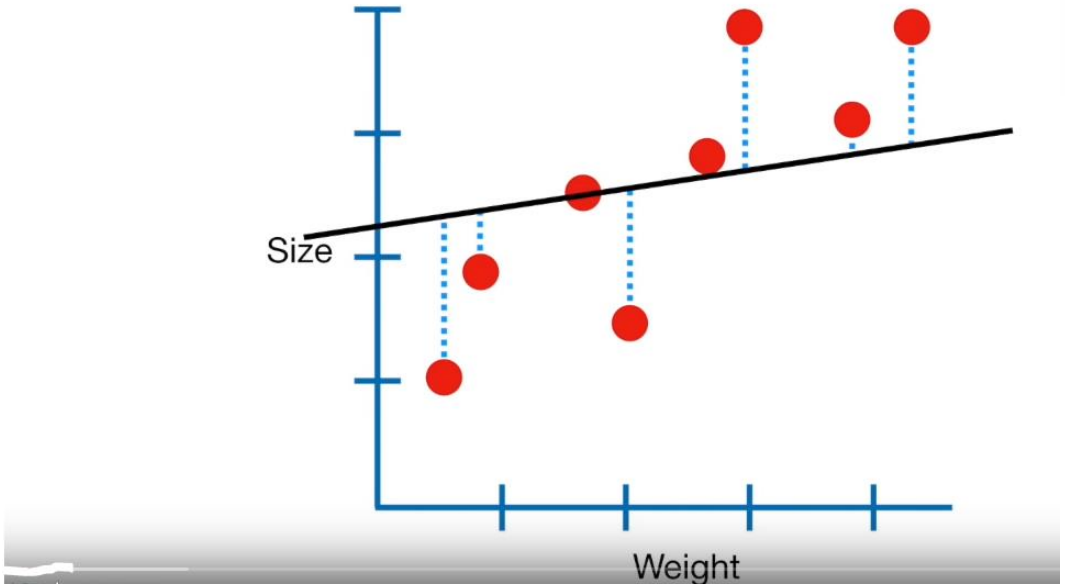
Let's start by collecting **Weight** and **Size** measurements from a bunch of mice…



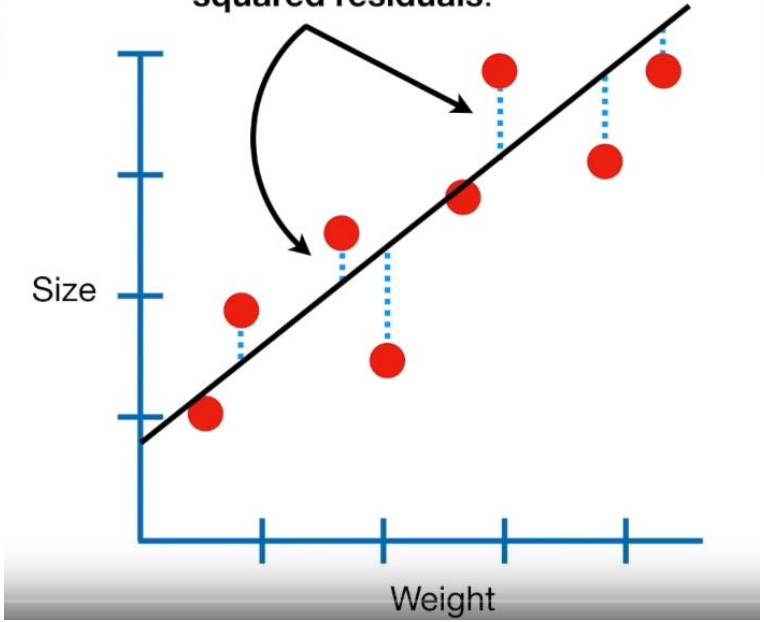
Since these data look relatively linear, we will use **Linear Regression**, AKA **Least Squares**, to model the relationship between **Weight** and **Size**.



So we'll fit a line to the data using **Least Squares**.

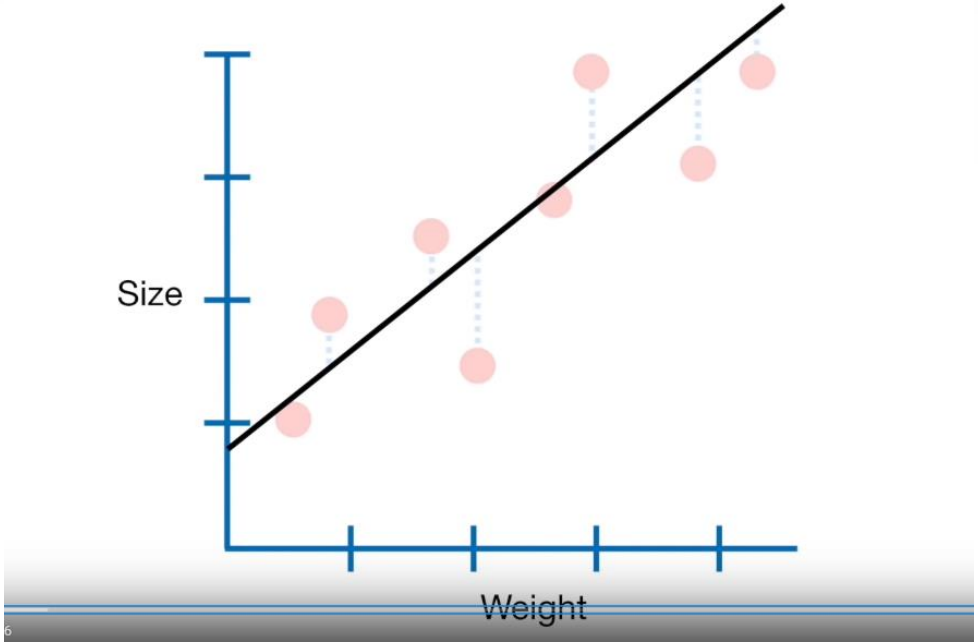


In other words, we find the line that results in the **minimum sum of squared residuals**.



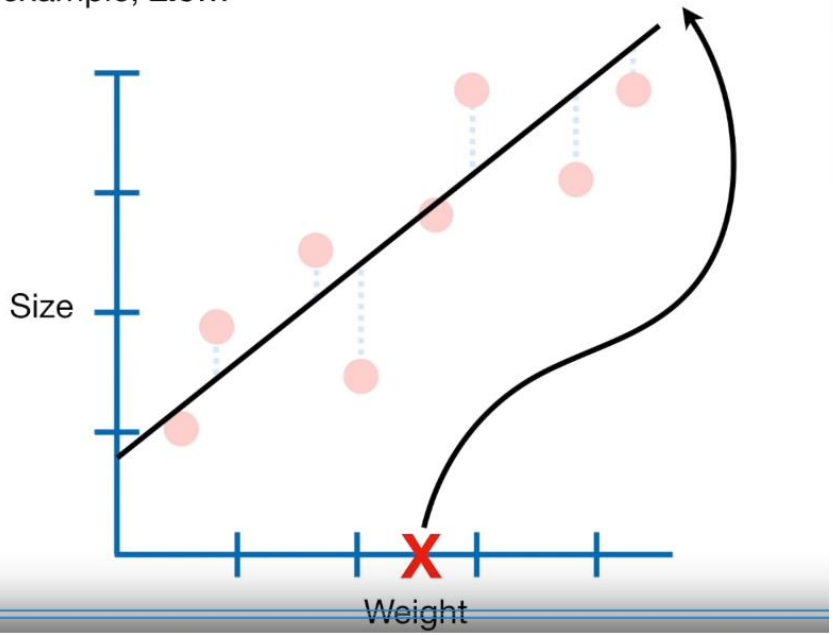
Ultimately, we end up with this equation for the line:

Size = 0.9 + 0.75 x **Weight**



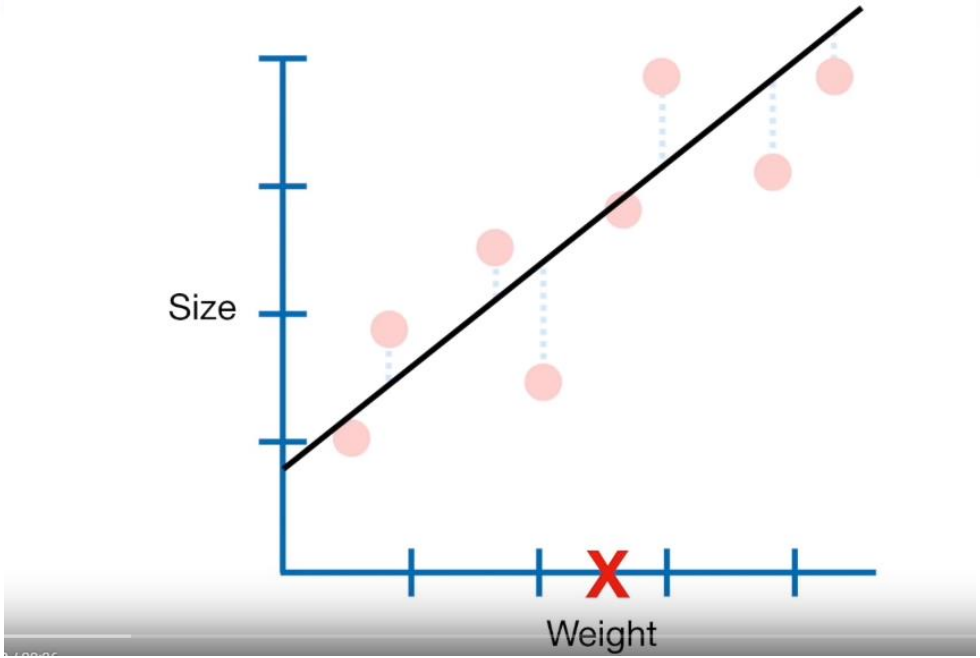
We can plug in a value for **Weight**, for example, **2.5**…

Size = 0.9 + 0.75 × **Weight**

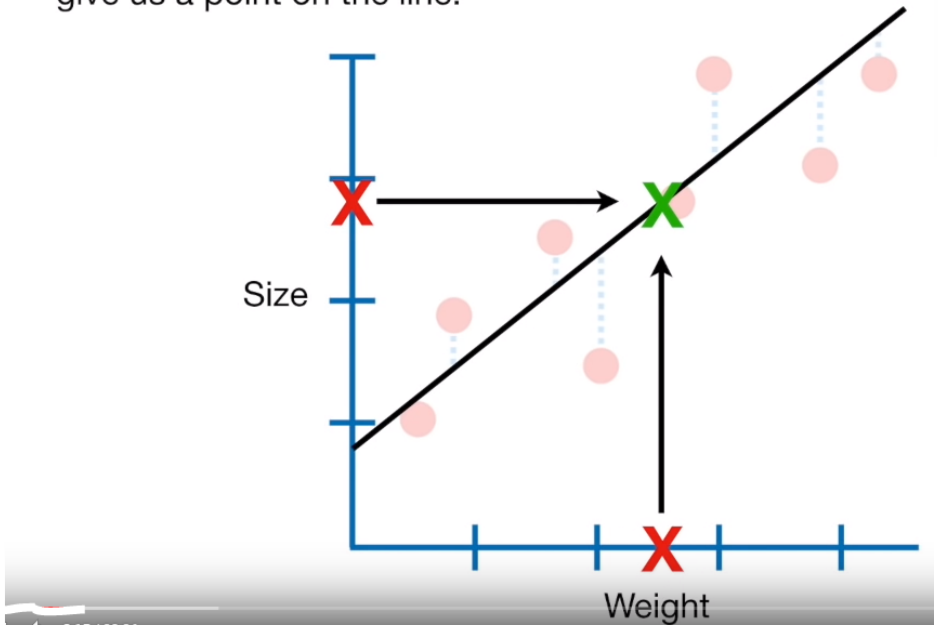


…and do the math…

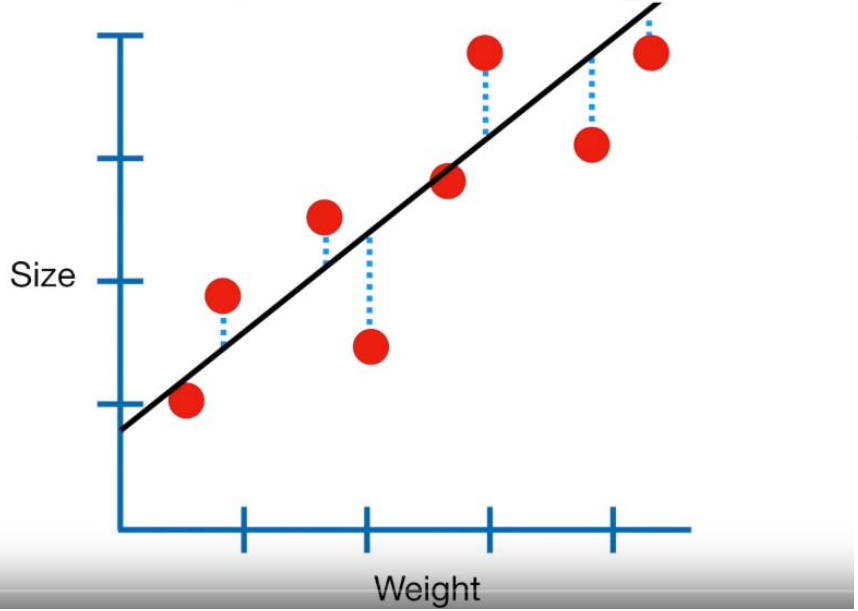
Size = 0.9 + 1.88

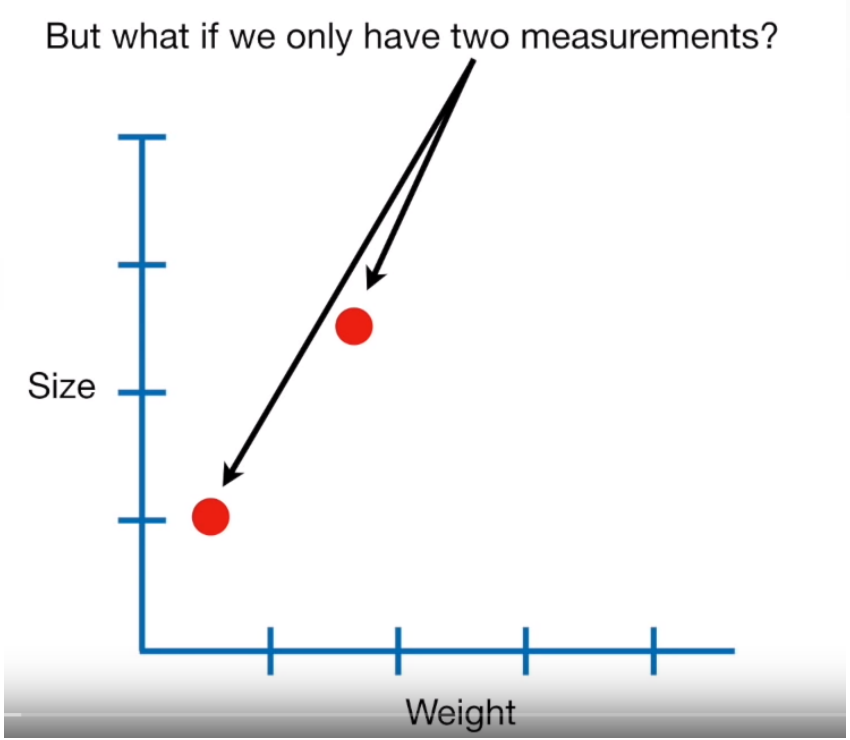


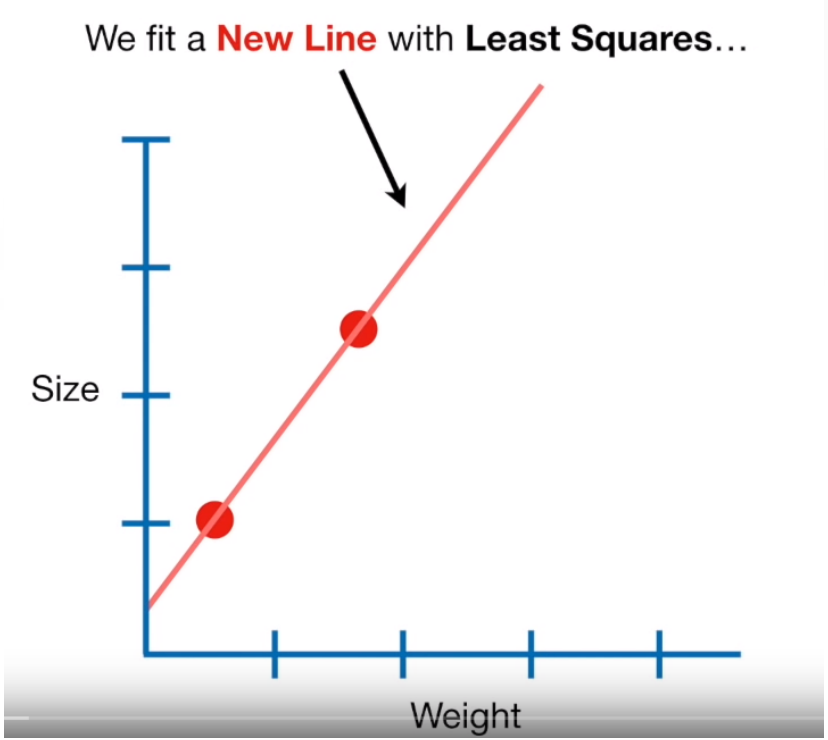
Together, the value for **Weight**, **2.5**, and the value for **Size**, **2.8**, give us a point on the line.

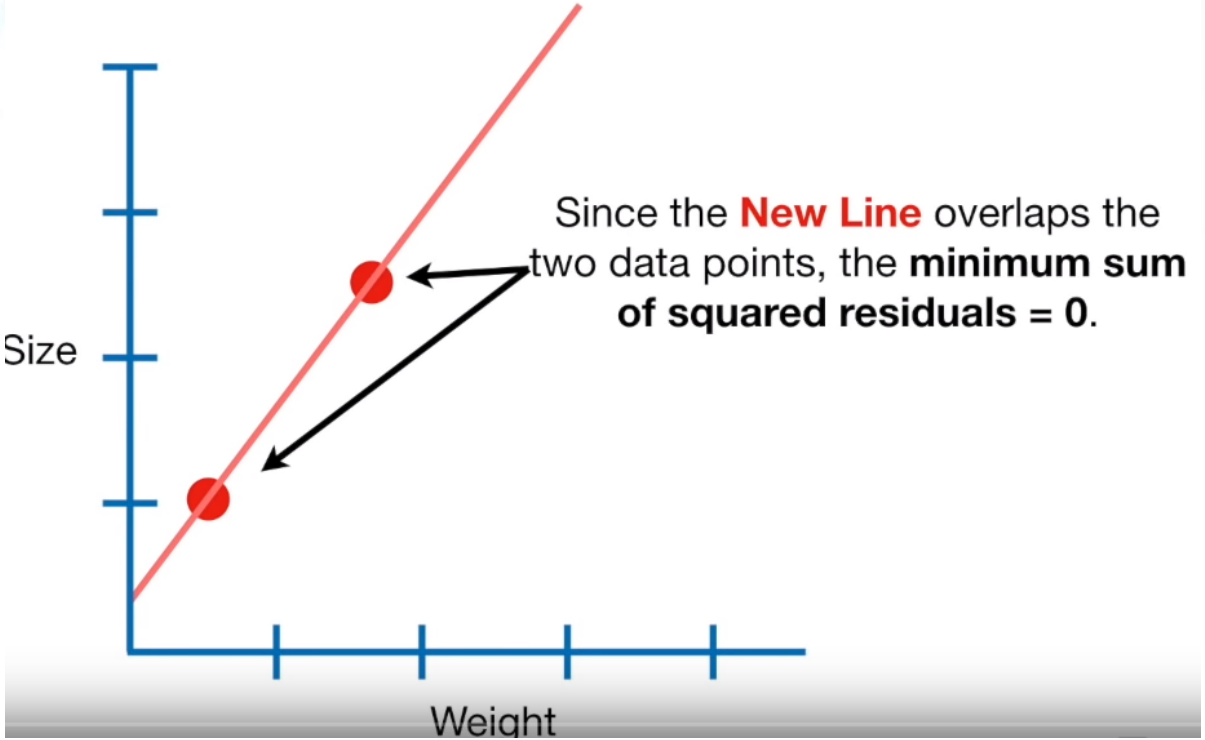


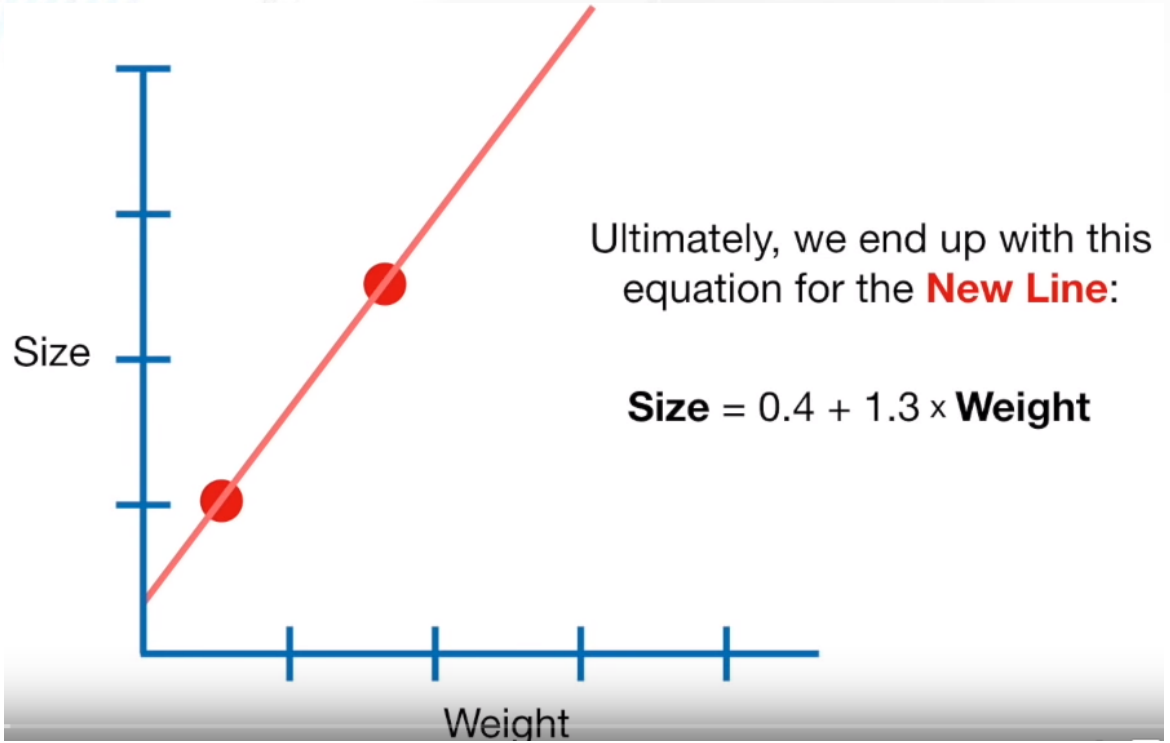
When we have a lot of measurements, we can be fairly confident that the **Least Squares** line accurately reflects the relationship between **Size** and **Weight**.




But what if we only have two measurements?
Size
Weight

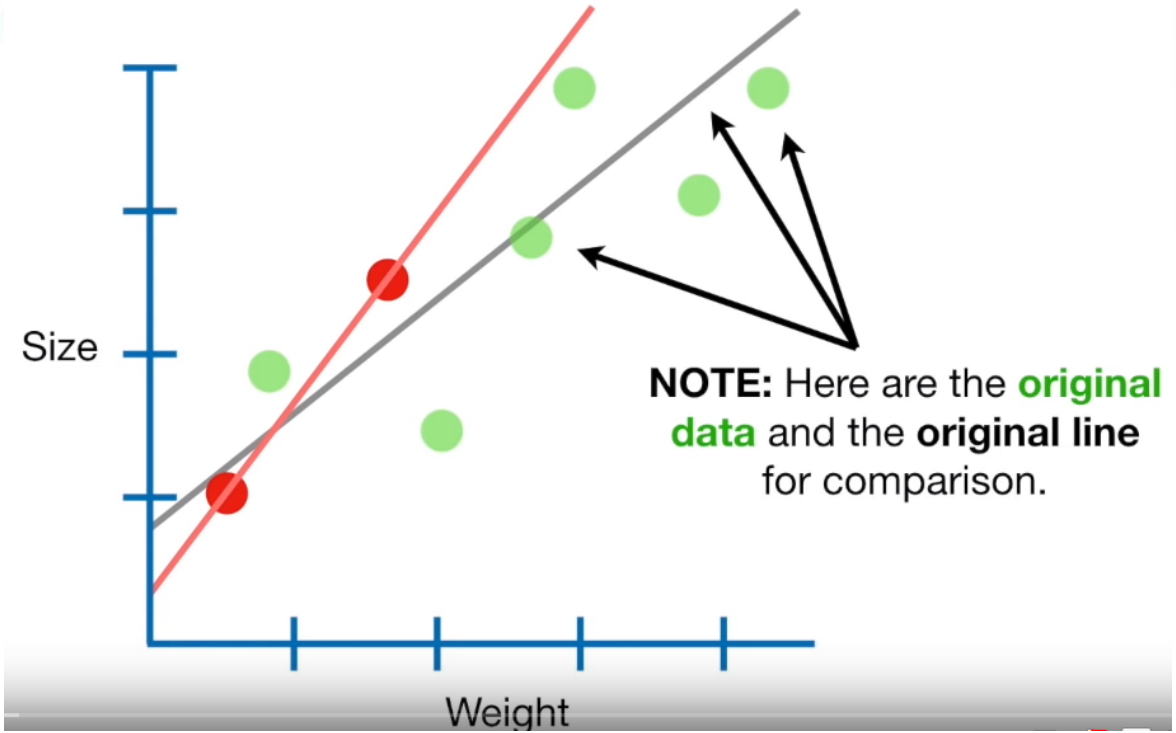

We fit a New Line with Least Squares…
Size
Weight


Since the New Line overlaps the two data points, the minimum sum of squared residuals = 0.
Size
Weight

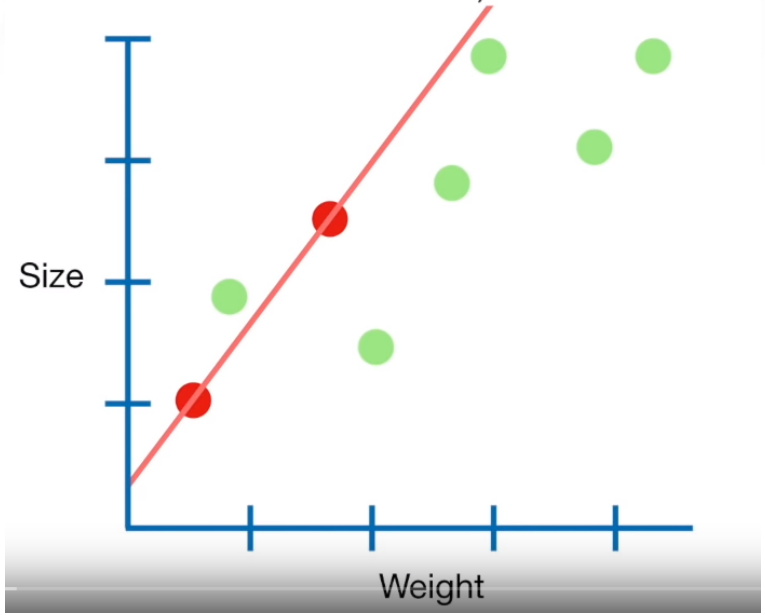


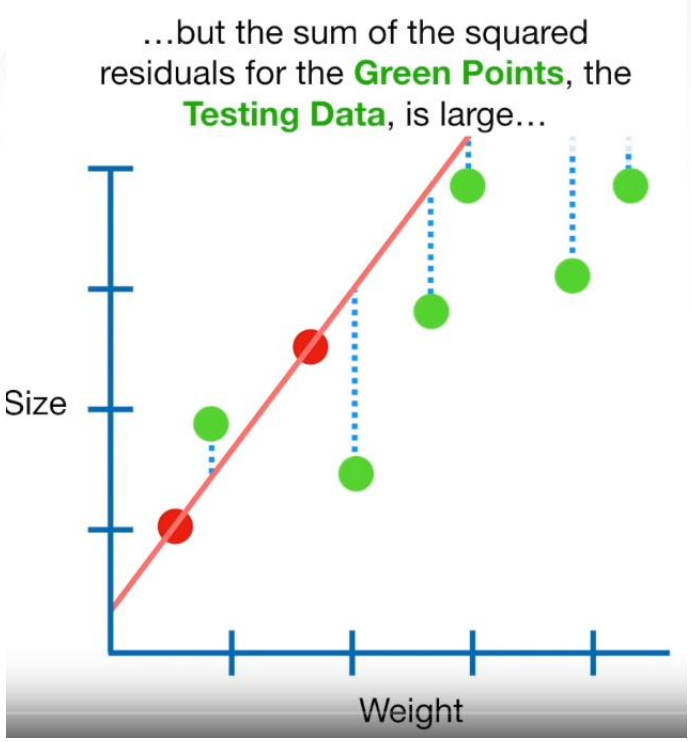
Ultimately, we end up with this equation for the **New Line**:

Size = 0.4 + 1.3 × **Weight**

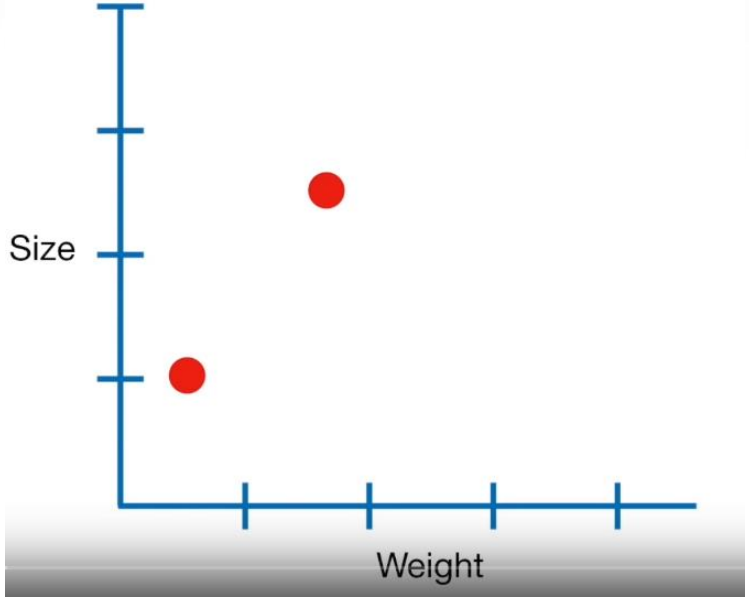

Size
Weight
NOTE: Here are the original data and the original line for comparison.

The sum of the squared residuals for just the **Two Red Points**, the **Training Data**, is small (in this case it is 0)…

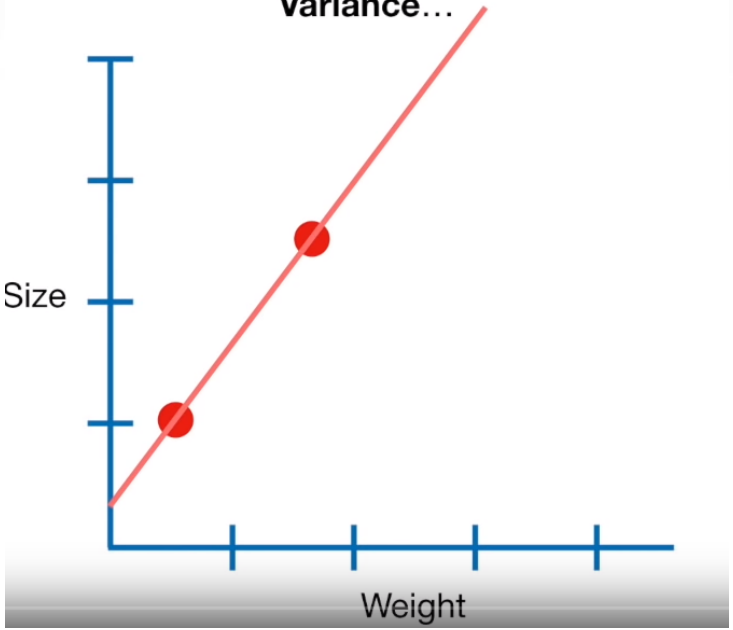



…but the sum of the squared residuals for the Green Points, the Testing Data, is large…
Size
Weight

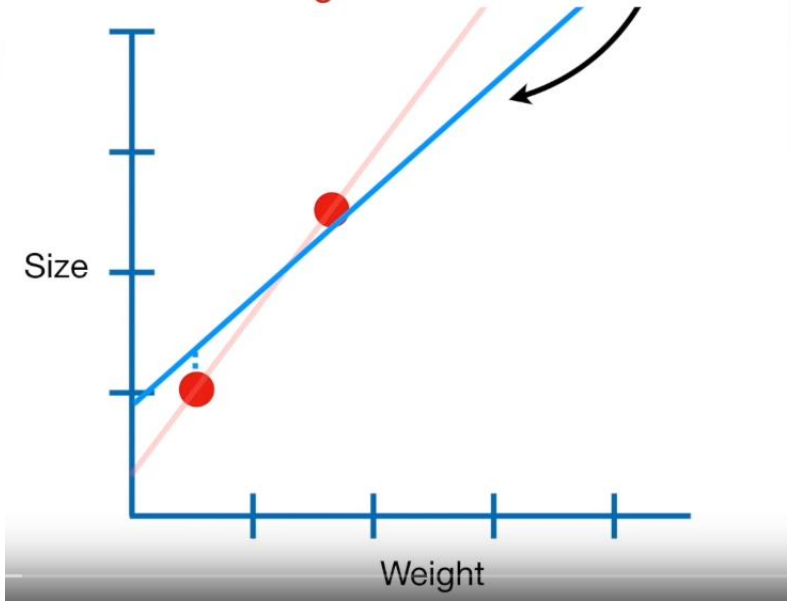
Now let's go back to just the **Training Data**…

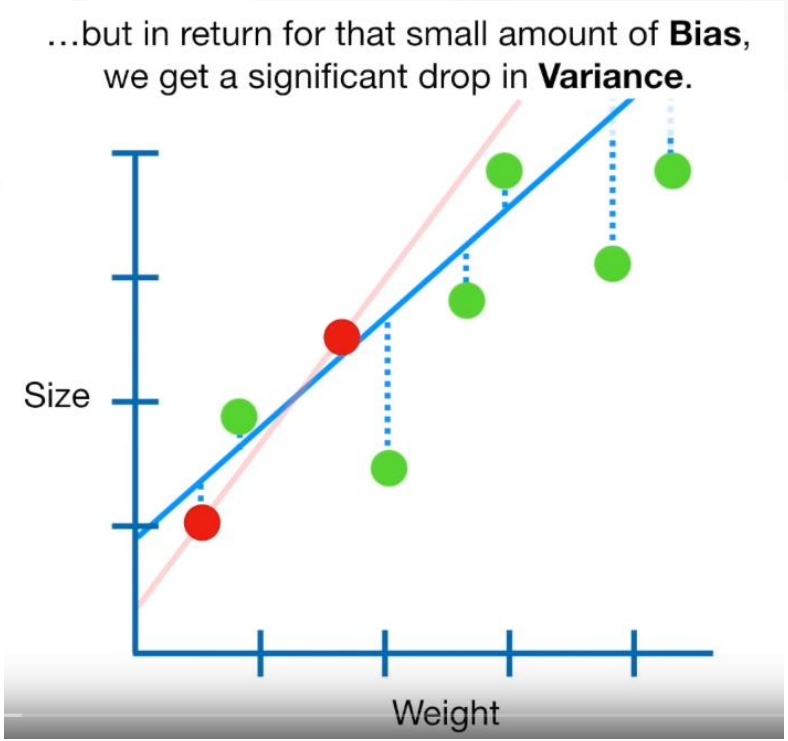


We just saw that **Least Squares** results in a **Line** that is **Over Fit** and has **High Variance**…

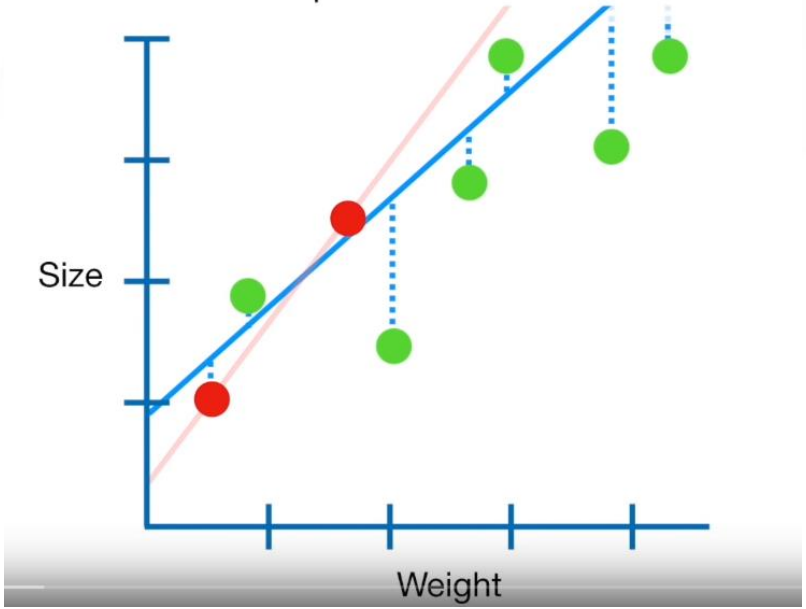


The main idea behind **Ridge Regression** is to find a **New Line** that doesn't fit the **Training Data** as well…

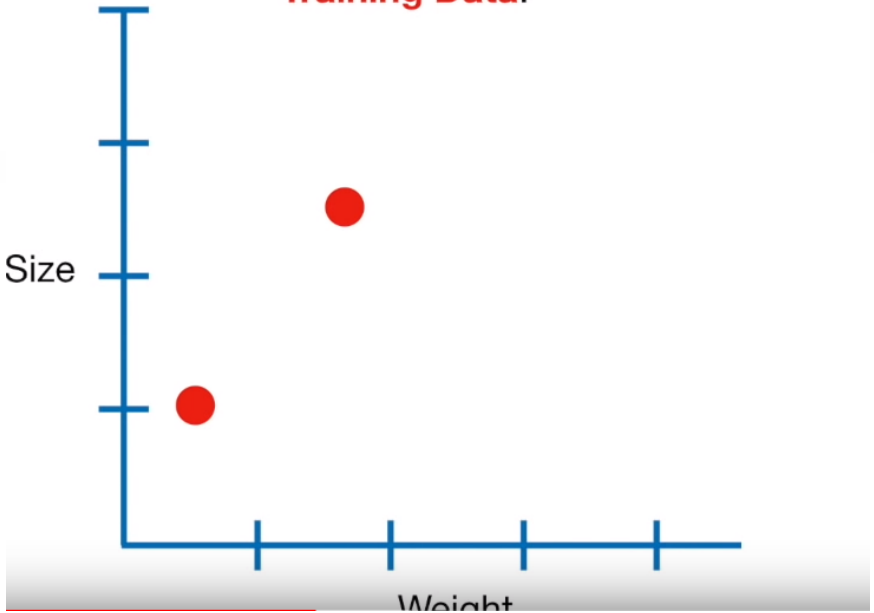



...but in return for that small amount of **Bias**,
we get a significant drop in **Variance**.
Size
Weight

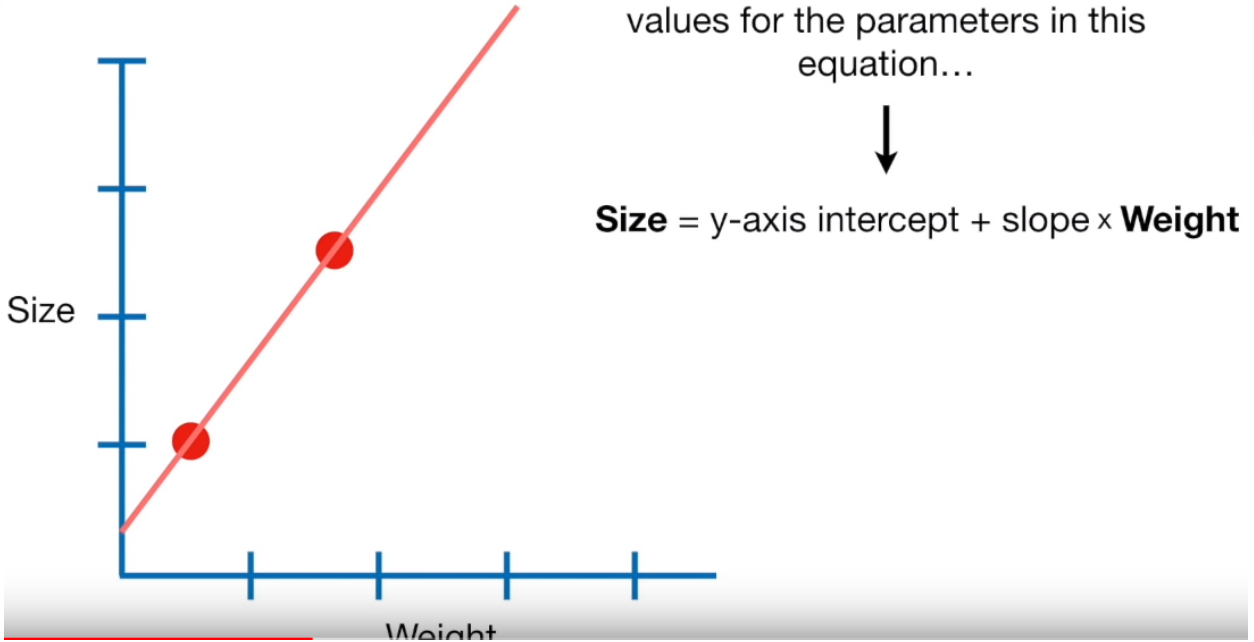
In other words, by starting with a slightly worse fit, **Ridge Regression** can provide better long term predictions.



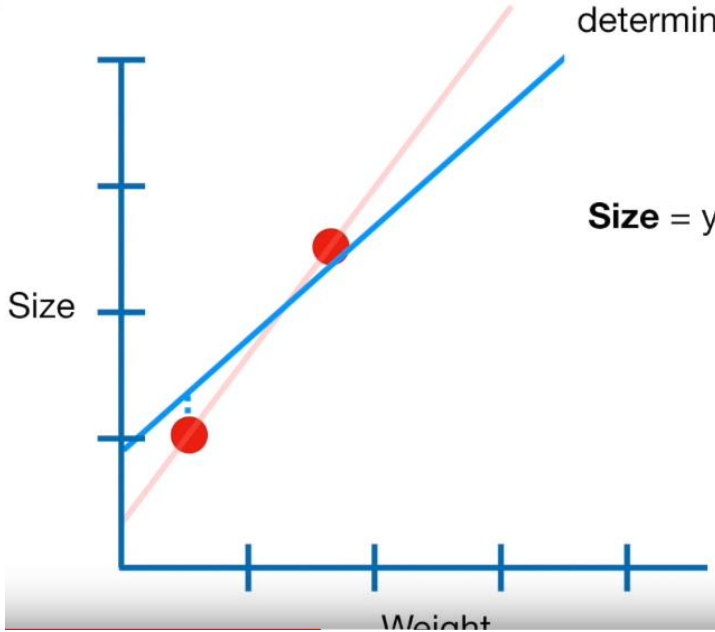
Let's go back to just the **Training Data**.



When **Least Squares** determines values for the parameters in this equation…

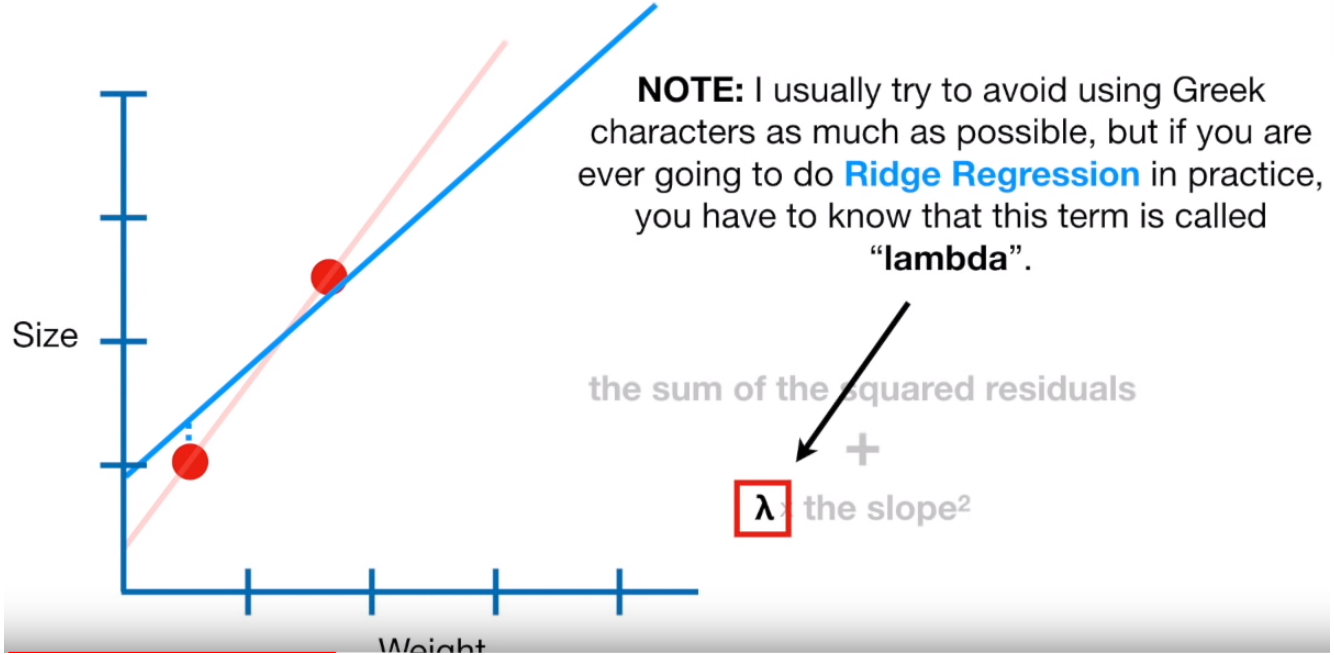


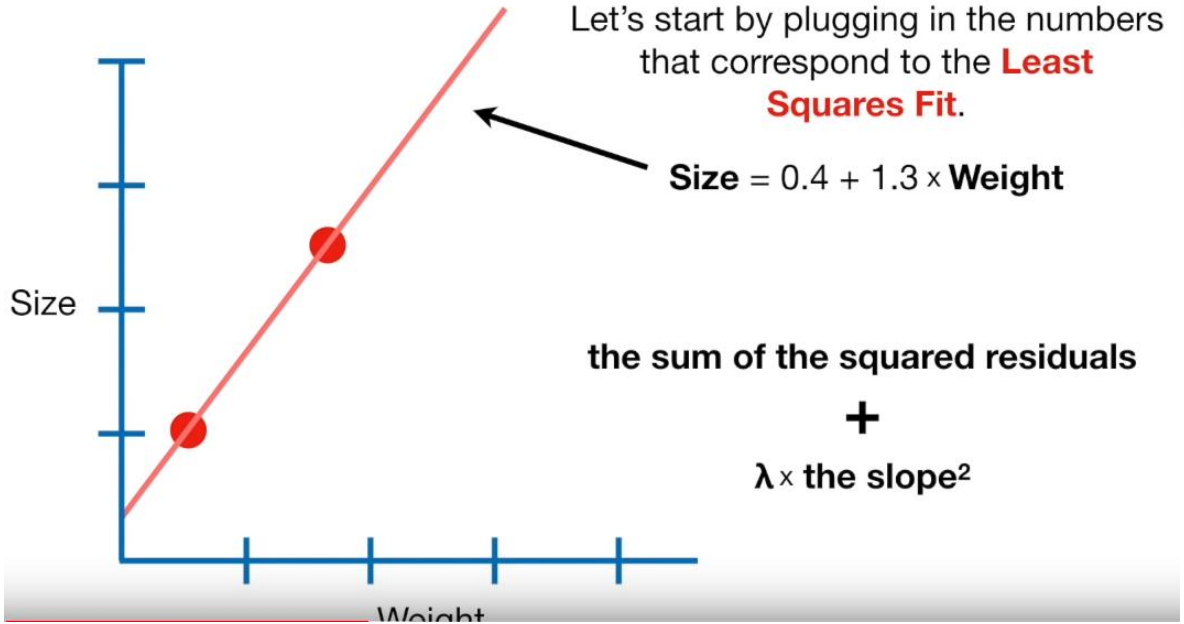
Size = y-axis intercept + slope x **Weight**

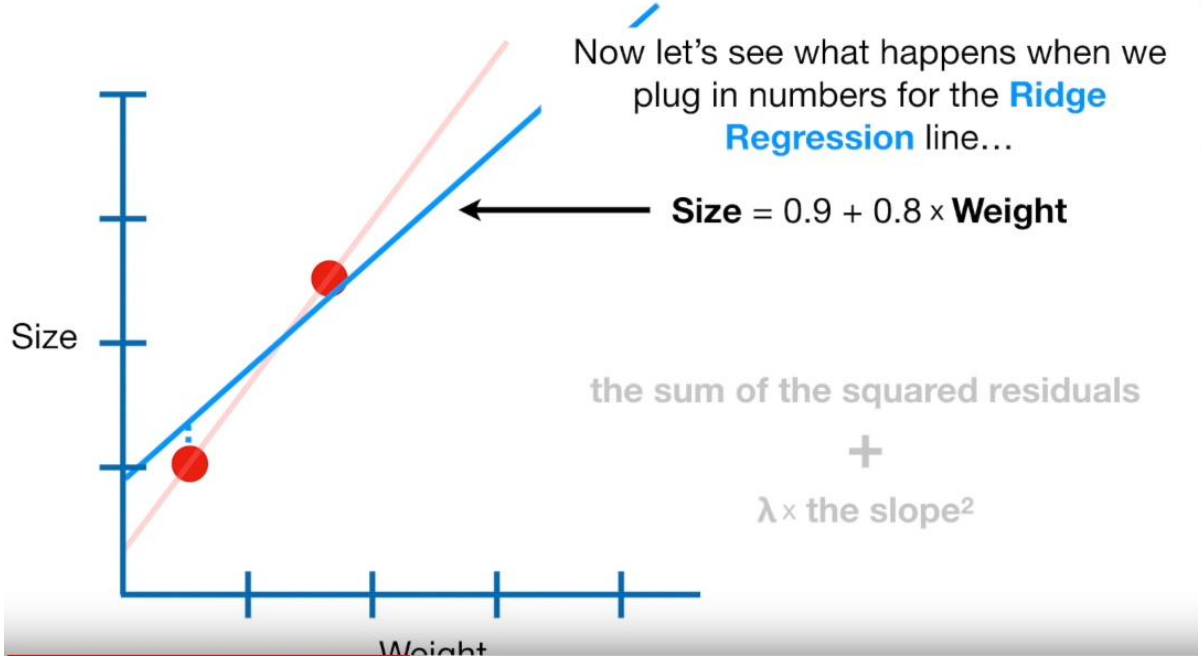


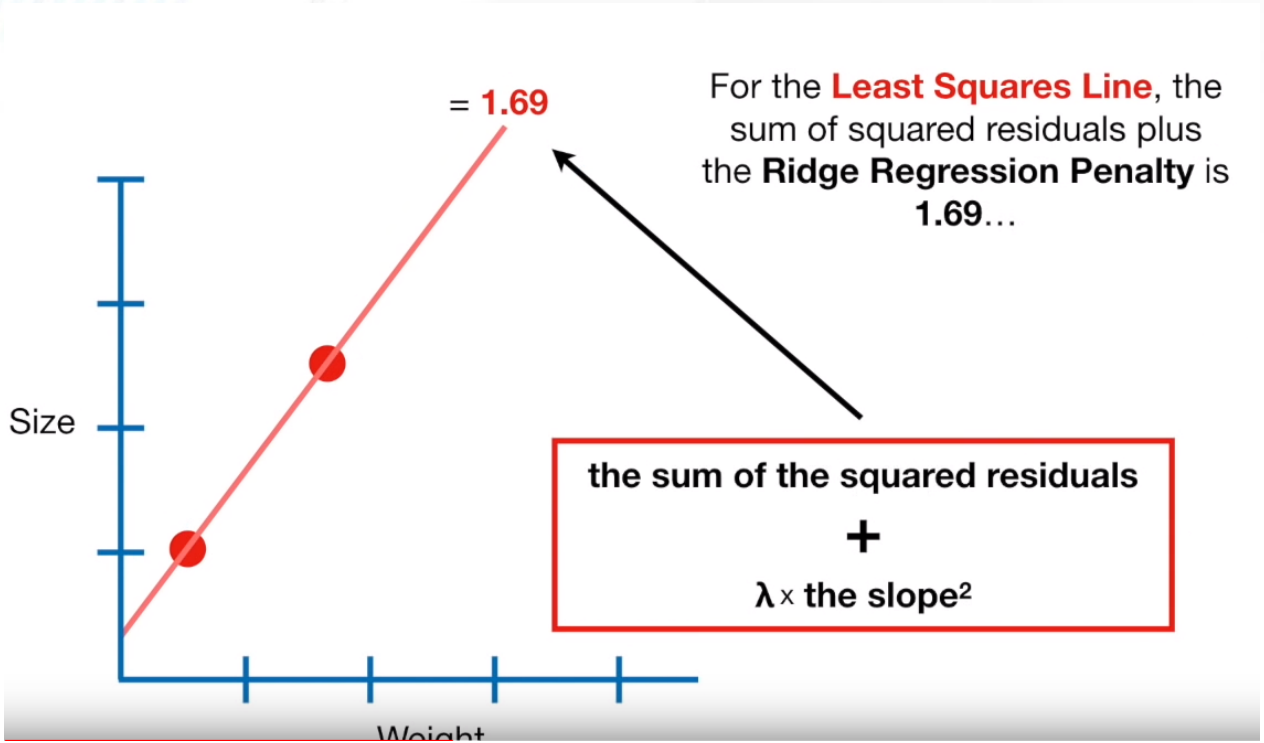
In contrast, when **Ridge Regression** determines values for the parameters in this equation…

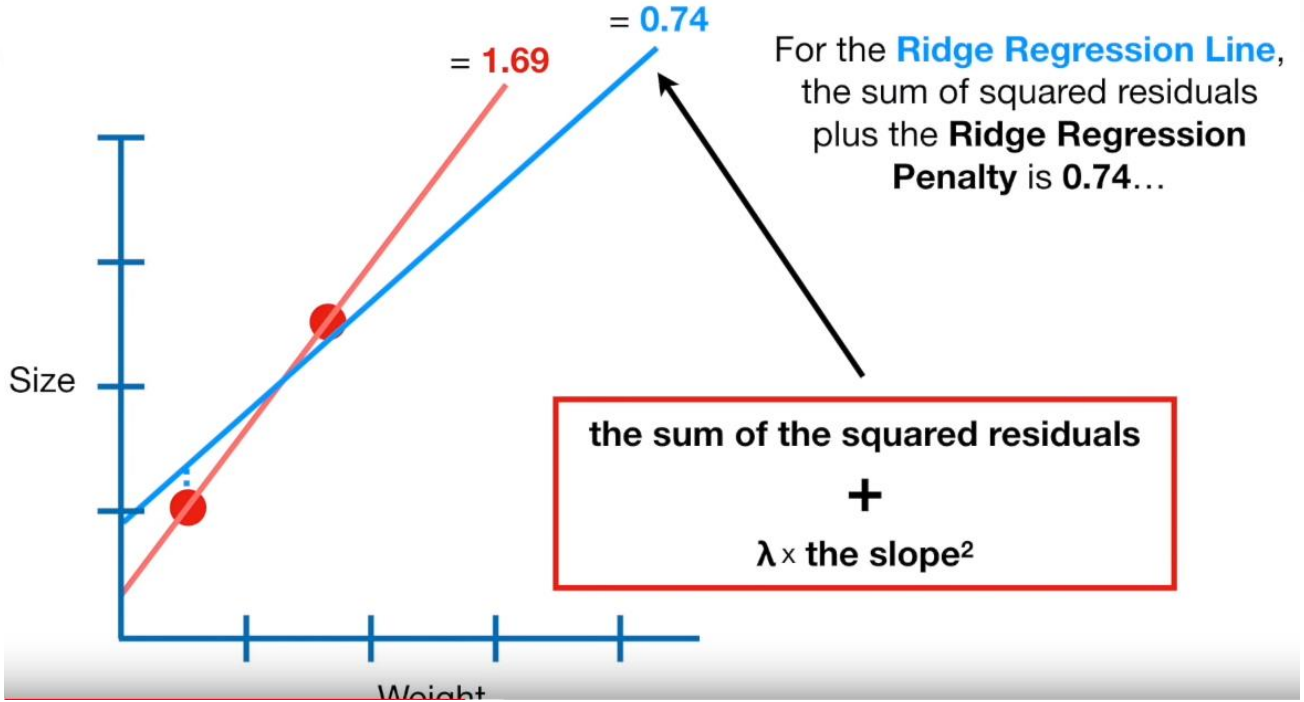
Size = y-axis intercept + slope × **Weight**

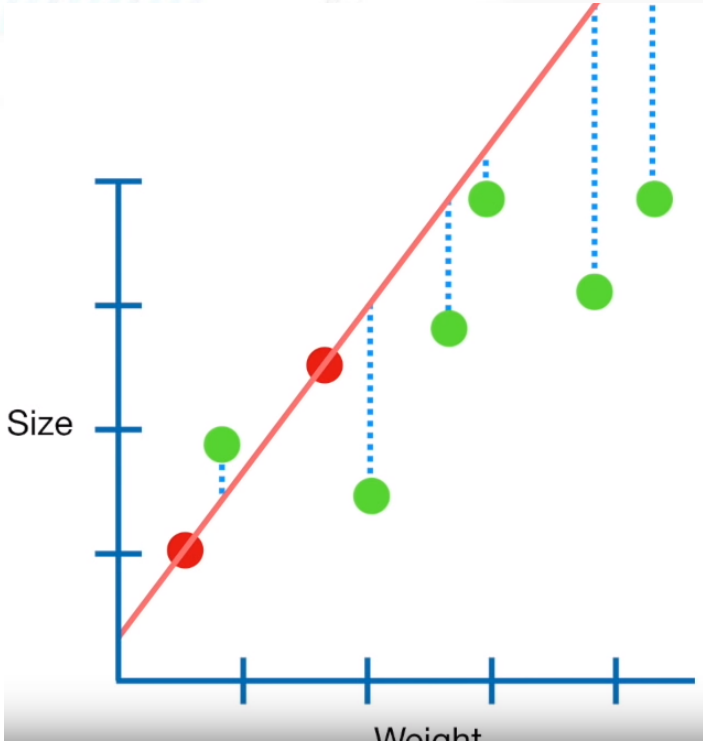

Size
Weight
NOTE: I usually try to avoid using Greek characters as much as possible, but if you are ever going to do Ridge Regression in practice, you have to know that this term is called "lambda".
the sum of the squared residuals
+
λ
the slope²


Let's start by plugging in the numbers that correspond to the Least Squares Fit.
Size = 0.4 + 1.3 x Weight
Size
the sum of the squared residuals
+
λ x the slope²
Weight

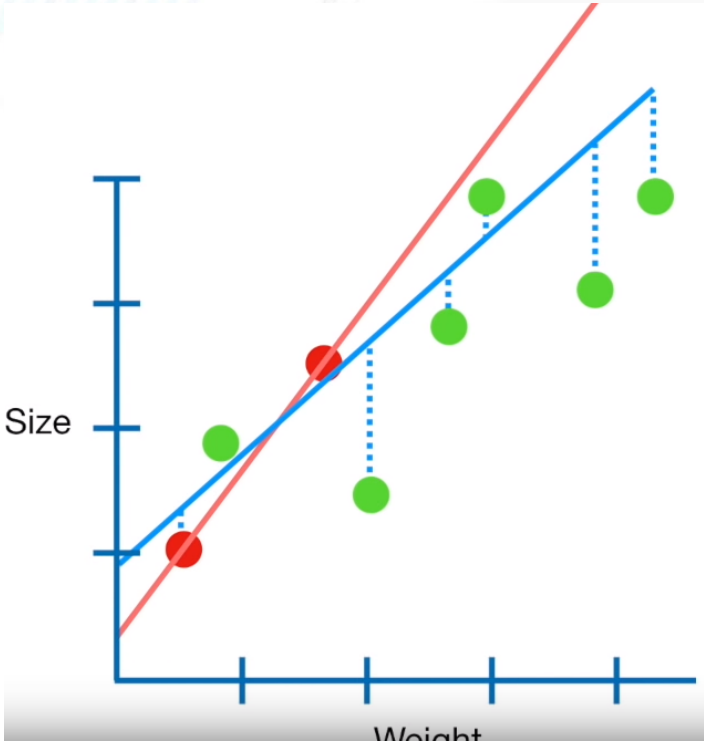

Now let's see what happens when we plug in numbers for the Ridge Regression line…
Size = 0.9 + 0.8 × Weight
Size
the sum of the squared residuals
+
λ × the slope²
Weight


= 1.69
For the Least Squares Line, the sum of squared residuals plus the Ridge Regression Penalty is 1.69…
Size
the sum of the squared residuals
+
λ x the slope²
Weight

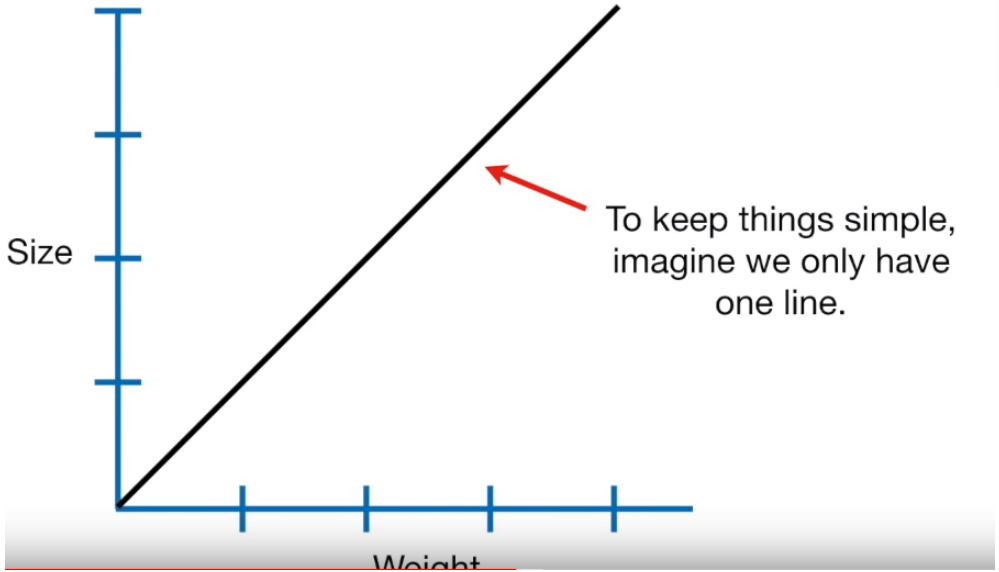

= 0.74
= 1.69
For the Ridge Regression Line, the sum of squared residuals plus the Ridge Regression Penalty is 0.74...
Size
the sum of the squared residuals
+
λ x the slope²
Weight

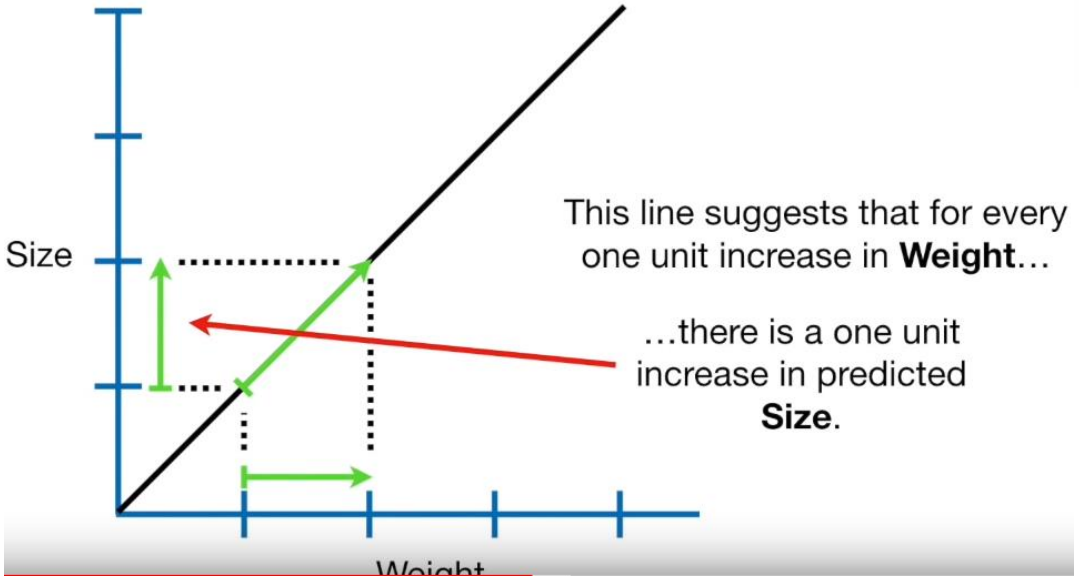


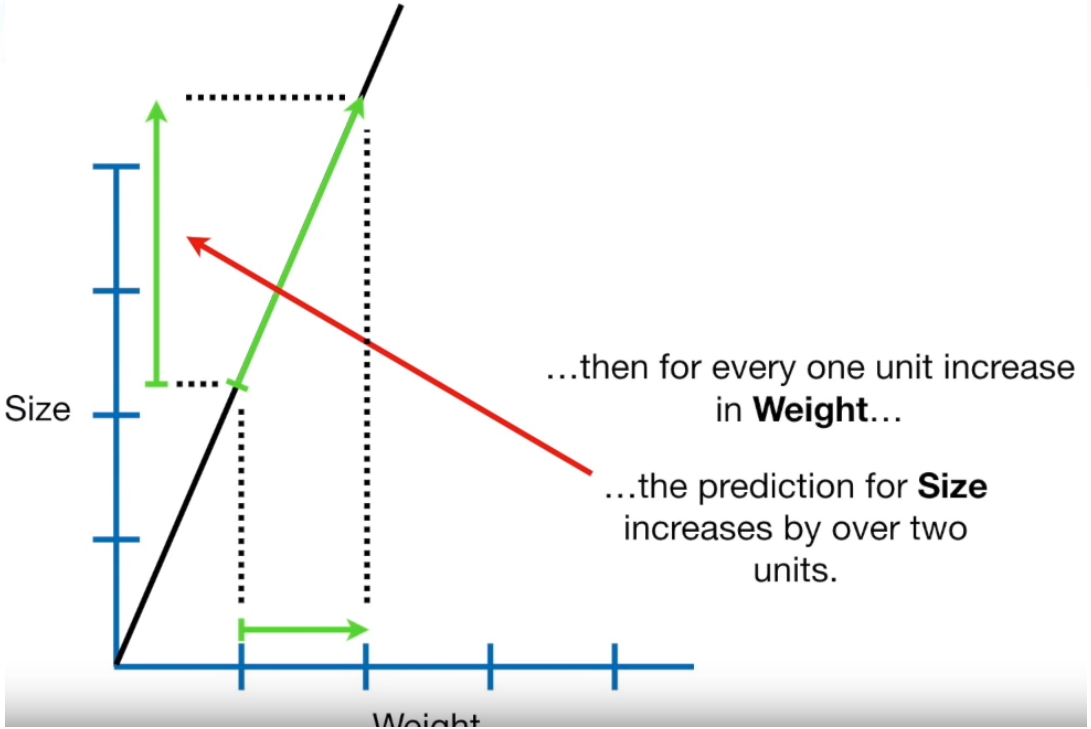
Without the small amount of **Bias** that the penalty creates, the **Least Squares Fit** has a large amount of **Variance**.

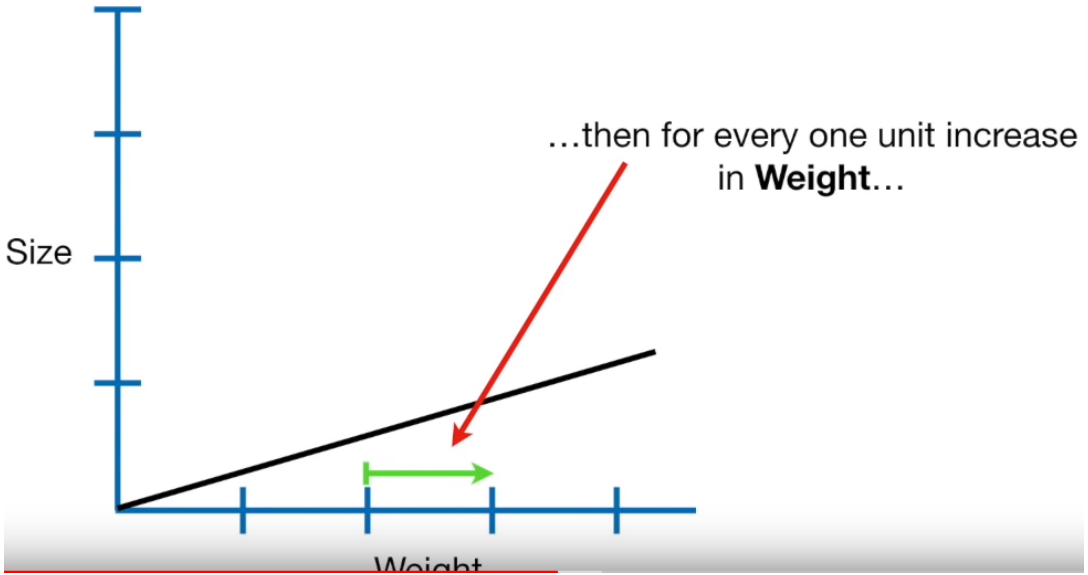


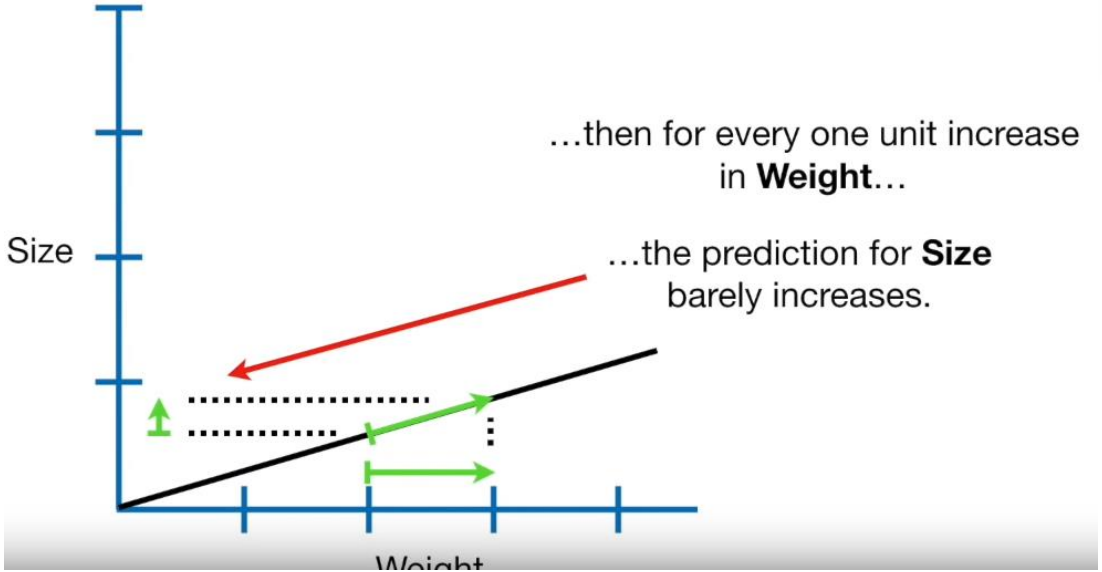
In contrast, the **Ridge Regression Line**, which has the small amount of **Bias** due to the penalty, has less **Variance**.

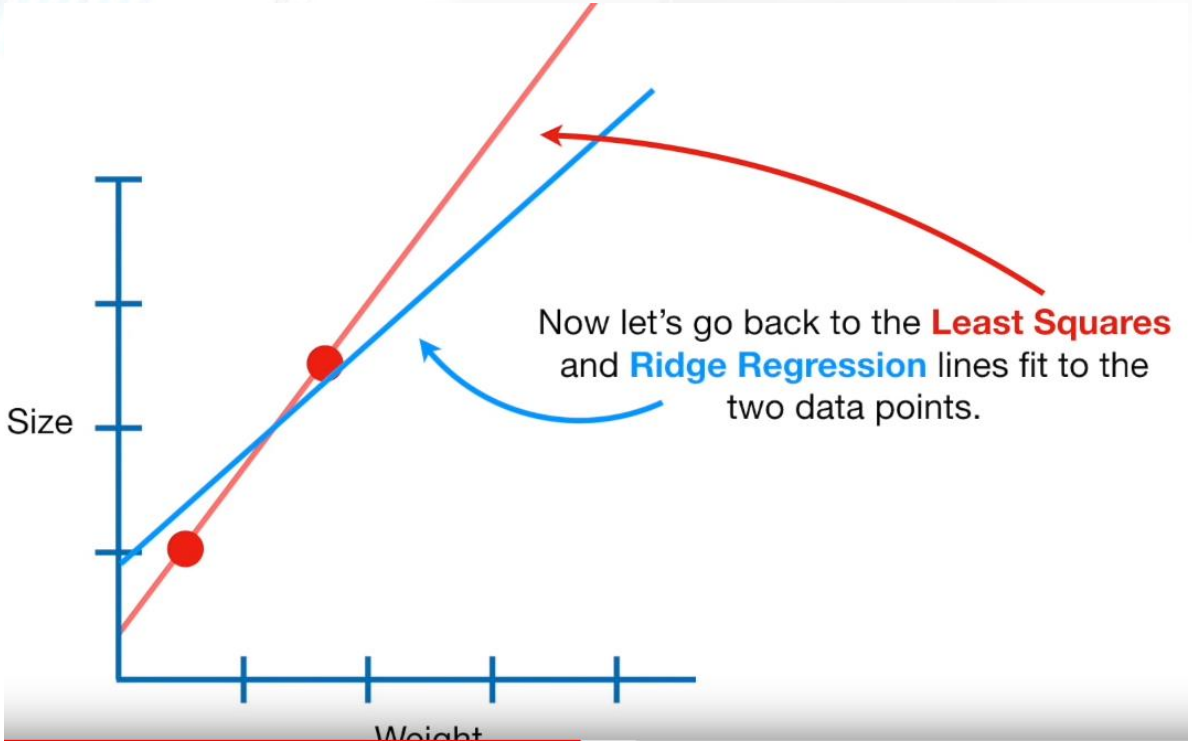

Size
To keep things simple, imagine we only have one line.
Weight

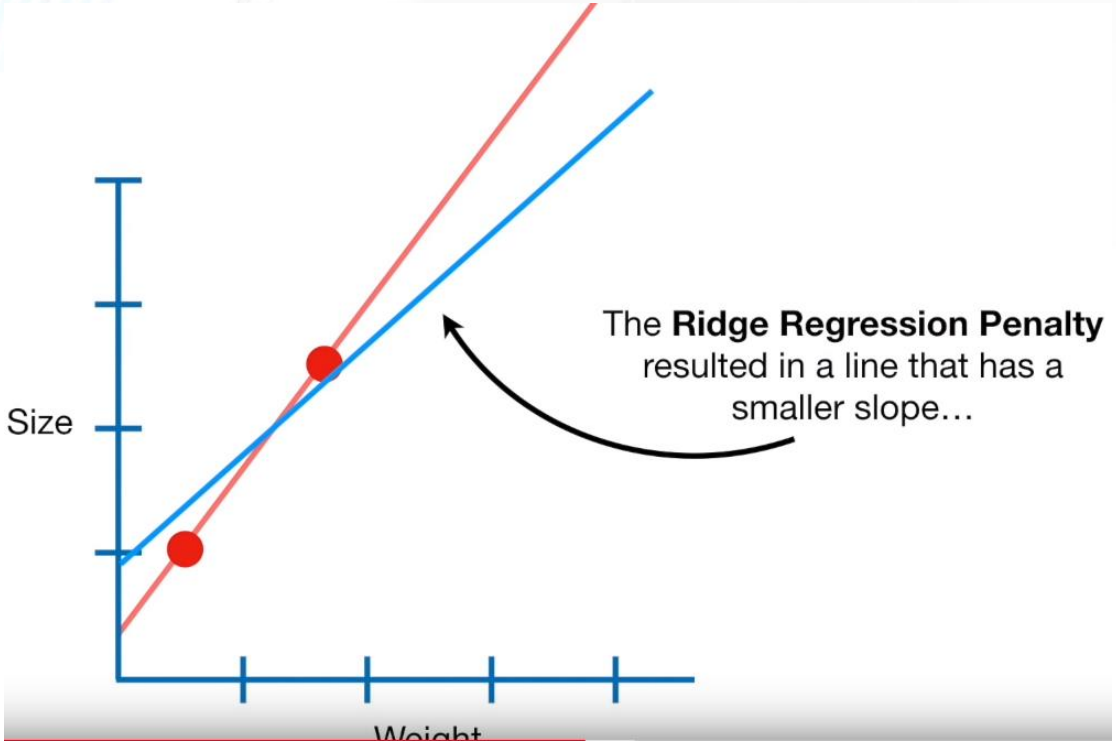

Size
This line suggests that for every one unit increase in **Weight**…
…there is a one unit increase in predicted **Size**.
Weight

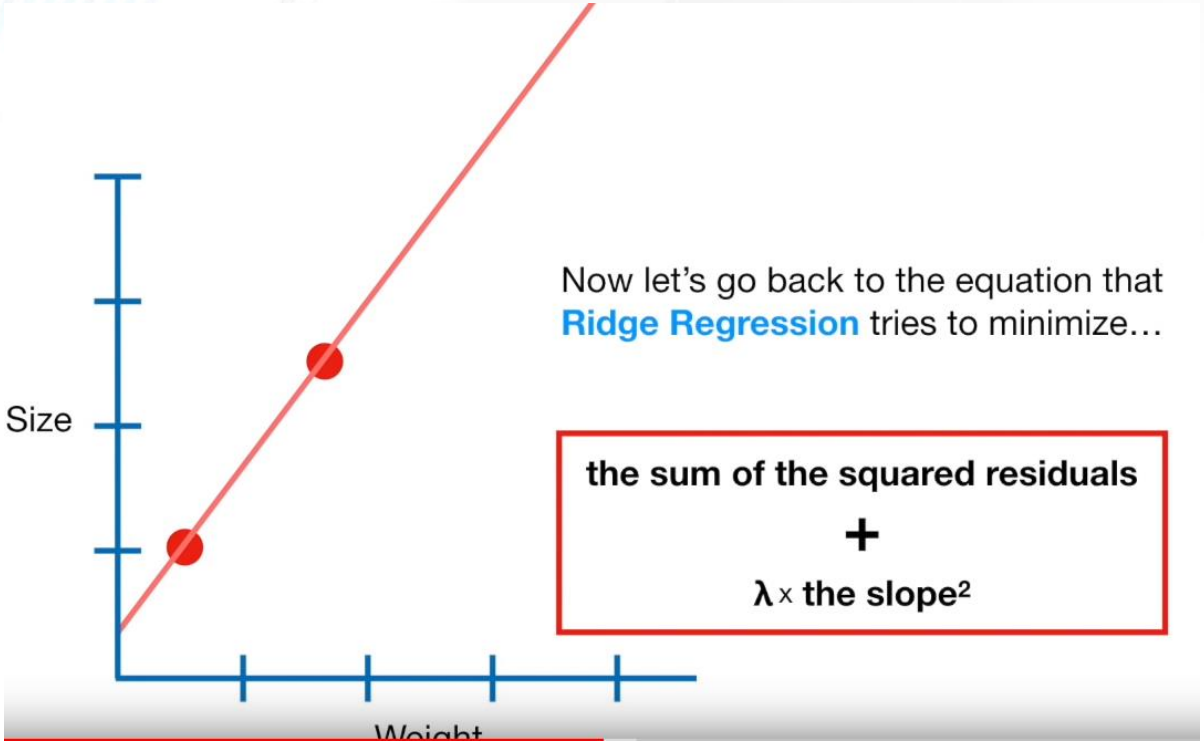

Size
…then for every one unit increase in **Weight**…
…the prediction for **Size** increases by over two units.
Weight


…then for every one unit increase
in **Weight**…
Size
Weight


…then for every one unit increase in **Weight**…
…the prediction for **Size** barely increases.
Size
Weight


Size
Weight
Now let's go back to the Least Squares and Ridge Regression lines fit to the two data points.


Size
Weight
The **Ridge Regression Penalty** resulted in a line that has a smaller slope…

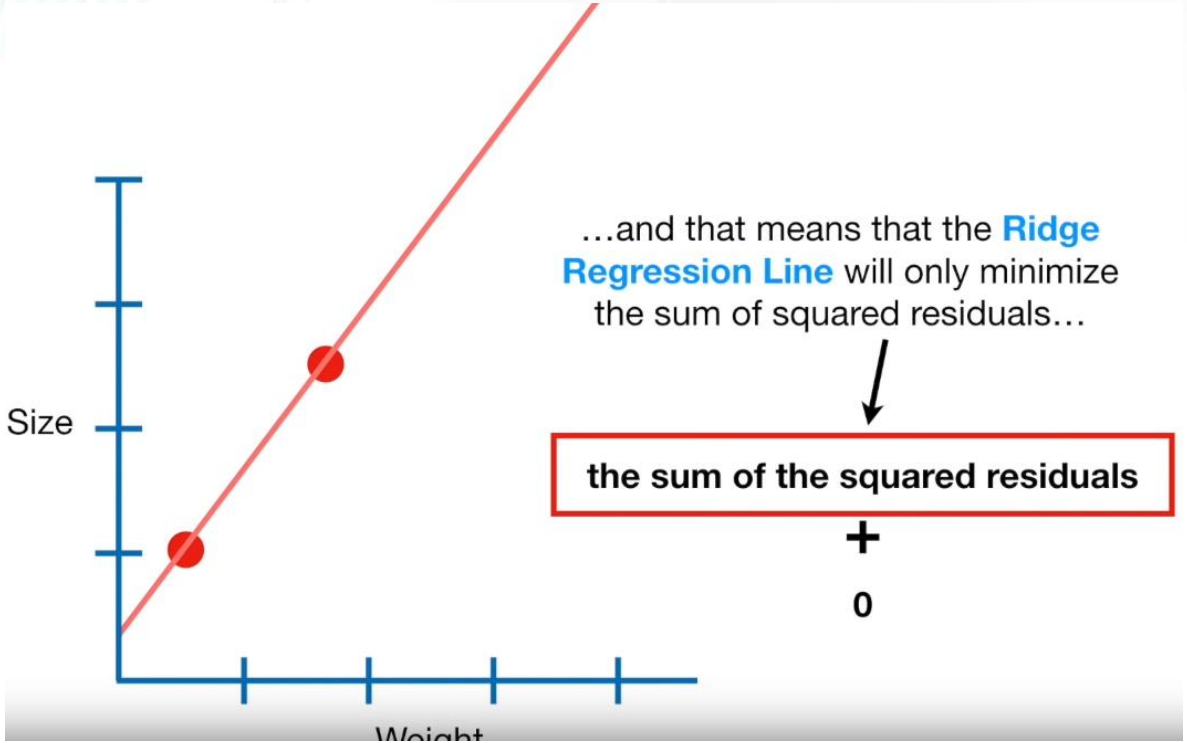


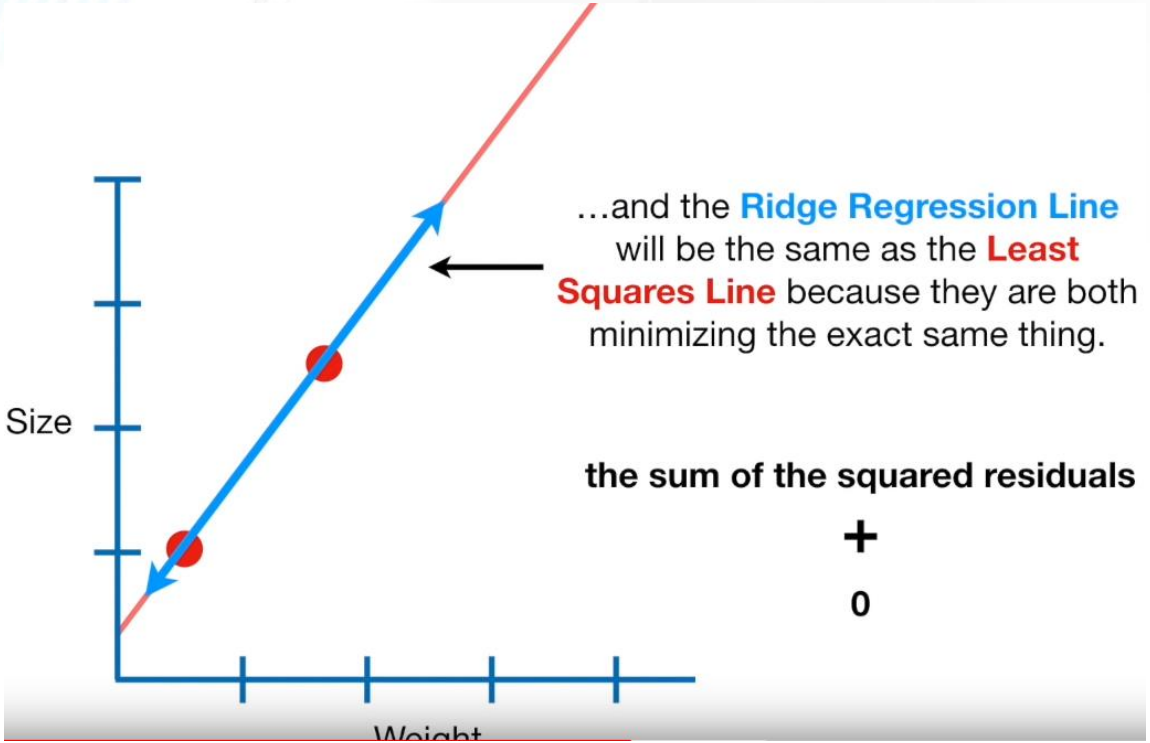
Now let's go back to the equation that **Ridge Regression** tries to minimize…

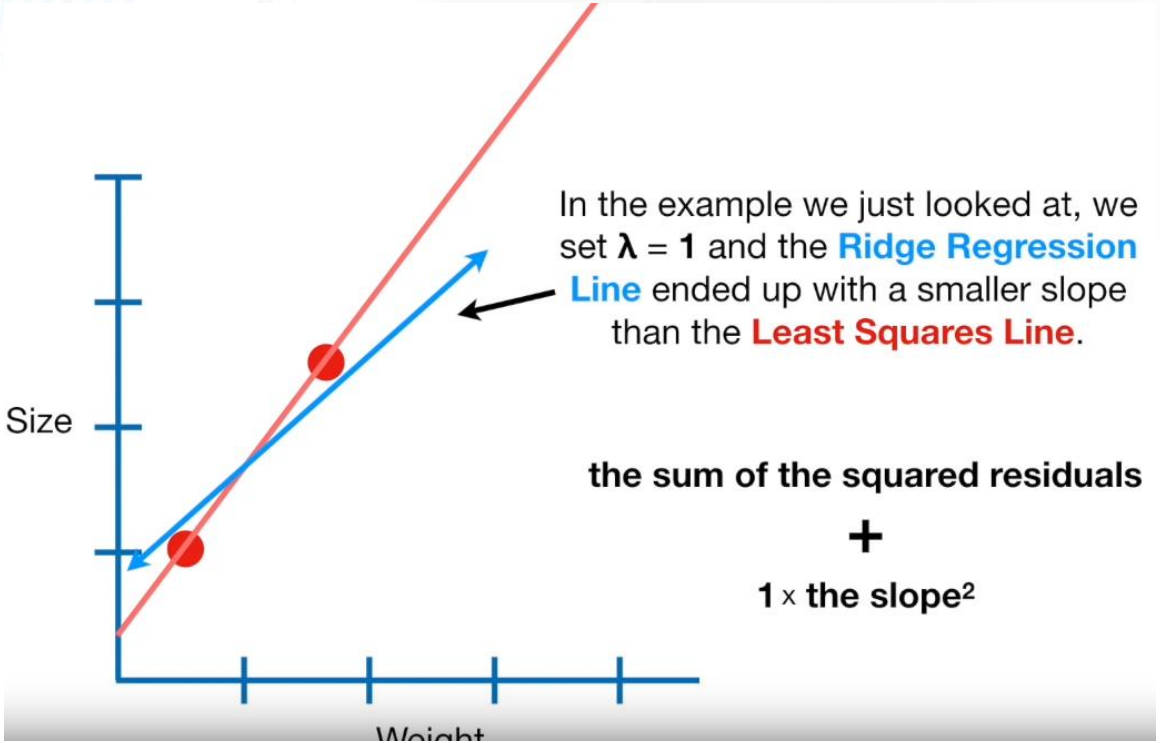
the sum of the squared residuals

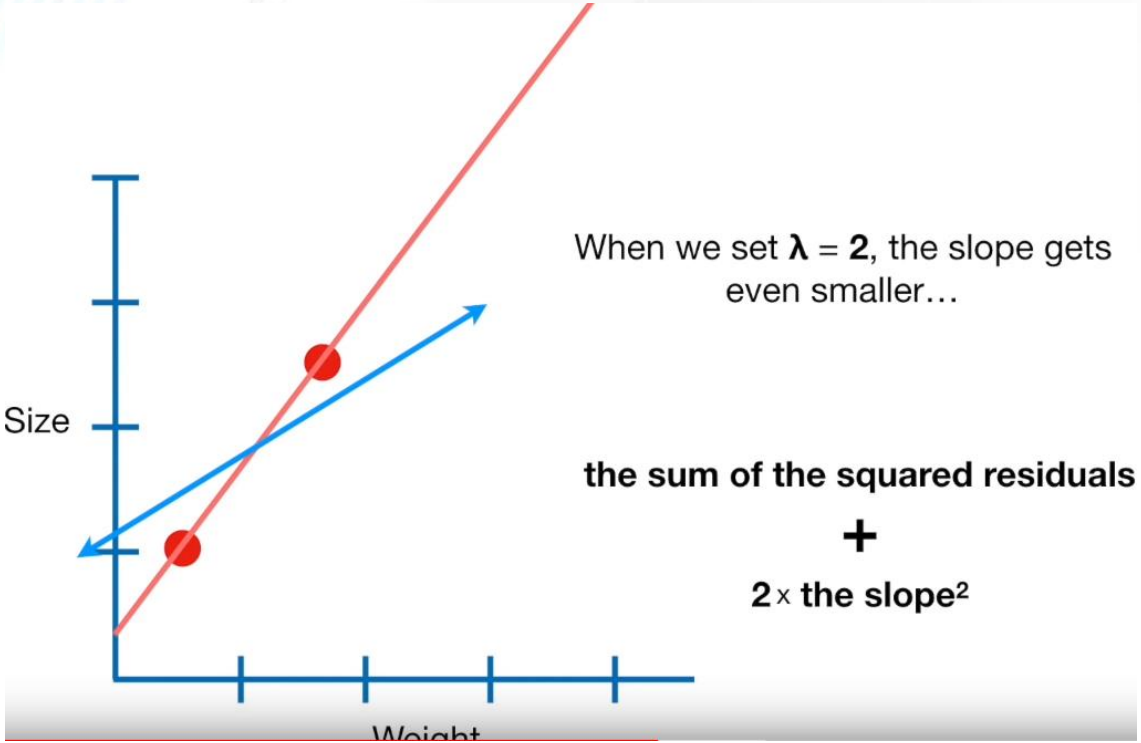
+

λ x the slope²


Size
Weight
…and that means that the Ridge Regression Line will only minimize the sum of squared residuals…
the sum of the squared residuals
+
0


Size
Weight
…and the Ridge Regression Line will be the same as the Least Squares Line because they are both minimizing the exact same thing.
the sum of the squared residuals
+
0


Size
Weight
In the example we just looked at, we set λ = 1 and the Ridge Regression Line ended up with a smaller slope than the Least Squares Line.
the sum of the squared residuals
+
1 x the slope²

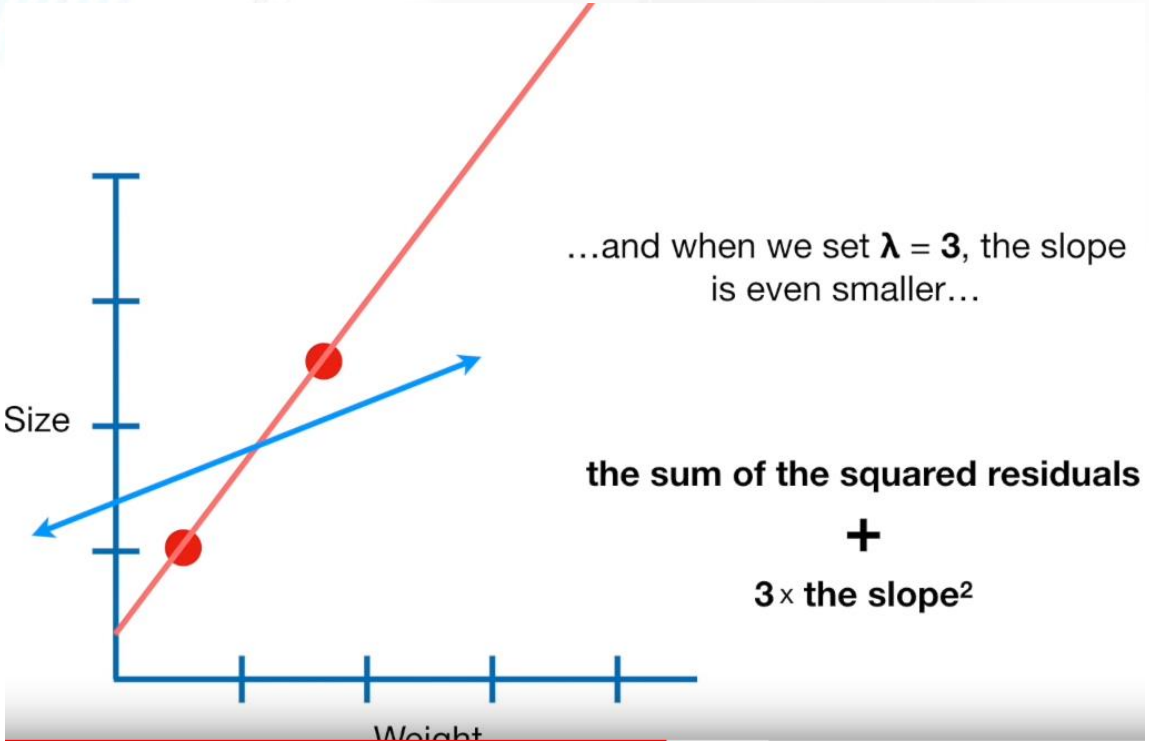


When we set **λ = 2**, the slope gets even smaller…

the sum of the squared residuals

+

2 x the slope²

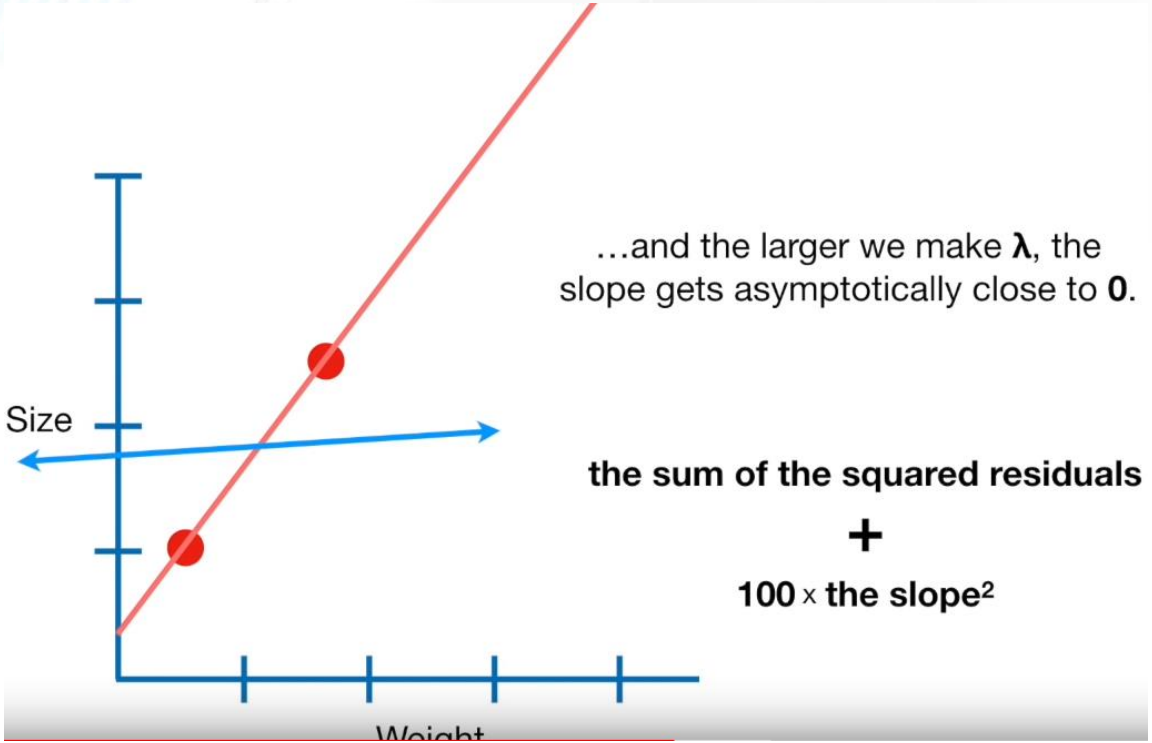


…and when we set **λ = 3**, the slope is even smaller…

the sum of the squared residuals

+

3 x the slope²

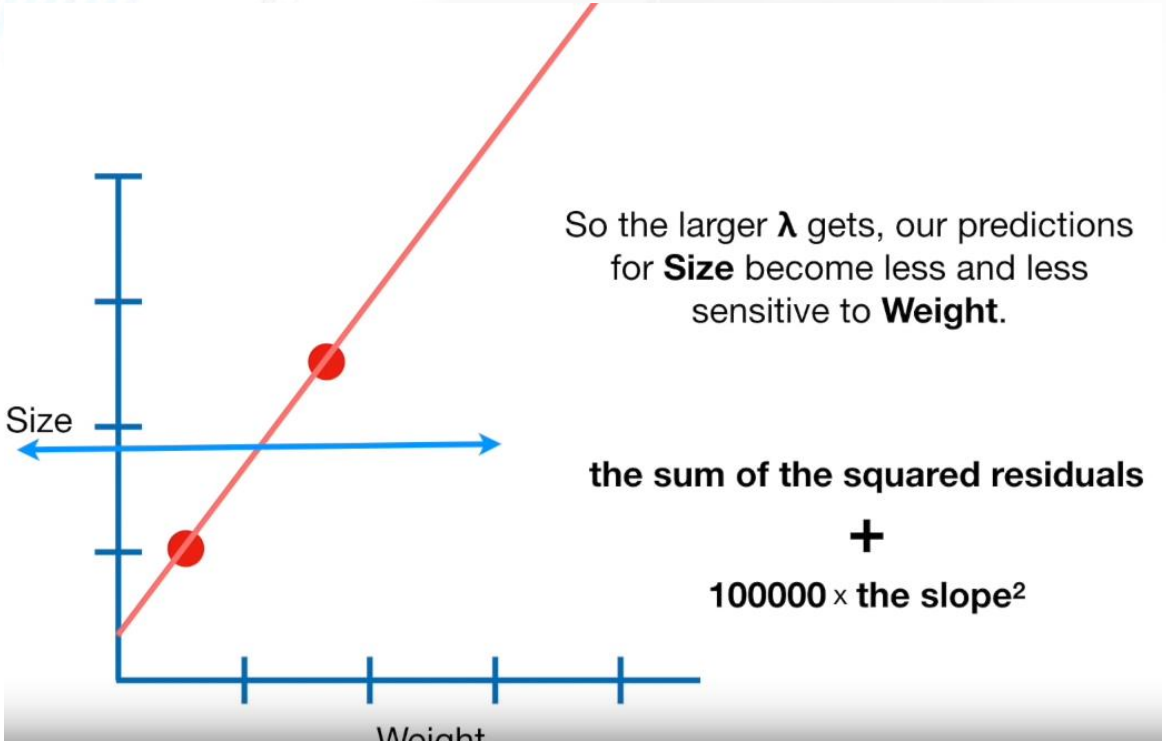


…and the larger we make **λ**, the slope gets asymptotically close to **0**.

the sum of the squared residuals

+

100 x **the slope²**

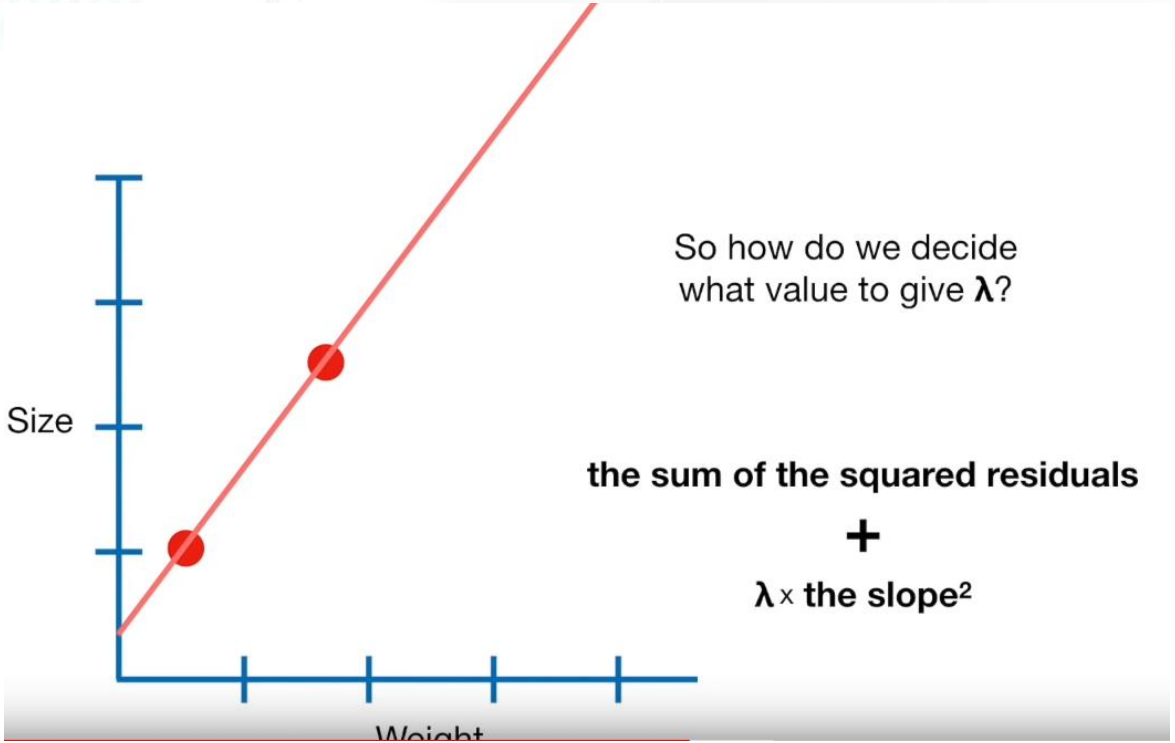


So the larger **λ** gets, our predictions for **Size** become less and less sensitive to **Weight**.

the sum of the squared residuals

+

100000 x the slope²

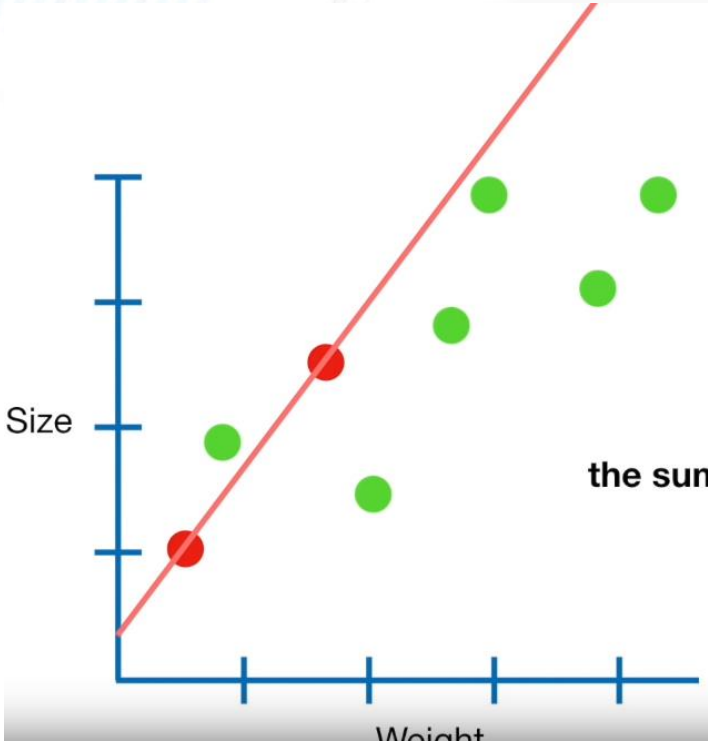


So how do we decide
what value to give **λ**?

the sum of the squared residuals

+

λ x **the slope²**



We just try a bunch of values for **λ** and use **Cross Validation**, typically **10-fold Cross Validation**, to determine which one results in the lowest **Variance**.

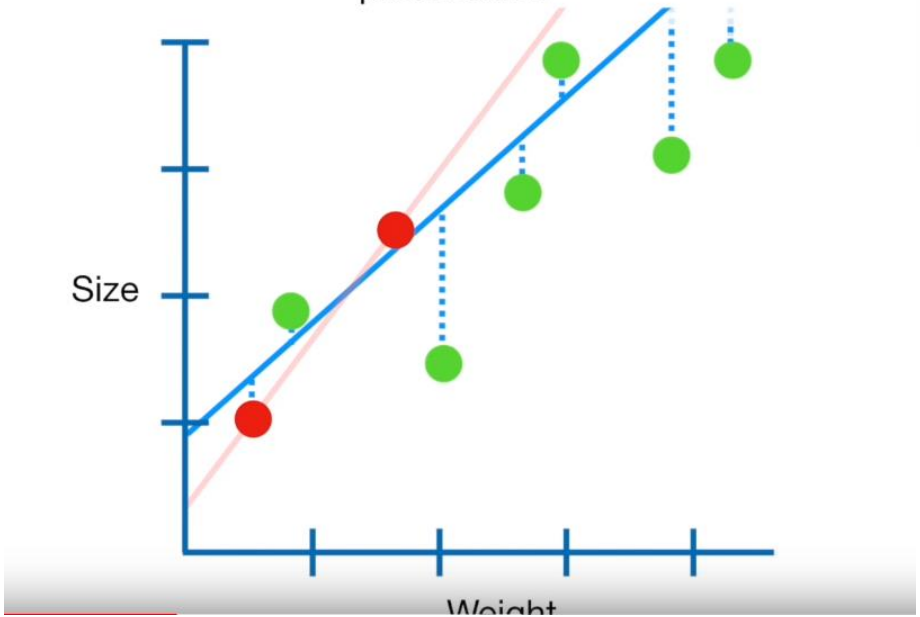
the sum of the squared residuals

+

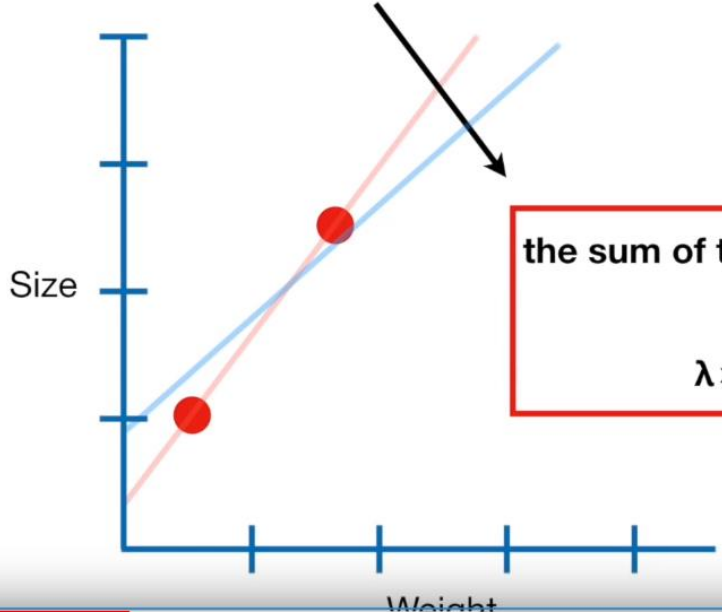
λ x **the slope²**

Lasso Regression

The main idea was that by starting with a slightly worse fit, **Ridge Regression** provided better long term predictions.



Now let's go back to the equation that **Ridge Regression** minimizes…

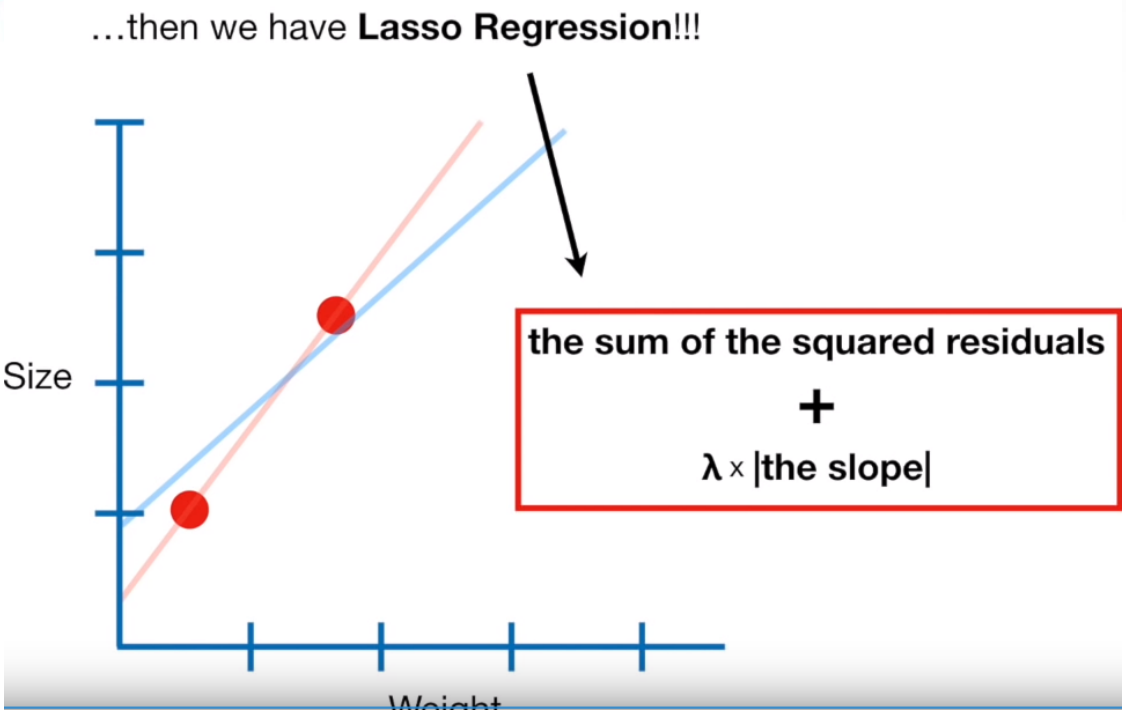


the sum of the squared residuals

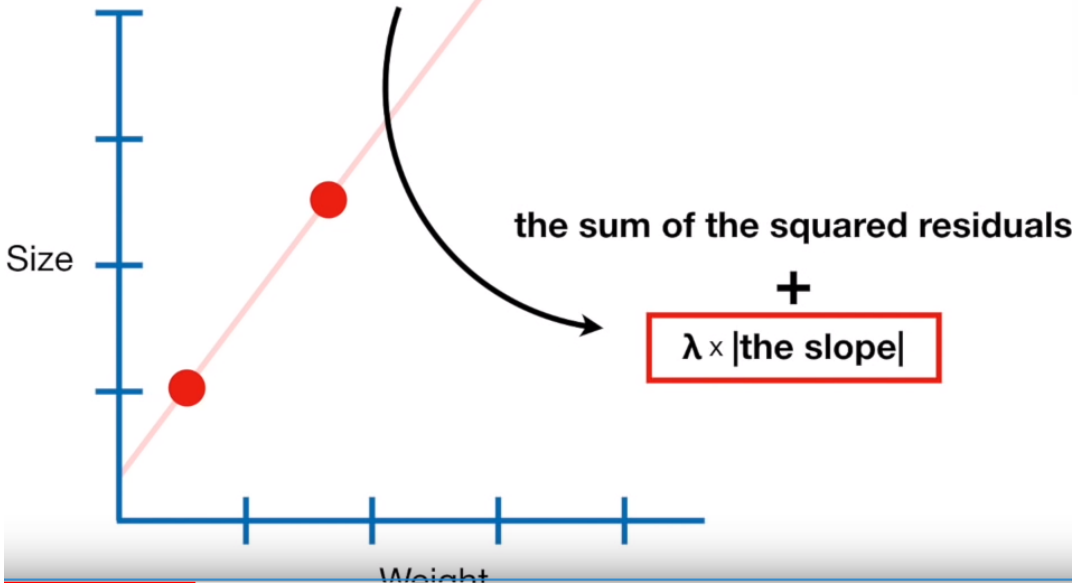
+

λ × the slope²

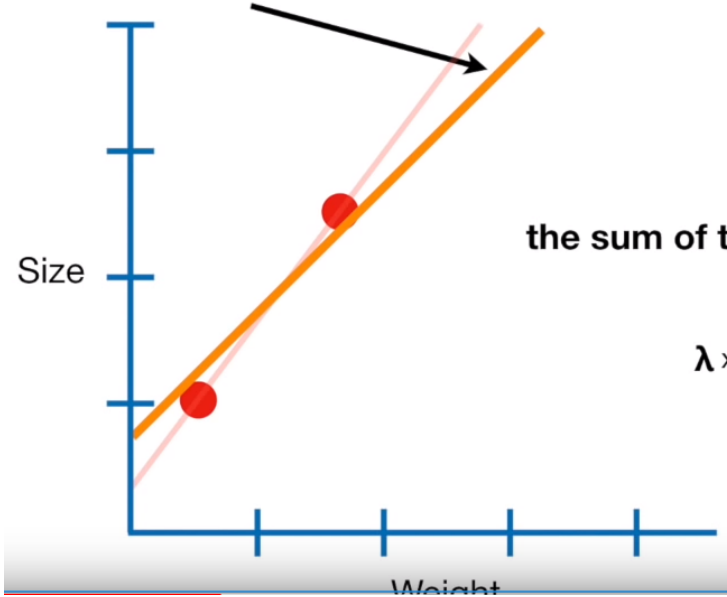
Lasso Regression



NOTE: Just like with **Ridge Regression**, **λ** can be any value from **0** to **positive infinity** and is determined using **Cross Validation**.



Like **Ridge Regression**, **Lasso Regression** (the **Orange Line**) results in a line with a little bit of **Bias**…

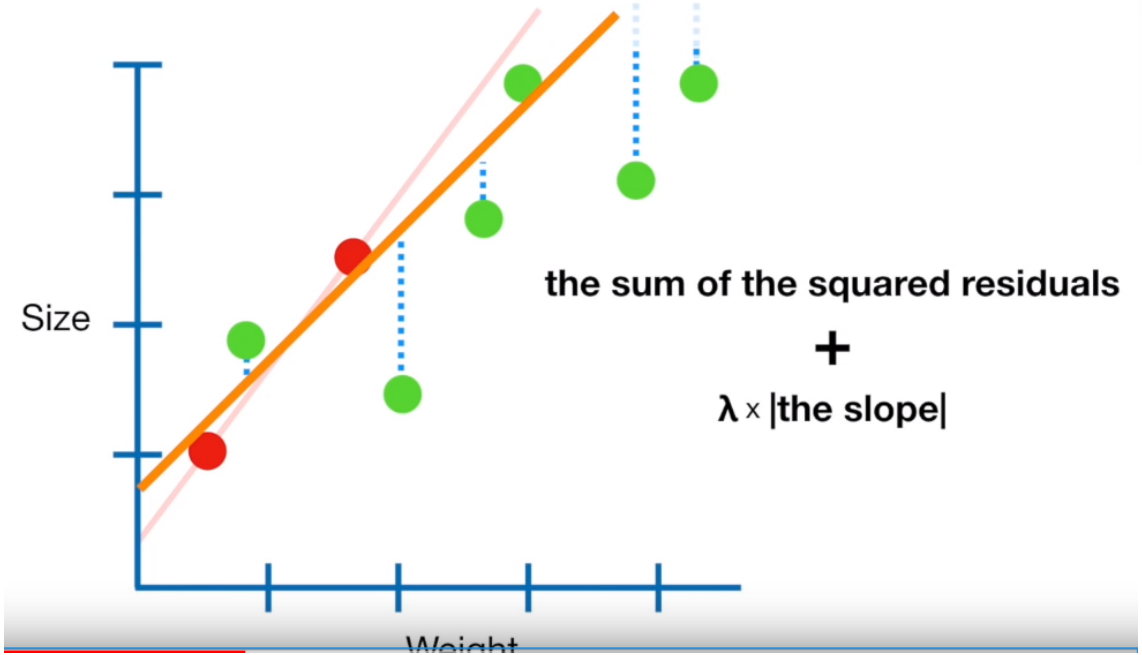


the sum of the squared residuals

+

λ × **|the slope|**

…but less **Variance** than **Least Squares**.



the sum of the squared residuals

+

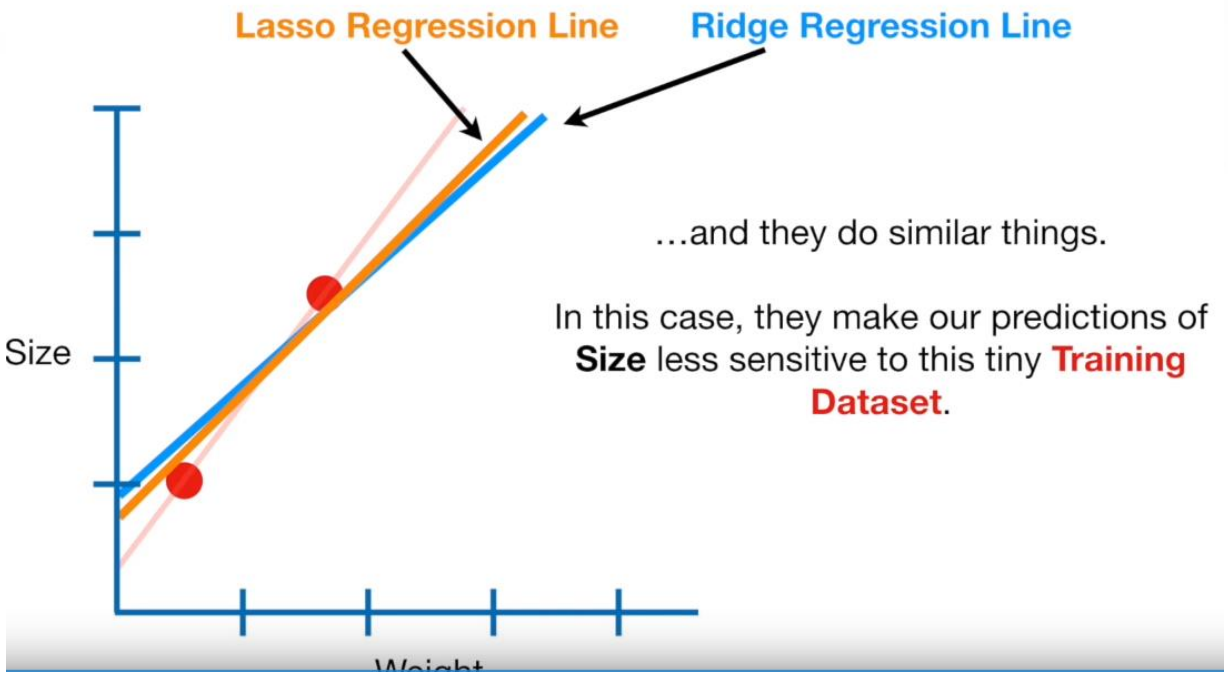
λ x the slope²

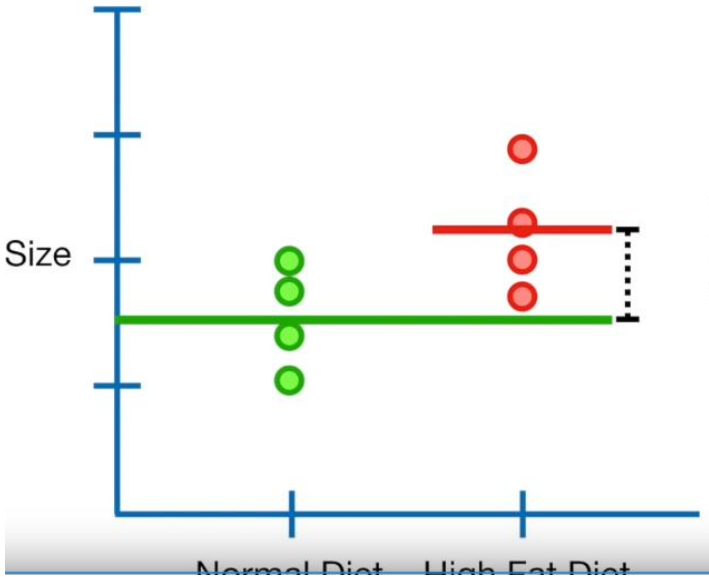
…look very similar….

the sum of the squared residuals

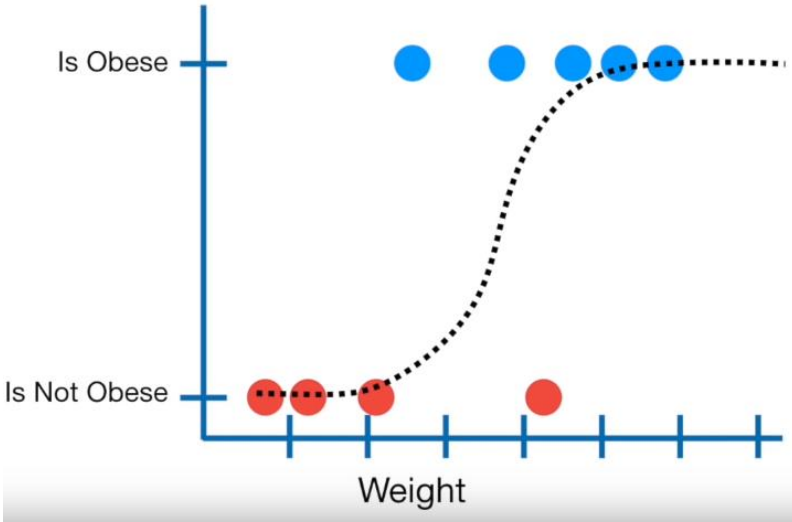
+

λ x |the slope|

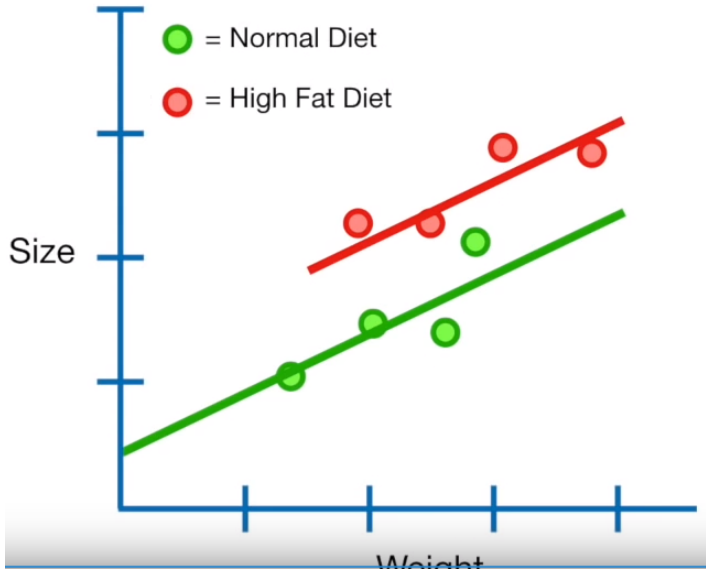

Lasso Regression Line
Ridge Regression Line
…and they do similar things.
In this case, they make our predictions of Size less sensitive to this tiny Training Dataset.
Size
Weight



Both **Ridge** and **Lasso Regression** can be applied in the same contexts, like this situation, where we are using two **Different Diets** to predict **Size**…

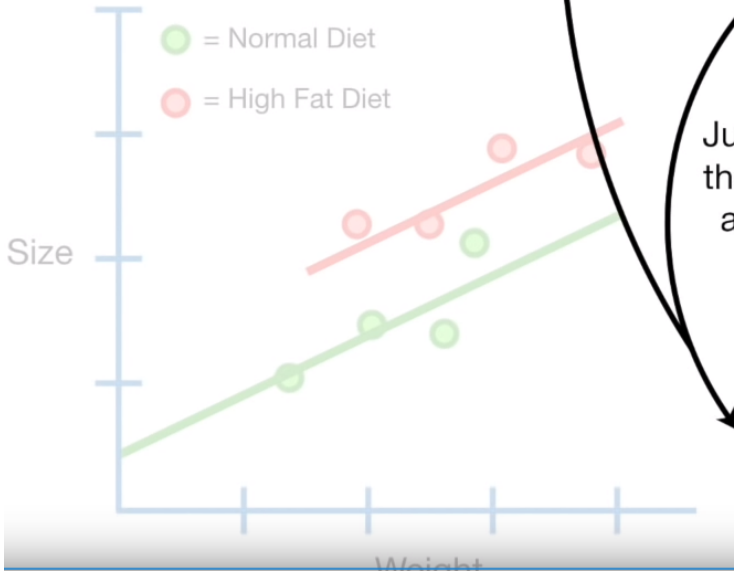


…or in a Logistic Regression setting where we use **Weight** to predict **Obesity**…



…and both **Ridge** and **Lasso Regression** can be applied to complicated models that combine different types of data.

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

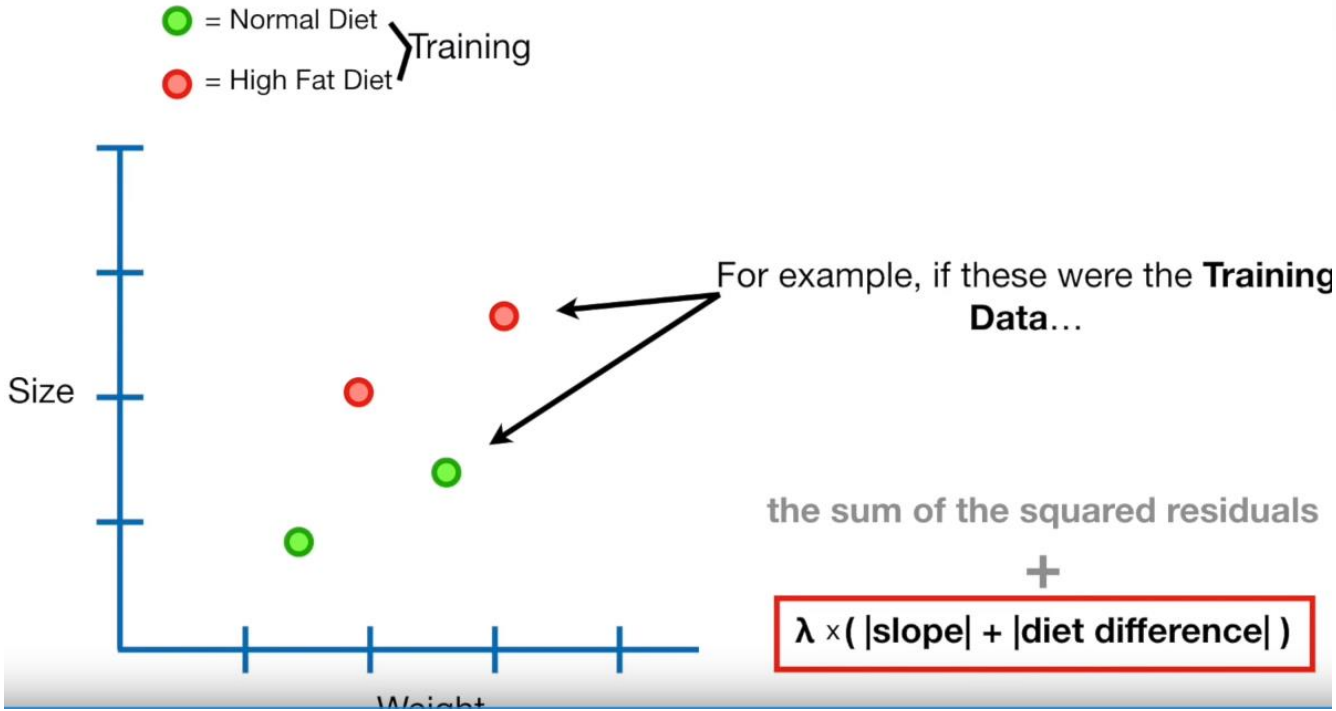


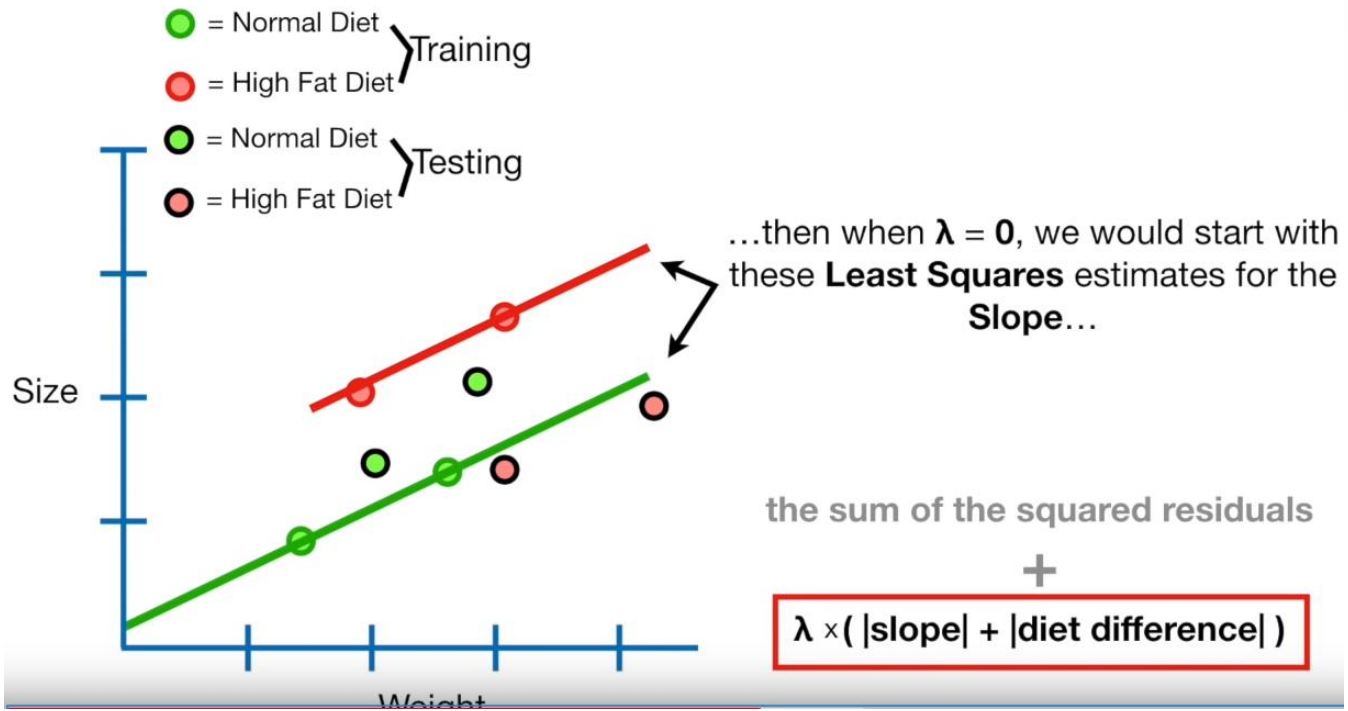
Just like the **Ridge Regression Penalty**, the **Lasso Regression Penalty** contains all of the estimated parameters except for the y-intercept…

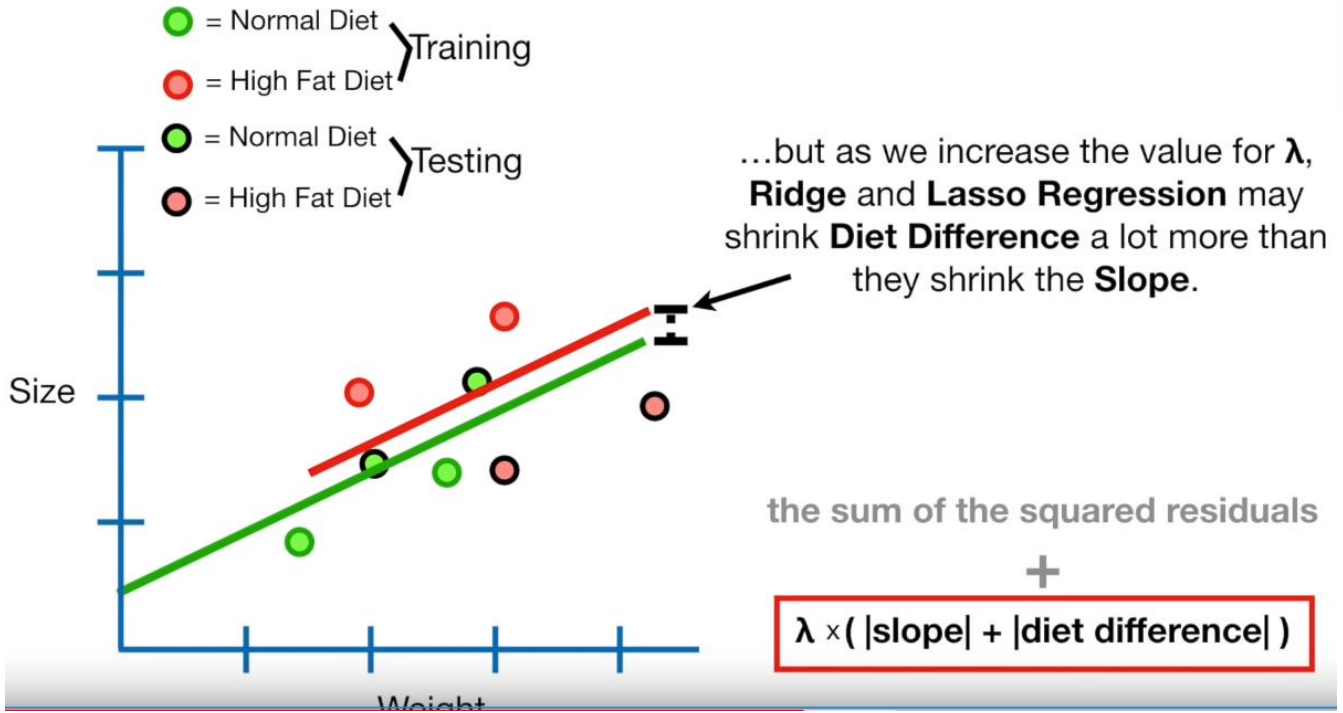
the sum of the squared residuals

+

λ x (|slope| + |diet difference|)


= Normal Diet
= High Fat Diet
Training
For example, if these were the Training Data…
Size
Weight
the sum of the squared residuals
+
λ × (|slope| + |diet difference|)

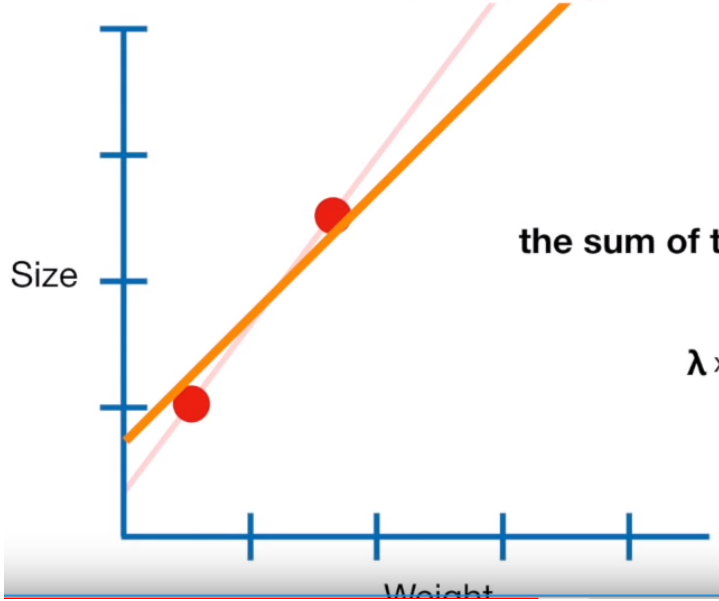

= Normal Diet
= High Fat Diet
Training
= Normal Diet
= High Fat Diet
Testing
Size
Weight
…then when λ = 0, we would start with these Least Squares estimates for the Slope…
the sum of the squared residuals
+
λ x (|slope| + |diet difference|)


= Normal Diet
= High Fat Diet
Training
= Normal Diet
= High Fat Diet
Testing
Size
Weight
…but as we increase the value for λ, Ridge and Lasso Regression may shrink Diet Difference a lot more than they shrink the Slope.
the sum of the squared residuals
+
λ x (|slope| + |diet difference|)

OK, we've seen how **Ridge** and **Lasso Regression** are similar.

Now let's talk about the big difference between them.

To see what makes **Lasso Regression** different from **Ridge Regression**, let's go back to the two sample **Training Data**.

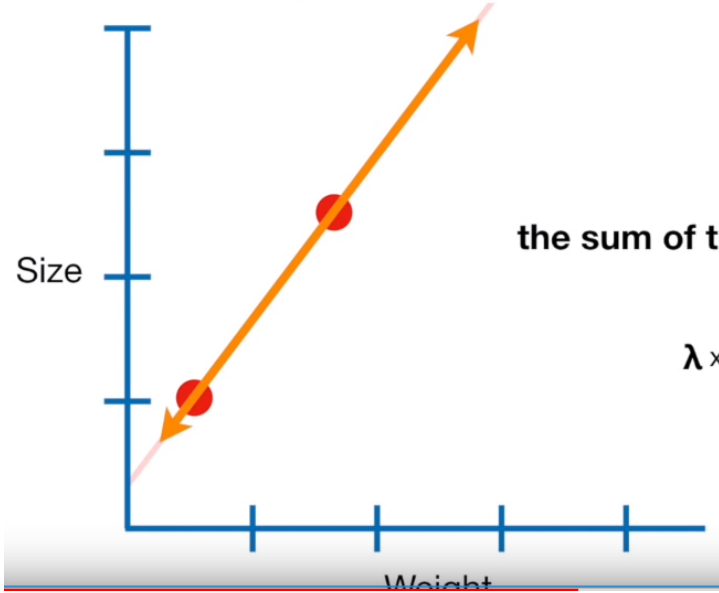


the sum of the squared residuals

+

$\lambda \times$ |the slope|

When **λ = 0**, then the **Lasso Regression Line** will be the same as the **Least Squares Line**…

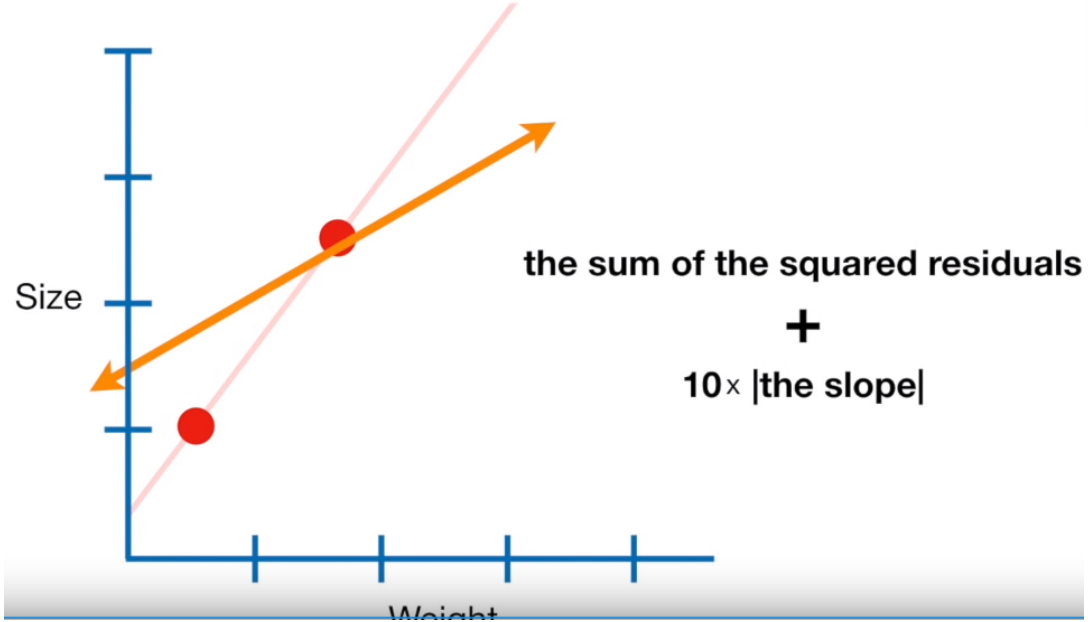


the sum of the squared residuals

+

λ × |the slope|

As **λ** increases in value, the slope gets smaller…

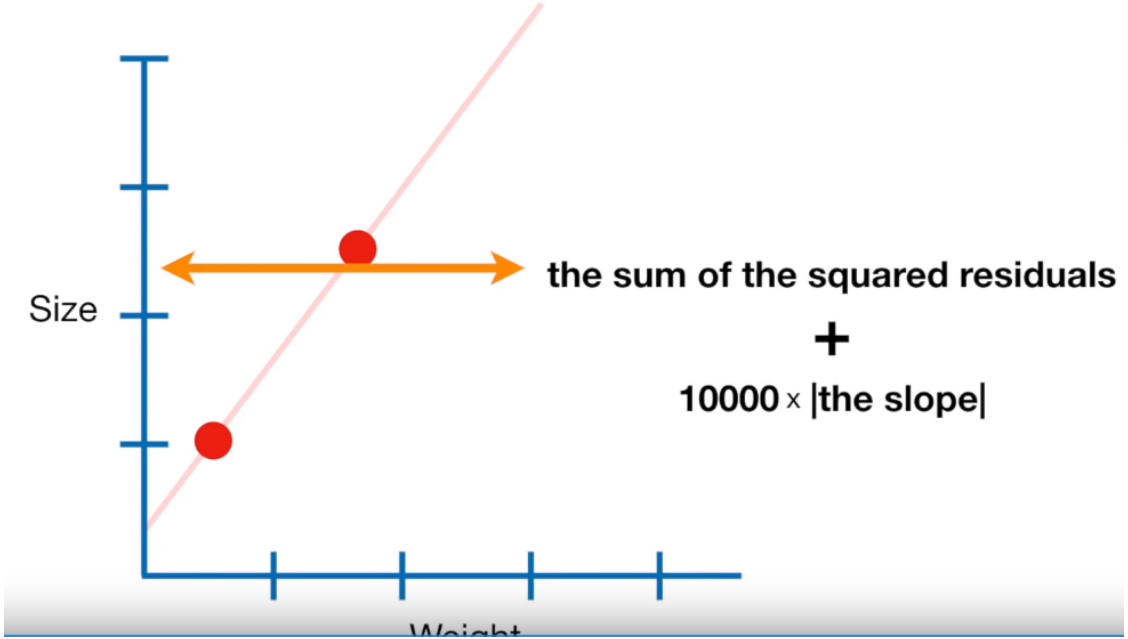


the sum of the squared residuals

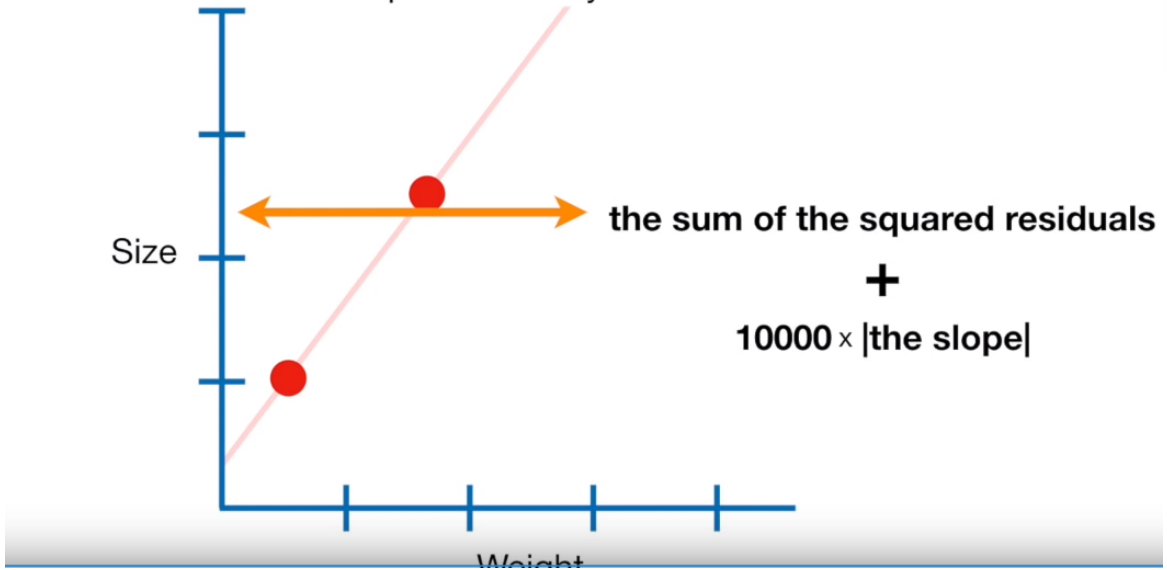
+

10 x **|the slope|**

…until the slope = **0**.



The big difference between **Ridge** and **Lasso Regression** is that **Ridge Regression** can only shrink the slope asymptotically close to **0** while **Lasso Regression** can shrink the slope all the way to **0**.



Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**

To appreciate this difference, let's look a big, huge, crazy equation…

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**

When we apply **Ridge Regression** to this equation, we find the minimal sum of the squared residuals plus the **Ridge Regression Penalty**...

↓

$$\lambda \times (\text{slope}^2 + \text{diet difference}^2 + \text{astrological offset}^2 + \text{airspeed scalar}^2)$$

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**

…and these parameters might shrink a lot, but they will never be equal to 0.

…and the larger we make **λ**…

λ (slope^2 + diet difference^2 + astrological offset^2 + airspeed scalar^2)

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**

In contrast, with **Lasso Regression**…

$$\lambda \times (|\text{slope}| + |\text{diet difference}| + |\text{astrological offset}| + |\text{airspeed scalar}|)$$

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**

…when we increase the value for **λ**…

…and these parameters will go all the way to 0…

λ (|slope| + |diet difference| + |astrological offset| + |airspeed scalar|)

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

…and we're left with a way to predict **Size** that only includes **Weight** and **Diet**…

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

~~+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**~~

Since **Lasso Regression** can exclude useless variables from equations, it is a little better than **Ridge Regression** at reducing the **Variance** in models that contain a lot of useless variables.

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

+ astrological offset x **Sign** + airspeed scalar x **Airspeed of Swallow**

In contrast, **Ridge Regression** tends to do a little better when most variables are useful.

Ridge Regression is very similar to… ⟶ **the sum of the squared residuals + λ x the slope²**

…Lasso Regression ⟶ **the sum of the squared residuals + λ x |the slope|**

The superficial difference is that **Ridge Regression** squares the variables…

the sum of the squared residuals

+

λ x the slope²

Size = y-intercept + slope x **Weight** + diet difference x **High Fat Diet**

But the big difference is that **Lasso Regression** can exclude useless variables from equations.

This makes the final equation simpler and easier to interpret.