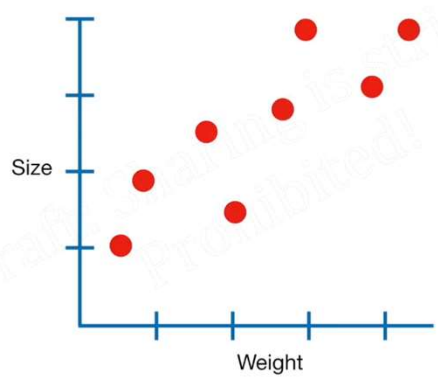
REGULARIZATION

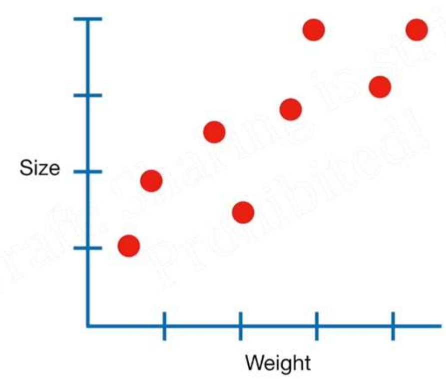
- Ridge Regression
- Lasso regression
- Elastic-Net

RIDGE REGRESSION

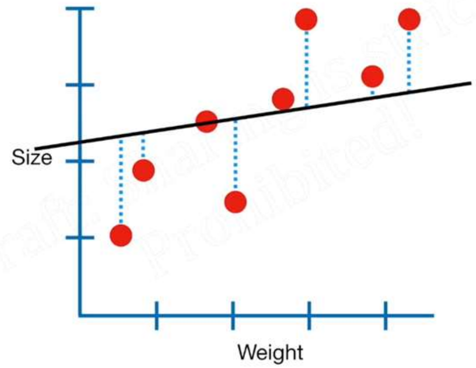
Let's start by collecting **Weight** and **Size** measurements from a bunch of mice…



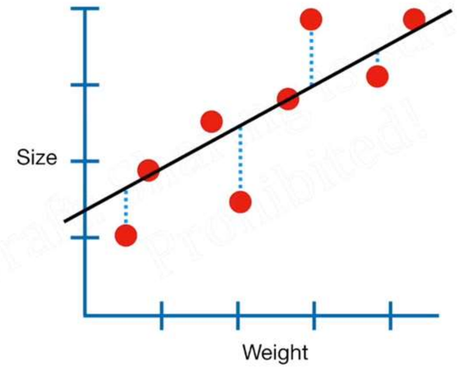
Since these data look relatively linear, we will use **Linear Regression**, AKA **Least Squares**, to model the relationship between **Weight** and **Size**.



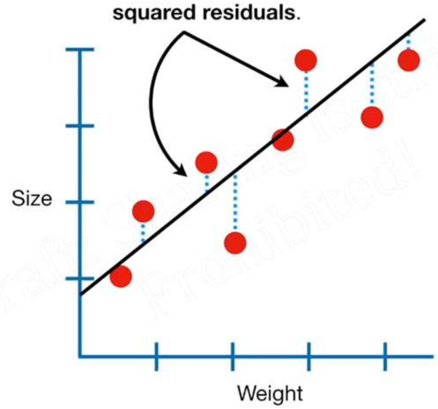
So we'll fit a line to the data using
Least Squares.



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

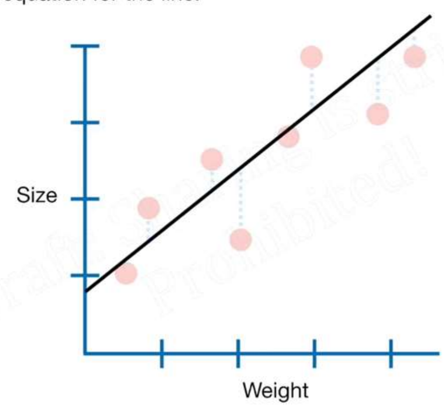


In other words, we find the line that results in the **minimum sum of squared residuals**.

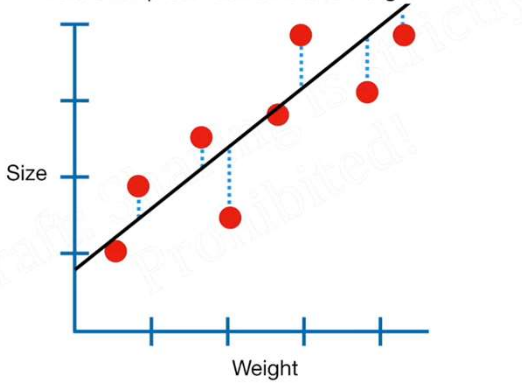


Ultimately, we end up with this equation for the line:

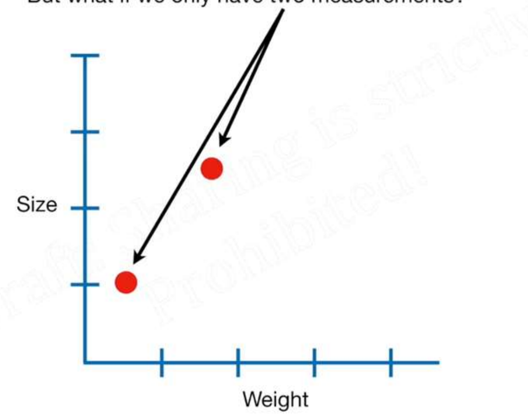
Size = 0.9 + 0.75 × **Weight**



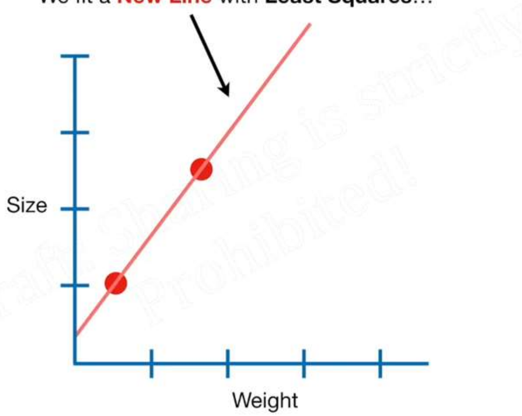
When we have a lot of measurements, we can be fairly confident that the **Least Squares** line accurately reflects the relationship between **Size** and **Weight**.

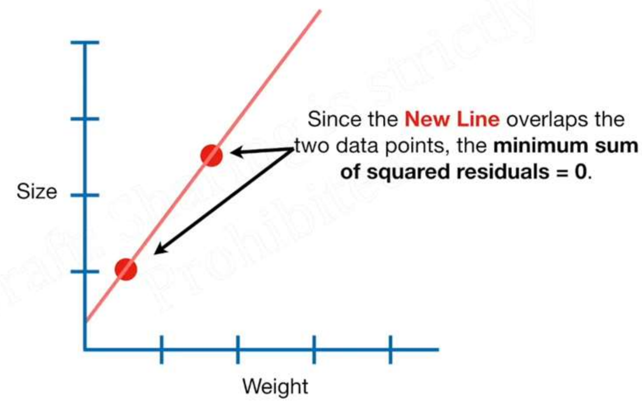


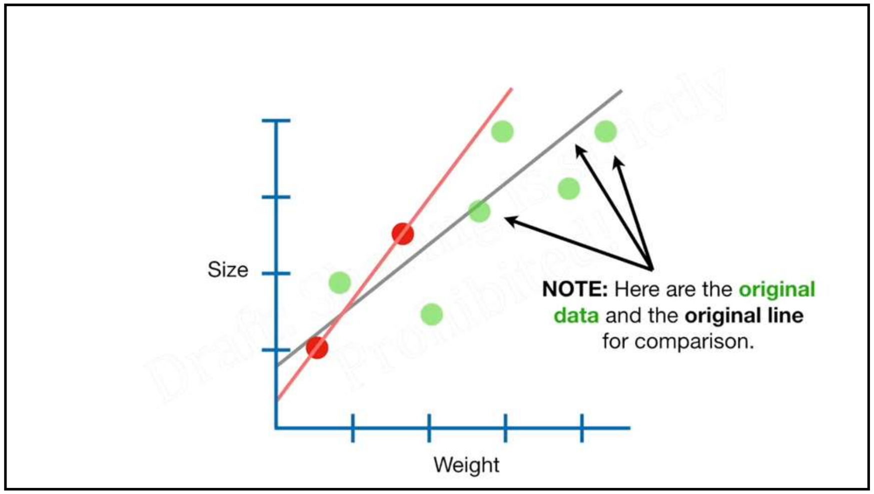
But what if we only have two measurements?

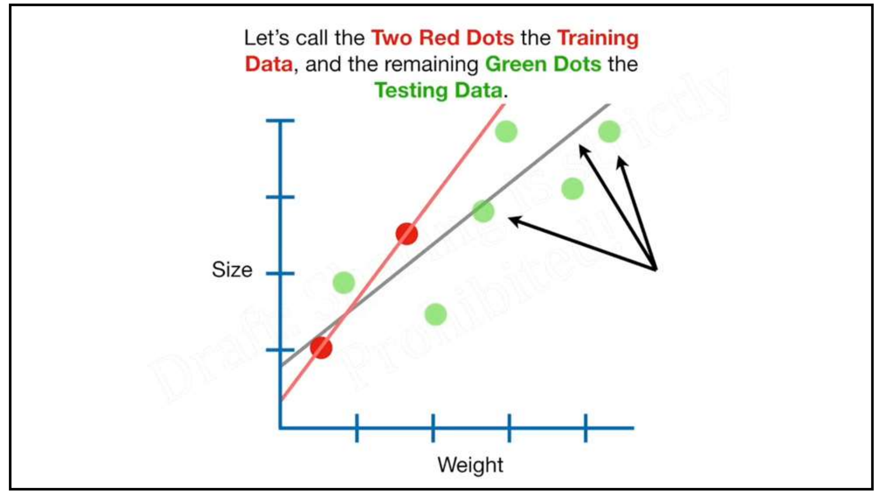


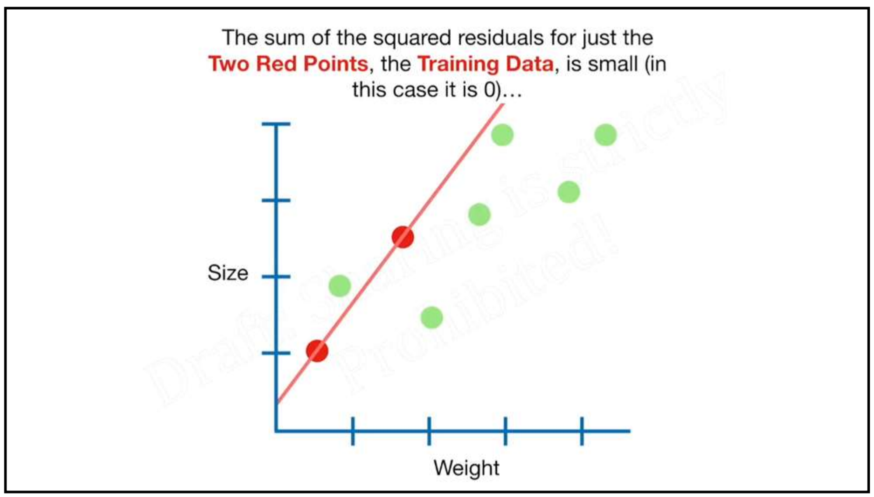
We fit a **New Line** with **Least Squares**…

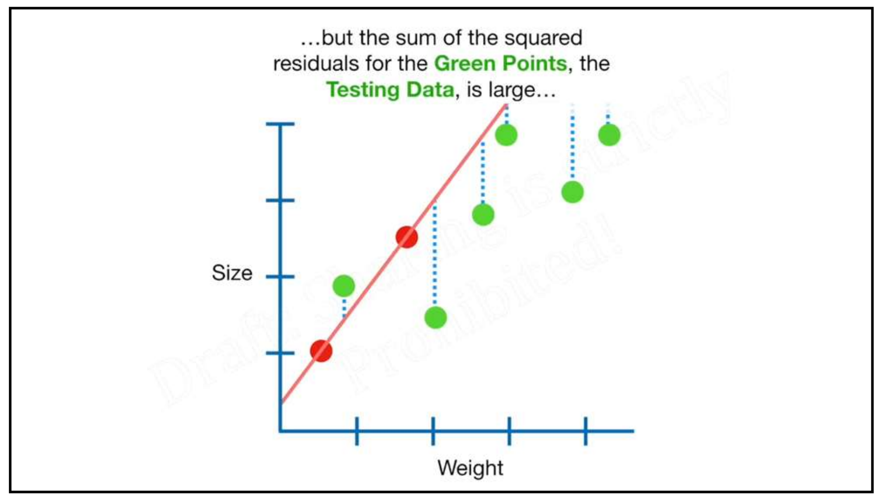


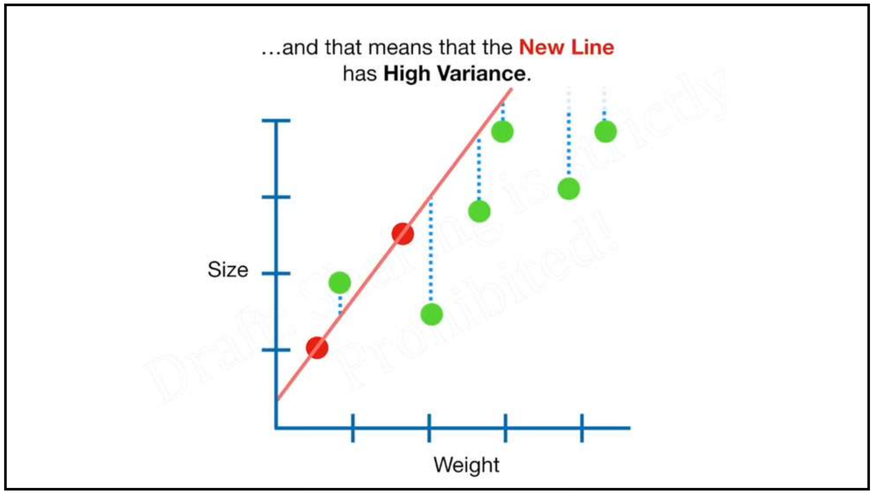


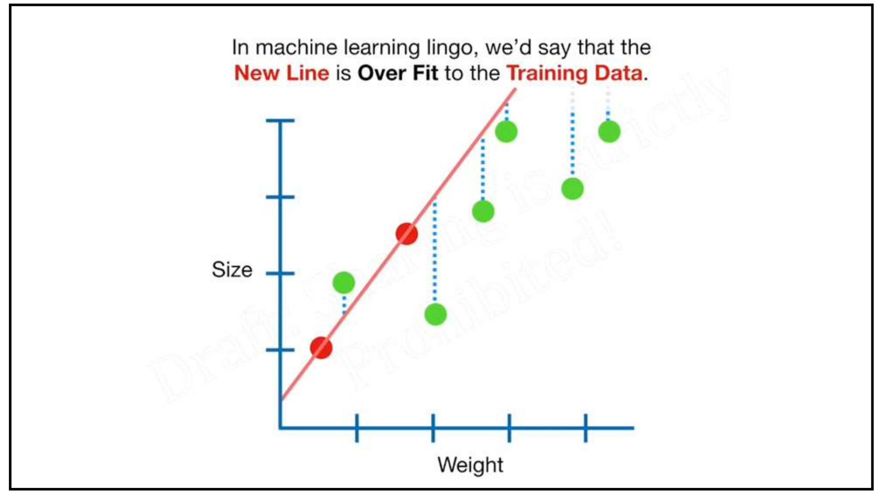

Size
Weight
NOTE: Here are the original data and the original line for comparison.

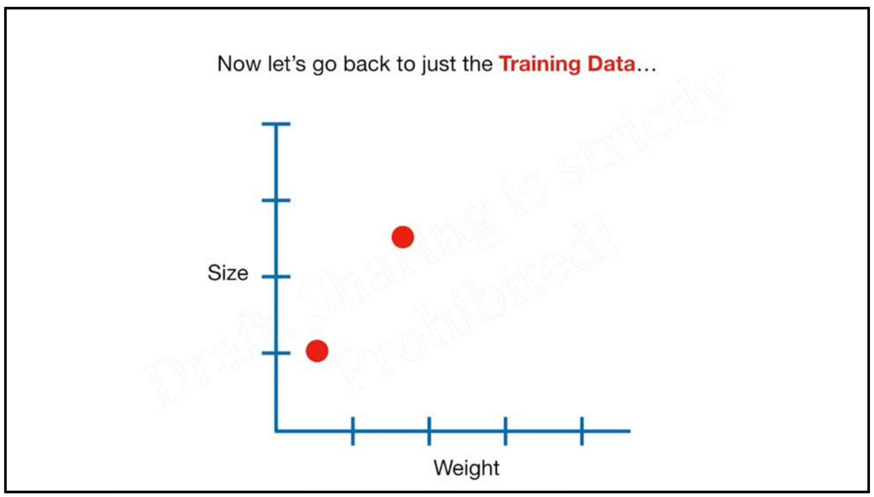

Let's call the Two Red Dots the Training Data, and the remaining Green Dots the Testing Data.
Size
Weight

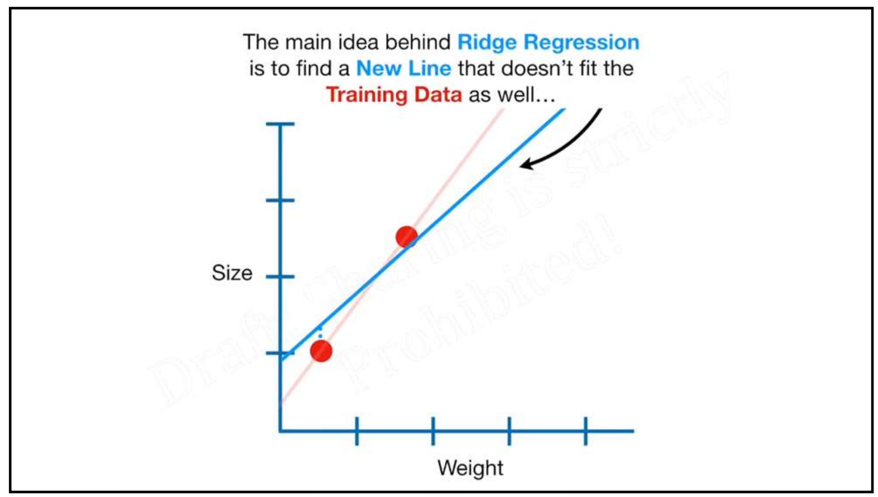

The sum of the squared residuals for just the Two Red Points, the Training Data, is small (in this case it is 0)…
Size
Weight

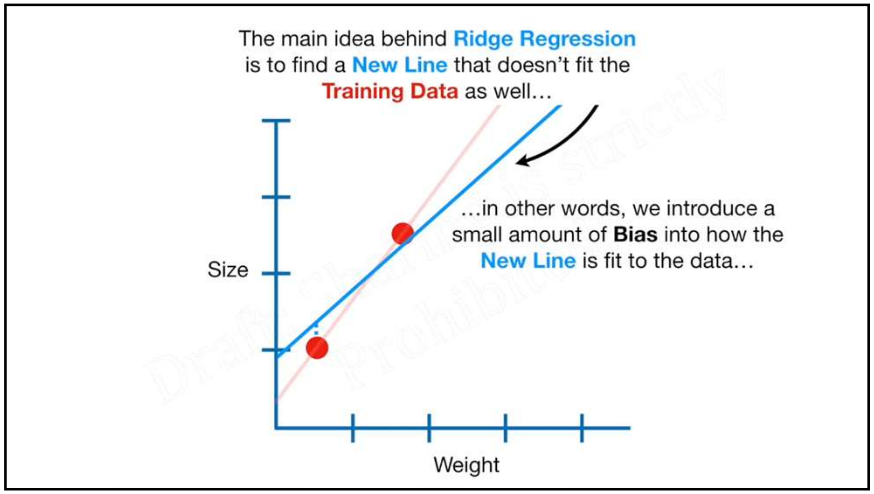

…but the sum of the squared residuals for the Green Points, the Testing Data, is large…
Size
Weight

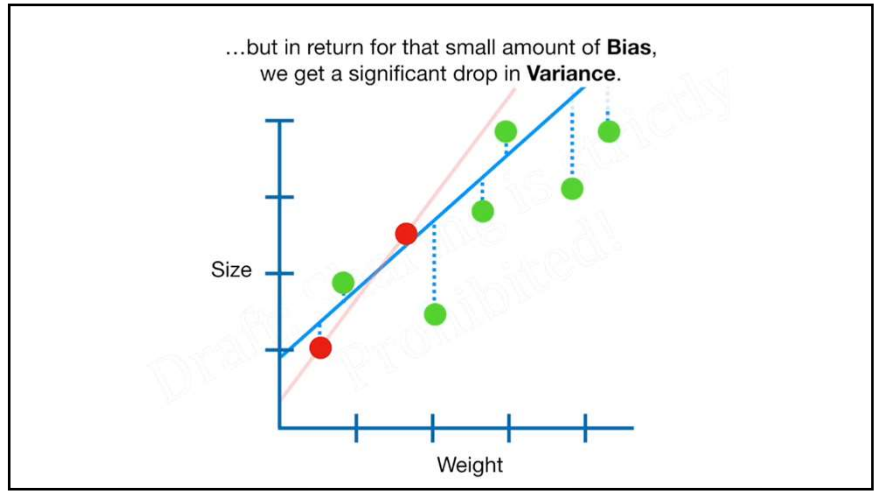

…and that means that the New Line
has High Variance.
Size
Weight

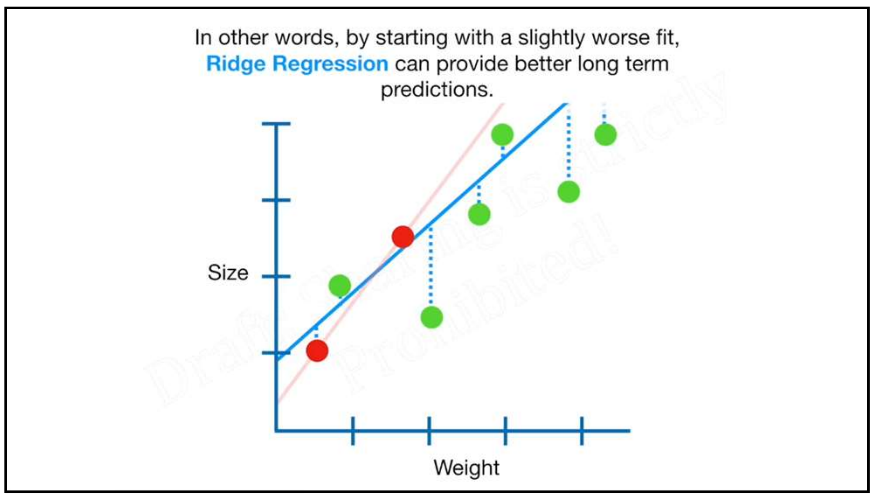

In machine learning lingo, we'd say that the
New Line is Over Fit to the Training Data.
Size
Weight

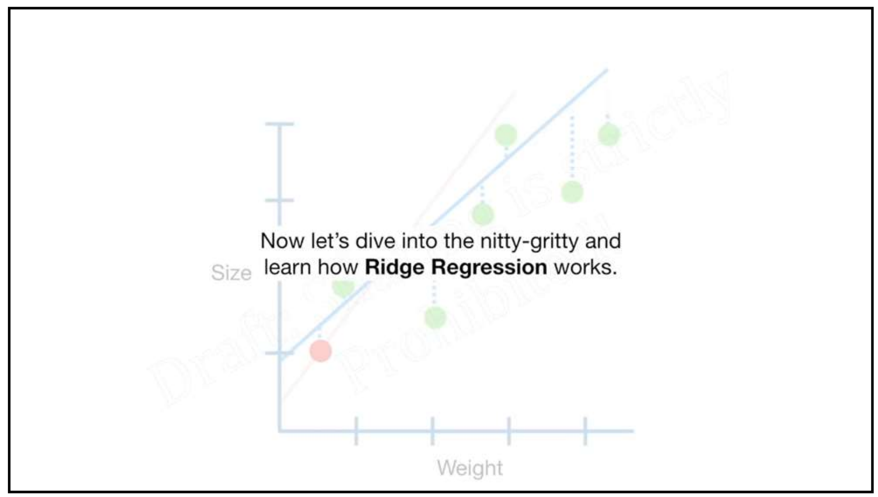

Now let's go back to just the Training Data…
Size
Weight

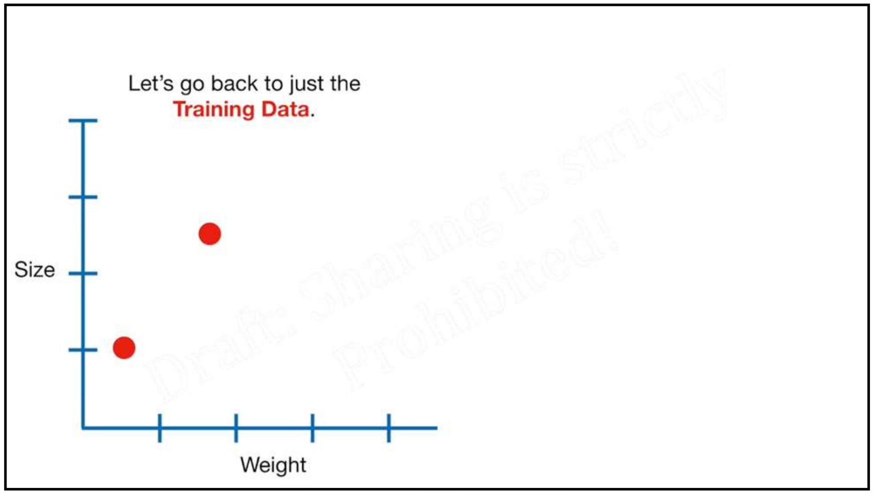

The main idea behind Ridge Regression
is to find a New Line that doesn't fit the
Training Data as well…
Size
Weight

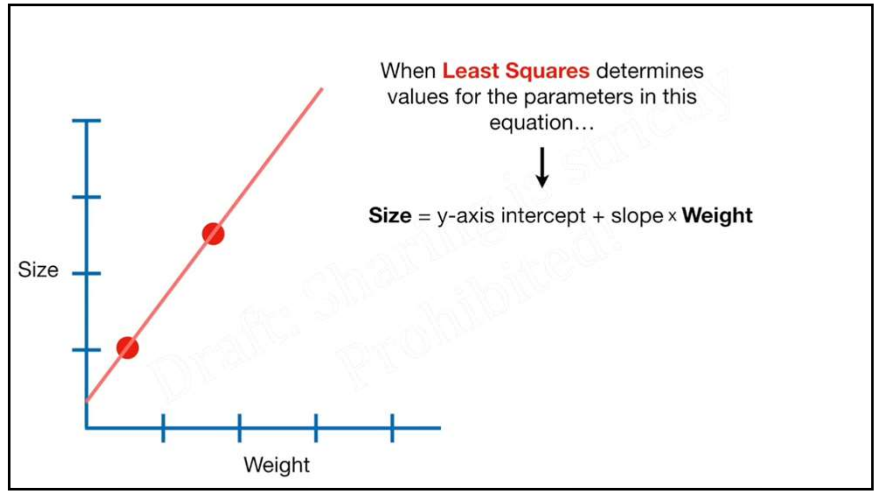

The main idea behind Ridge Regression
is to find a New Line that doesn't fit the
Training Data as well…
…in other words, we introduce a
small amount of Bias into how the
New Line is fit to the data…
Size
Weight

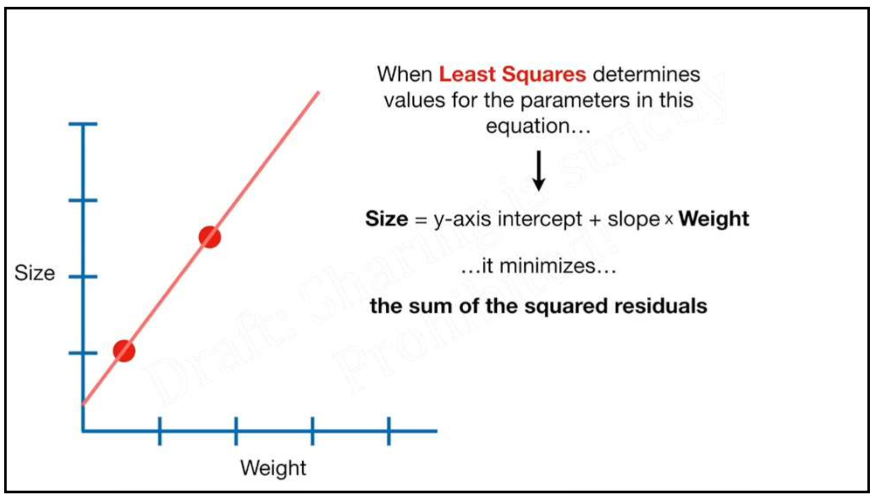

…but in return for that small amount of Bias,
we get a significant drop in Variance.
Size
Weight

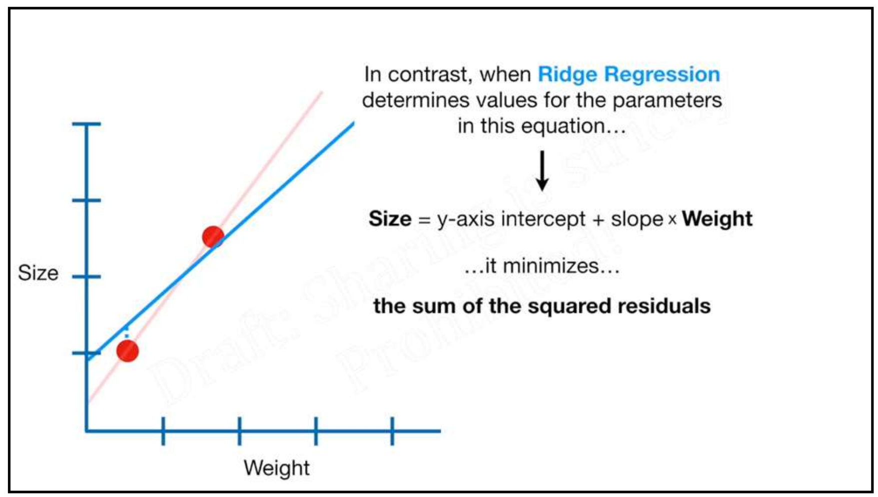

In other words, by starting with a slightly worse fit,
Ridge Regression can provide better long term
predictions.
Size
Weight

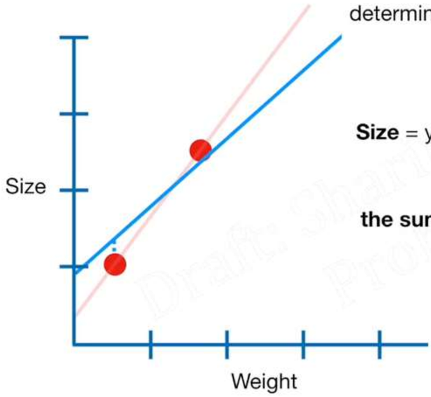

Now let's dive into the nitty-gritty and
learn how Ridge Regression works.
Size
Weight


Let's go back to just the
Training Data.
Size
Weight


When Least Squares determines
values for the parameters in this
equation…
Size = y-axis intercept + slope x Weight
Size
Weight


When Least Squares determines
values for the parameters in this
equation…
Size = y-axis intercept + slope x Weight
…it minimizes…
the sum of the squared residuals
Size
Weight


In contrast, when Ridge Regression
determines values for the parameters
in this equation…
Size = y-axis intercept + slope x Weight
…it minimizes…
the sum of the squared residuals
Size
Weight



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

↓

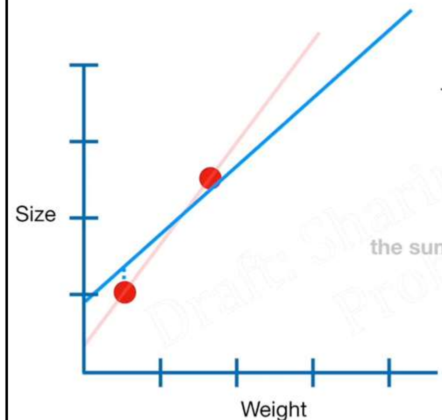
Size = y-axis intercept + slope × **Weight**

…it minimizes…

the sum of the squared residuals

+

λ × the slope^2

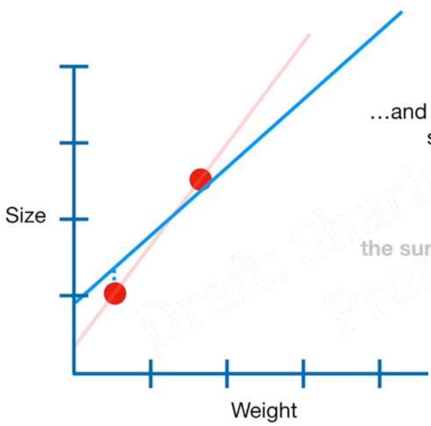


This part of the equation adds a penalty to the traditional **Least Squares** method…

the sum of the squared residuals

+

λ × $\boxed{\text{the slope}^2}$

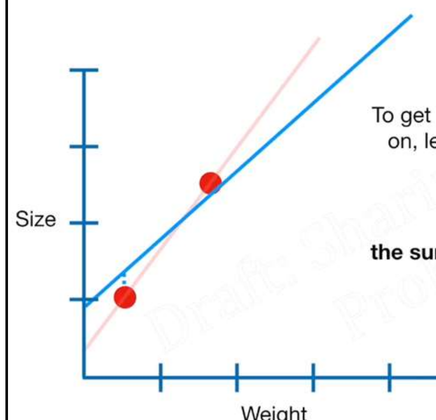


…and **λ** (lambda) determines how severe that penalty is.

the sum of the squared residuals

+

$\boxed{\lambda}$ × the slope^2

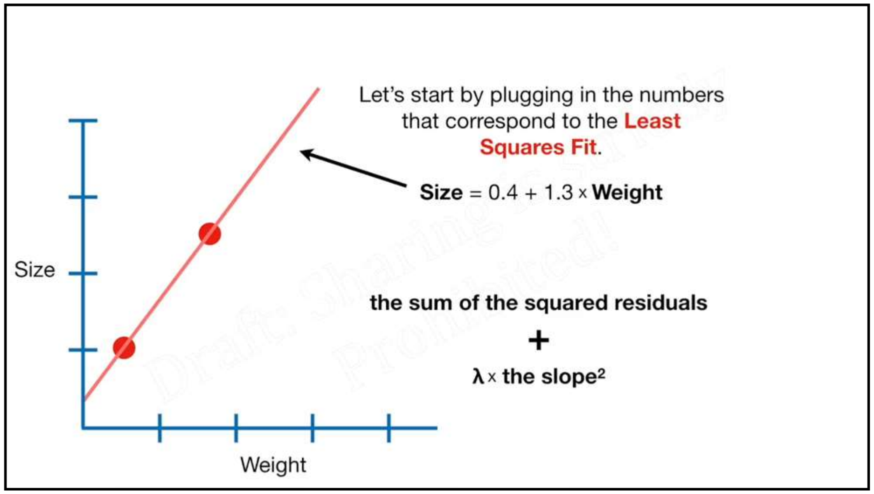


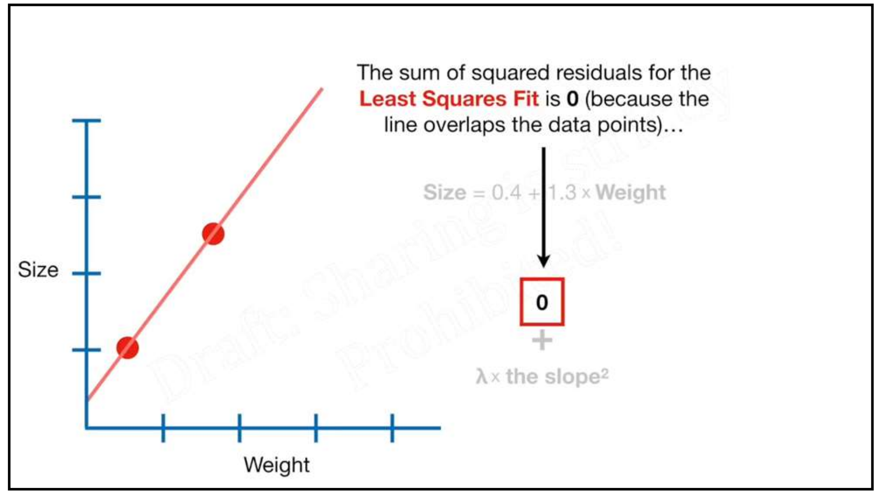
To get a better idea of what's going on, lets plug in some numbers!!!

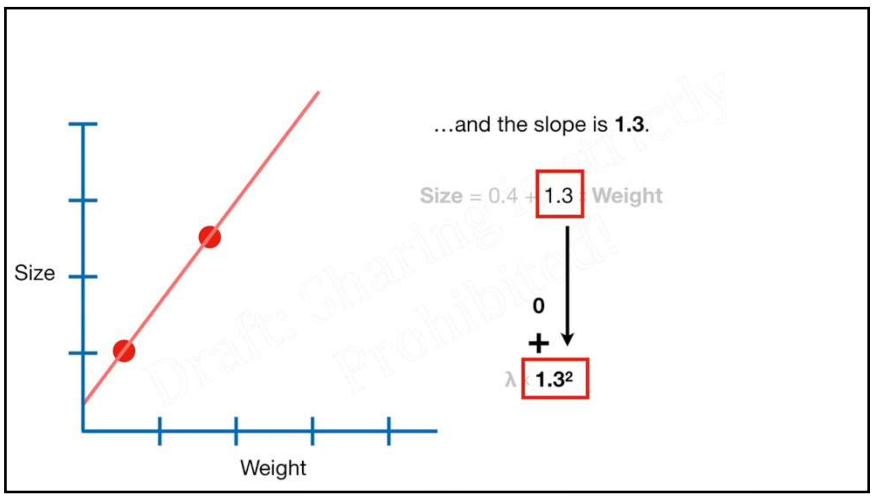
the sum of the squared residuals

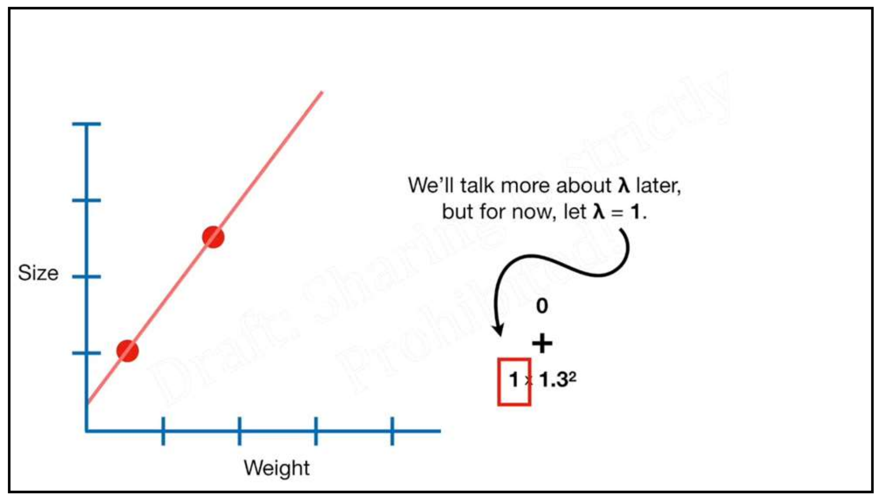
+

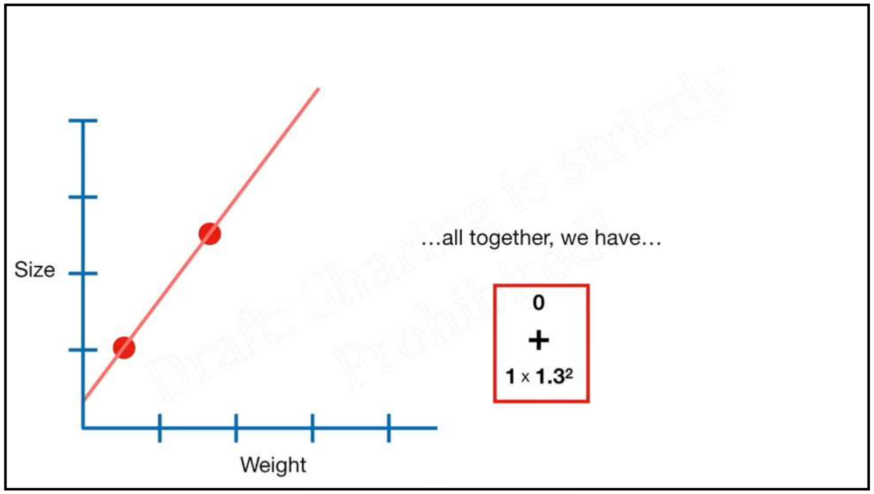
λ × the slope^2

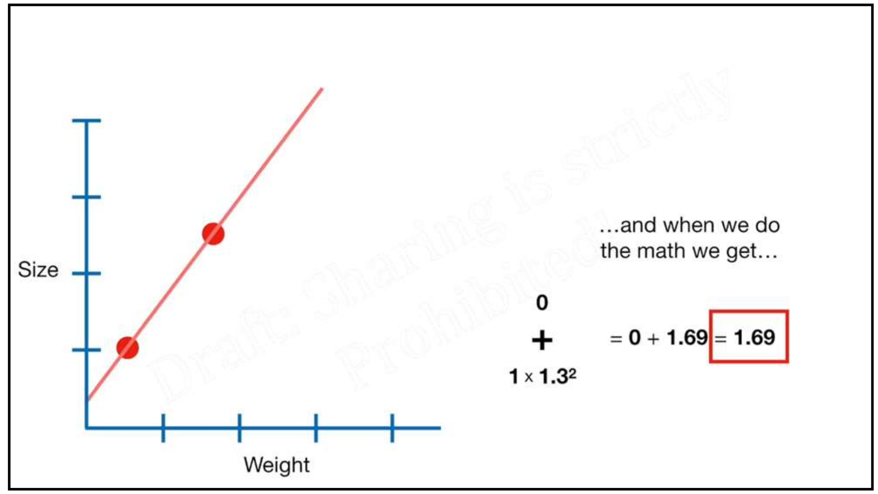

Let's start by plugging in the numbers that correspond to the Least Squares Fit.
Size = 0.4 + 1.3 x Weight
the sum of the squared residuals
+
λ x the slope²
Size
Weight

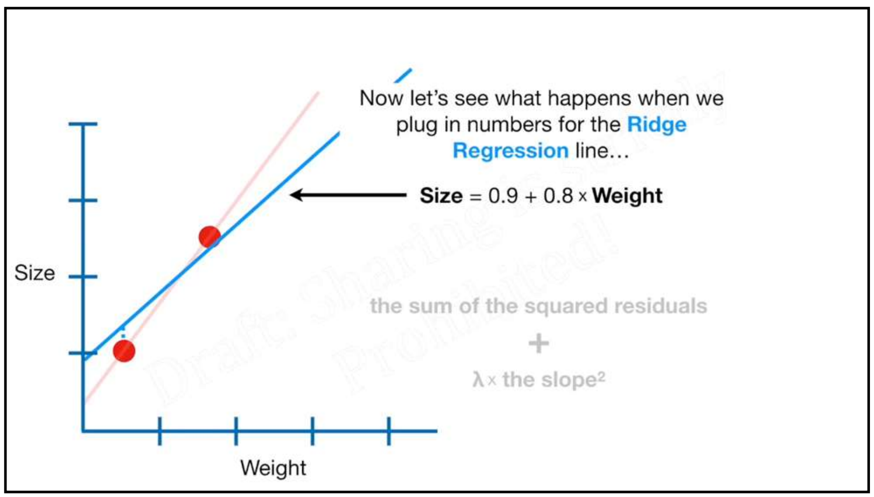

The sum of squared residuals for the Least Squares Fit is 0 (because the line overlaps the data points)…
Size = 0.4 + 1.3 x Weight
0
+
λ x the slope²
Size
Weight

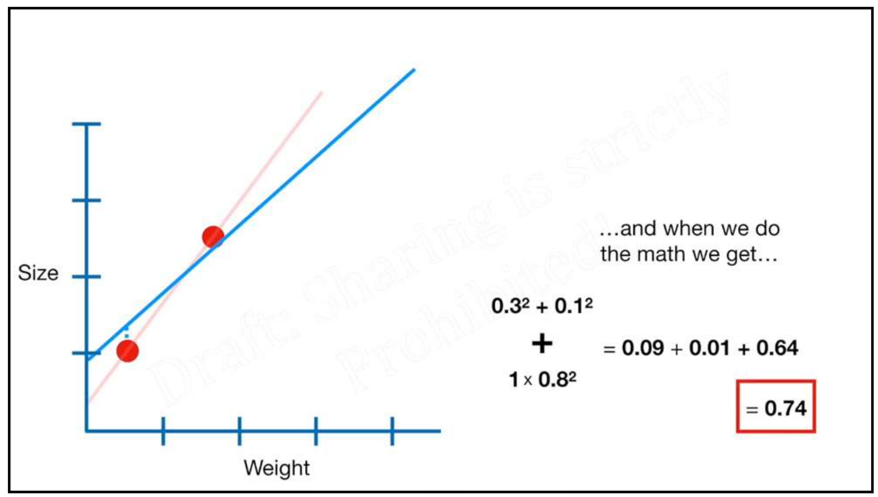

…and the slope is 1.3.
Size = 0.4 + 1.3 x Weight
0
+
λ x 1.3²
Size
Weight

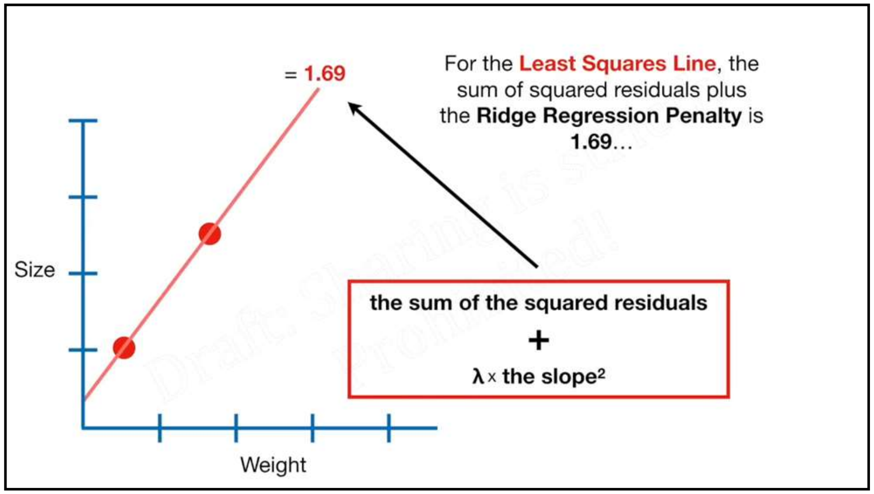

We'll talk more about λ later, but for now, let λ = 1.
0
+
1 x 1.3²
Size
Weight

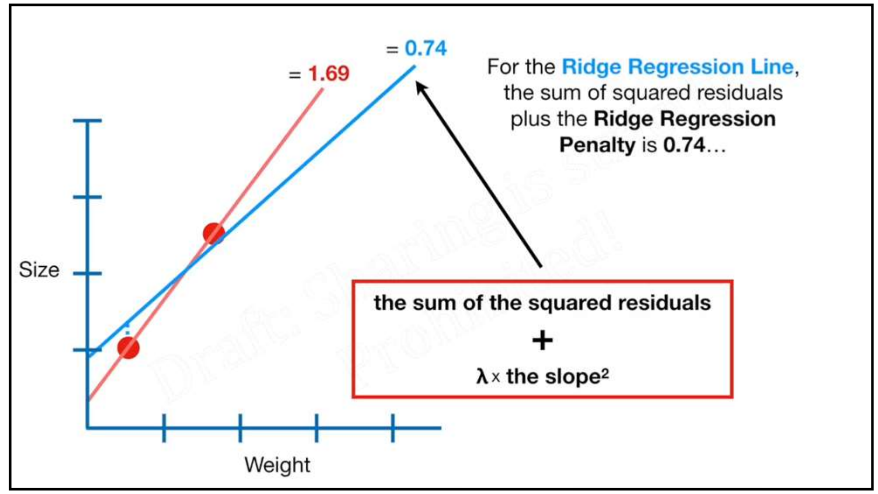

Size
Weight
...all together, we have...
0
+
1 x 1.3²

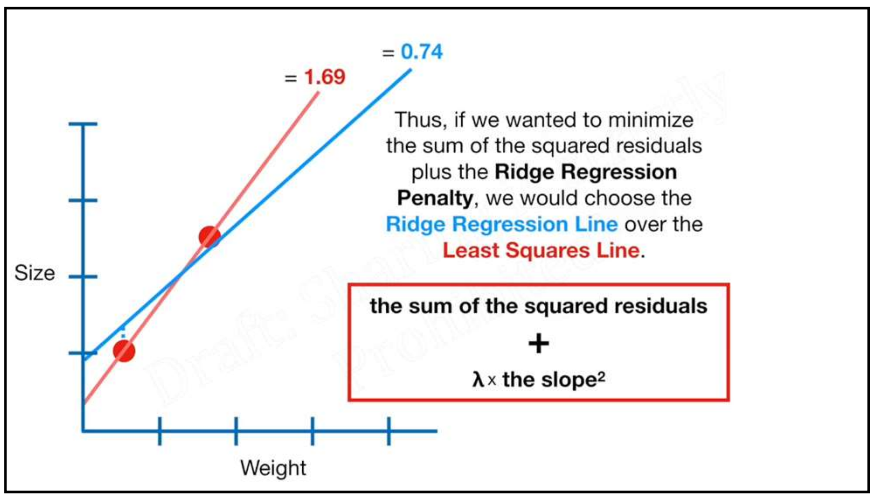

Size
Weight
...and when we do the math we get...
0
+
1 x 1.3²
= 0 + 1.69 = 1.69

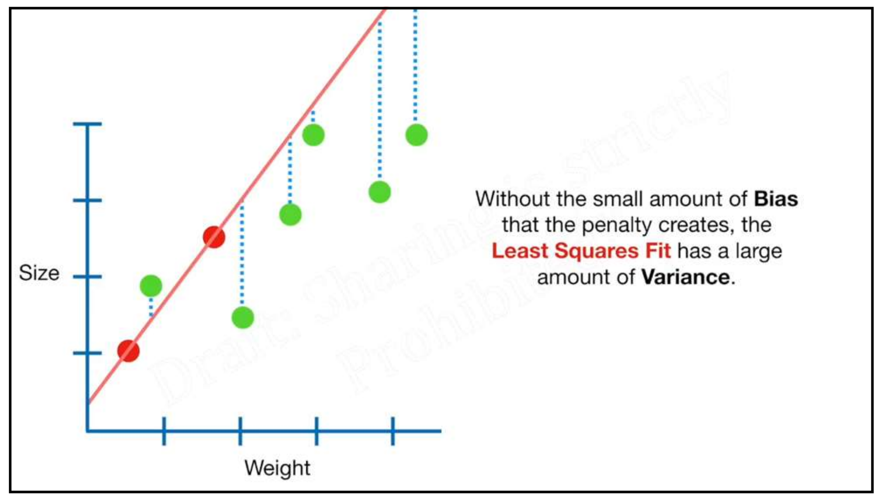

Size
Weight
Now let's see what happens when we plug in numbers for the Ridge Regression line...
Size = 0.9 + 0.8 x Weight
the sum of the squared residuals
+
λ x the slope²

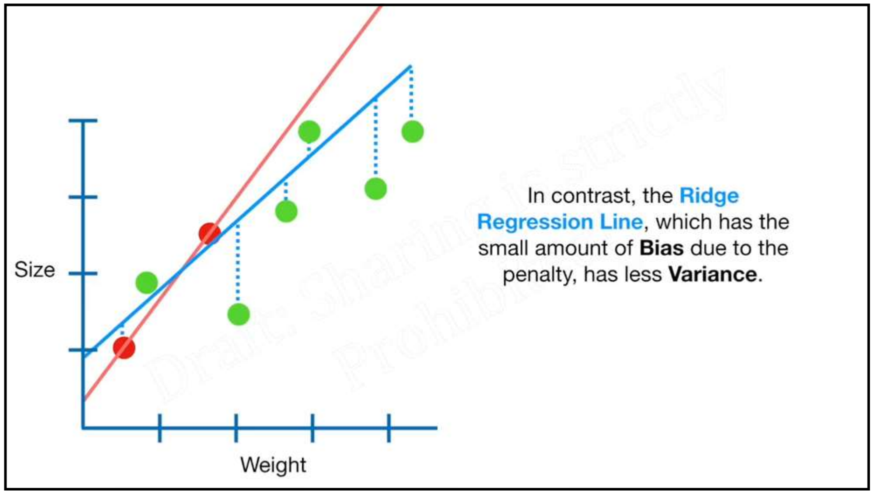

Size
Weight
...and when we do the math we get...
0.3² + 0.1²
+
1 x 0.8²
= 0.09 + 0.01 + 0.64
= 0.74

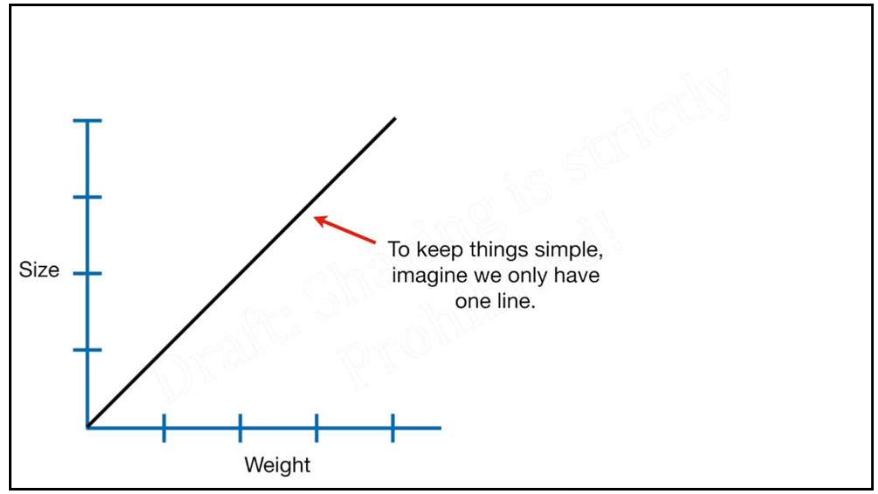

= 1.69
For the Least Squares Line, the sum of squared residuals plus the Ridge Regression Penalty is 1.69…
the sum of the squared residuals
+
λ x the slope²
Size
Weight

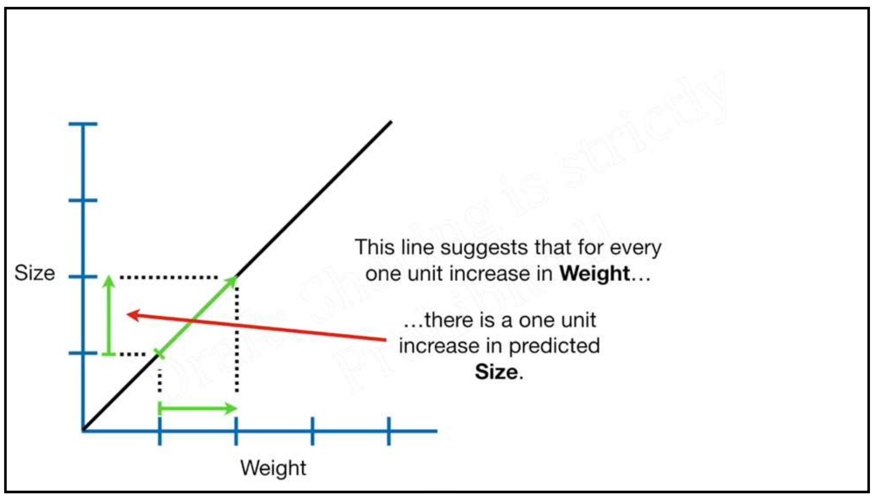

= 1.69
= 0.74
For the Ridge Regression Line, the sum of squared residuals plus the Ridge Regression Penalty is 0.74…
the sum of the squared residuals
+
λ x the slope²
Size
Weight

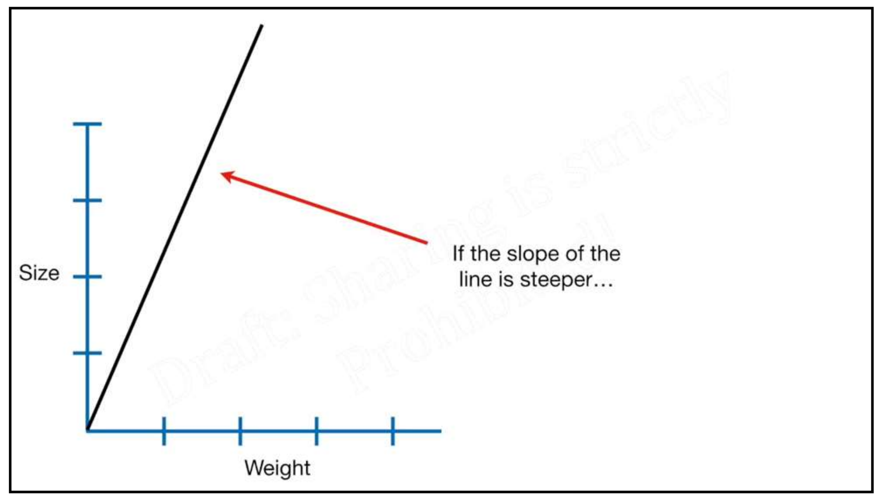

= 1.69
= 0.74
Thus, if we wanted to minimize the sum of the squared residuals plus the Ridge Regression Penalty, we would choose the Ridge Regression Line over the Least Squares Line.
the sum of the squared residuals
+
λ x the slope²
Size
Weight

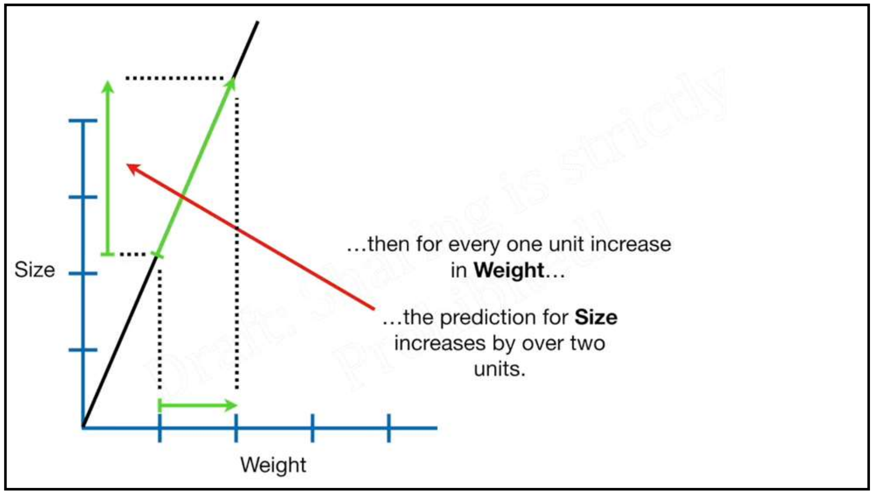

Without the small amount of Bias that the penalty creates, the Least Squares Fit has a large amount of Variance.
Size
Weight

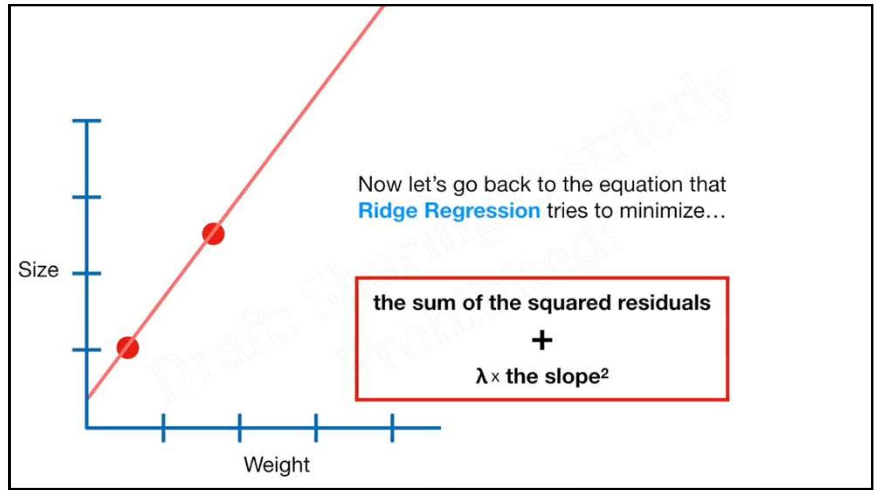

In contrast, the Ridge Regression Line, which has the small amount of Bias due to the penalty, has less Variance.
Size
Weight

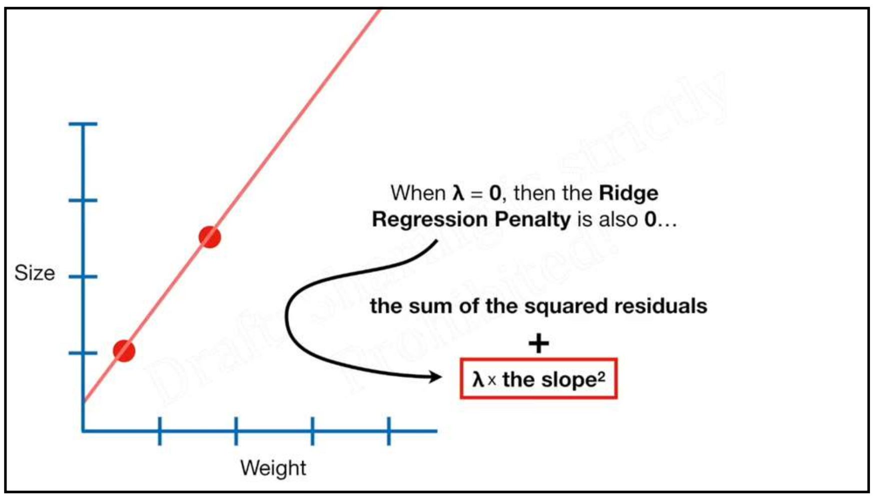

To keep things simple, imagine we only have one line.
Size
Weight

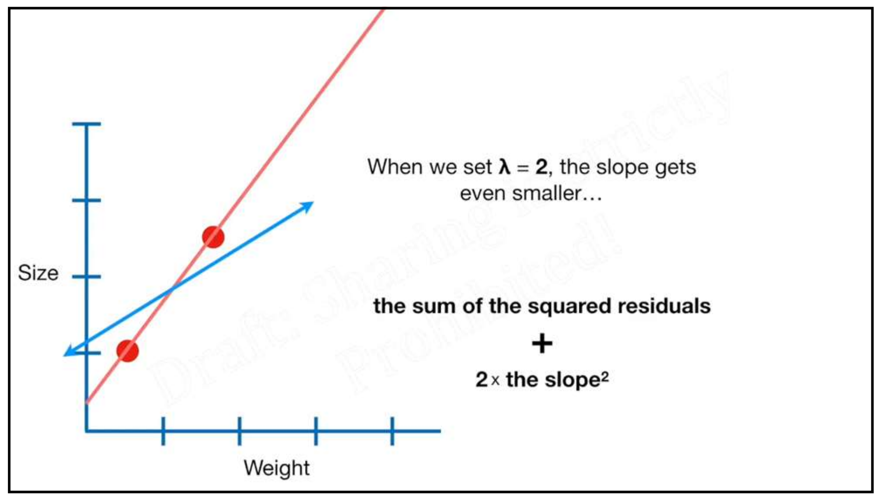

This line suggests that for every one unit increase in Weight…
…there is a one unit increase in predicted Size.
Size
Weight

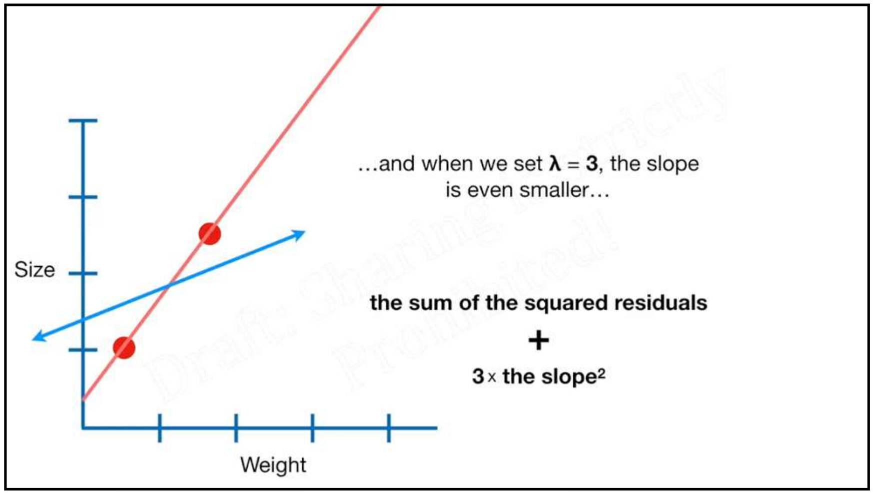

If the slope of the line is steeper…
Size
Weight

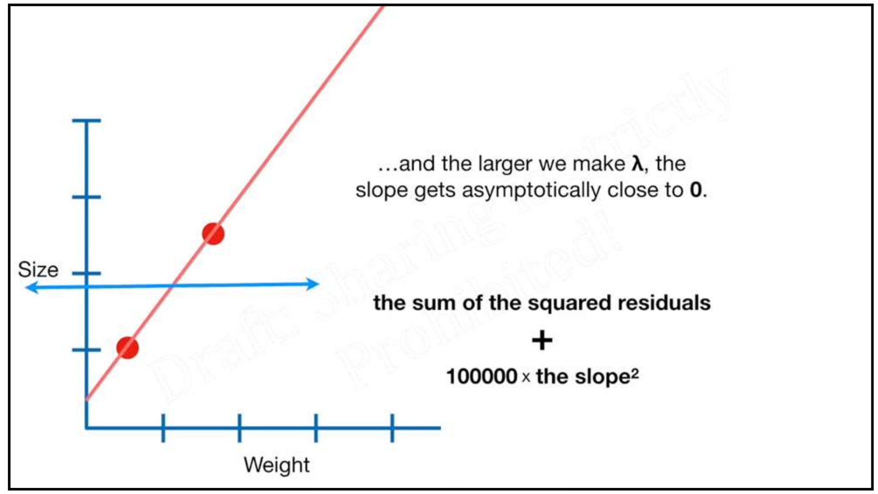

Size
Weight
…then for every one unit increase in **Weight**…
…the prediction for **Size** increases by over two units.

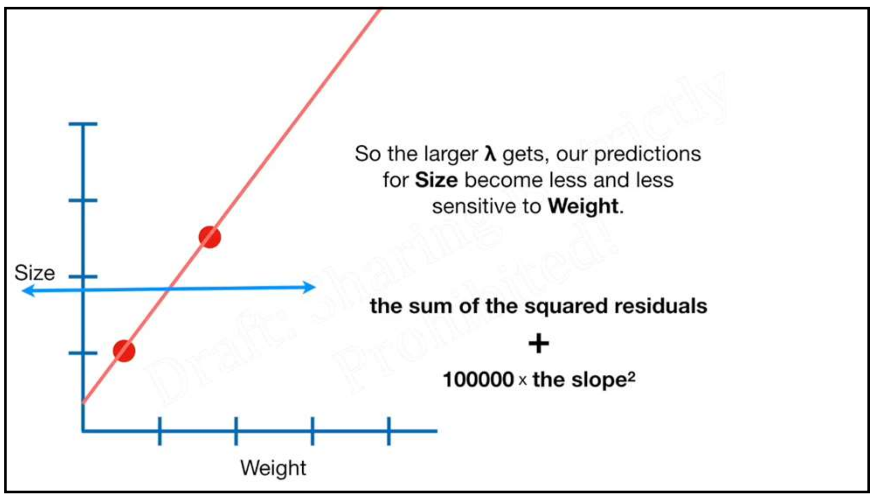

Size
Weight
Now let's go back to the equation that **Ridge Regression** tries to minimize…
the sum of the squared residuals
+
λ × the slope²

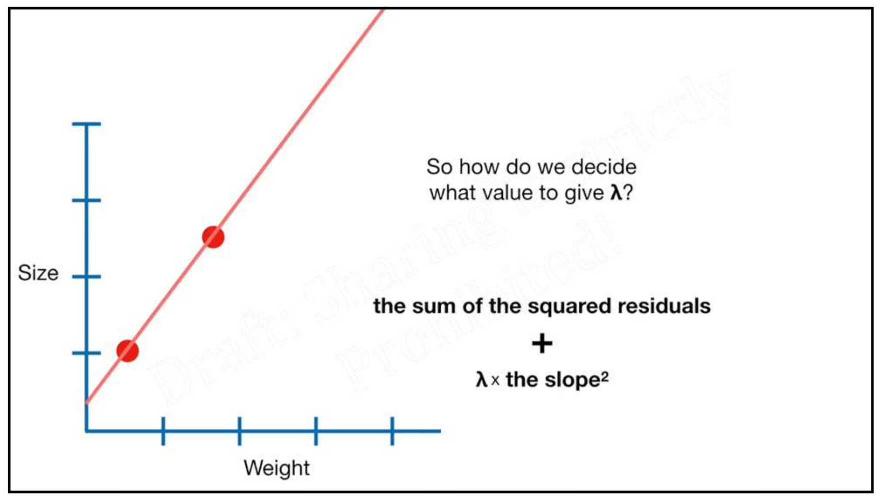

Size
Weight
When **λ** = **0**, then the **Ridge Regression Penalty** is also **0**…
the sum of the squared residuals
+
λ × the slope²

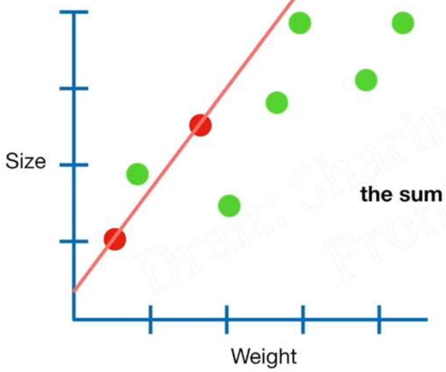

Size
Weight
When we set **λ** = **2**, the slope gets even smaller…
the sum of the squared residuals
+
2 × the slope²


Size
Weight
…and when we set λ = 3, the slope is even smaller…
the sum of the squared residuals
+
3 x the slope²


Size
Weight
…and the larger we make λ, the slope gets asymptotically close to 0.
the sum of the squared residuals
+
100000 x the slope²


Size
Weight
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


Size
Weight
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.**

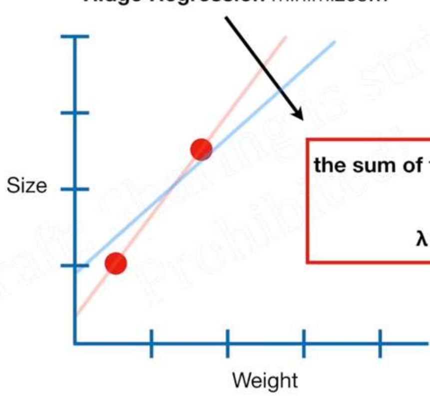
the sum of the squared residuals

+

λ × the slope²

LASSO REGRESSION

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

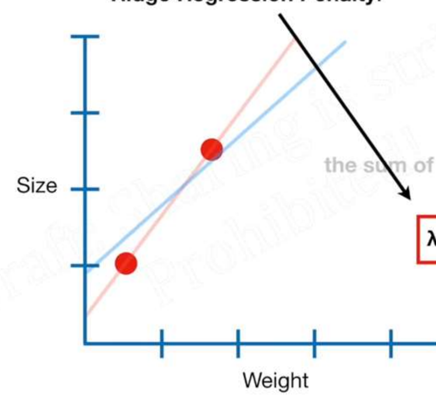


the sum of the squared residuals

+

λ × the slope²

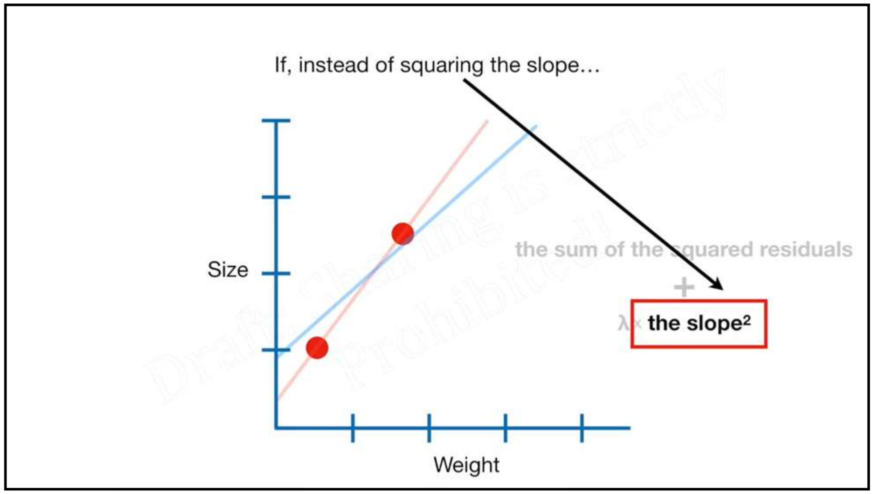
…and focus on the **Ridge Regression Penalty**.

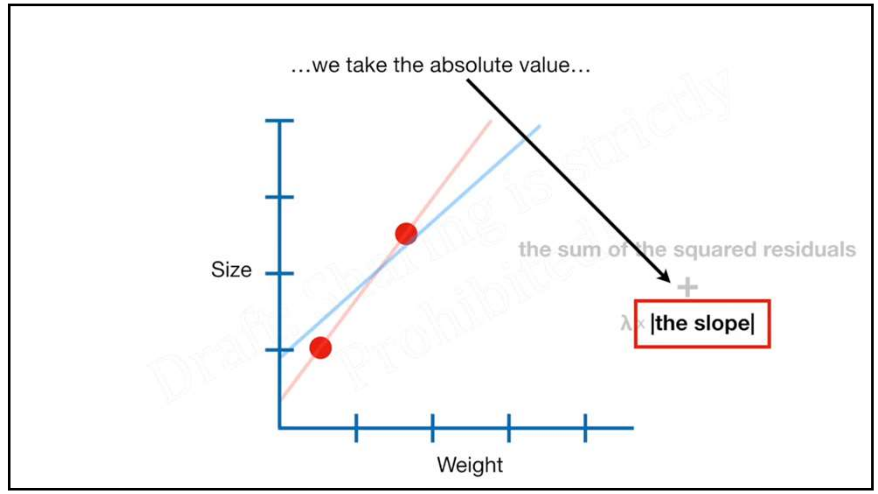


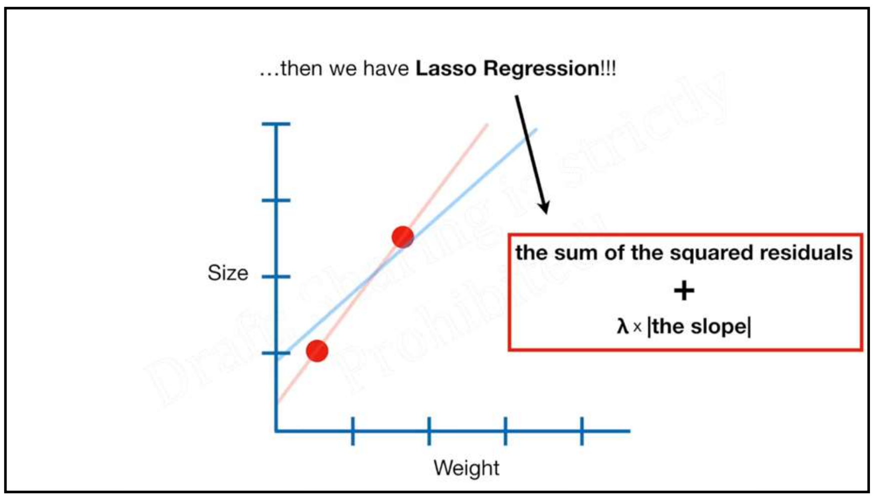
the sum of the squared residuals

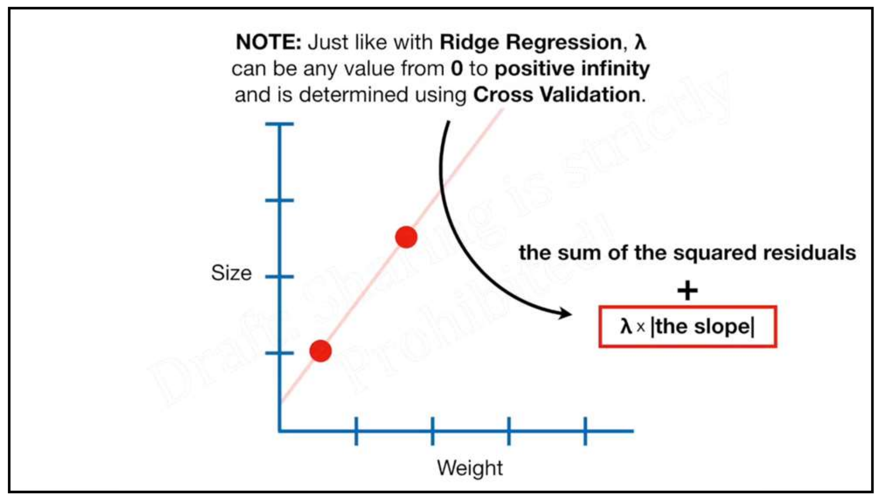
+

λ × the slope²

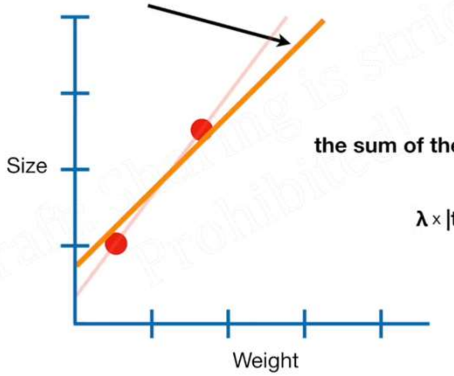

If, instead of squaring the slope…
Size
Weight
the sum of the squared residuals
+
λ × the slope²


…we take the absolute value…
Size
Weight
the sum of the squared residuals
+
λ × |the slope|


…then we have **Lasso Regression**!!!
Size
Weight
the sum of the squared residuals
+
λ × |the slope|


NOTE: Just like with **Ridge Regression**, **λ** can be any value from **0** to **positive infinity** and is determined using **Cross Validation**.
Size
Weight
the sum of the squared residuals
+
λ × |the slope|

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

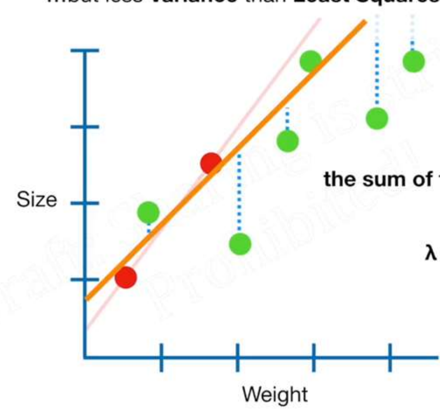


the sum of the squared residuals

+

λ × |the slope|

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



the sum of the squared residuals

+

λ × |the slope|

Ridge Regression…

the sum of the squared residuals

+

λ × the slope^2

the sum of the squared residuals

+

λ × the slope^2

…and **Lasso Regression…**

the sum of the squared residuals

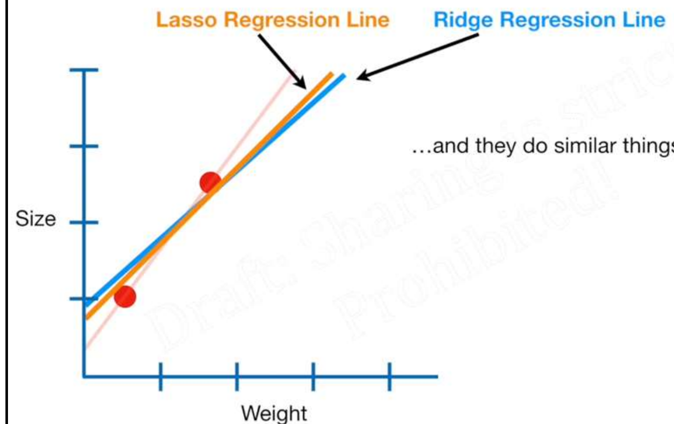
+

λ × |the slope|

…look very similar….

the sum of the squared residuals
+
$\lambda \times$ the slope^2

the sum of the squared residuals
+
$\lambda \times$ |the slope|

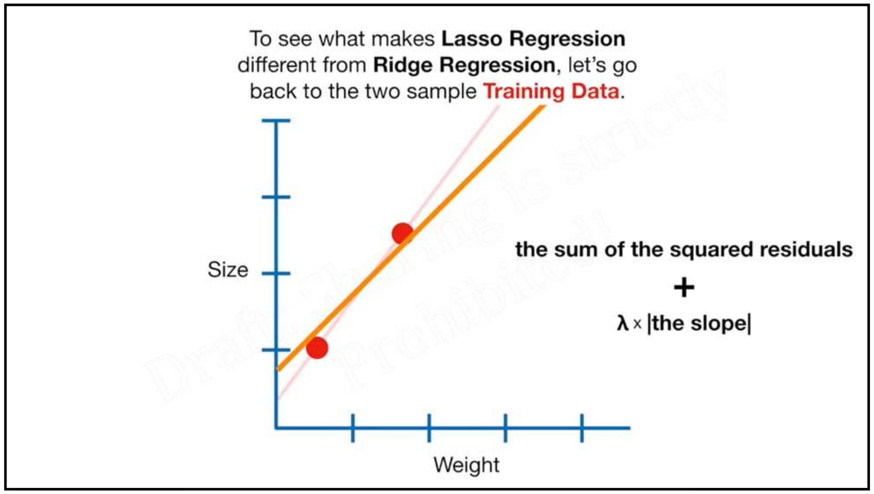


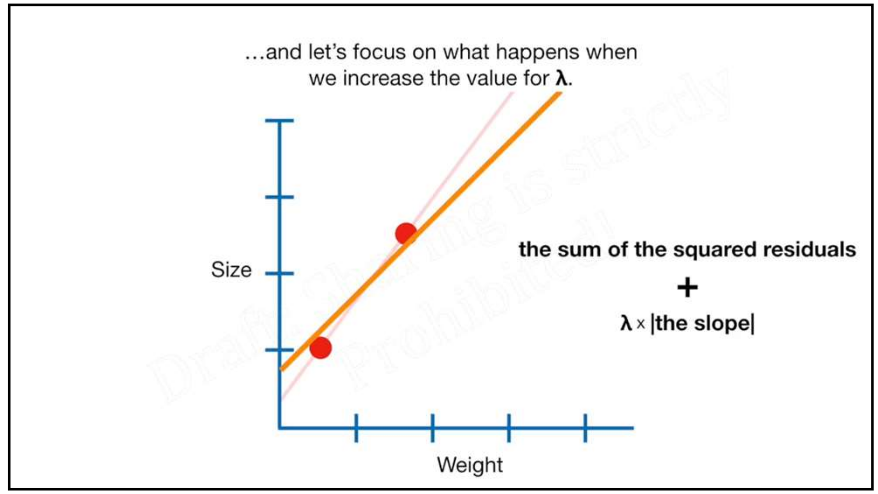
…and they do similar things.

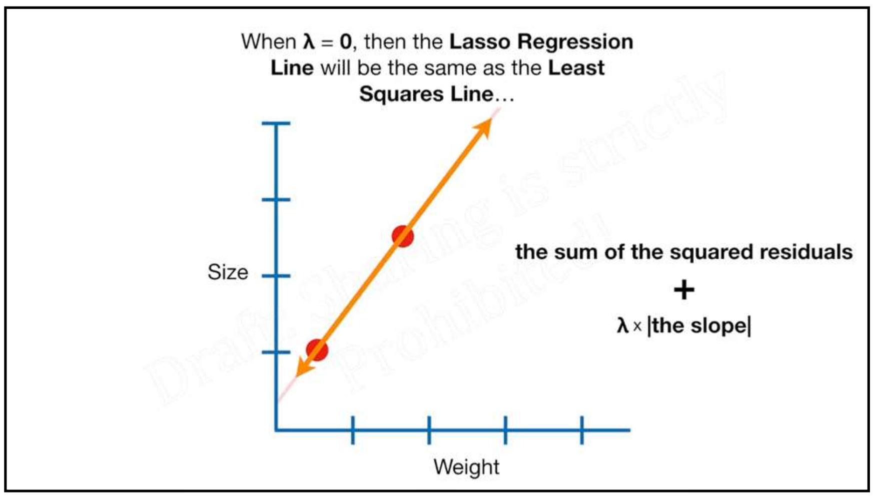
OK, we've seen how **Ridge** and **Lasso Regression** are similar.

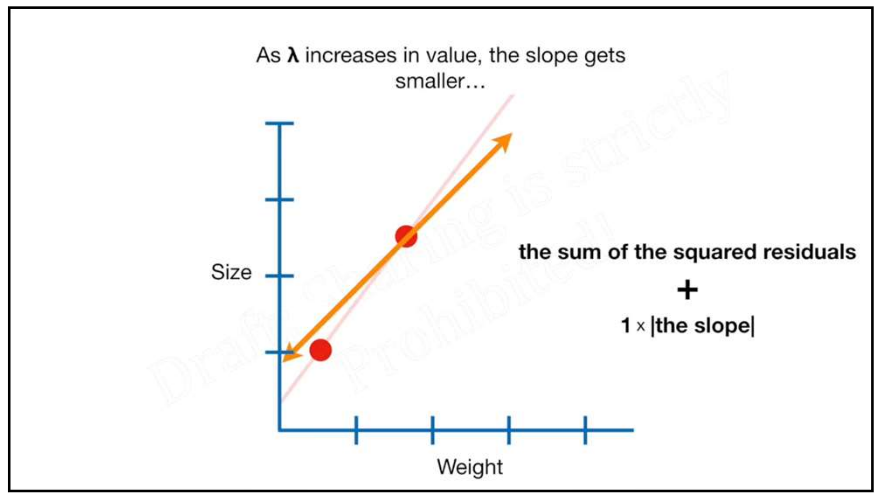
OK, we've seen how **Ridge** and **Lasso Regression** are similar.

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

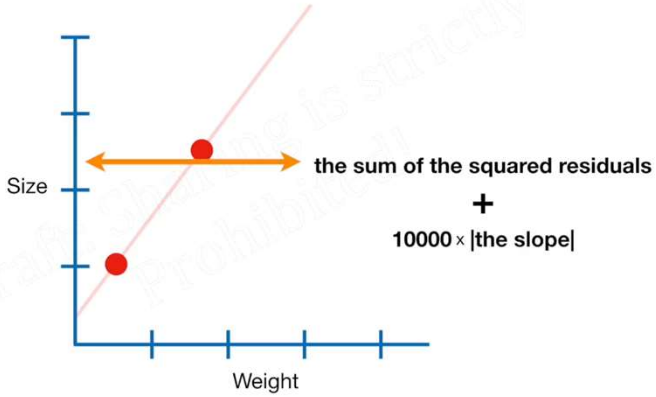

To see what makes **Lasso Regression** different from **Ridge Regression**, let's go back to the two sample **Training Data**.
Size
Weight
the sum of the squared residuals
+
λ × **|the slope|**


…and let's focus on what happens when we increase the value for **λ**.
Size
Weight
the sum of the squared residuals
+
λ × **|the slope|**

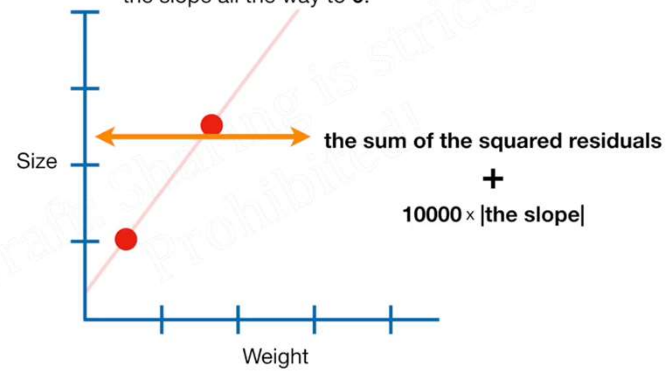

When **λ** = **0**, then the **Lasso Regression Line** will be the same as the **Least Squares Line**…
Size
Weight
the sum of the squared residuals
+
λ × **|the slope|**


As **λ** increases in value, the slope gets smaller…
Size
Weight
the sum of the squared residuals
+
1 × **|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**

The goal of this equation is to predict **Size**.

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

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

The terms for **Weight** and **High Fat Diet**, are both reasonable things to use to predict **Size**…

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

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

…but the **Astrological Sign** and the **Airspeed of a Swallow** (African or European) are terrible ways to predict **Size**.

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**…

λ x (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**

…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**

…and the larger we make **λ**…

…these parameters might shrink a little bit…

λ (slope² + diet difference² + astrological offset² + airspeed scalar²)

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

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

…and the larger we make **λ**…

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

λ (slope² + diet difference² + astrological offset² + airspeed scalar²)

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**…

λ x(|slope| + |diet difference| + |astrological offset| + |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**…

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

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

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

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

…then these parameters will shrink a little bit…

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

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

+ astrological offset × **Sign** + airspeed scalar × **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 × **Weight** + diet difference × **High Fat Diet**

+ 0 × **Sign** + 0 × **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 × **Weight** + diet difference × **High Fat Diet**

+ ~~0 × **Sign**~~ + ~~0 × **Airspeed of Swallow**~~

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

…and these terms go away…

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

$\boxed{\textbf{Size} = \text{y-intercept} + \text{slope} \times \textbf{Weight} + \text{diet difference} \times \textbf{High Fat Diet}}$

…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**~~

…and excludes all of the silly stuff!!!

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**

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** **+** $\boxed{\lambda \times \text{the slope}^2}$

The superficial difference is that **Ridge Regression** squares the variables… → **the sum of the squared residuals** **+** $\boxed{\lambda \times \text{the slope}^2}$

…and **Lasso Regression** takes the absolute value. → **the sum of the squared residuals** **+** $\boxed{\lambda \times |\text{the slope}|}$

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

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

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

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.

ELASTIC-NET

…but what do we do when we have a model that includes tons more variables?

Size = y-intercept + slope_1 x **Weight** + diet difference x **High Fat Diet**

+ slope_2 x **Age** + slope_3 x **Size of Father** + …. + **tons more variables…**

…and when you have millions of parameters, then you will almost certainly need to use some sort of regularization to estimate them.

Size = y-intercept + $slope_1$ x **Weight** + diet difference x **High Fat Diet** + $slope_2$ x **Age** + $slope_3$ x **Size of Father** + …. + **tons more variables**…

However, the variables in those models might be useful or useless. We don't know in advance.

Size = y-intercept + $slope_1$ x **Weight** + diet difference x **High Fat Diet** + $slope_2$ x **Age** + $slope_3$ x **Size of Father** + …. + **tons more variables**…

So how do you choose if you should use **Lasso** or **Ridge Regression**?

Size = y-intercept + $slope_1$ x **Weight** + diet difference x **High Fat Diet** + $slope_2$ x **Age** + $slope_3$ x **Size of Father** + …. + **tons more variables**…

The good news is that you don't have to choose, instead, you use **Elastic-Net Regression!!!**

Size = y-intercept + $slope_1$ x **Weight** + diet difference x **High Fat Diet** + $slope_2$ x **Age** + $slope_3$ x **Size of Father** + …. + **tons more variables**…

Elastic-Net Regression sounds super fancy, but if you already know about **Lasso** and **Ridge Regression**, it's super simple.

Just like **Lasso** and **Ridge Regression, Elastic-Net Regression** starts with **Least Squares**…

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X|$ + $\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

…then it combines the **Lasso Regression Penalty**…

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X|$ + $\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

…with the **Ridge Regression Penalty**.

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X|$ **+** $\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

Altogether, **Elastic Net Regression** combines the strengths of **Lasso** and **Ridge Regression**.

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_x|$ **+** $\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_x^2$

NOTE: The **Lasso Regression Penalty** and the **Ridge Regression Penalty** get their own **λs**.

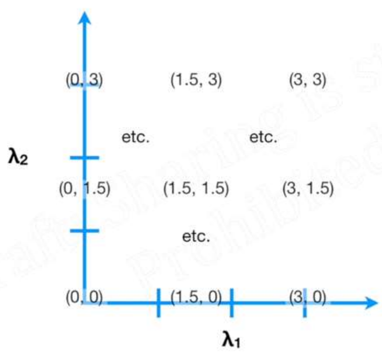
λ_1 for **Lasso**… …and **λ_2** for **Ridge**.

the sum of the squared residuals

+

$\boxed{\lambda_1} \times |\text{variable}_1| + \ldots + |\text{variable}_x|$ + $\boxed{\lambda_2} \times \text{variable}_1^2 + \ldots + \text{variable}_x^2$

We use **Cross Validation** on different combinations of **λ_1** and **λ_2** to find the best values.



The hybrid **Elastic-Net Regression** is especially good at dealing with situations when there are correlations between parameters.

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_x|$ **+** $\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_x^2$

This is because on it's own, **Lasso Regression** tends to pick just one of the correlated terms and eliminates the others…

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X|$

…whereas **Ridge Regression** tends to shrink all of the parameters for the correlated variables together.

the sum of the squared residuals

+

$\lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

By combining **Lasso** and **Ridge Regression**, **Elastic-Net Regression** groups and shrinks the parameters associated with the correlated variables and leaves them in equation or removes them all at once.

the sum of the squared residuals

+

$\lambda_1 \times |\text{variable}_1| + \ldots + |\text{variable}_X| + \lambda_2 \times \text{variable}_1^2 + \ldots + \text{variable}_X^2$

THANK YOU

StatQuest with Josh Starmer
https://www.youtube.com/watch?v=Q81RR3yKn30