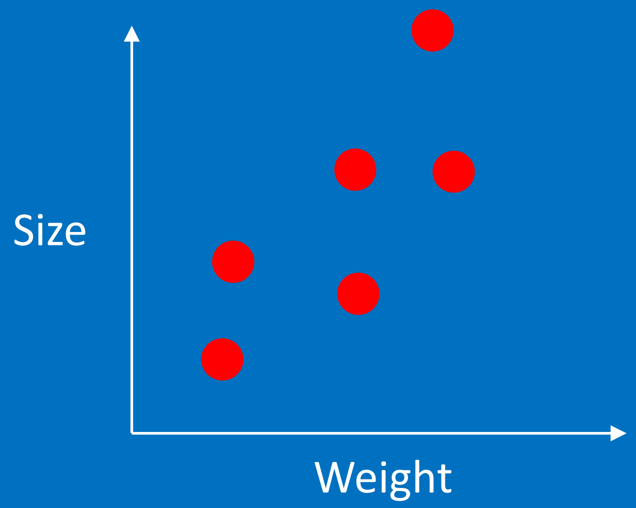
Regularization

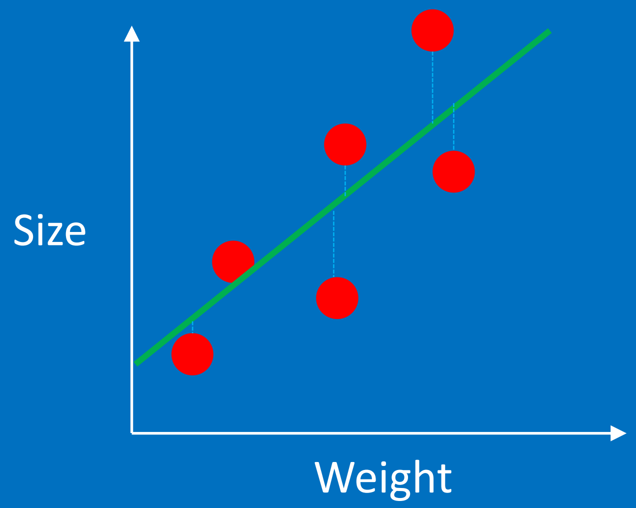
Let’s use Linear regression as an example

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



Let's use Linear regression as an example

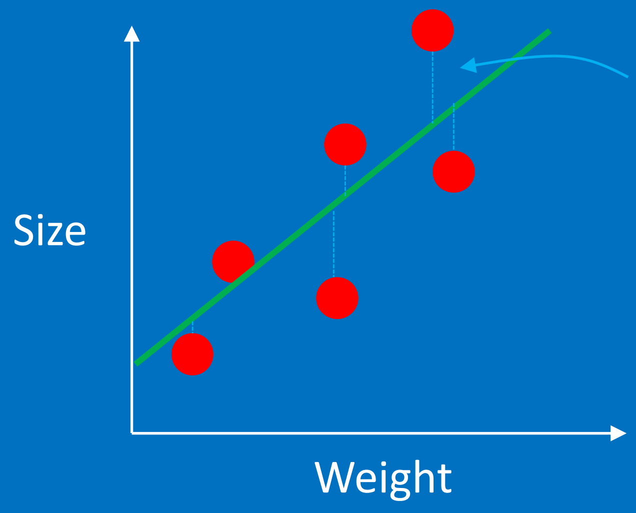
Let's start by collecting Weight and Size measurements from a bunch of mice …



We can use Linear regression to model the relationship between weight and size

Let's use Linear regression as an example

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

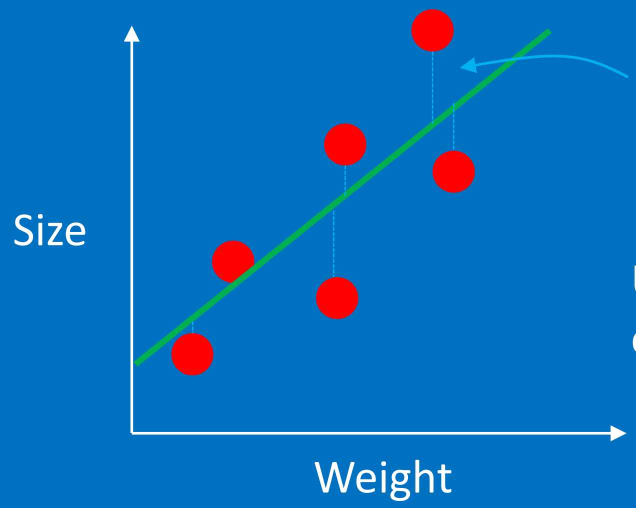


In other words, we find the line that results in the minimum sum of squared residuals

We can use Linear regression to model the relationship between weight and size

Let's use Linear regression as an example

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



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

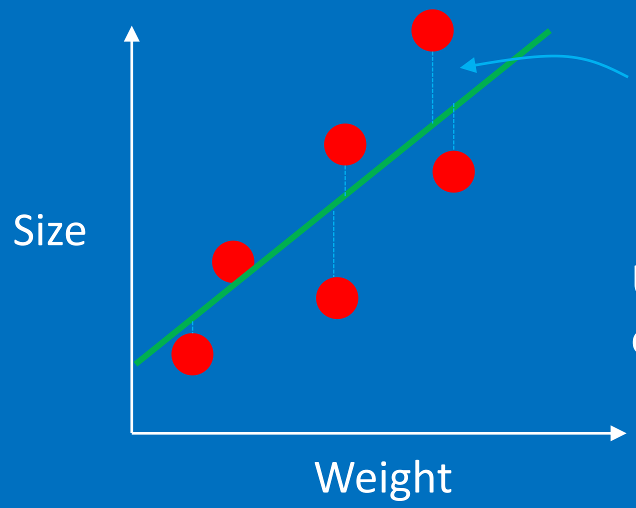
Ultimately, we end up with this equation for the line

$$size = 0.9 + 0.75 \times weight$$

We can use Linear regression to model the relationship between weight and size

Let's use Linear regression as an example

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



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

$$size = 0.9 + 0.75 \times weight$$

Y-axis intercept

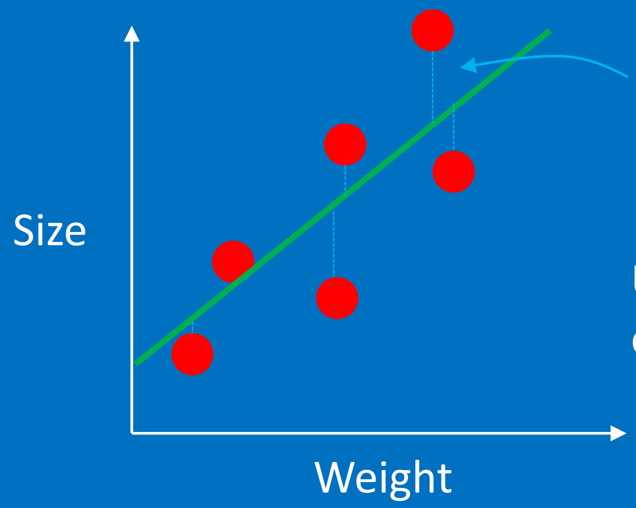
slop

The equation has two parameters

We can use Linear regression to model the relationship between weight and size

Let's use Linear regression as an example

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



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

$$size = 0.9 + 0.75 \times weight$$

Y-axis intercept

slop

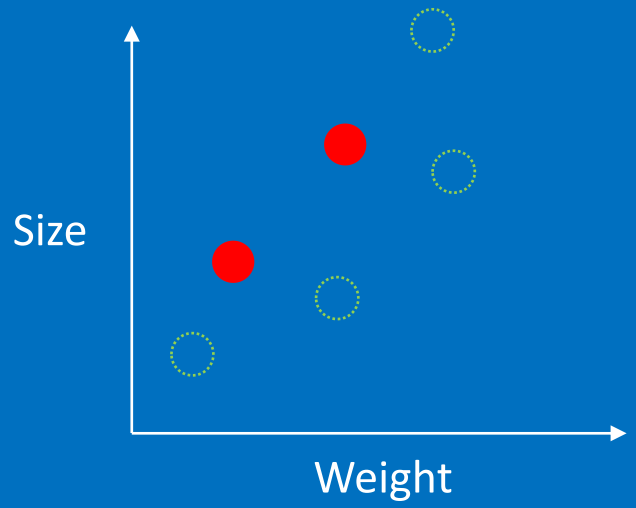
The equation has two parameters

When we have many measurements, we can be fairly confident that the linear regression line accurately reflects the relationship between size and weight

We can use Linear regression to model the relationship between weight and size

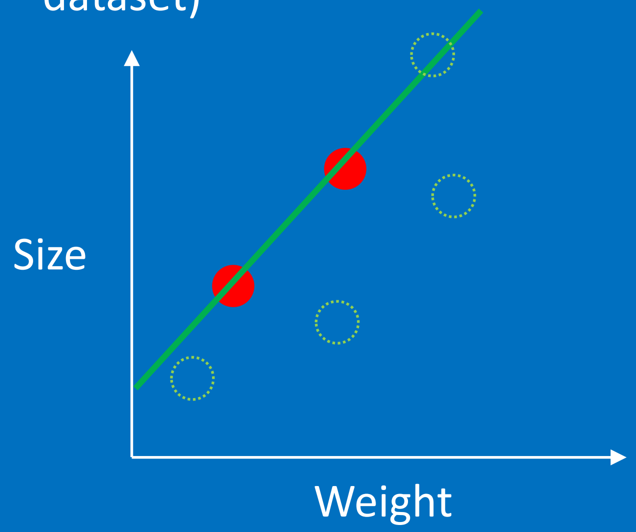
Let's use Linear regression as an example

However, what if we have very limited measurements, for example, if only two measurements (selected from the original dataset)



Let's use Linear regression as an example

However, what if we have very limited measurements, for example, if only two measurements (selected from the original dataset)

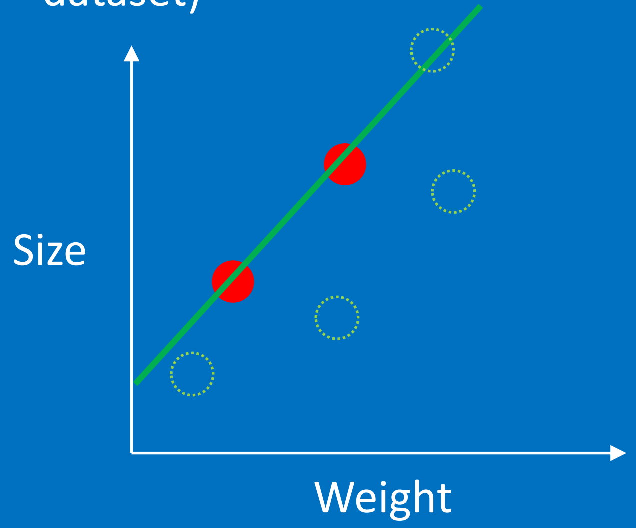


We fit a new line with linear regression

This line overlaps the two data points, the minimum sum of squared residuals = 0

Let's use Linear regression as an example

However, what if we have very limited measurements, for example, if only two measurements (selected from the original dataset)



Ultimately, we end up with this equation for the line

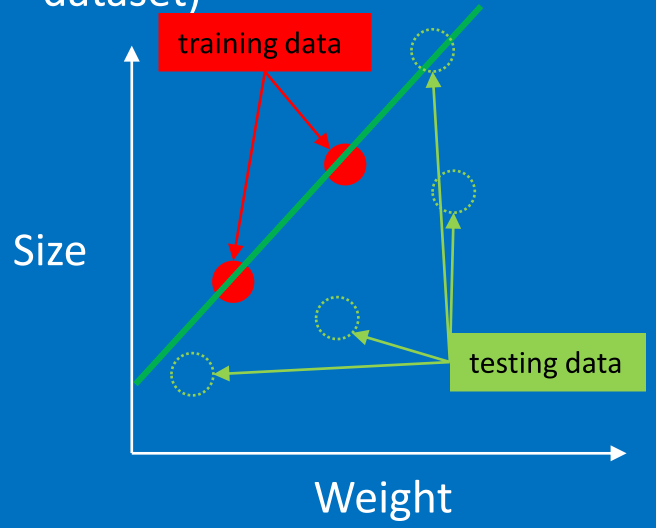
$$size = 0.4 + 1.3 \times weight$$

We fit a new line with linear regression

This line overlaps the two data points, the minimum sum of squared residuals = 0

Let's use Linear regression as an example

However, what if we have very limited measurements, for example, if only two measurements (selected from the original dataset)



Ultimately, we end up with this equation for the line

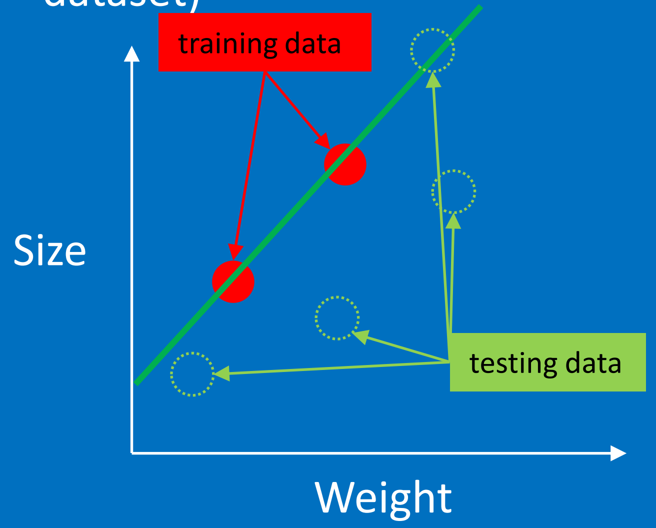
$$size = 0.4 + 1.3 \times weight$$

We fit a new line with linear regression

This line overlaps the two data points, the minimum sum of squared residuals = 0

Let's use Linear regression as an example

However, what if we have very limited measurements, for example, if only two measurements (selected from the original dataset)



Ultimately, we end up with this equation for the line

$$size = 0.4 + 1.3 \times weight$$

- The residuals for the training data site is prefect (in this case it's 0)
- However the residuals for the testing data is huge

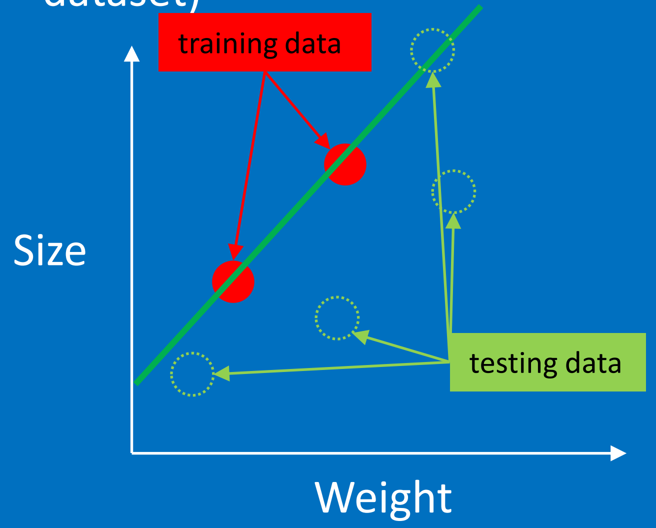
We fit a new line with linear regression

This line overlaps the two data points, the minimum sum of squared residuals = 0

This means that the results have (1) low bias, but (2) high variance

Let's use Linear regression as an example

However, what if we have very limited measurements, for example, if only two measurements (selected from the original dataset)



Ultimately, we end up with this equation for the line

$$size = 0.4 + 1.3 \times weight$$

- The residuals for the training data site is prefect (in this case it's 0)
- However the residuals for the testing data is huge

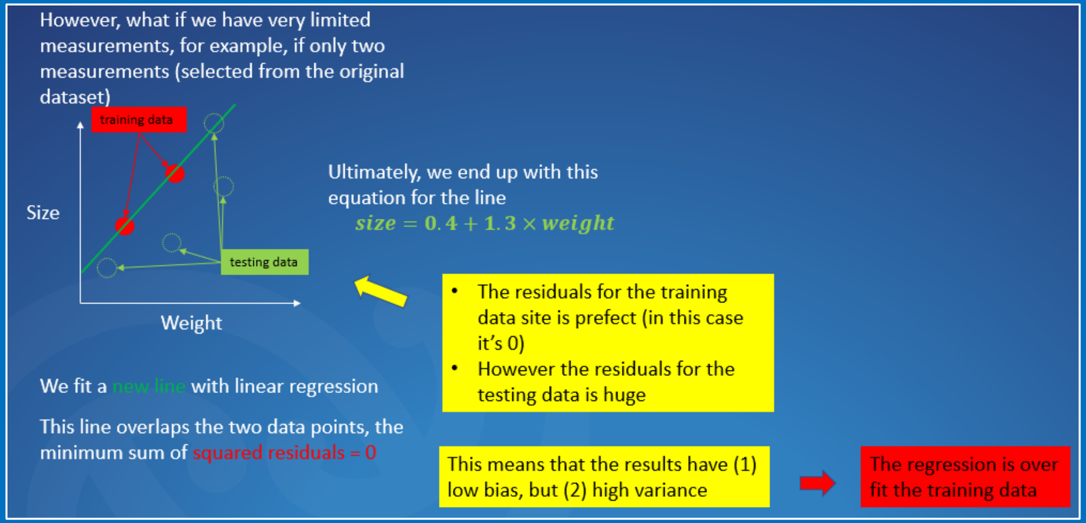
We fit a new line with linear regression

This line overlaps the two data points, the minimum sum of squared residuals = 0

This means that the results have (1) low bias, but (2) high variance

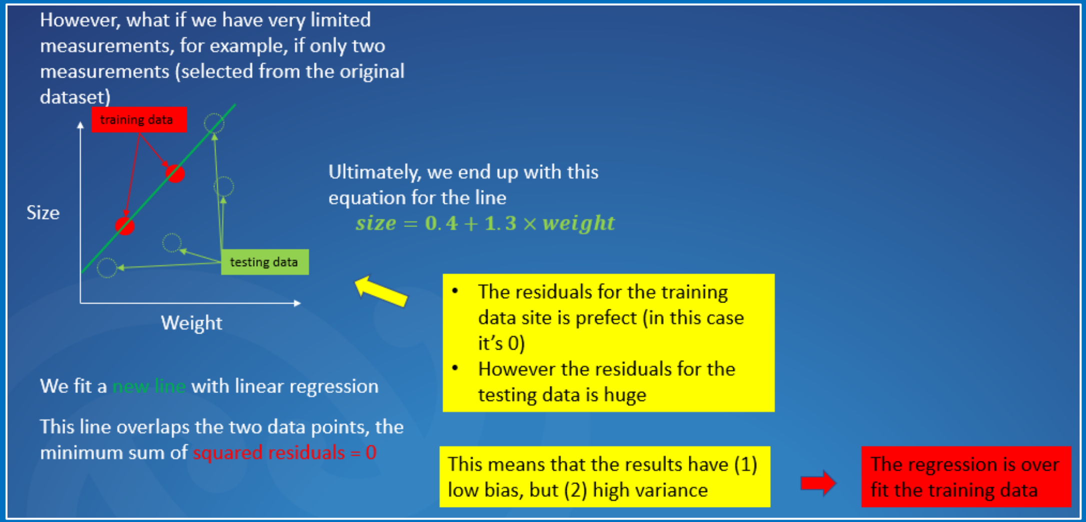
The regression is over fit the training data

Let’s use Linear regression as an example



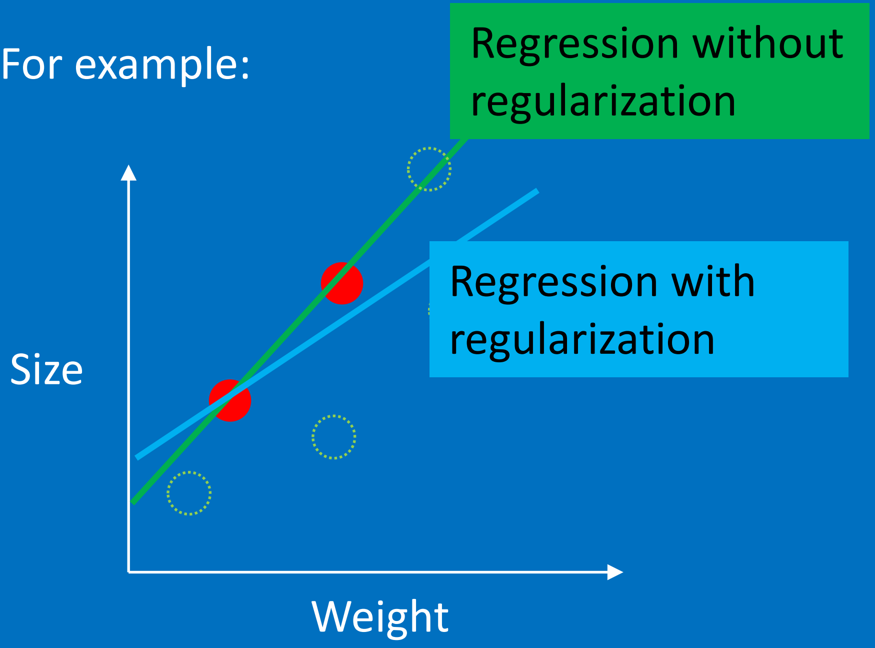
The main idea behind Ridge regularization is to find a new line that does not fit the training data so we

Let's use Linear regression as an example



The main idea behind Ridge regularization is to find a new line that does not fit the training data so we

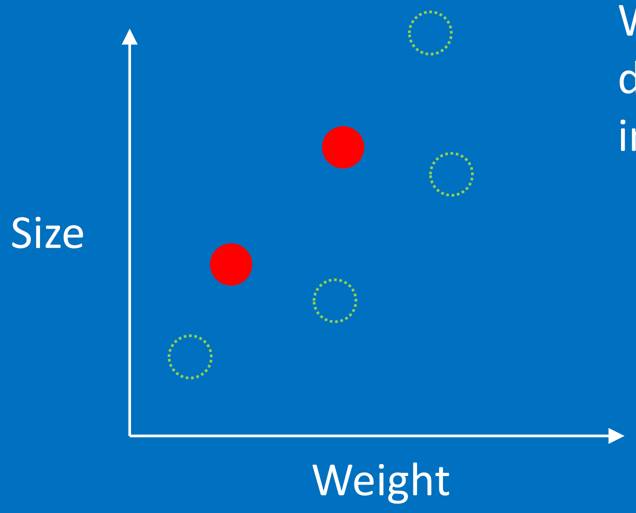
For example:



In other words, we introduce a bit "bias" to the fitted line, but in return we get a significant drop in "variance"

Let's use Linear regression as an example

Ridge regularization step-by-step



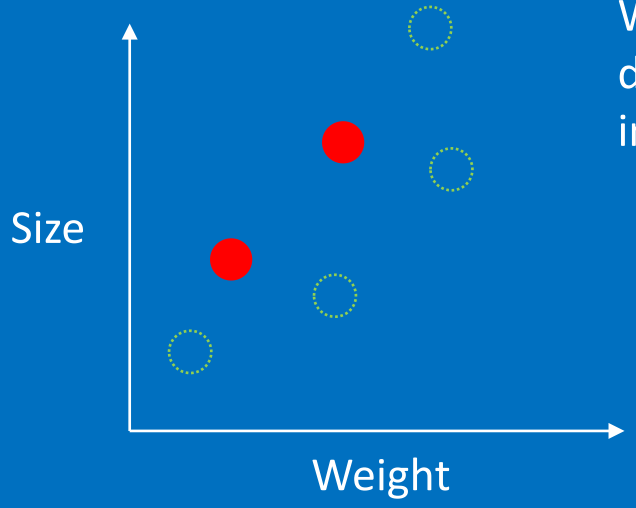
Let's go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = \boxed{intercept} + \boxed{slop} \times weight$$

Let's use Linear regression as an example

Ridge regularization step-by-step



Let's go back to just the training data (two red dots)

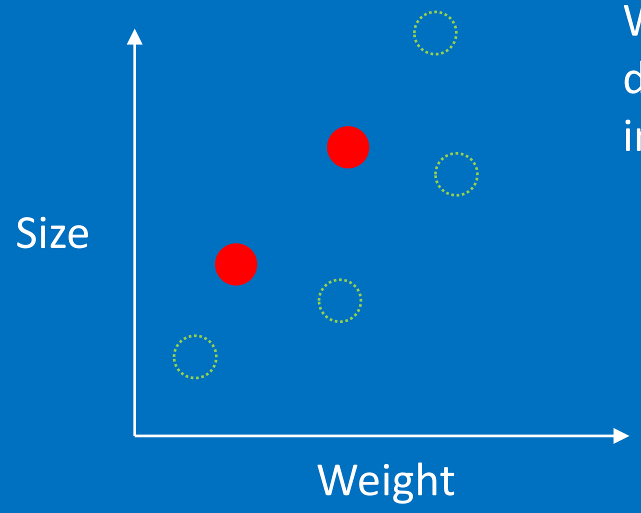
When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

Without regularization, the cost function minimizes "the sum of the squared residuals"

Let's use Linear regression as an example

Ridge regularization step-by-step



Let's go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

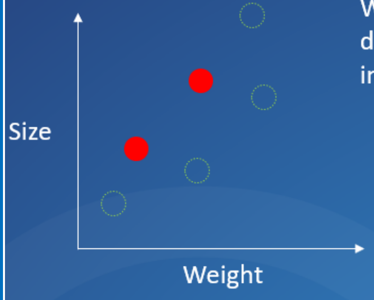
$$size = intercept + slop \times weight$$

Without regularization, the cost function minimizes "the sum of the squared residuals"

With regularization, the cost function minimizes "the sum of the squared residuals" + $\lambda \times slope^2$

Let’s use Linear regression as an example

Ridge regularization step-by-step



Let’s go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

Without regularization, the cost function minimizes “the sum of the squared residuals”

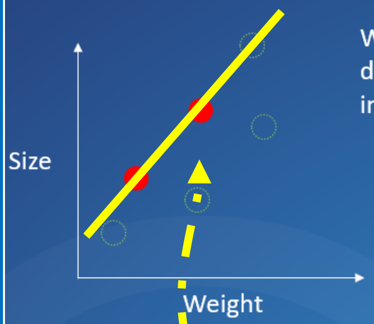
With regularization, the cost function minimizes “the sum of the squared residuals” + $\lambda \times slope^2$

Assuming that the regression equation to calculate the size is

$\underline{size = 0.4 + 1.3 \times weight}$

Let's use Linear regression as an example

Ridge regularization step-by-step



Let's go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

Without regularization, the cost function minimizes "the sum of the squared residuals"

With regularization, the cost function minimizes "the sum of the squared residuals" + $\lambda \times slope^2$

The line overlap the data points

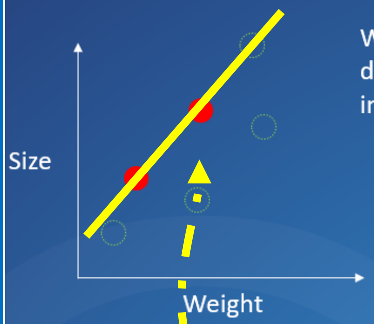
Assuming that the regression equation to calculate the size is $\underline{size = 0.4 + 1.3 \times weight}$

The sum of the squared residuals=0

Let's use Linear regression as an example

Ridge regularization step-by-step

Let's go back to just the training data (two red dots)



When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

The line overlap the data points

Without regularization, the cost function minimizes "the sum of the squared residuals"

With regularization, the cost function minimizes "the sum of the squared residuals" + $\lambda \times slope^2$

Assuming that the regression equation to calculate the size is

$\underline{size = 0.4 + 1.3 \times weight}$

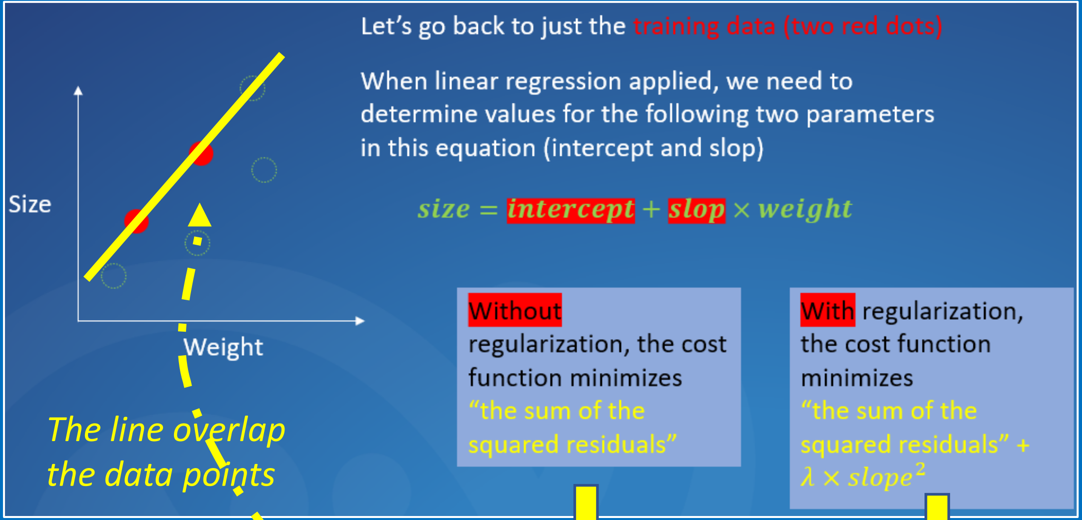
Error = The sum of the squared residuals=0

Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

Let’s use Linear regression as an example

Ridge regularization step-by-step



Assuming that the regression equation to calculate the size is

$\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

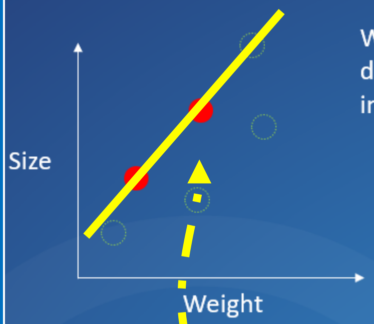
Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

In the left example, we use $\lambda = 1.0$. But how λ would affect the results ?

Let’s use Linear regression as an example

Ridge regularization step-by-step



Let’s go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

The line overlap the data points

Without regularization, the cost function minimizes “the sum of the squared residuals”

With regularization, the cost function minimizes “the sum of the squared residuals” + $\lambda \times slope^2$

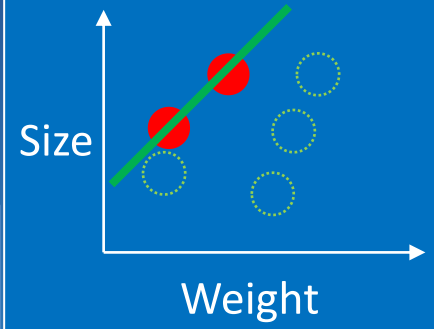
Assuming that the regression equation to calculate the size is $\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

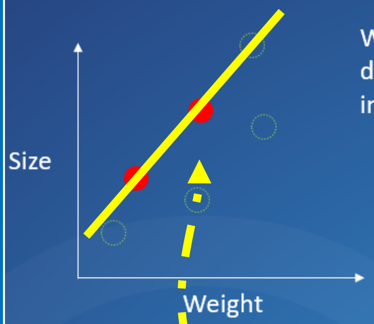
In the left example, we use $\lambda = 1.0$. But how λ would affect the results ?



When $\lambda = 0.0$, the error is the same to the one without any regularization

Let's use Linear regression as an example

Ridge regularization step-by-step



Let's go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

The line overlap the data points

Without regularization, the cost function minimizes "the sum of the squared residuals"

With regularization, the cost function minimizes "the sum of the squared residuals" + $\lambda \times slope^2$

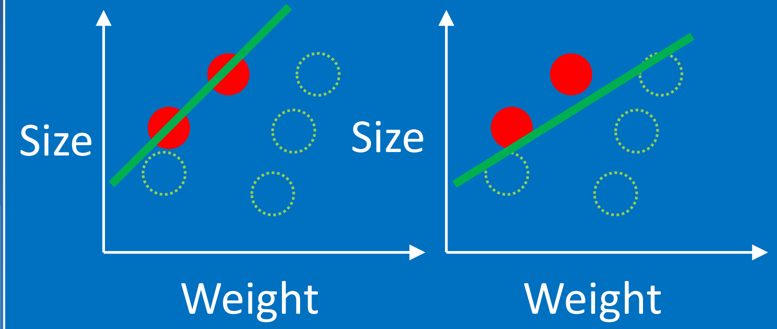
Assuming that the regression equation to calculate the size is $\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

In the left example, we use $\lambda = 1.0$. But how λ would affect the results ?

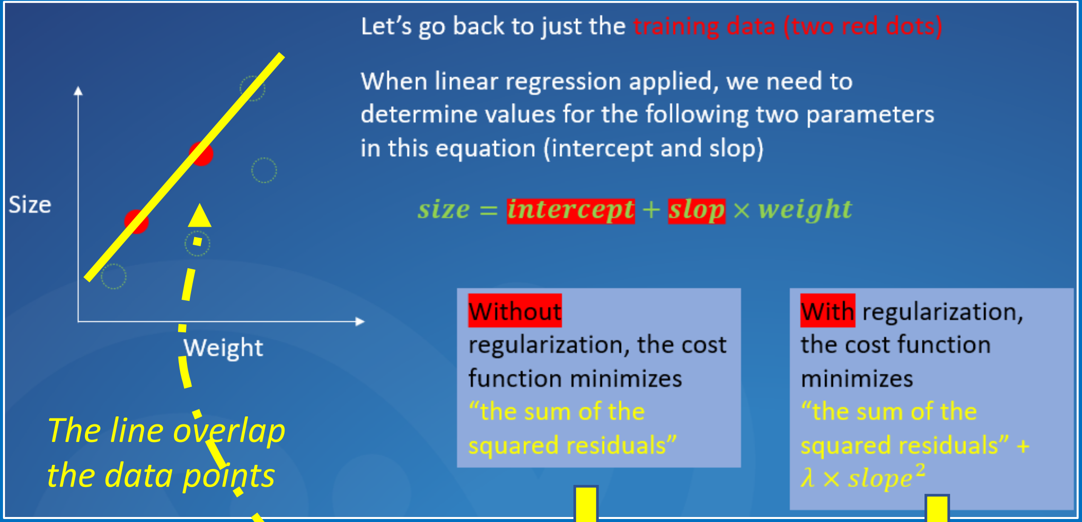


When $\lambda = 0.0$, the error is the same to the one without any regularization

When $\lambda = 1.0$, the fitted line will be less steep

Let’s use Linear regression as an example

Ridge regularization step-by-step



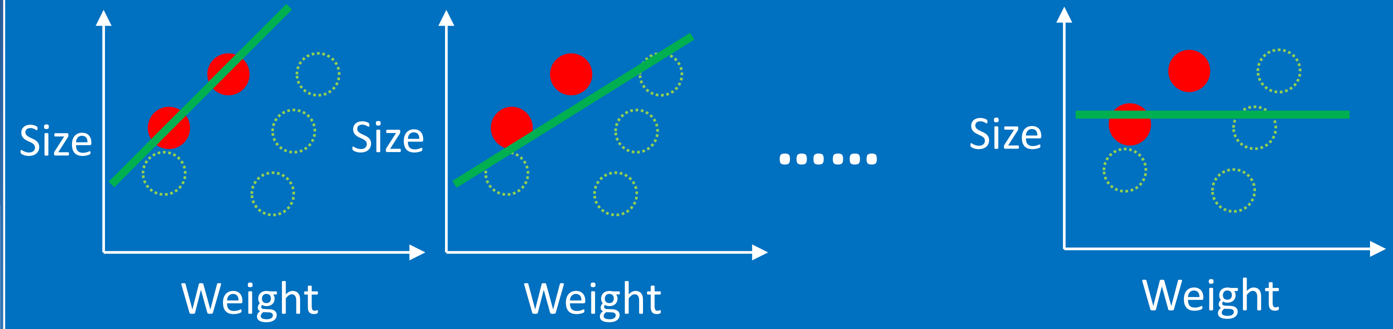
Assuming that the regression equation to calculate the size is $\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

In the left example, we use $\lambda = 1.0$. But how λ would affect the results ?



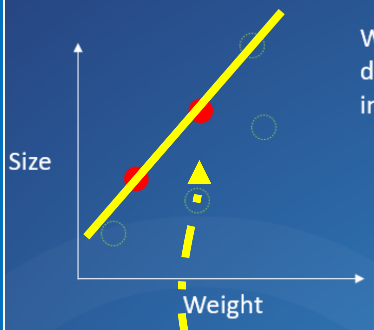
When $\lambda = 0.0$, the error is the same to the one without any regularization

When $\lambda = 1.0$, the fitted line will be less steep

When we keep increasing $\lambda = 10000 +$, the fitted line will be less and less steep …

Let’s use Linear regression as an example

Ridge regularization step-by-step



Let’s go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$size = intercept + slop \times weight$

The line overlap the data points

Without regularization, the cost function minimizes “the sum of the squared residuals”

With regularization, the cost function minimizes “the sum of the squared residuals” + $\lambda \times slope^2$

Assuming that the regression equation to calculate the size is

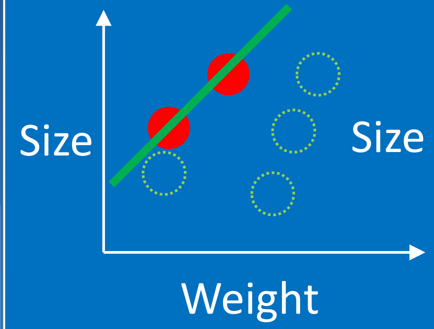
$\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

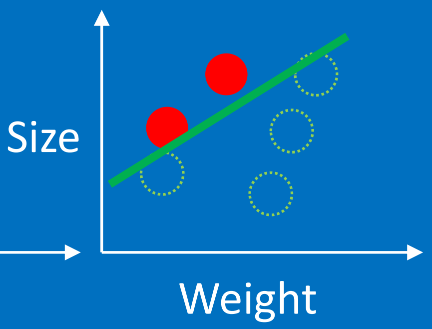
Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

In the left example, we use $\lambda = 1.0$. But how λ would affect the results ?

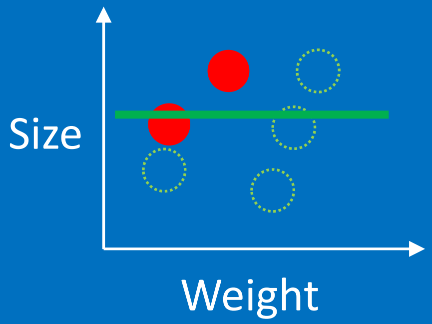


When $\lambda = 0.0$, the error is the same to the one without any regularization



When $\lambda = 1.0$, the fitted line will be less steep

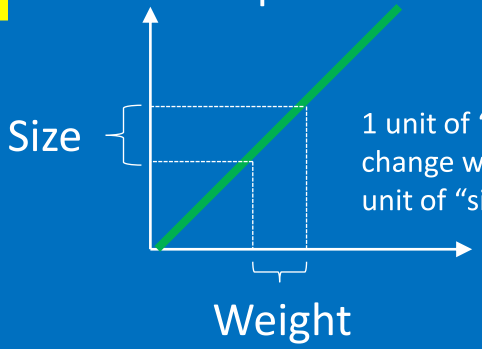
……



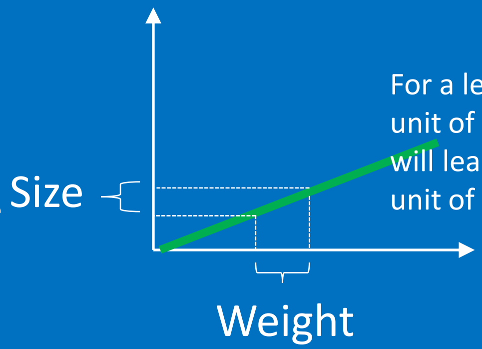
When we keep increasing $\lambda = 10000 +$, the fitted line will be less and less steep …

This means that the prediction (“size”) is getting less dependant on the dependencies (e.g., “weight”)

For example:



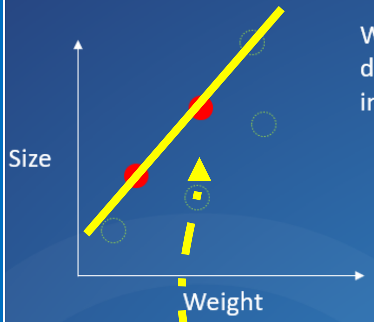
1 unit of “weight” change will lead to 1 unit of “size” change



For a less steep line: 1 unit of “weight” change will lead to less than 1 unit of “size” change

Let’s use Linear regression as an example

Ridge regularization step-by-step



Let’s go back to just the training data (two red dots)

When linear regression applied, we need to determine values for the following two parameters in this equation (intercept and slop)

$$size = intercept + slop \times weight$$

The line overlap the data points

Without regularization, the cost function minimizes “the sum of the squared residuals”

With regularization, the cost function minimizes “the sum of the squared residuals” + $\lambda \times slope^2$

Assuming that the regression equation to calculate the size is

$\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

Error = $0 + 1.0 \times 1.3^2 = 1.69$

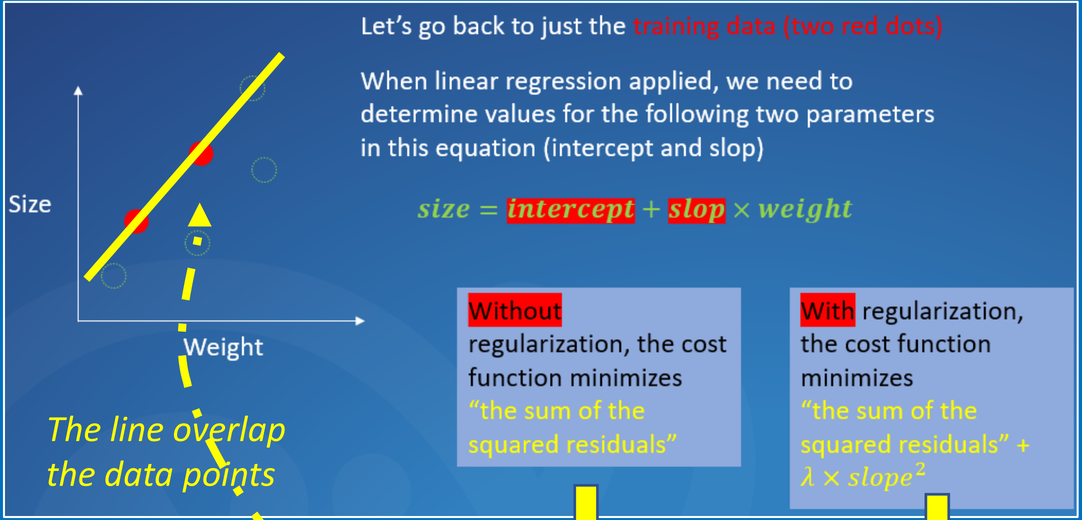
Assuming $\lambda = 1.0$

So what we do in practical is that

- We try a bunch of λ value, and
- then create corresponding fitted lines (regression equations)
- then compare them (through cross validation) and
- get the best one

Let’s use Linear regression as an example

Ridge regularization step-by-step



Assuming that the regression equation to calculate the size is

$\underline{size = 0.4 + 1.3 \times weight}$

Error = The sum of the squared residuals=0

Error = $0 + 1.0 \times 1.3^2 = 1.69$

Assuming $\lambda = 1.0$

So what we do in practical is that

- We try a bunch of λ value, and
- then create corresponding fitted lines (regression equations)
- then compare them (through cross validation) and
- get the best one

Regularization uses a new error to make the prediction being less dependant on the dependants (“shrinking dependants”), and therefore reducing the impact of “overfitting”

Let’s use Linear regression as an example

There are other regularization approach, e.g., Lasso regularization

Let's use Linear regression as an example

There are other regularization approach, e.g., Lasso regularization

- For Ridge regularization, the error is

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

Let's use Linear regression as an example

There are other regularization approach, e.g., Lasso regularization

- For Ridge regularization, the error is

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

- For Lasso regularization, the error is

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

Let's use Linear regression as an example

There are other regularization approach, e.g., Lasso regularization

- For Ridge regularization, the error is

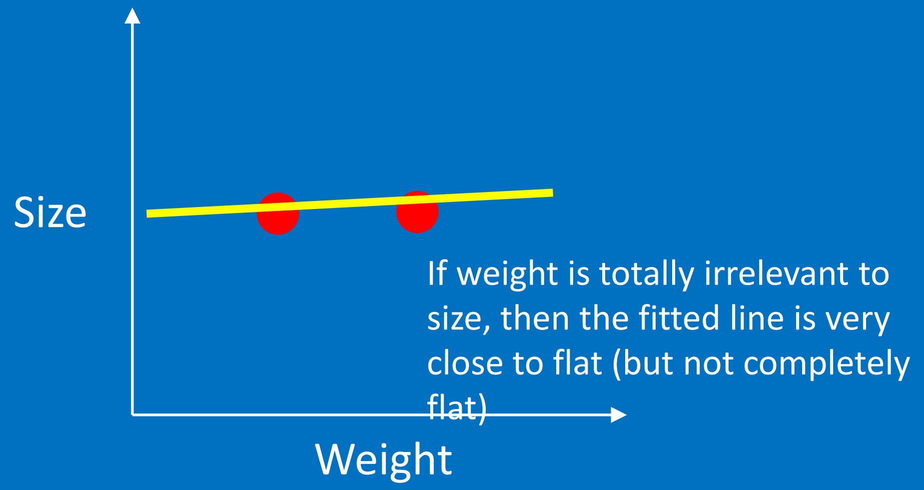
  “the sum of the squared residuals” + $\lambda \times slope^2$

- For Lasso regularization, the error is

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

The big difference here is that

For Ridge regularization



Let’s use Linear regression as an example

There are other regularization approach, e.g., Lasso regularization

- For Ridge regularization, the error is

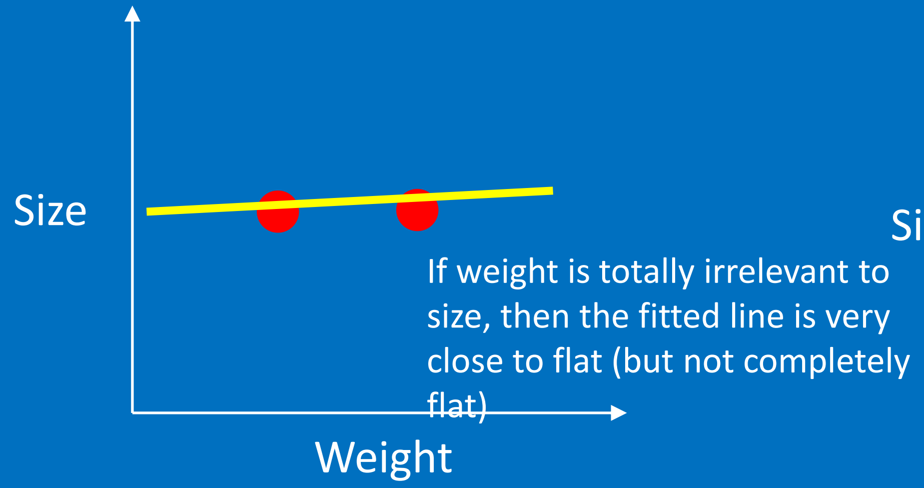
  “the sum of the squared residuals” + $\lambda \times slope^2$

- For Lasso regularization, the error is

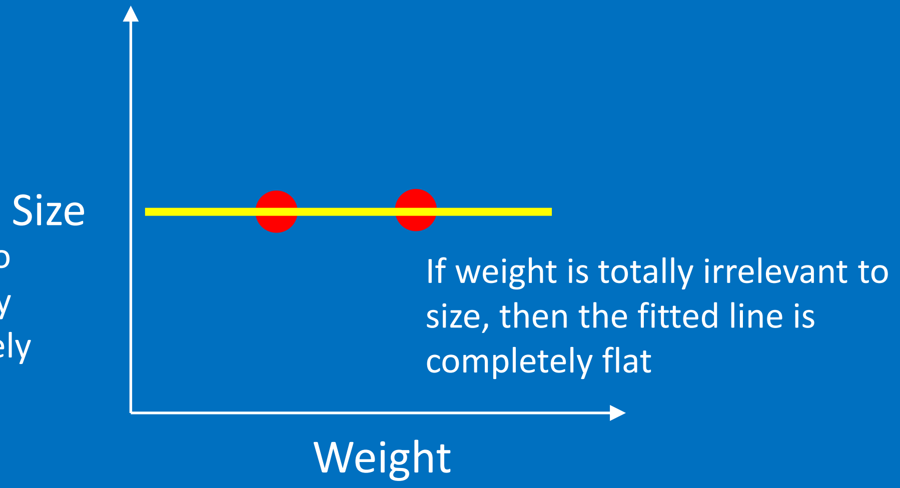
  “the sum of the squared residuals” + $\lambda \times |slope|$

The big difference here is that

For Ridge regularization



For Lasso regularization



Let’s use Linear regression as an example

There are other regularization approach, e.g., Lasso regularization

- For Ridge regularization, the error is

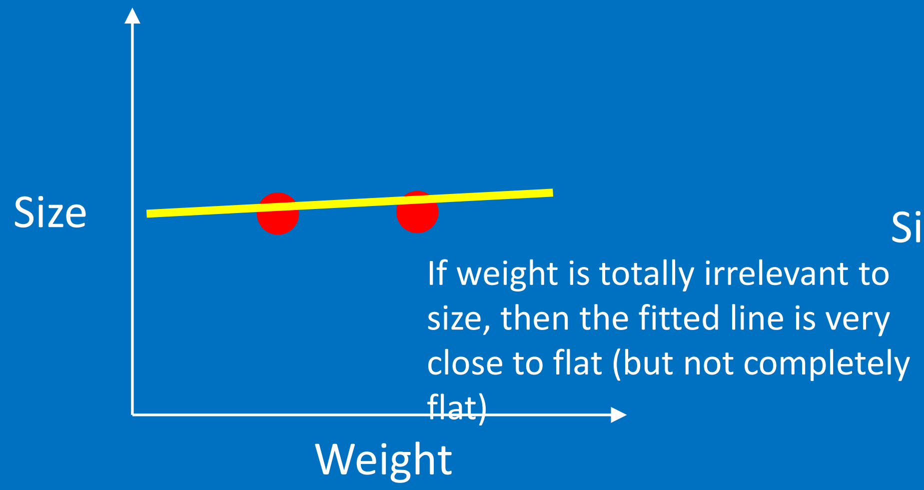
  “the sum of the squared residuals” + $\lambda \times slope^2$

- For Lasso regularization, the error is

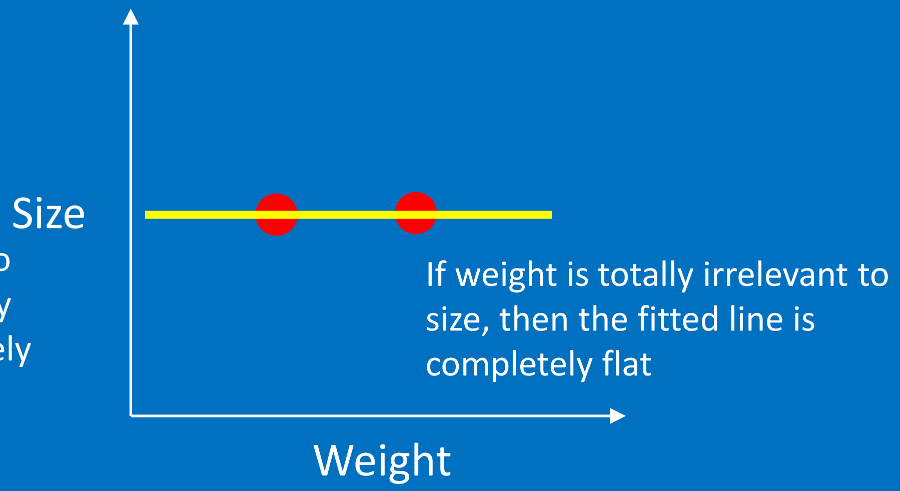
  “the sum of the squared residuals” + $\lambda \times |slope|$

The big difference here is that

For Ridge regularization



For Lasso regularization



This means that Lasso regularization is better at removing irrelevant dependants than Ridge regularization