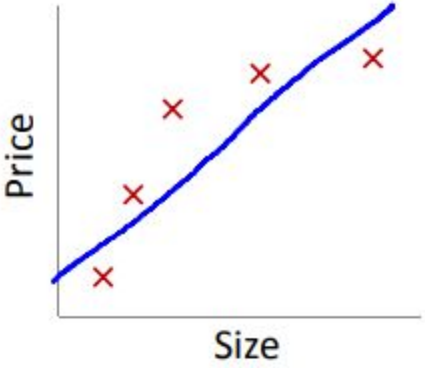
Regularization : Rationale

- Regularization helps to solve overfitting problem in machine learning.
- Simple model will be a very poor generalization of data. At the same time, complex model may not perform well in test data due to overfitting.
- Need to choose the right model in between simple and complex model.
- Regularization helps to choose preferred model complexity, so that model is better at predicting.
- Regularization is nothing but adding a penalty term to the objective function and control the model complexity using that penalty term. It can be used for many machine learning algorithms.

What is overfitting

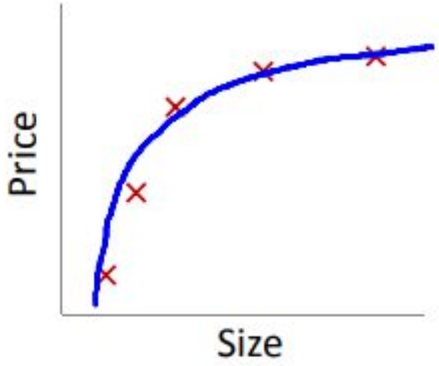
- Building a model that matches the training data "too closely".
- Learning from the error/disturbance/noise in the data, rather than just the true values.

Example: Linear regression (housing prices)



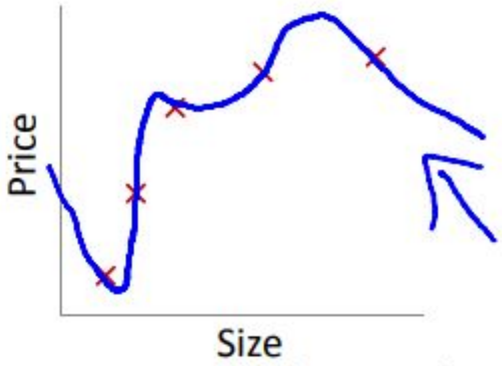
→ $\theta_0 + \theta_1 x$

"Underfit" "High bias"



→ $\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

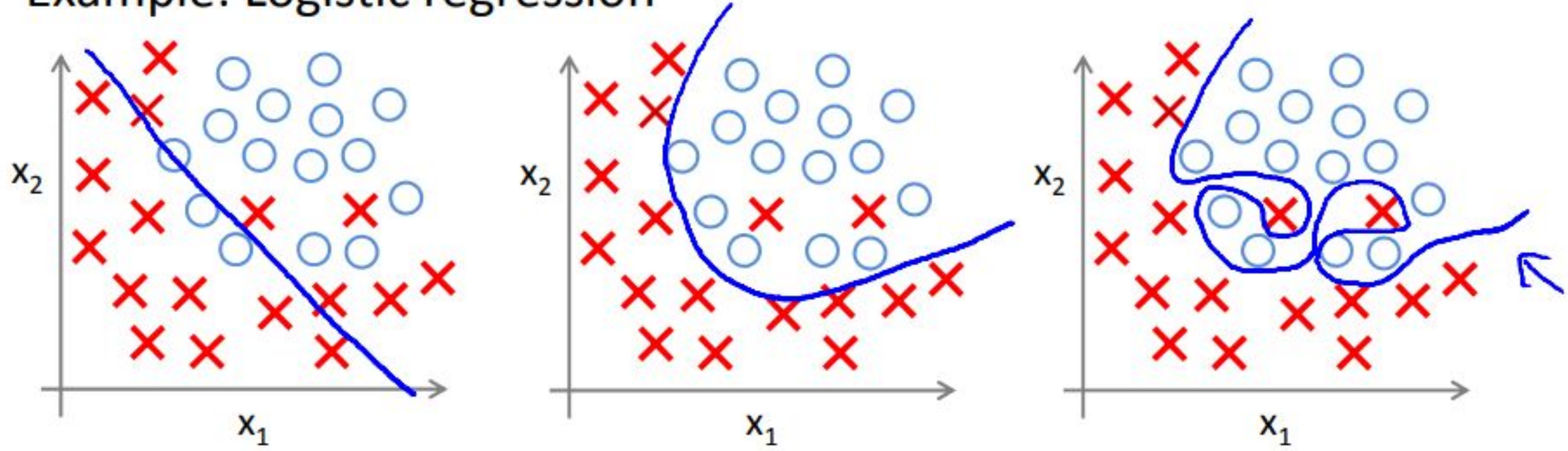


→ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 \underline{x^4}$

"Overfit" "High variance"

Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($\underline{J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \approx 0}$), but fail to generalize to new examples (predict prices on new examples).

Example: Logistic regression



$$h_\theta(x) = g(\underline{\theta_0 + \theta_1 x_1 + \theta_2 x_2})$$

(g = sigmoid function)

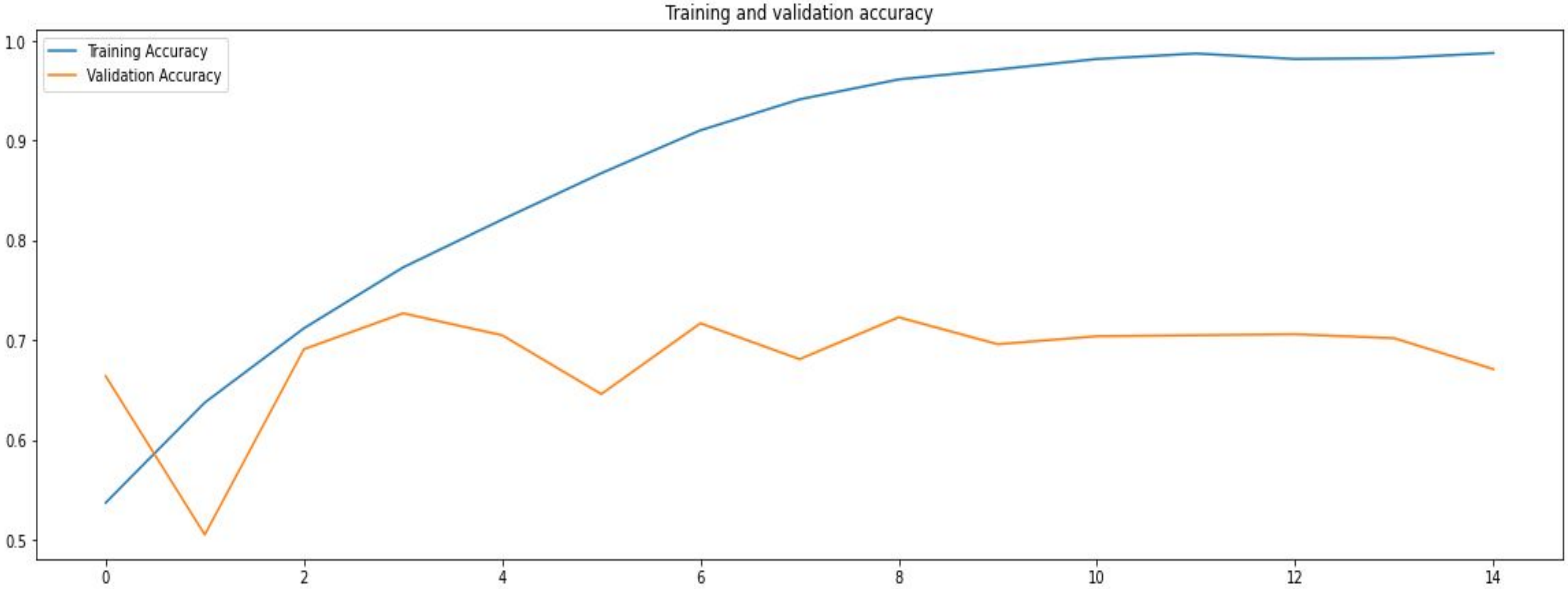
"Underfit"

$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 \underline{x_1^2} + \theta_4 \underline{x_2^2} + \theta_5 \underline{x_1 x_2})$$

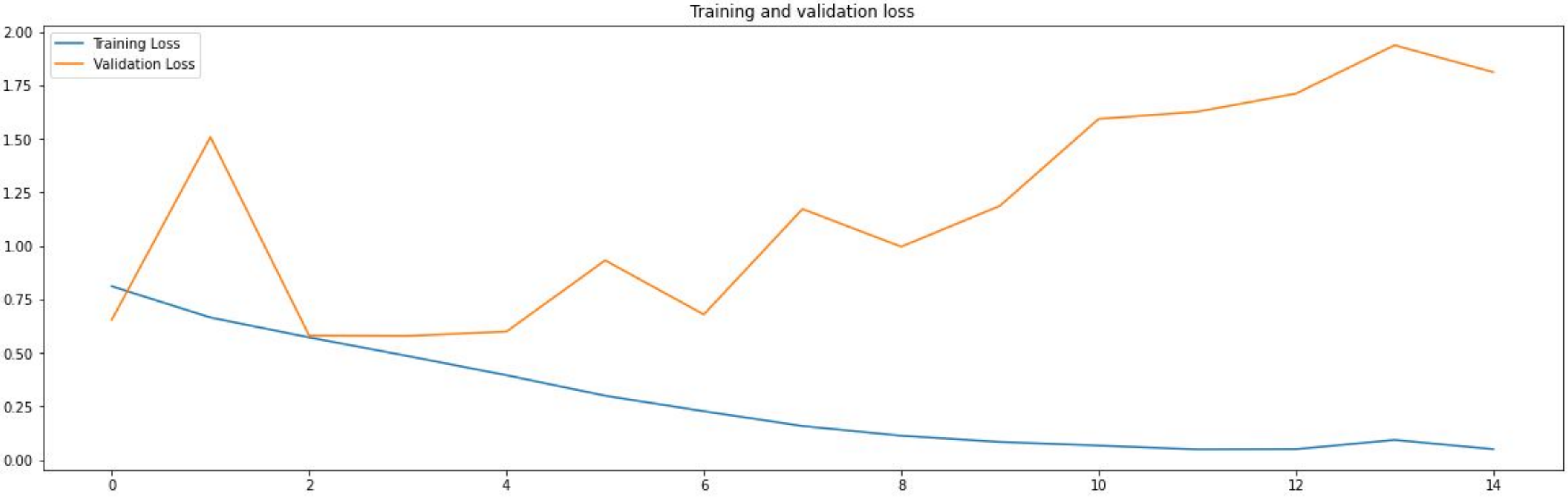
$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 \underline{x_1^2 x_2} + \theta_4 \underline{x_1^2 x_2^2} + \theta_5 \underline{x_1^2 x_2^3} + \theta_6 \underline{x_1^3 x_2} + \ldots)$$

"Overfit"

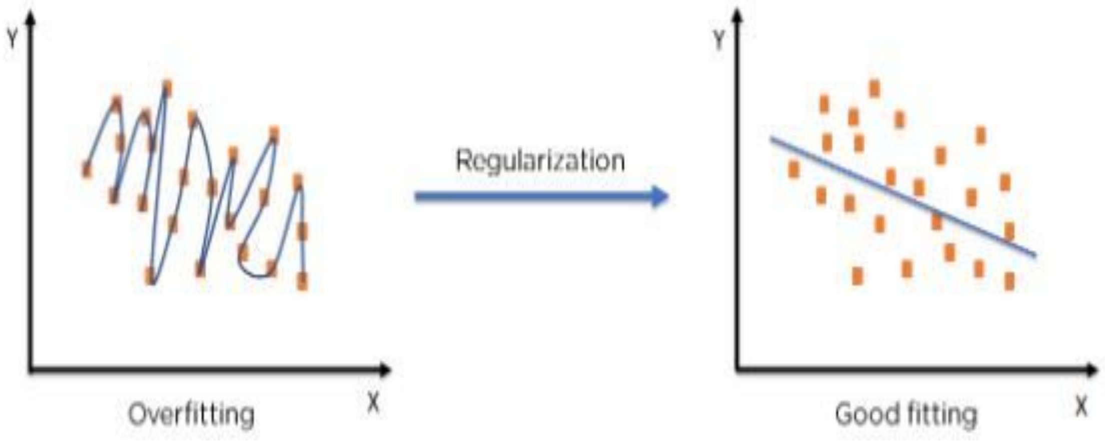
Overfitted Model? *Training and Validation Accuracy*



Overfitted Model?



Regularization on an over-fitted model



How does overfitting occur?

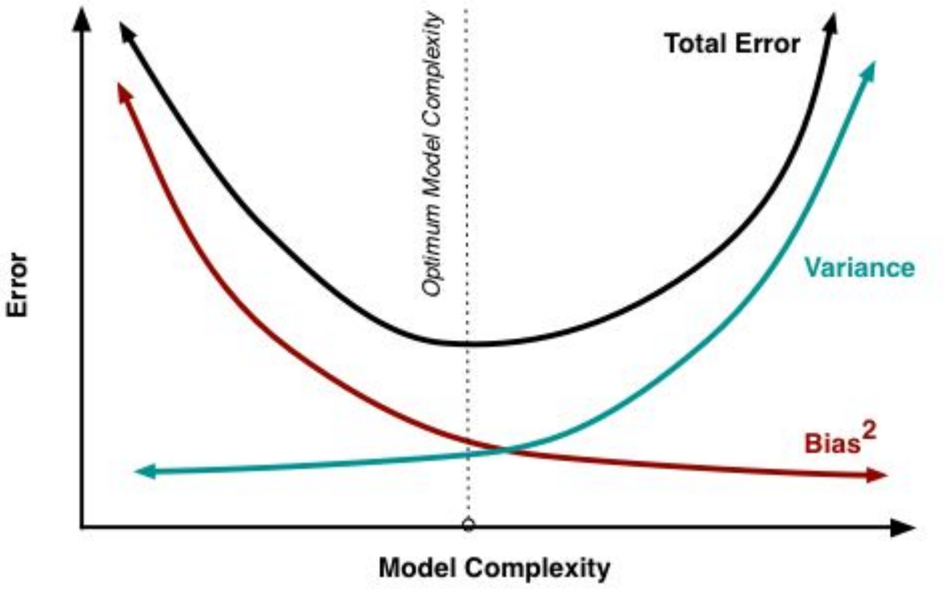
- Evaluating a model by testing it on the same data that was used to train it.
- Creating a model that is "too complex".
    a. Irrelevant features
    b. Correlated features(Muticollinearity)
    c. Large coefficients

What is the impact of overfitting?

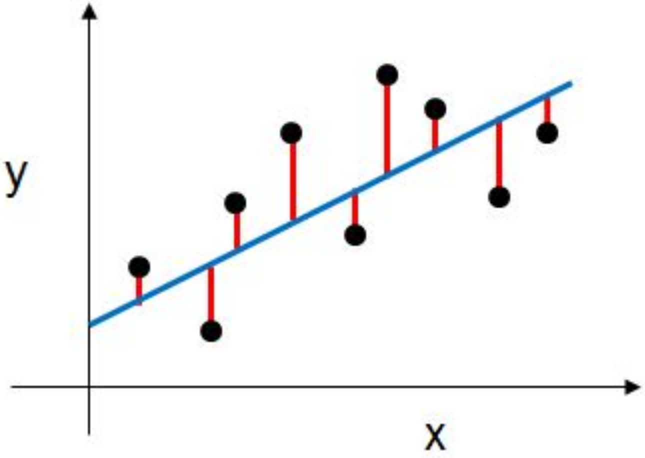
- Model will do well on the training data, but won't generalize to out-of-sample data i.e., test
- Model will have low bias, but high variance.

REgularization help!

Our aim is to locate the **optimum model complexity**, and thus regularization is useful when we believe our model is too complex.



Normal linear Regression

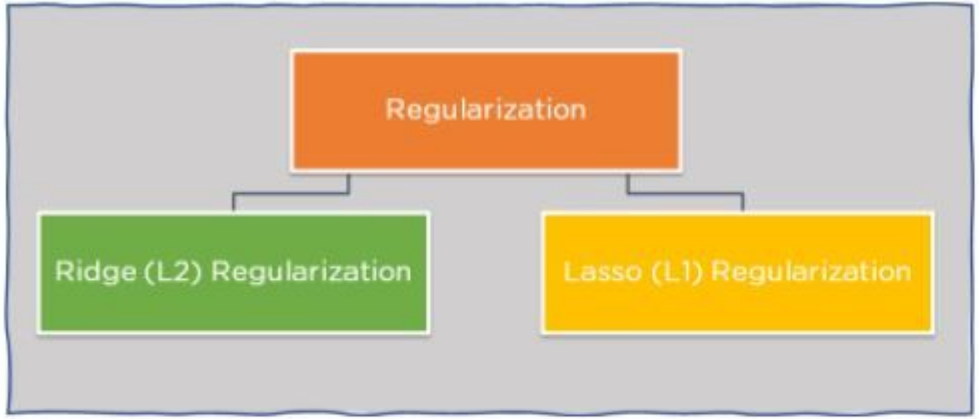


$$SS_{residuals} = \sum_{i=1}^{N} (\overset{\text{Model Prediction}}{\hat{y}_i} - \underset{\text{Observed Result}}{y_i})^2$$

Techniques of Regularization

Mainly, there are two types of regularization techniques,below:

- Ridge Regression ("L2 regularization")
- Lasso Regression (“L1 regularization")



Ridge Regression ("L2 regularization")

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

- Ridge regression is one of the types of linear regression in which we introduce a small amount of bias, known as Ridge regression penalty so that we can get better long-term predictions.
- In this technique, the cost function is altered by adding the penalty term (shrinkage term), which multiplies the lambda with the squared weight of each individual feature.
- In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the magnitudes of the coefficients that help to decrease the complexity of the model.

Lasso Regression (“L1 regularization")

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

- It stands for Least Absolute Shrinkage and Selection Operator
- It adds L1 the penalty
- L1 is the sum of the absolute value of the beta coefficients
- In this technique, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero which means there is a complete removal of some of the features for model evaluation when the tuning parameter λ is sufficiently large. Therefore, the lasso method also performs Feature selection and is said to yield sparse models.

Limitation of Lasso Regression

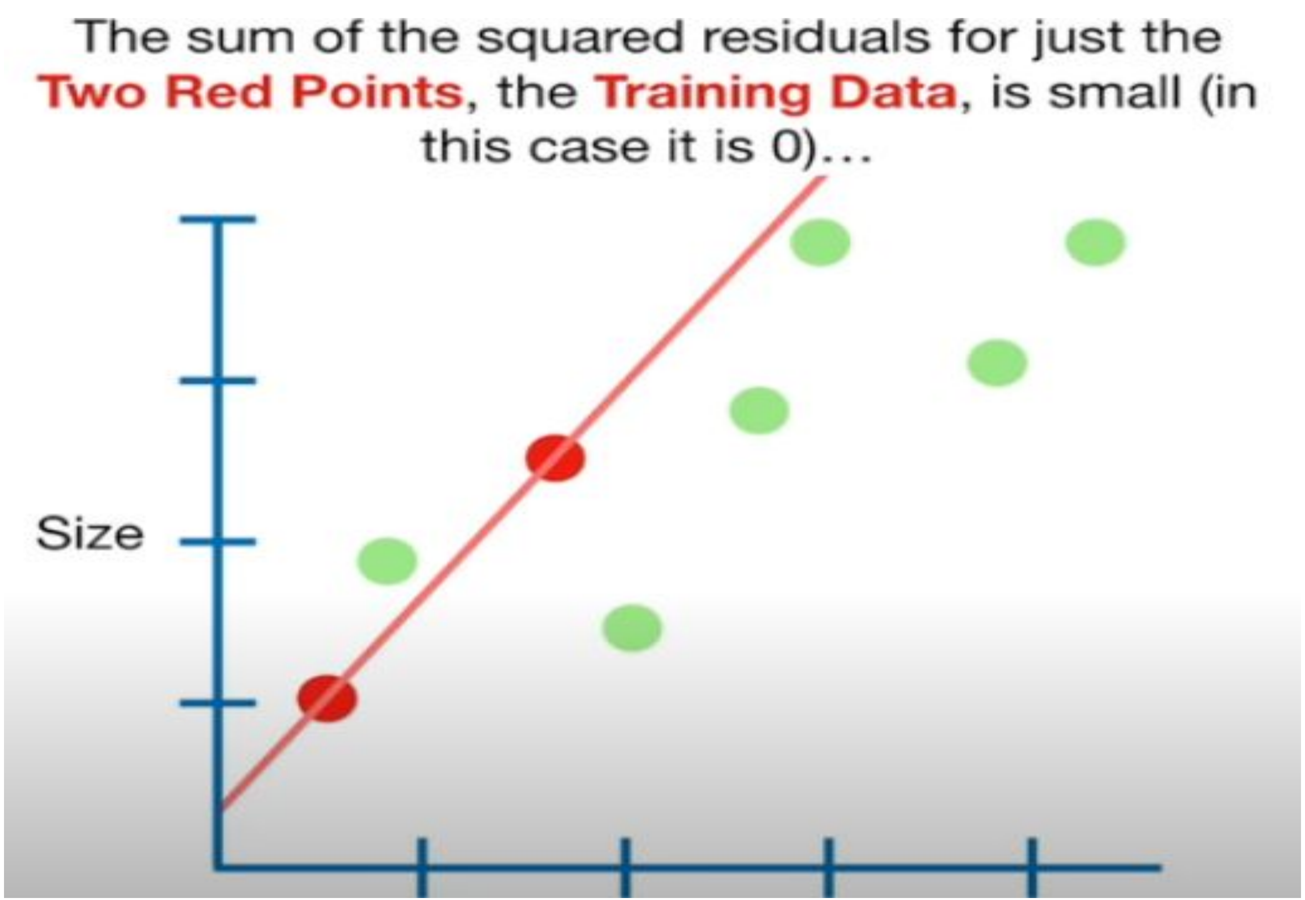
- Problems with some types of Dataset: If the number of predictors is greater than the number of data points, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant.
- Multicollinearity Problem: If there are two or more highly collinear variables then LASSO regression selects one of them randomly which is not good for the interpretation of our model.

LAsso vs Ridge

- **Lasso regression** shrinks coefficients all the way to zero, thus removing them from the model
- **Ridge regression** shrinks coefficients toward zero, but they rarely reach zero
- Ridge regression helps us to reduce only the overfitting in the model while keeping all the features present in the model. It reduces the complexity of the model by shrinking the coefficients whereas Lasso regression helps in reducing the problem of overfitting in the model as well as automatic feature selection.

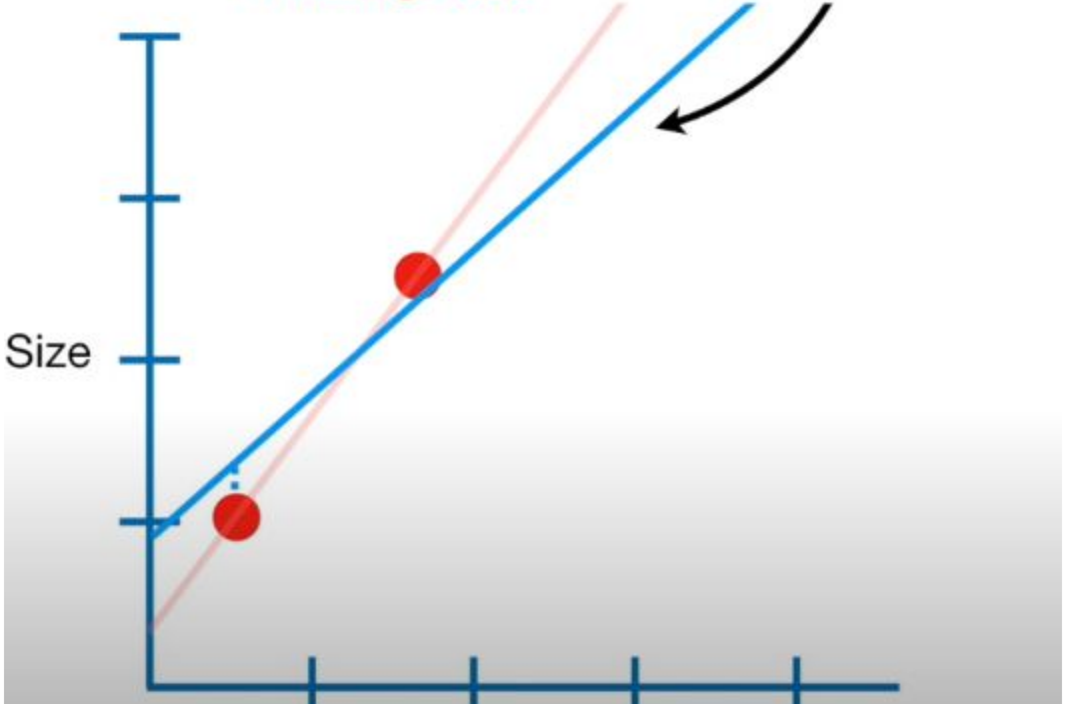
Important points about λ:

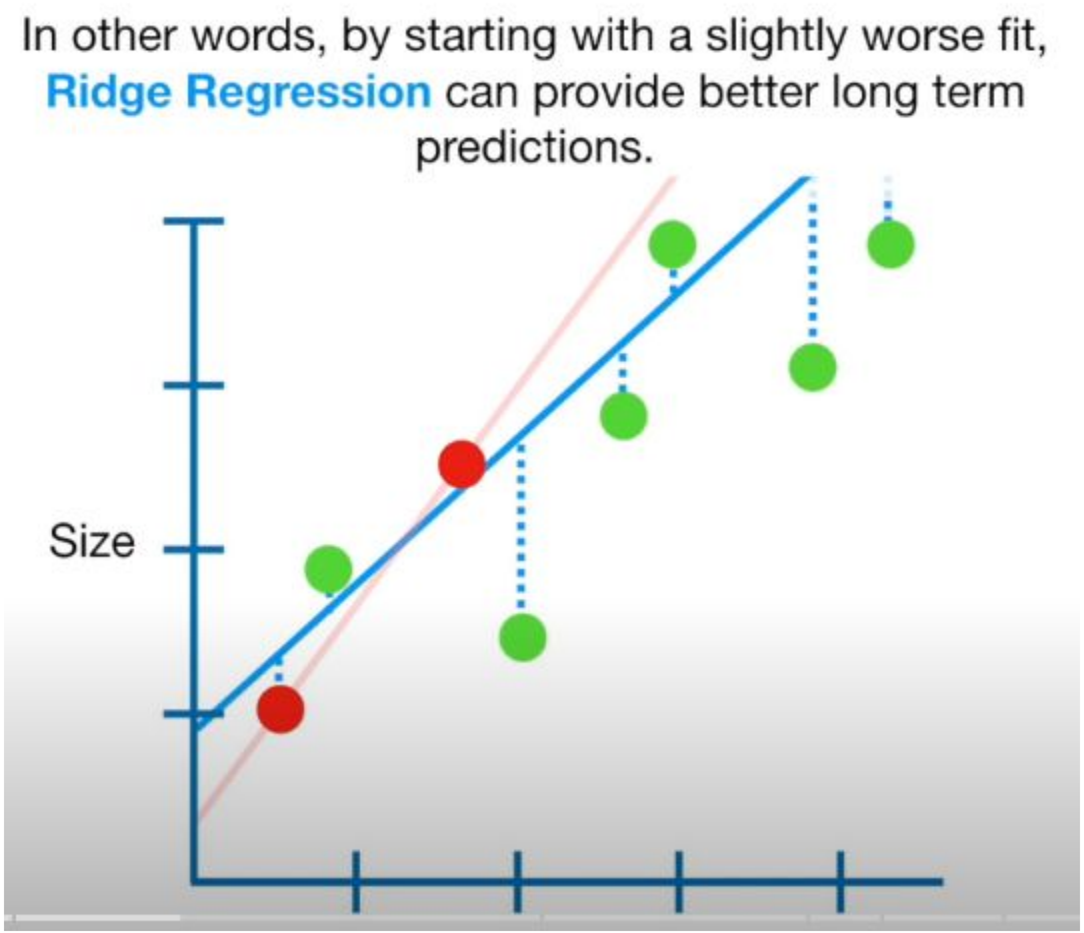
- λ is the tuning parameter used in regularization that decides how much we want to penalize the flexibility of our model i.e, controls the impact on bias and variance.
- When λ = 0, the penalty term has no effect, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ i.e, λ=0, the model will resemble the linear regression model. So, the estimates produced by ridge regression will be equal to least squares.
- However, as λ→∞ (tends to infinity), the impact of the shrinkage penalty increases, and the ridge regression coefficient estimates will approach zero.



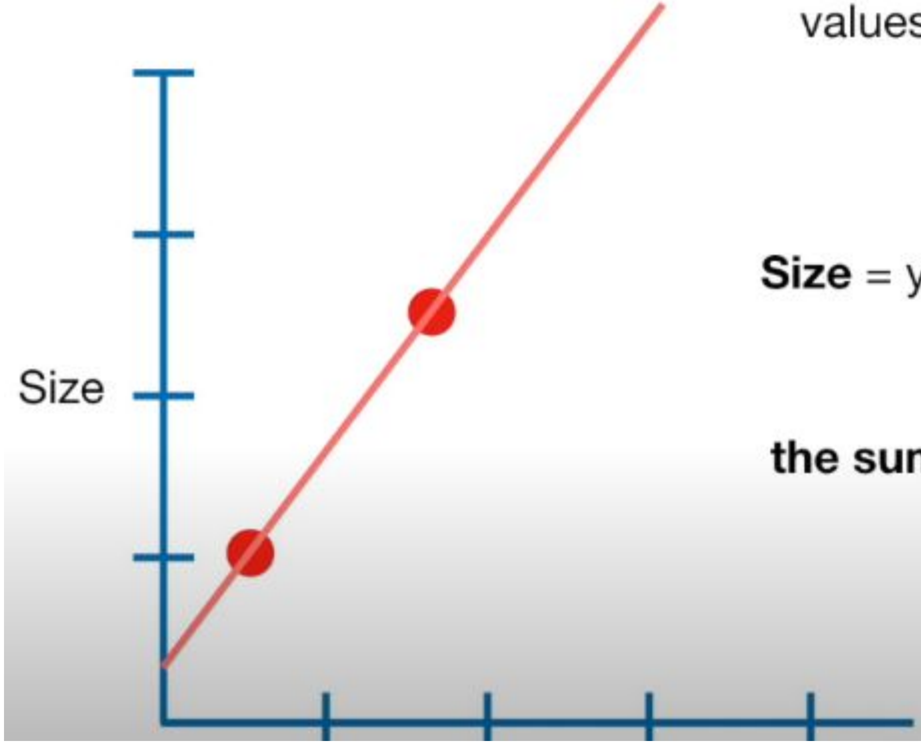
The sum of the squared residuals for just the **Two Red Points**, the **Training Data**, is small (in this case it is 0)...

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





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



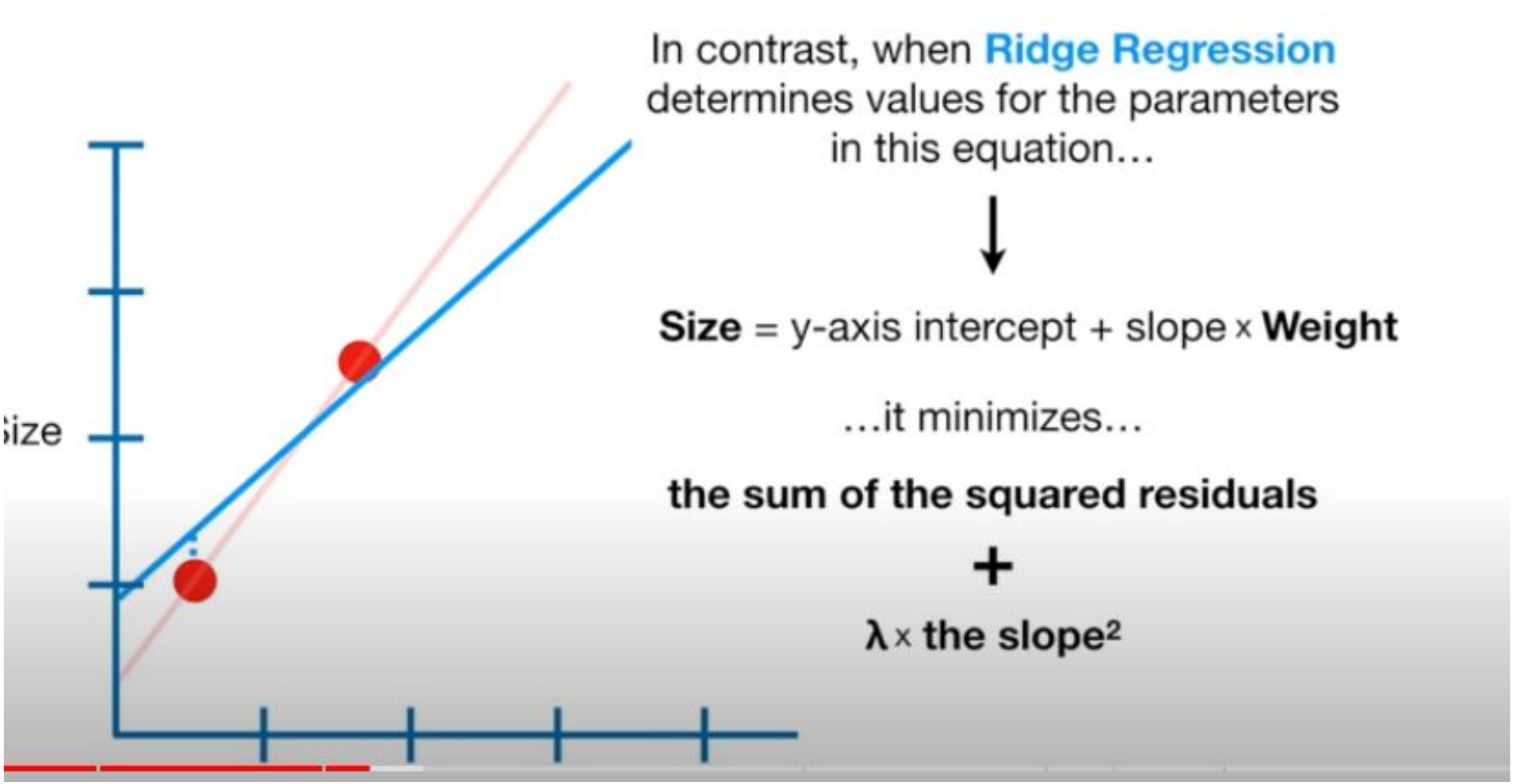
When **Least Squares** determines values for the parameters in this equation…

↓

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

…it minimizes…

the sum of the squared residuals



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

↓

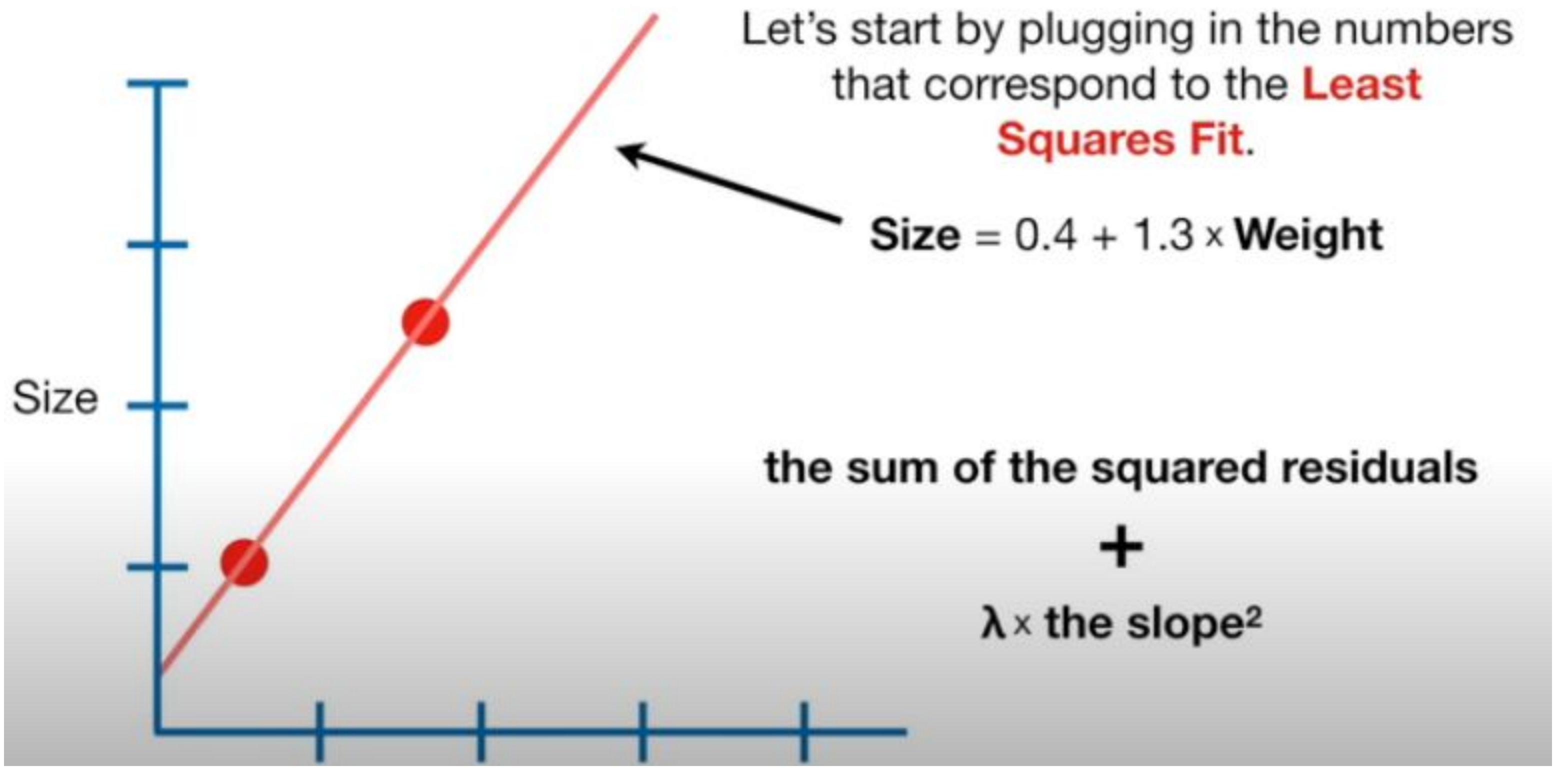
Size = y-axis intercept + slope × **Weight**

…it minimizes…

the sum of the squared residuals

+

λ × the slope²



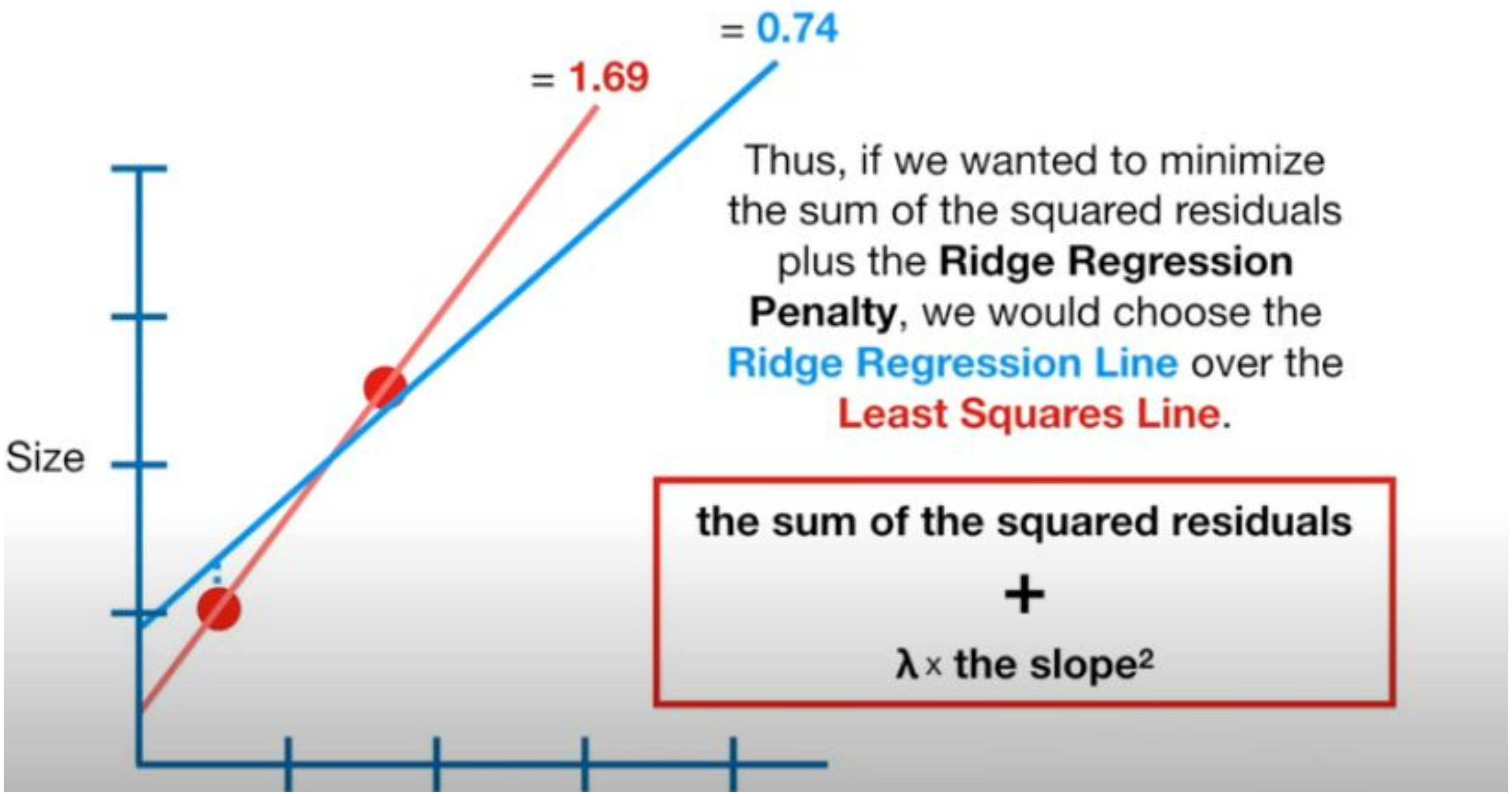
Let's start by plugging in the numbers that correspond to the **Least Squares Fit**.

Size = 0.4 + 1.3 × **Weight**

the sum of the squared residuals

+

λ × the slope^2

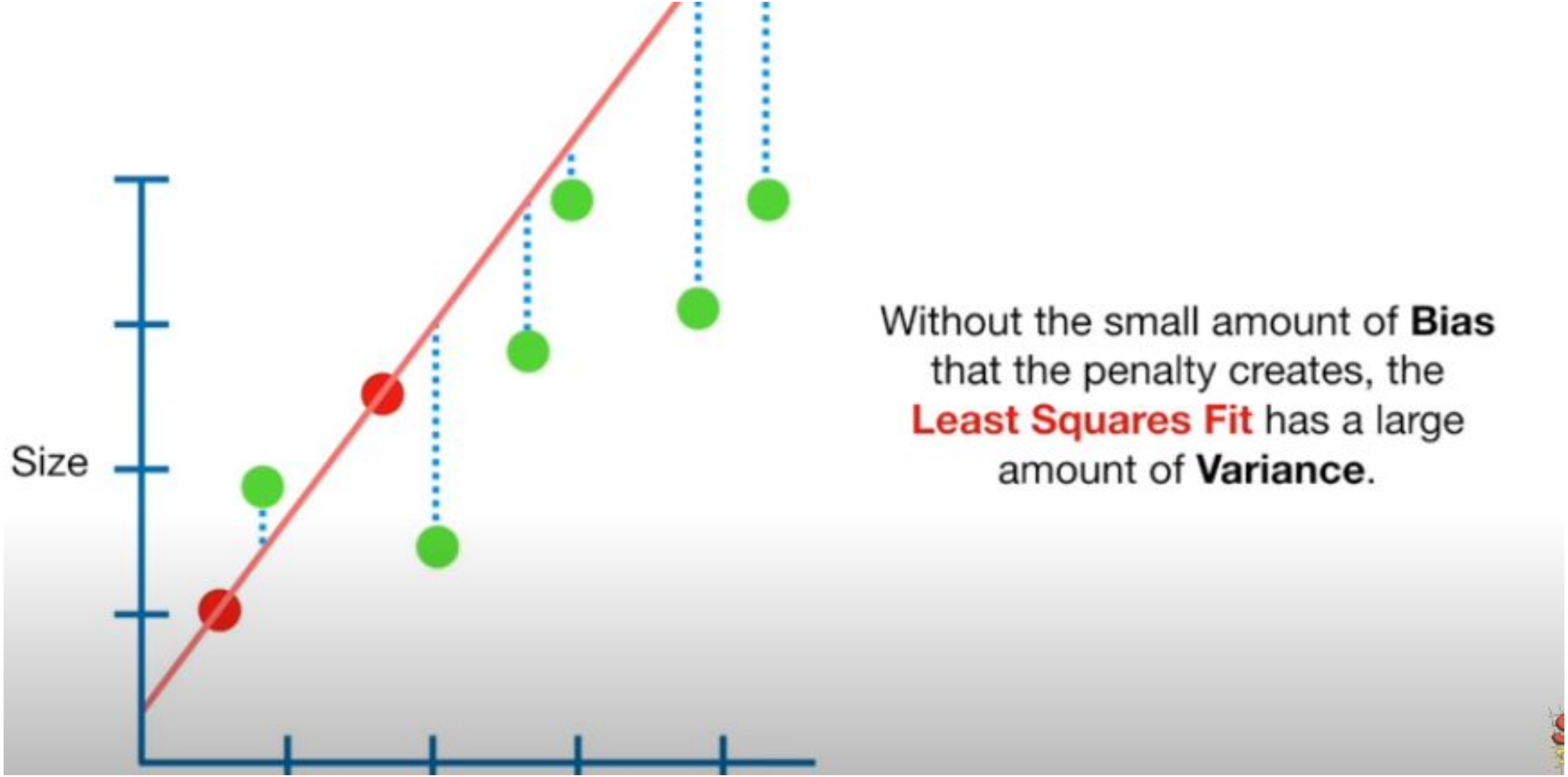


Thus, if we wanted to minimize the sum of the squared residuals plus the **Ridge Regression Penalty**, we would choose the **Ridge Regression Line** over the **Least Squares Line**.

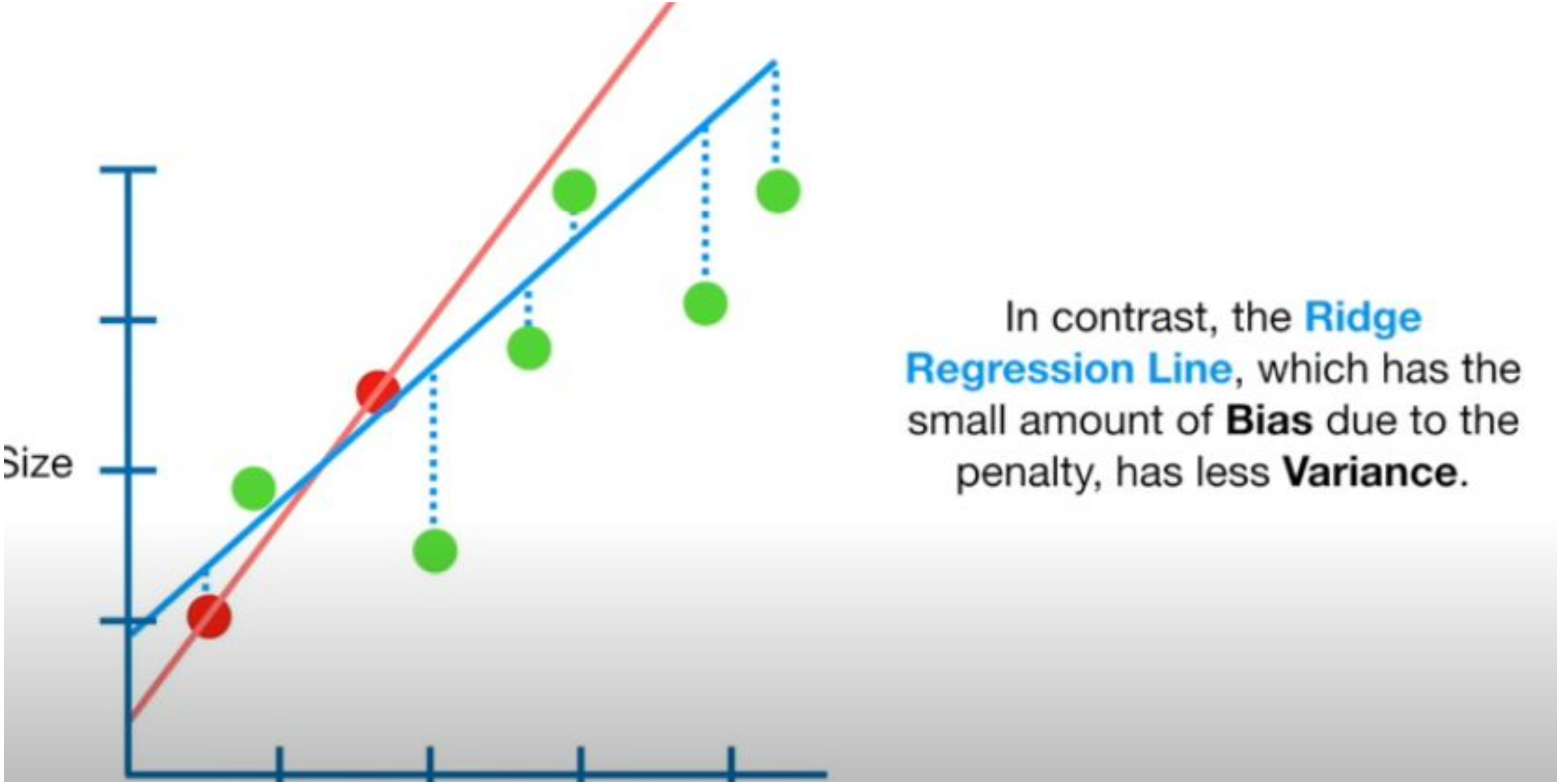
the sum of the squared residuals

+

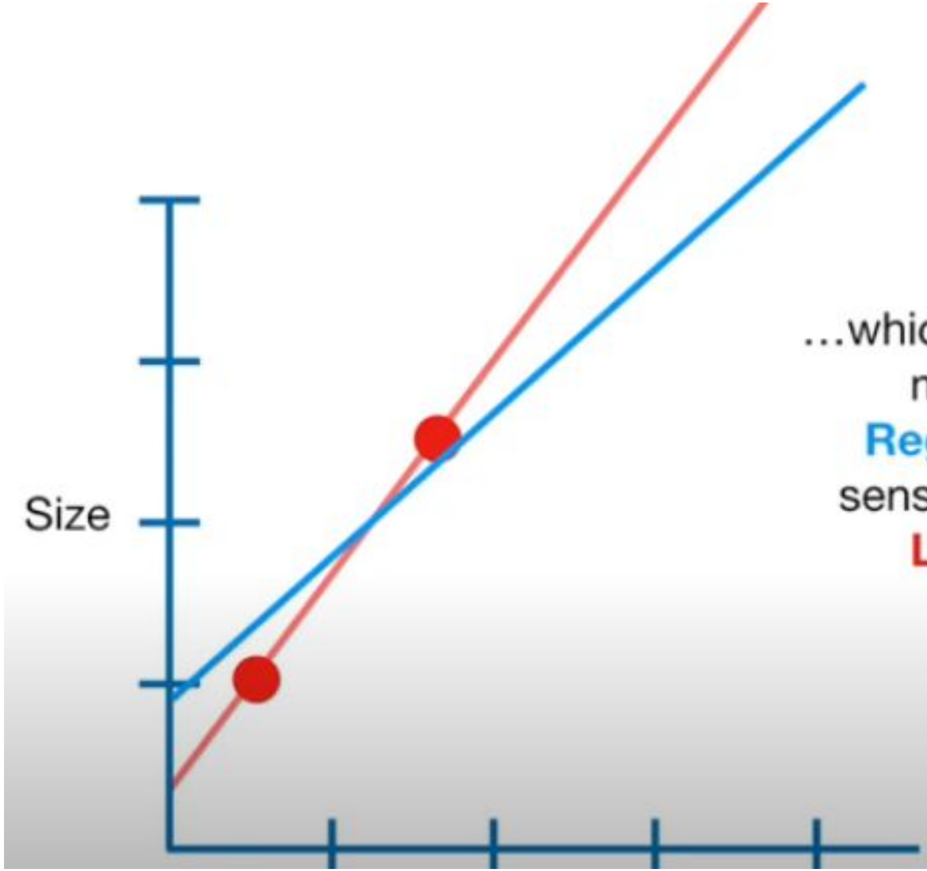
λ × the slope²



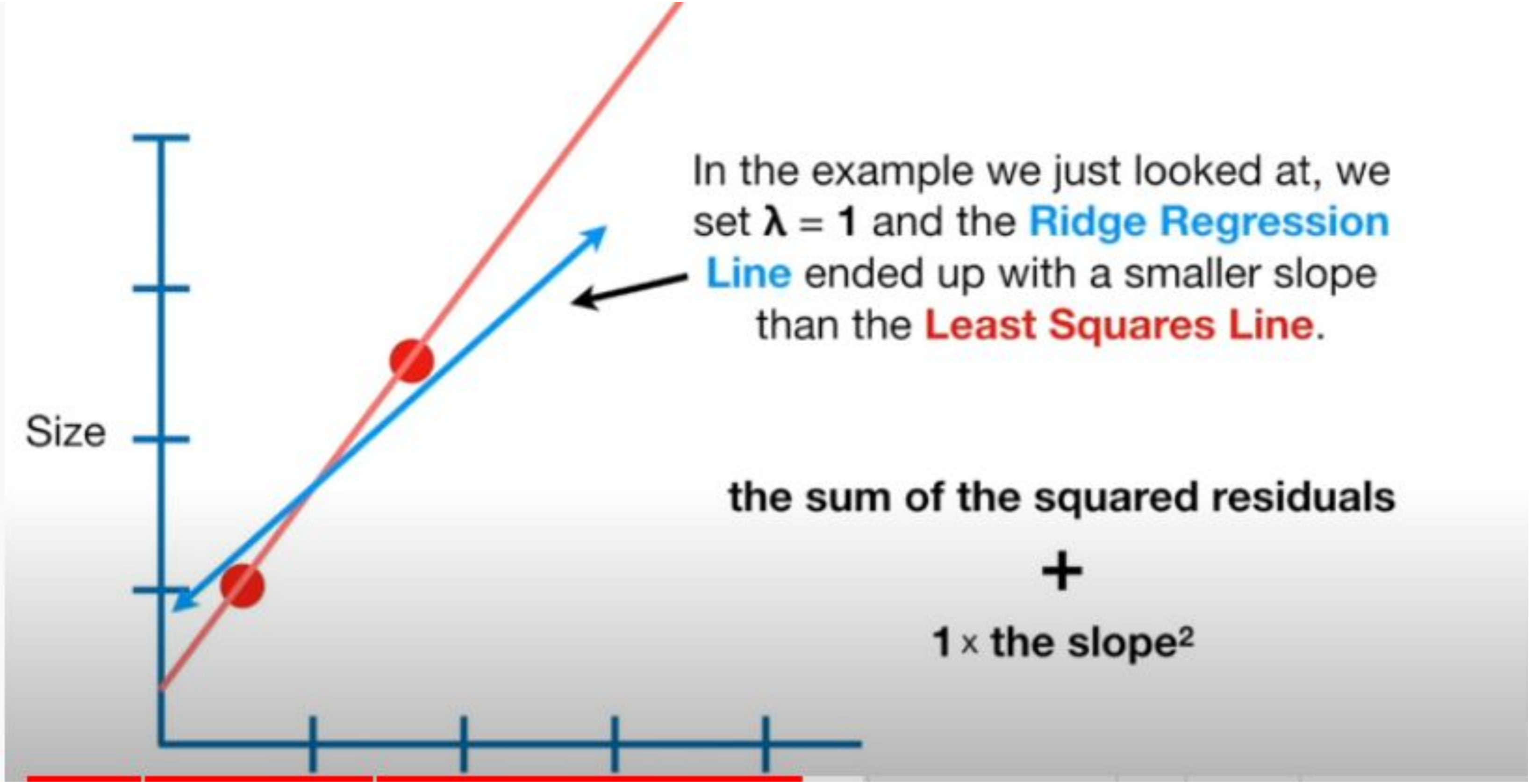
Without the small amount of **Bias** that the penalty creates, the **Least Squares Fit** has a large amount of **Variance**.



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



...which means that predictions made with the **Ridge Regression Line** are less sensitive to **Weight** than the **Least Squares Line**.

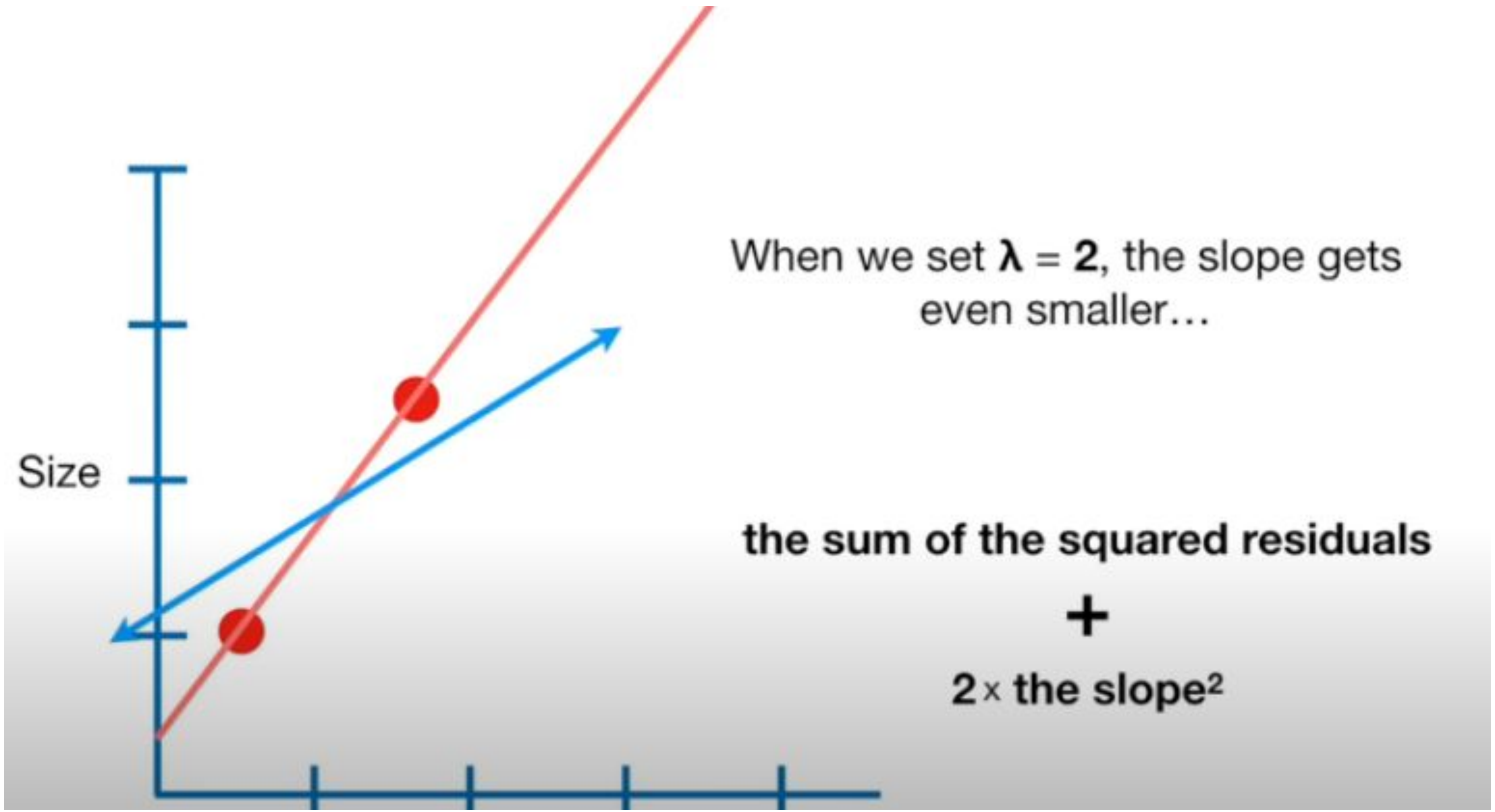


In the example we just looked at, we set $\lambda = 1$ and the **Ridge Regression Line** ended up with a smaller slope than the **Least Squares Line**.

the sum of the squared residuals

+

1 × the slope^2

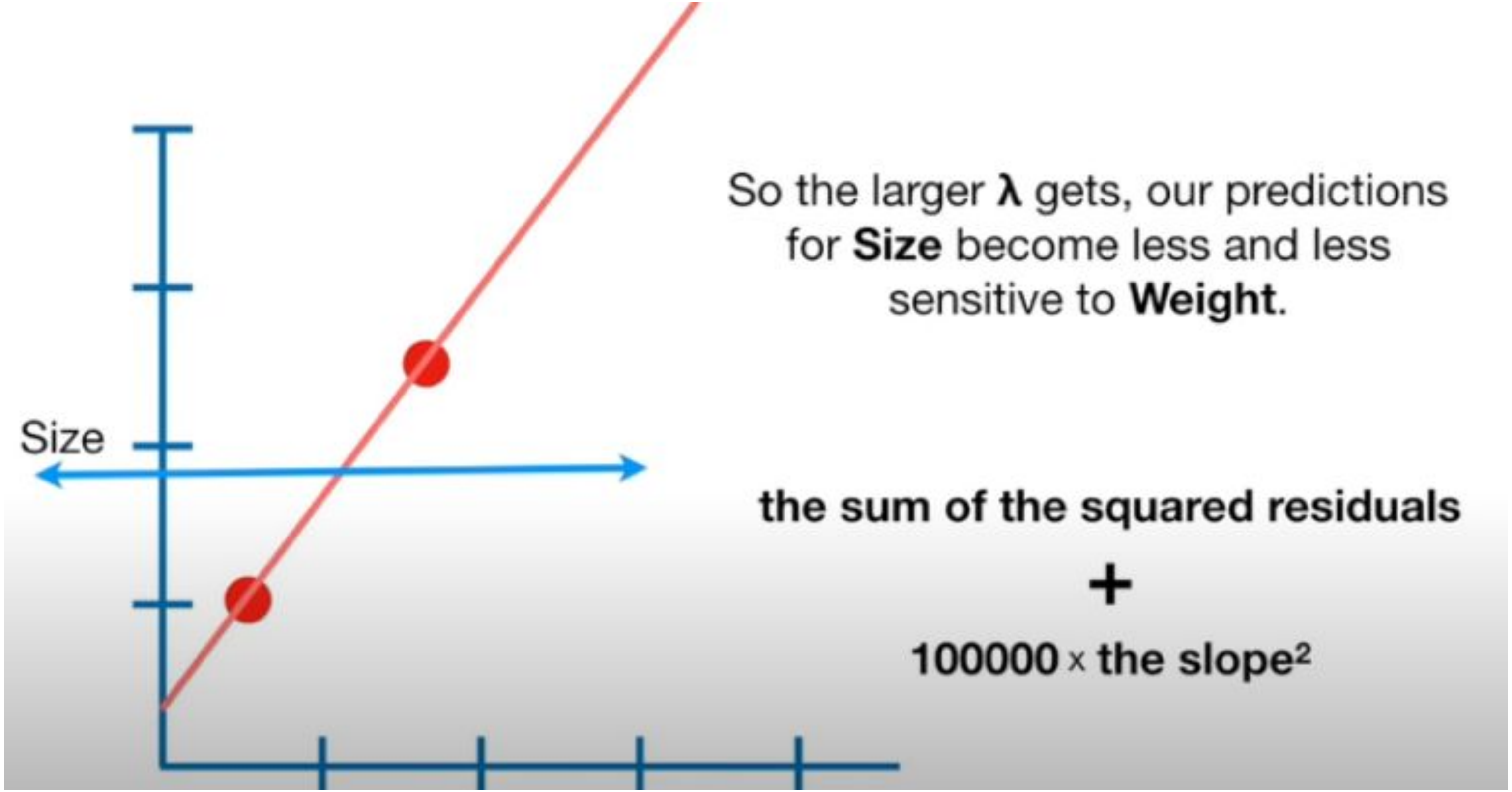


When we set **λ** = **2**, the slope gets even smaller…

the sum of the squared residuals

+

2 × **the slope²**

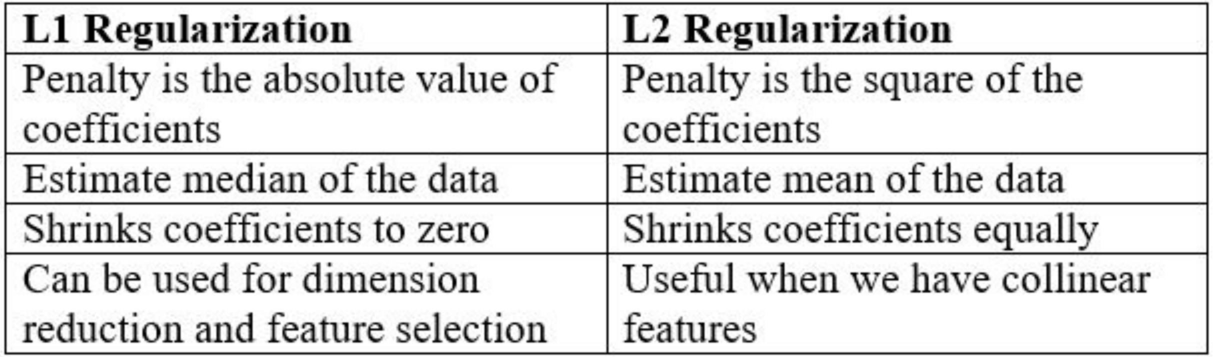


So the larger **λ** gets, our predictions for **Size** become less and less sensitive to **Weight**.

the sum of the squared residuals

+

100000 × the slope²



L1 Regularization	**L2 Regularization**
Penalty is the absolute value of coefficients	Penalty is the square of the coefficients
Estimate median of the data	Estimate mean of the data
Shrinks coefficients to zero	Shrinks coefficients equally
Can be used for dimension reduction and feature selection	Useful when we have collinear features

Conclusion

Regularization tries to reduce the variance of the model, without a substantial increase in the bias.