Regularization for Linear Regression

Data Set Up

03-Regularization-Ridge-Lasso-ElasticNet[LECs8-9].ipynb

Ridge Regression

Theory and Intuition

Ridge Regression

- Ridge Regression is a regularization technique that works by helping reduce the potential for overfitting to the training data.
- It does this by adding in a penalty term to the error that is based on the squared value of the coefficients.

Ridge Regression

- Ridge Regression is a regularization method for Linear Regression.
- Relevant Reading in ISLR:
    - Section 6.2.1
- Let’s explore the main concepts behind how Ridge Regression works…

Ridge Regression

- Recall the general formula for the regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Ridge Regression

- These Beta coefficients were solved by minimizing the residual sum of squares (RSS).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Ridge Regression

- These Beta coefficients were solved by minimizing the residual sum of squares (RSS).

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Ridge Regression

- We could substitute our regression equation for ŷ:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Ridge Regression

- We could substitute our regression equation for **ŷ:**

$$
\begin{aligned}
\mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2
\end{aligned}
$$

Ridge Regression

- We can then summarize RSS as:

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

Ridge Regression

- The goal of Ridge Regression is to help prevent overfitting by adding an additional penalty term.

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

Ridge Regression

- Ridge Regression adds a **shrinkage penalty:**

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

Ridge Regression

- Ridge Regression seeks to minimize this entire error term **RSS + Penalty**.

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

Ridge Regression

- **Shrinkage penalty** based off the squared coefficient:

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

Ridge Regression

- **Shrinkage penalty** has a **tunable lambda parameter!**

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

Ridge Regression

- Lambda determines how severe the penalty is.

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

Ridge Regression

- In theory it can be any value from 0 to positive infinity.

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

Ridge Regression

- If it is zero, then it is simply back to RSS.

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

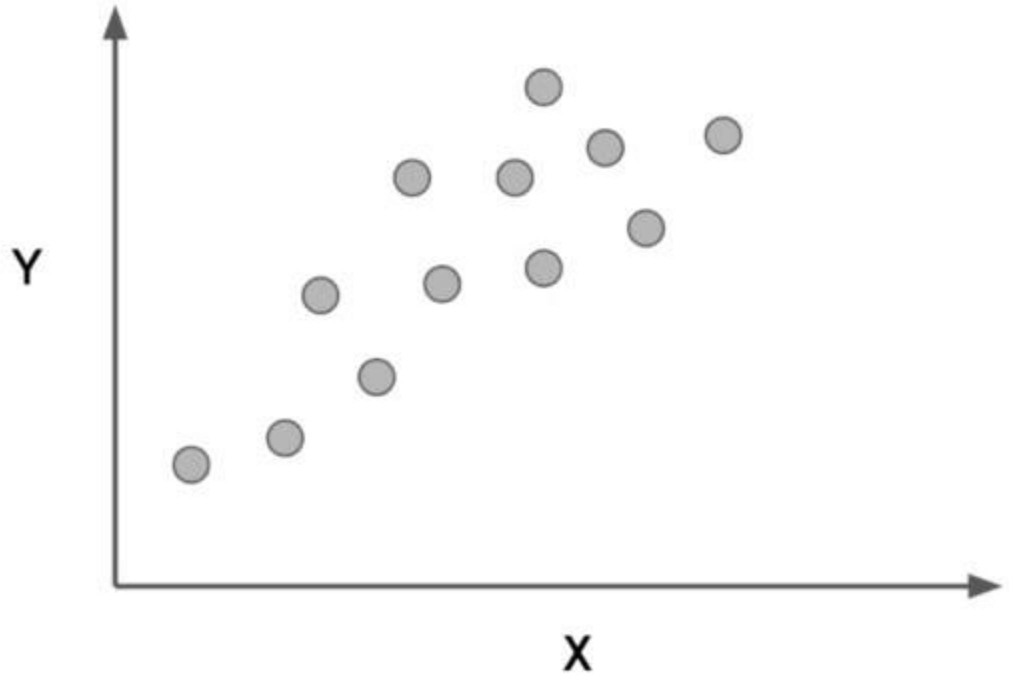
Ridge Regression

- Let’s explore a simple thought experiment to get an intuition behind Ridge Regression...

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

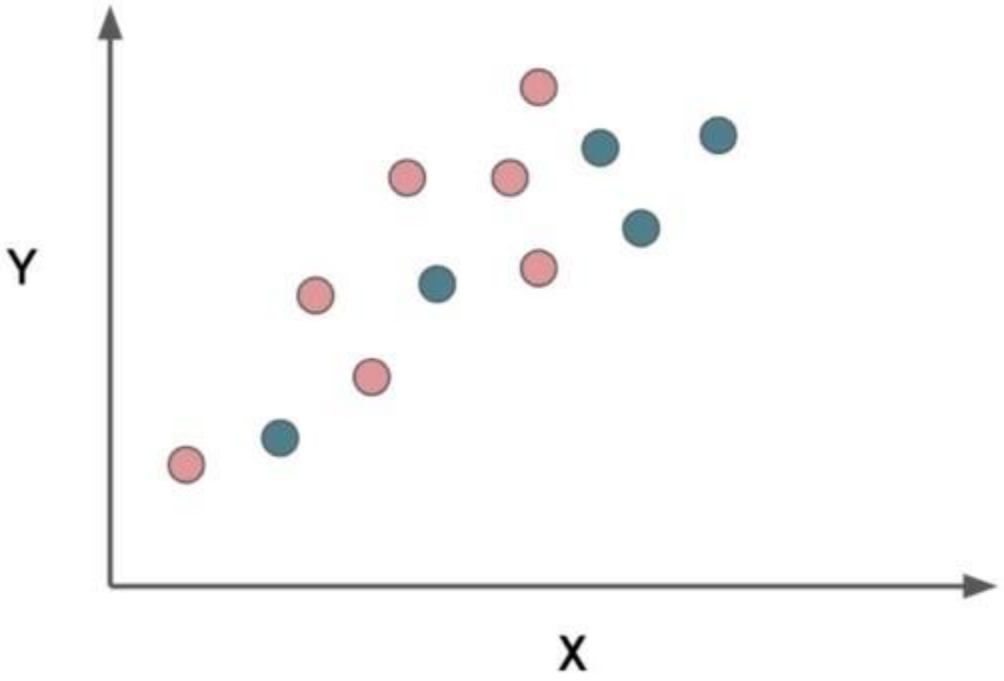
Ridge Regression

- Imagine the following data set.



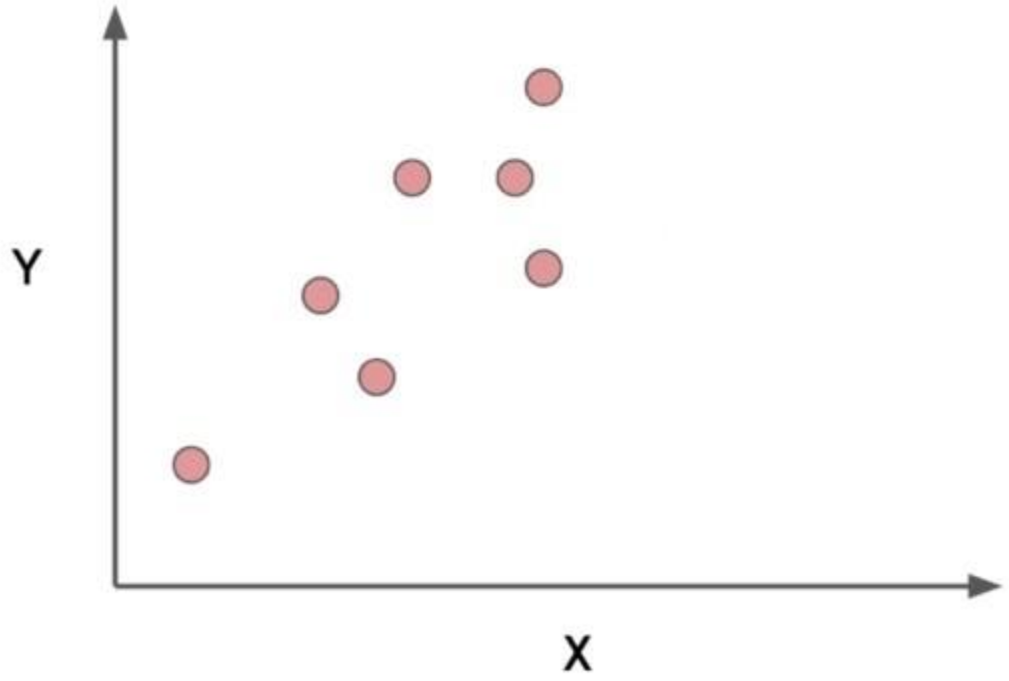
Ridge Regression

- We can split it into a training set and test set:



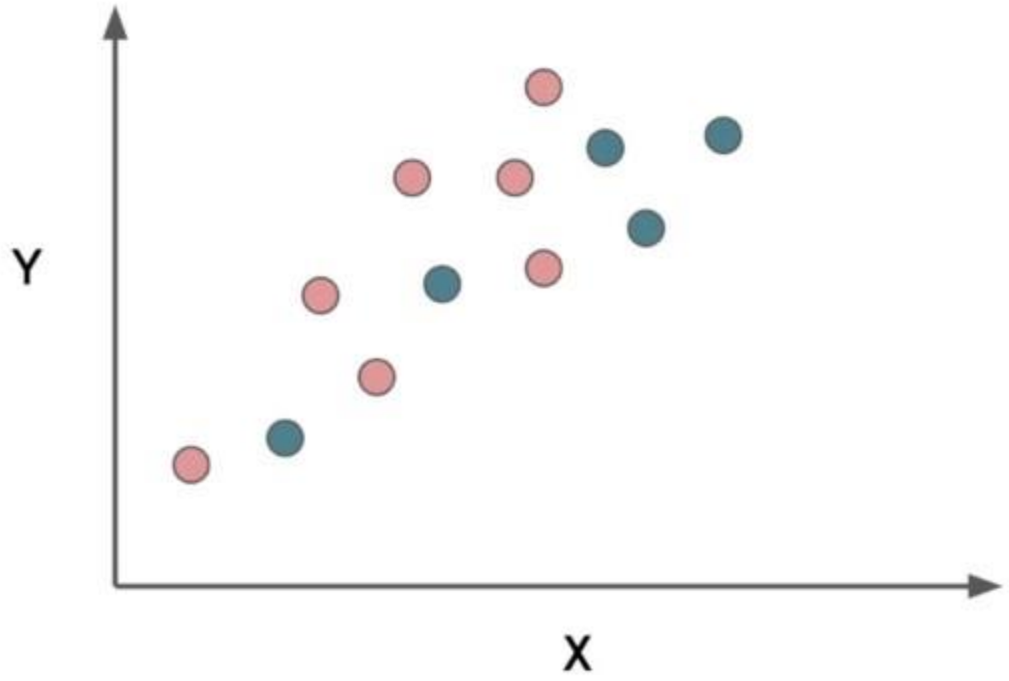
Ridge Regression

- Now we can fit on the training data to produce the line: $\hat{y} = \beta_1 x + \beta_0$



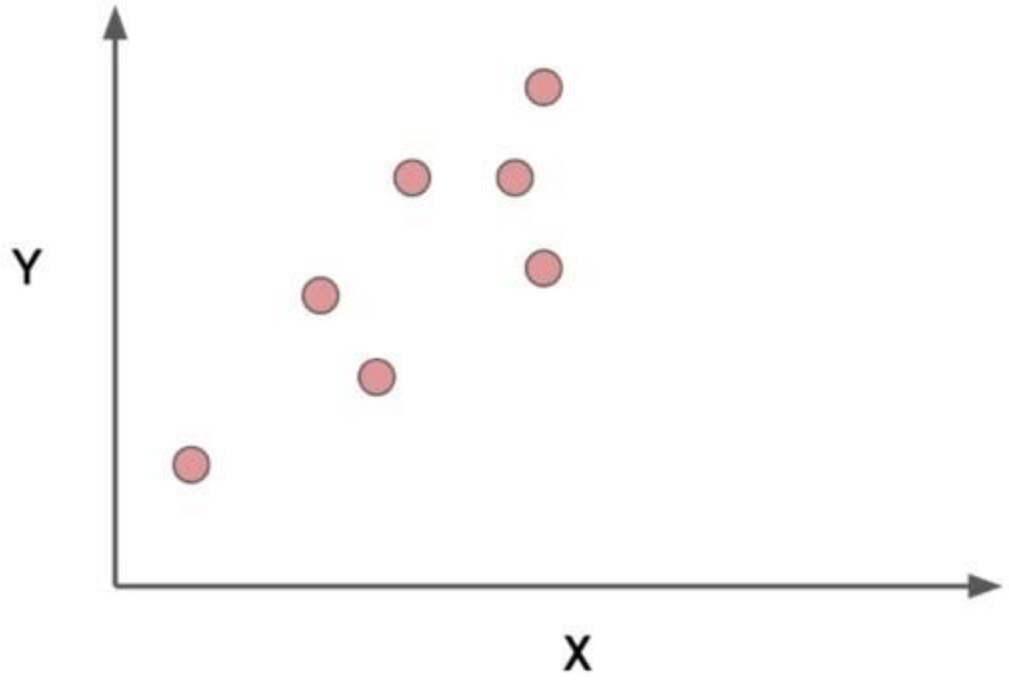
Ridge Regression

- We can split it into a training set and test set:



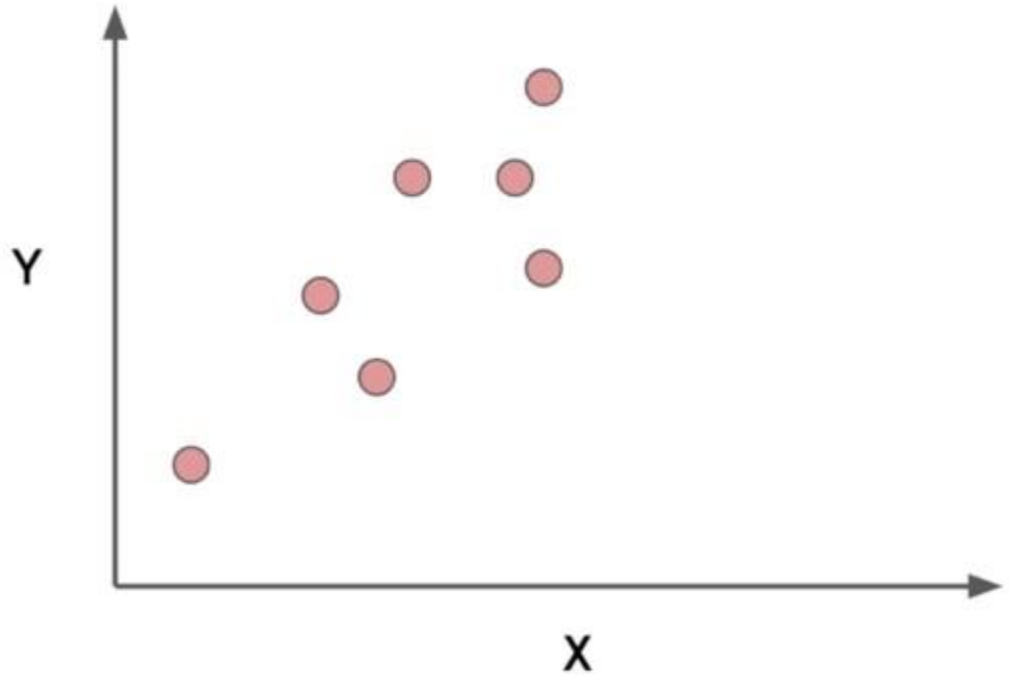
Ridge Regression

- Regardless of RSS or Ridge error, we're still trying to create a line: $\hat{y} = \beta_1 x + \beta_0$



Ridge Regression

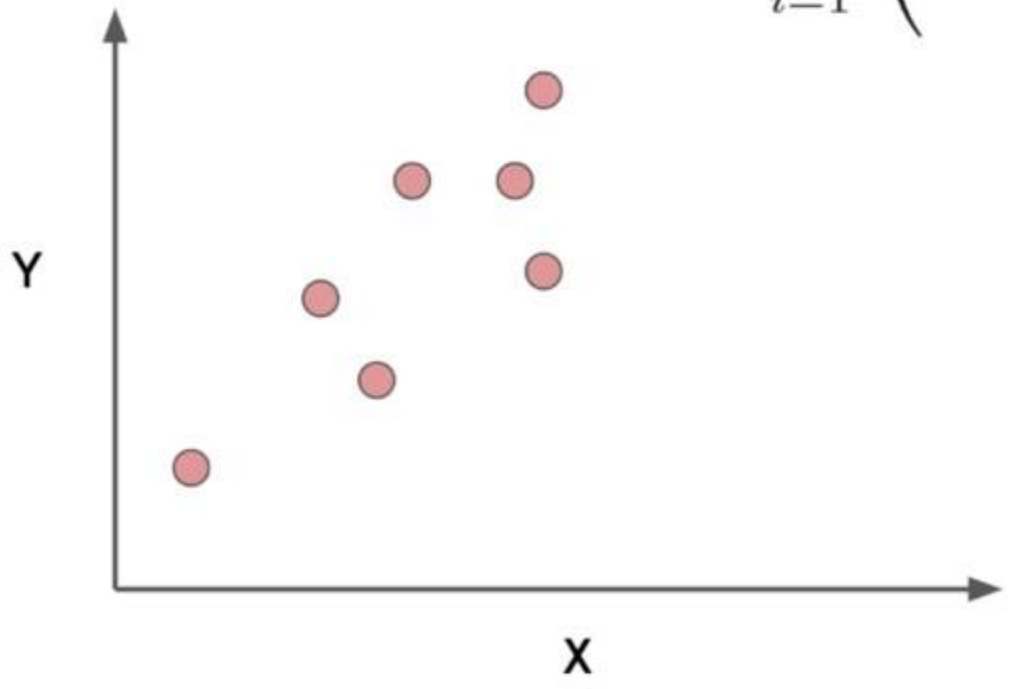
- The only difference would be the coefficients found.



Ridge Regression

- First let's fit using only RSS...

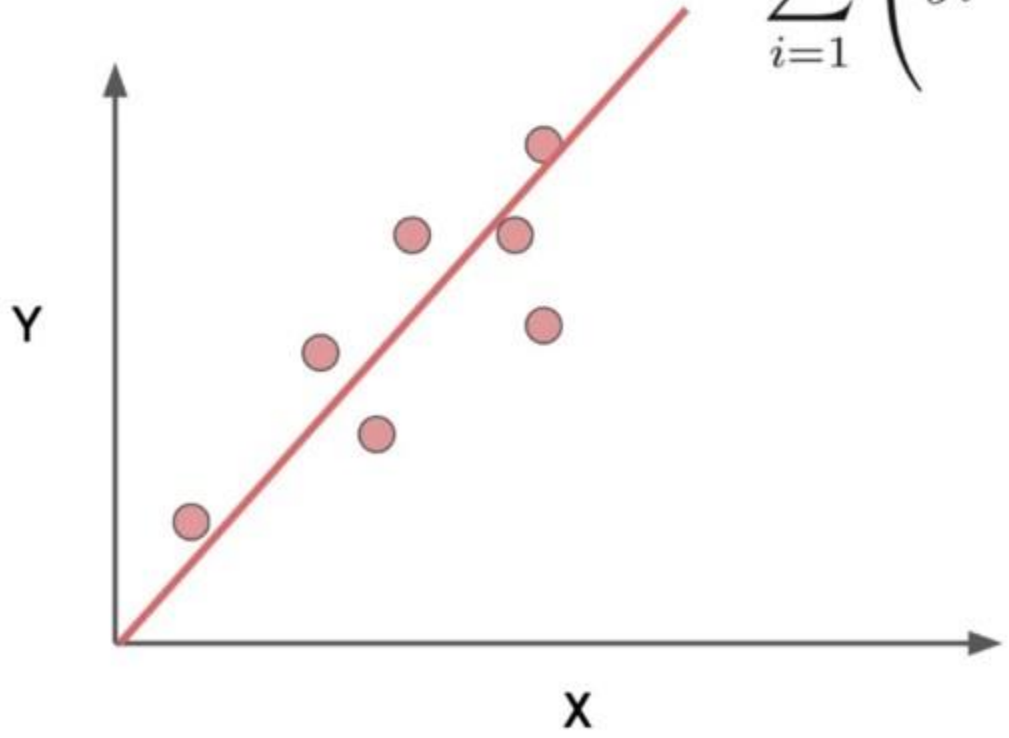
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$



Ridge Regression

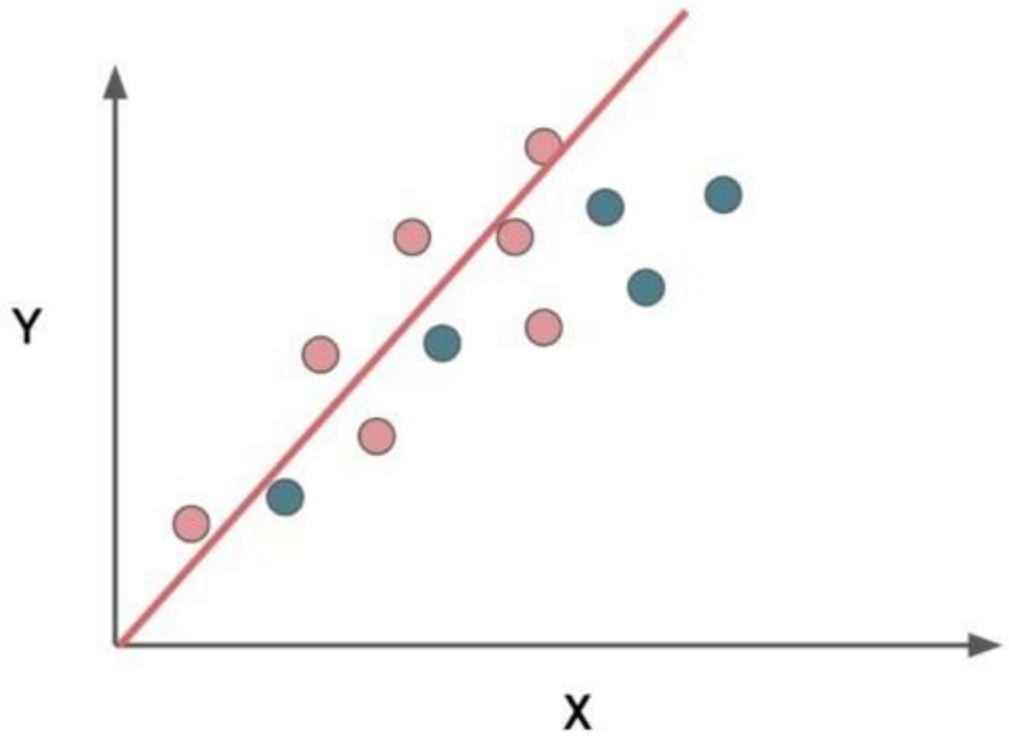
- Our fitted $\hat{y} = \beta_1 x + \beta_0$

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



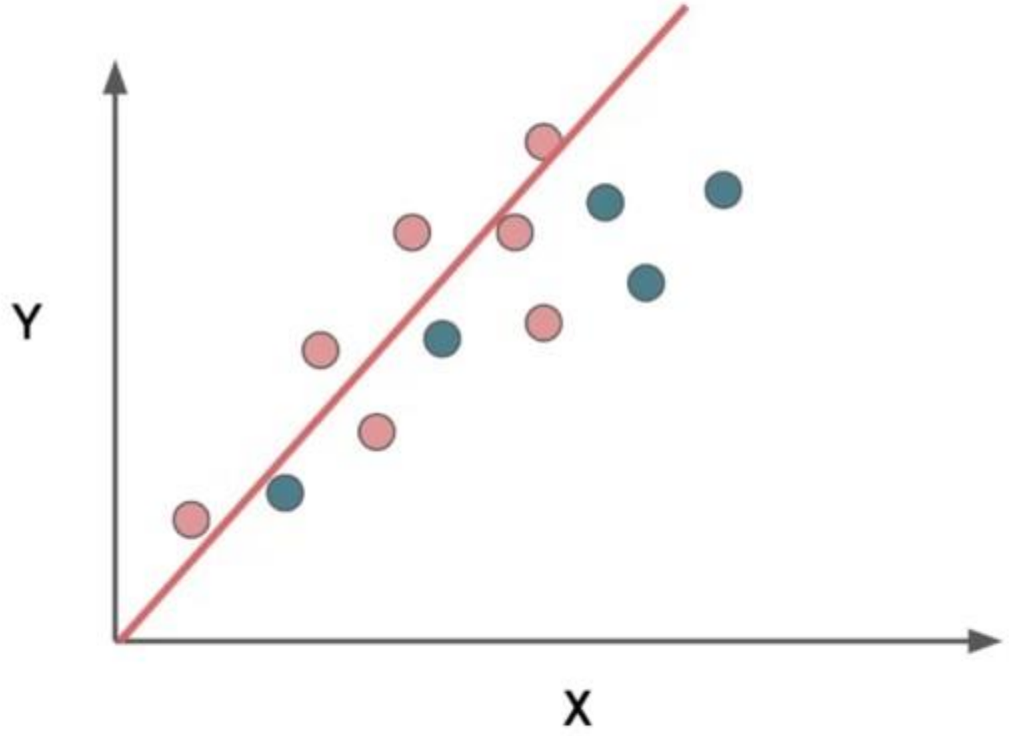
Ridge Regression

- Appears to have over fit to training data.



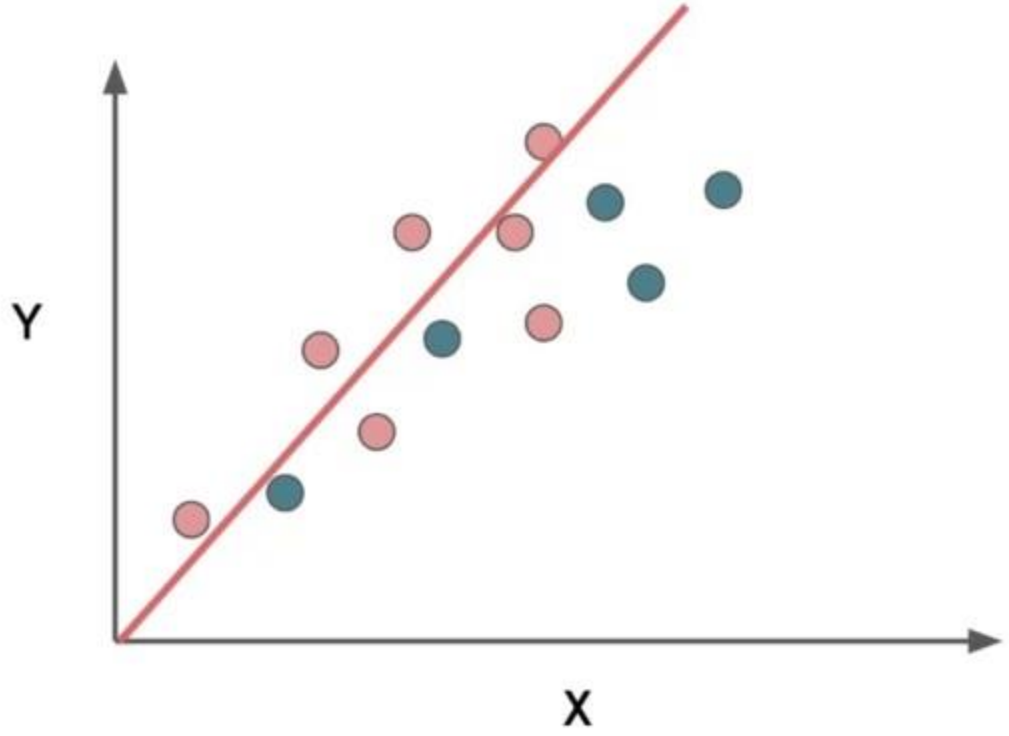
Ridge Regression

- This means we have high **variance.**



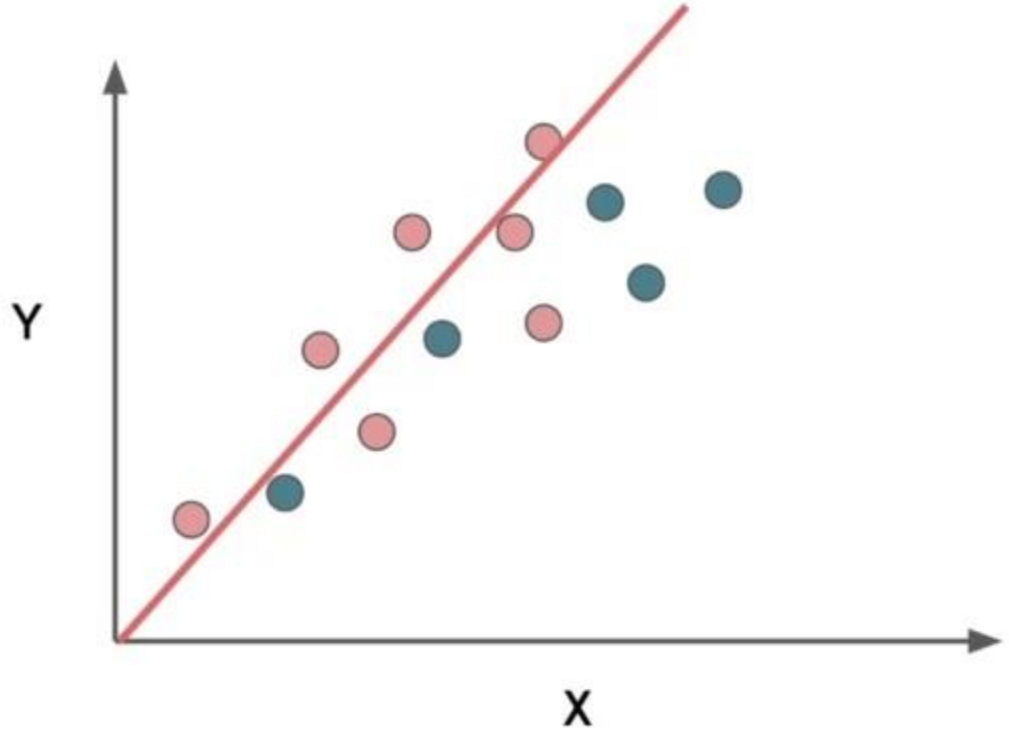
Ridge Regression

- We know there is a **bias-variance** trade-off.



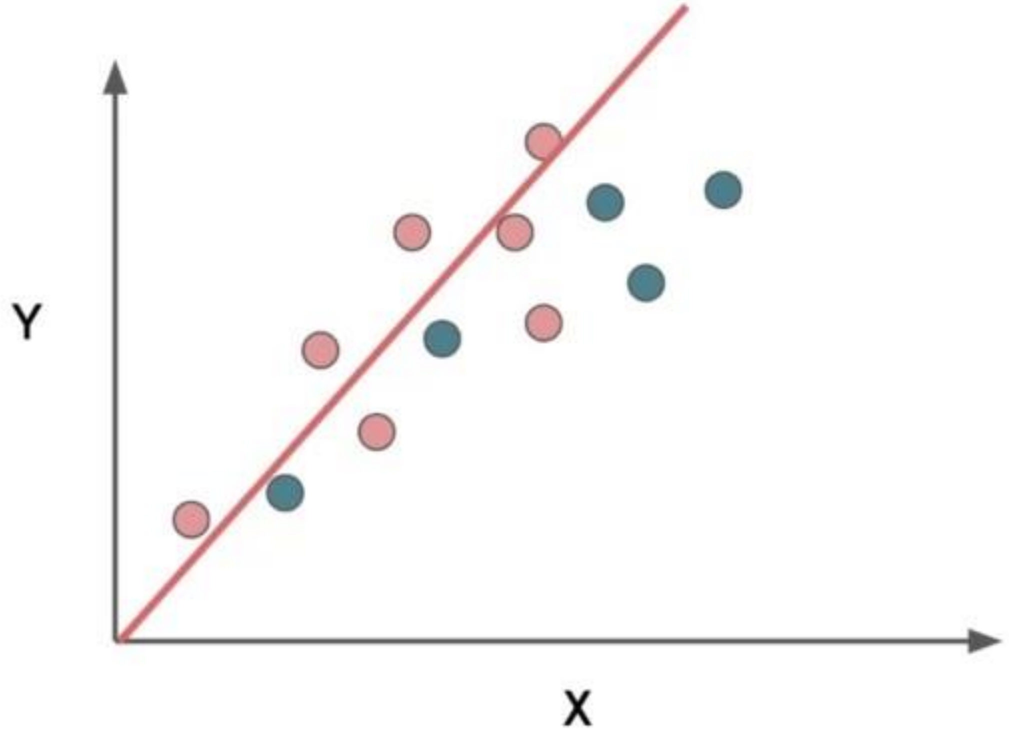
Ridge Regression

- But could we introduce a little more **bias** to significantly **reduce** variance?



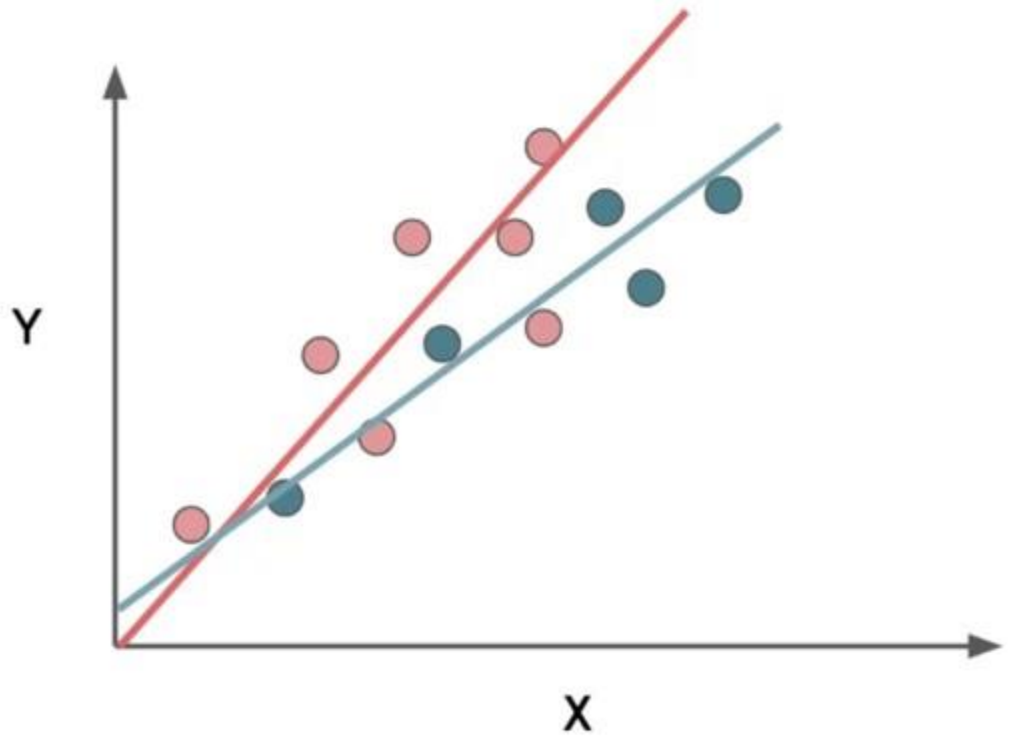
Ridge Regression

- Would adding the penalty term help generalize with more **bias**?



Ridge Regression

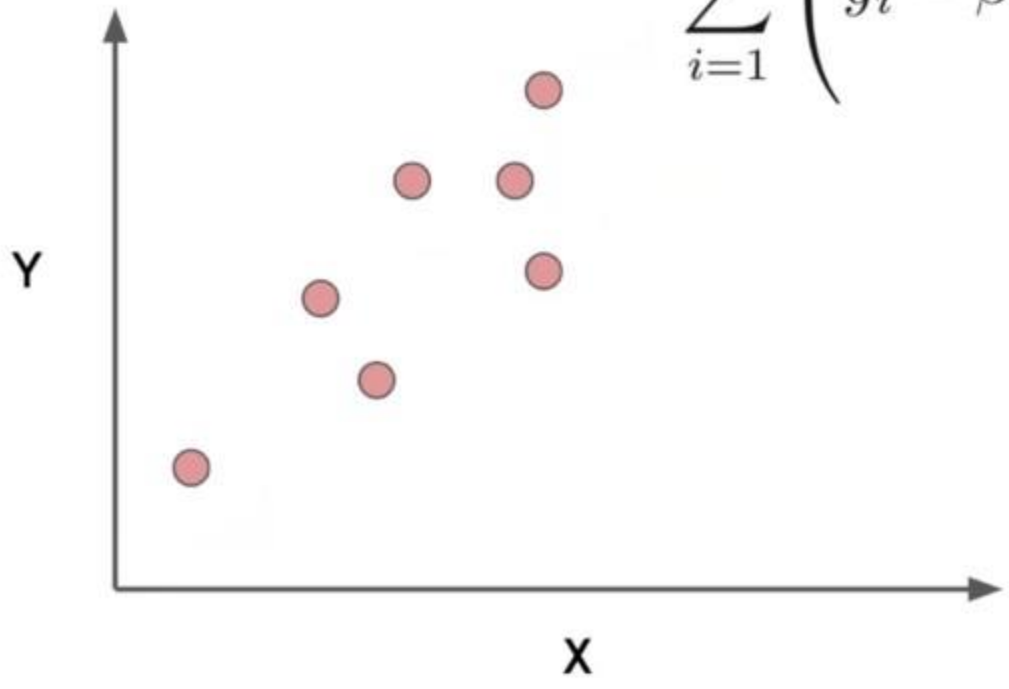
- Adding bias can help generalize $\hat{y} = \beta_1 x + \beta_0$



Ridge Regression

- Let's imagine trying to reduce the Ridge Regression error term:

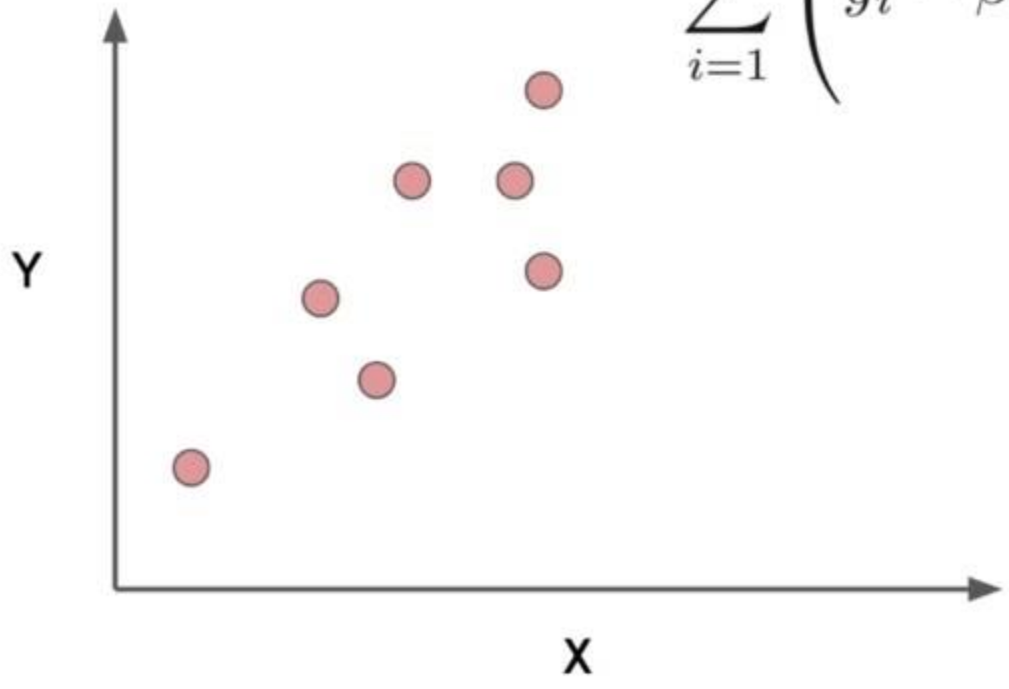
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$



Ridge Regression

- There is λ and the squared slope coefficient.

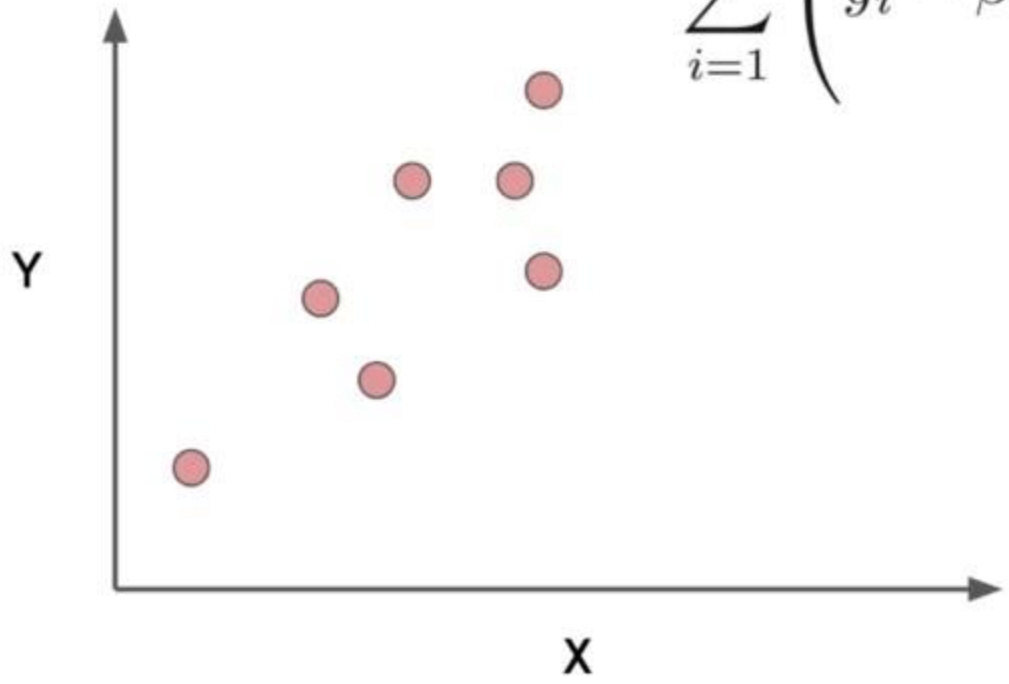
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda\sum_{j=1}^{p}\beta_j^2}$$



Ridge Regression

- In the case of $\hat{y} = \beta_1 x + \beta_0$

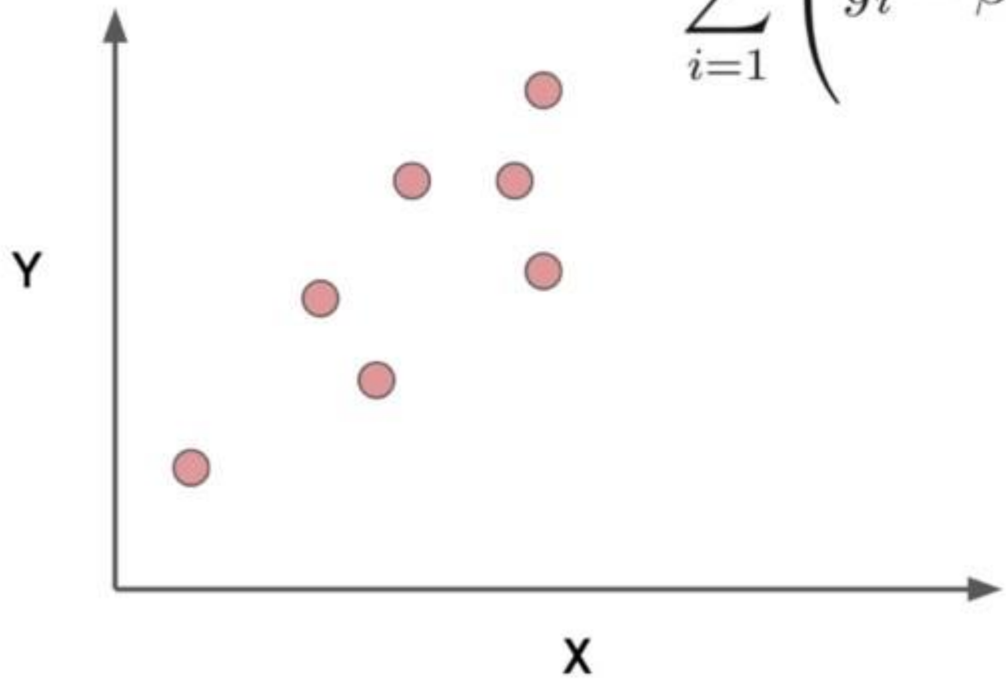
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$



Ridge Regression

- Let's assume λ = 1

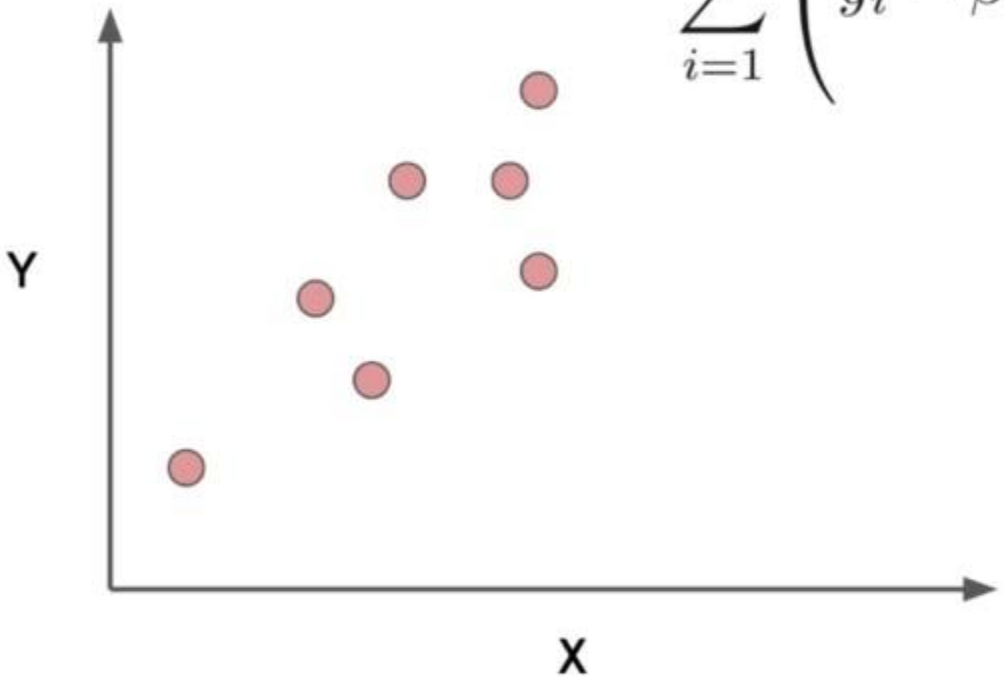
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$



Ridge Regression

- This punishes a large slope for $\hat{y} = \boxed{\beta_1} x + \beta_0$

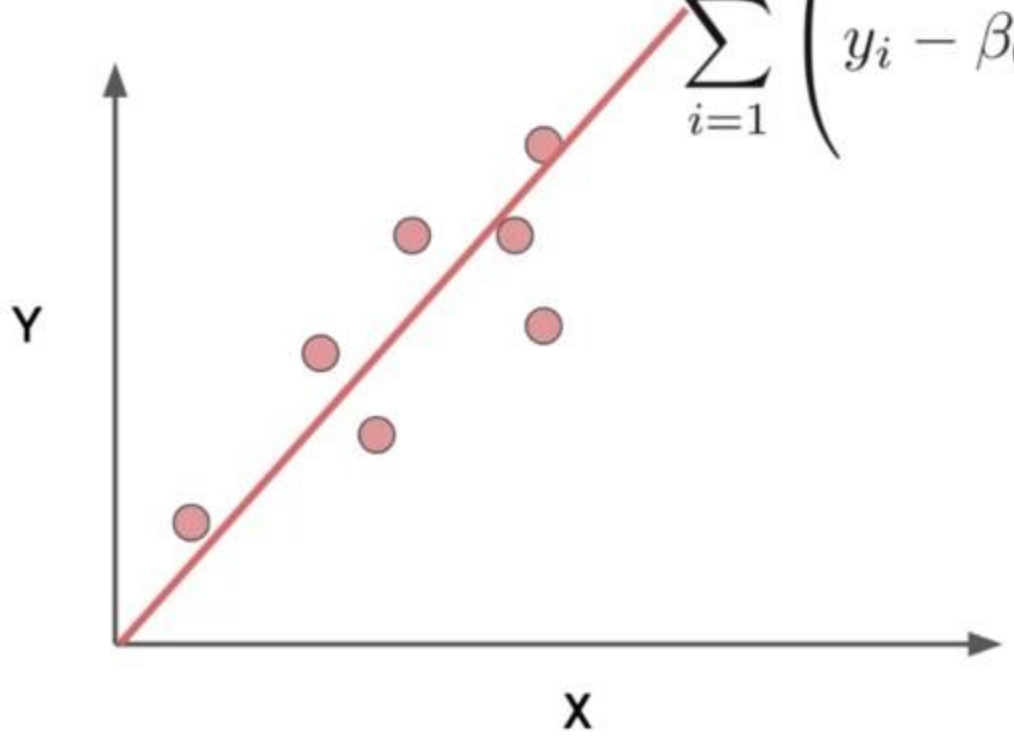
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda \sum_{j=1}^{p}\beta_j^2}$$



Ridge Regression

- For single feature this lowers slope

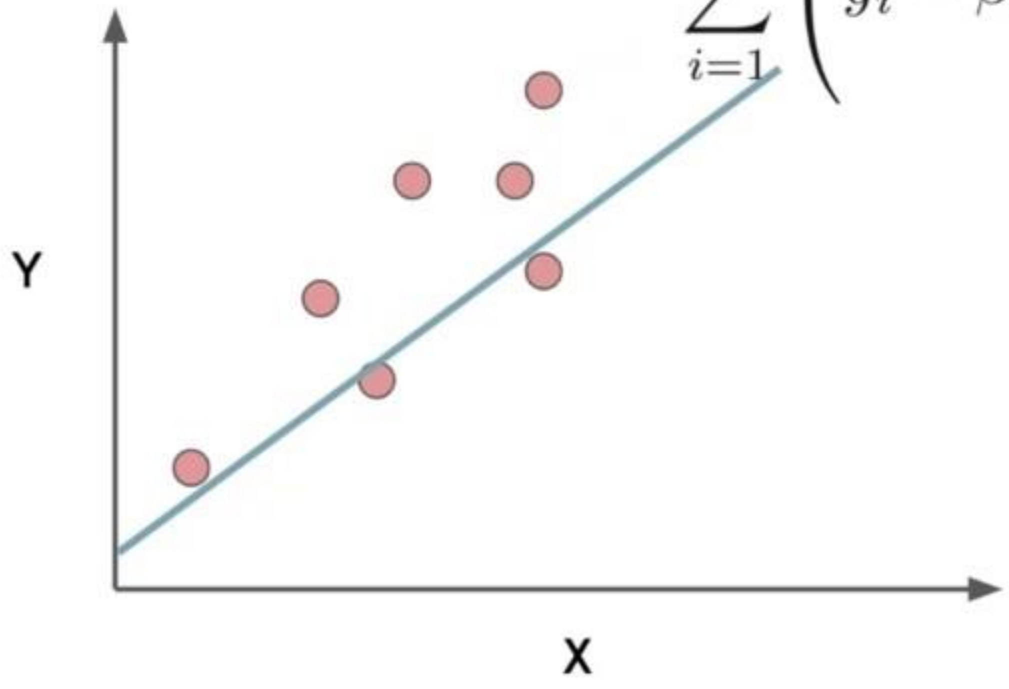
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\boxed{\lambda\sum_{j=1}^{p}\beta_j^2}$$



Ridge Regression

- For single feature this lowers slope

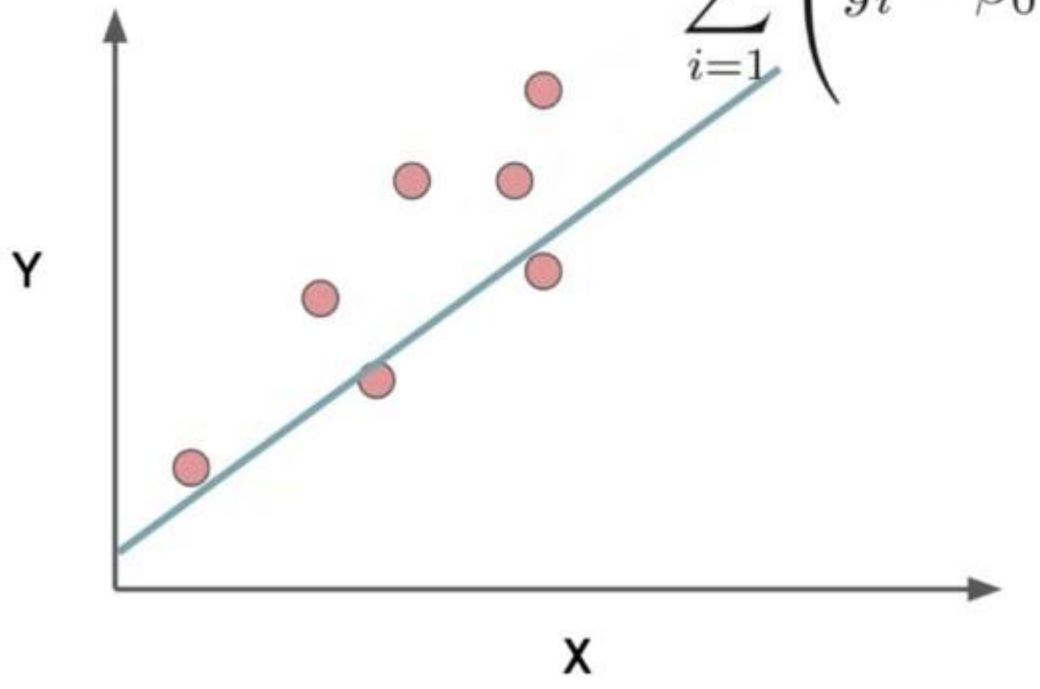
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda \sum_{j=1}^{p}\beta_j^2}$$



Ridge Regression

- At the cost of some additional bias (error in training set)

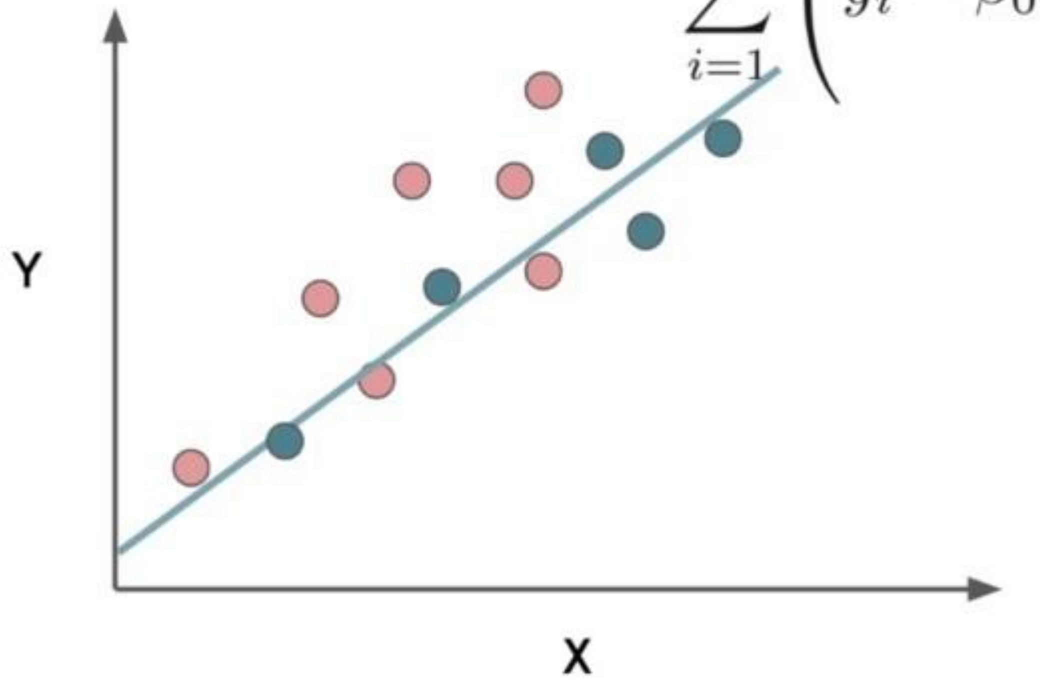
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda \sum_{j=1}^{p}\beta_j^2}$$



Ridge Regression

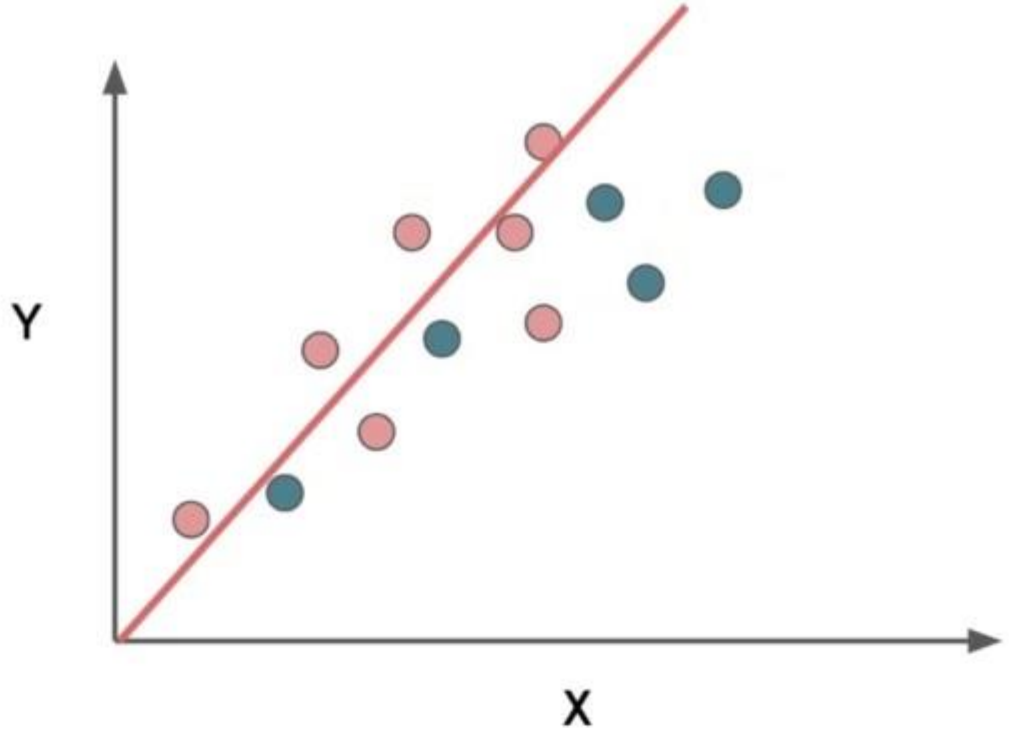
- We generalize better to unseen data

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



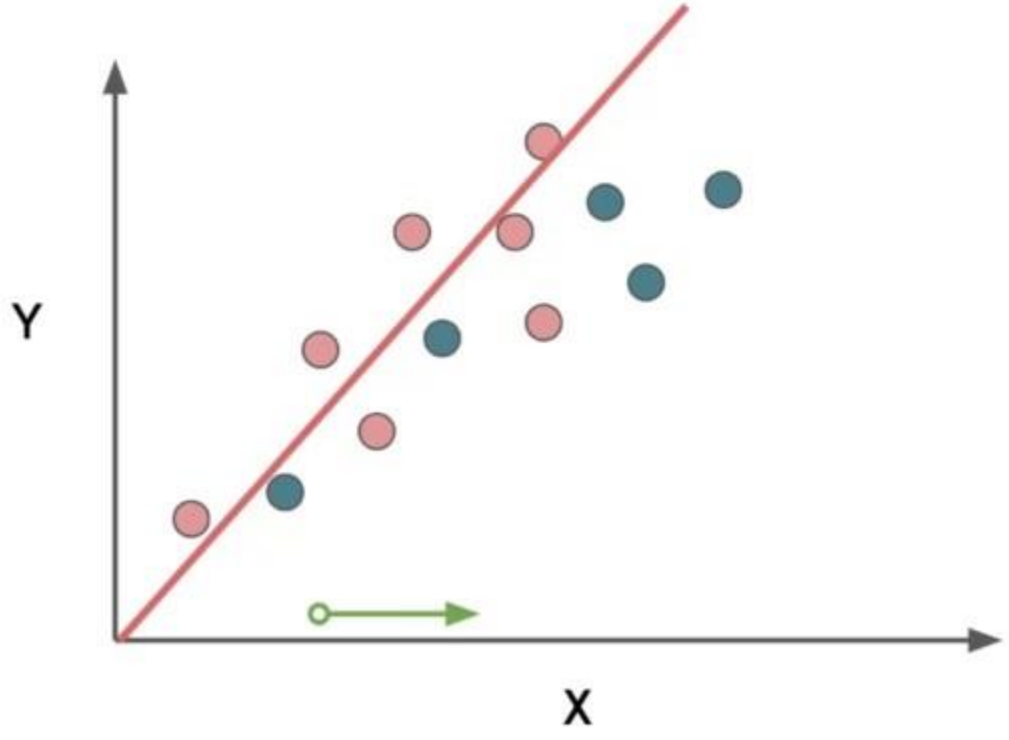
Ridge Regression

- Consider overfitting to training set:



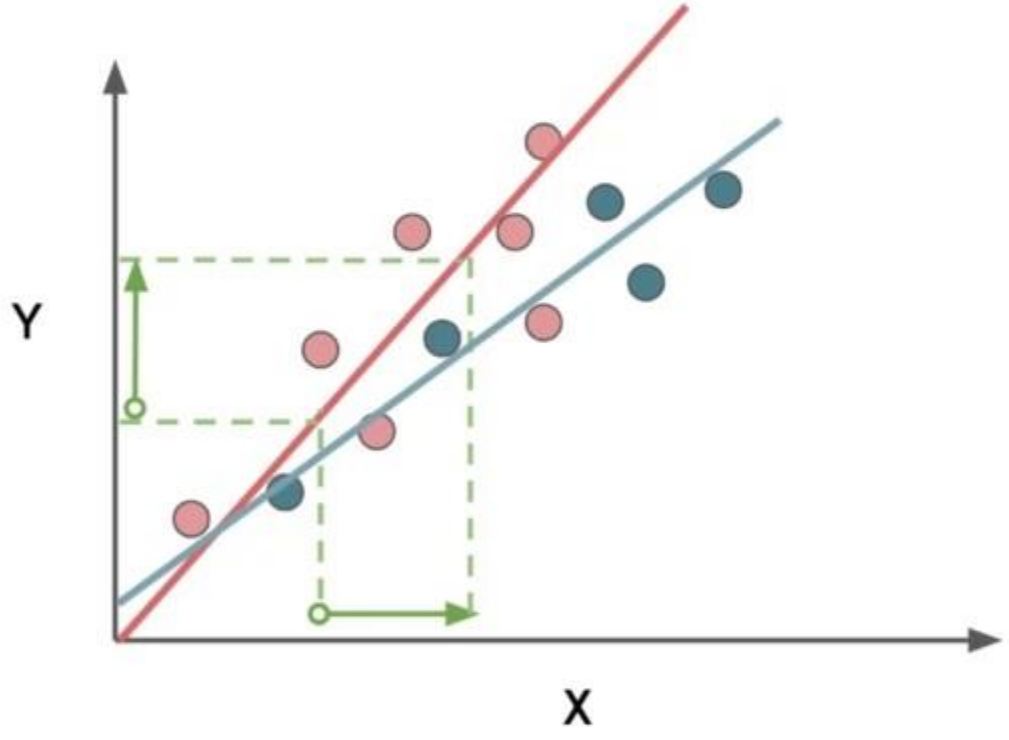
Ridge Regression

- An increase in X results in a greater y response:



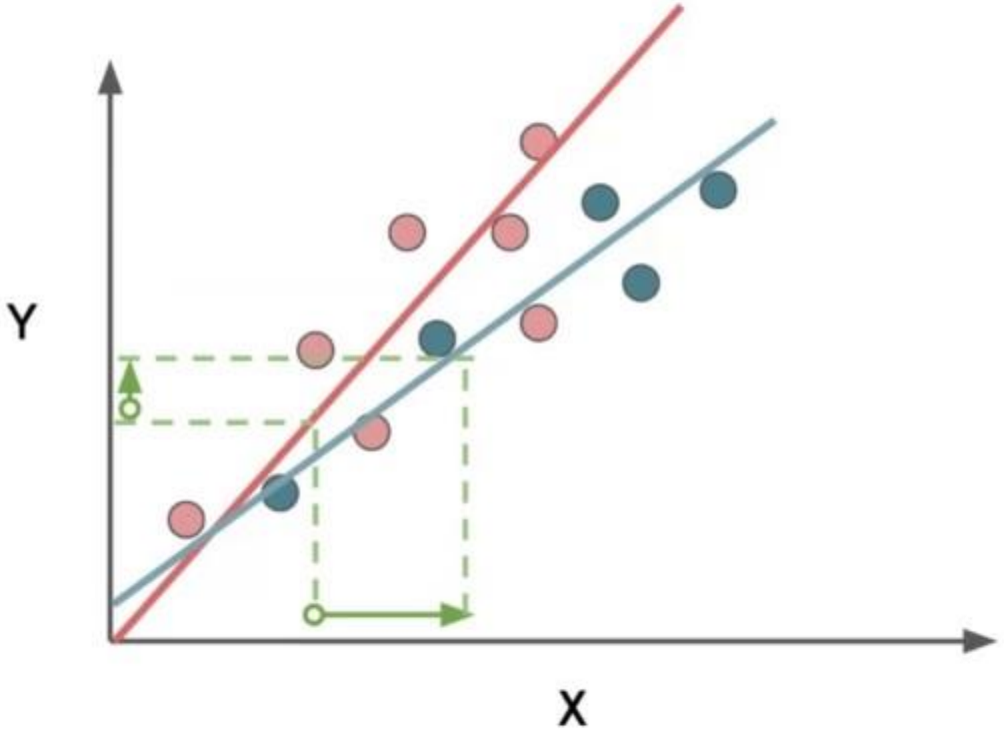
Ridge Regression

- Compare to a more generalized model that used Ridge Regression:



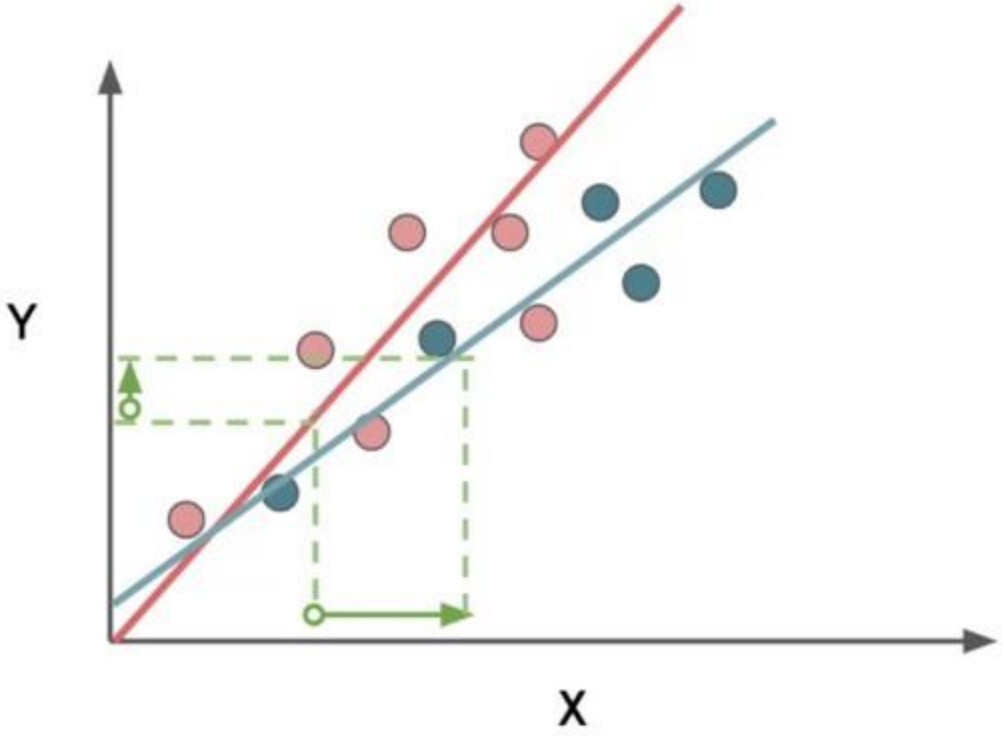
Ridge Regression

- Same feature change does not produce as much y response:



Ridge Regression

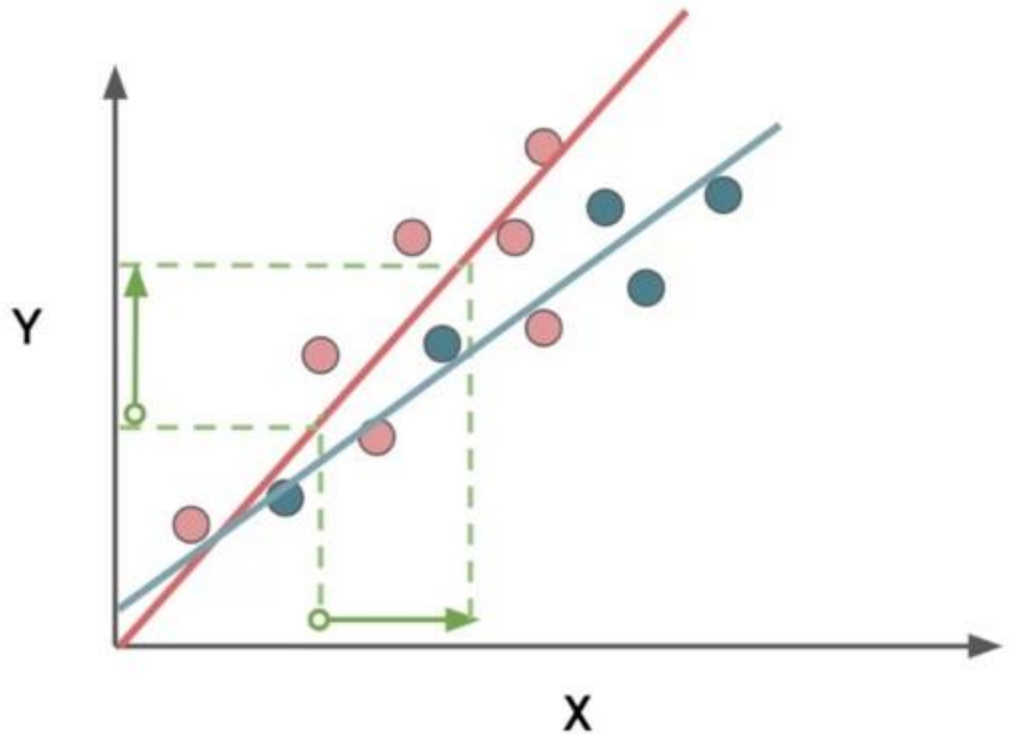
- Trying to minimize a squared Beta term leads us to punish larger coefficients.



$$\lambda \sum_{j=1}^{p} \beta_j^2$$

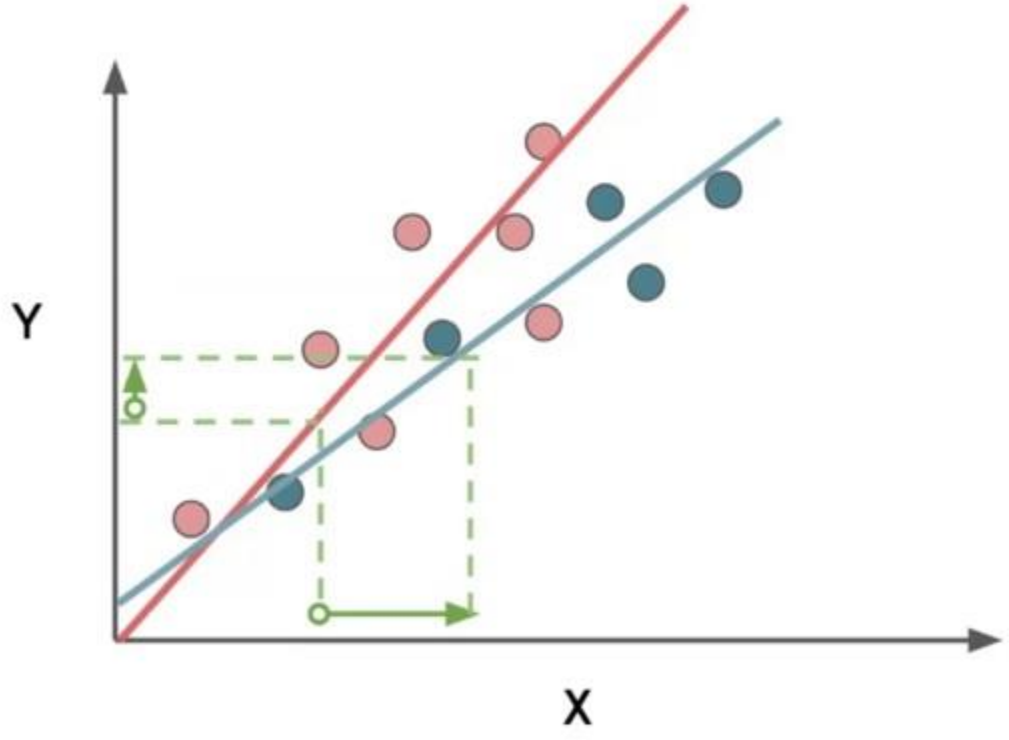
Ridge Regression

- In the case of a single feature, a larger Beta means a steeper sloped line.



Ridge Regression

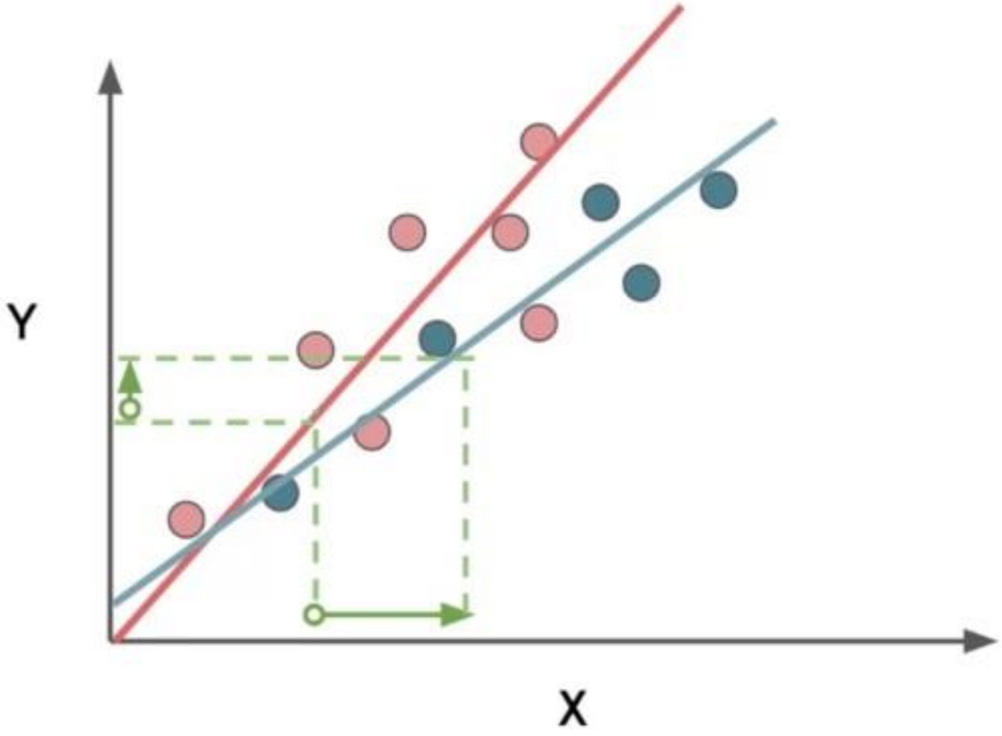
- A steeper sloped line would mean more response per increase in X value.



$$\lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge Regression

- What about the lambda term? How much should we punish these larger coefficients?



$$\lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge Regression

- We simply use cross-validation to explore multiple lambda options and then choose the best one!

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

Ridge Regression

L2 Regularization

Ridge Regression

- Important Note!
  - Sklearn refers to lambda as alpha within the class call!

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

Ridge Regression

- Important Note!
    - For cross validation metrics, sklearn uses a “scorer object”.
    - All scorer objects follow the convention that **higher** return values are **better** than lower return values.

Ridge Regression

- Important Note!
  - For example, obviously higher accuracy is better.
  - But higher RMSE is actually worse!
  - So Scikit-Learn fixes this by using a **negative** RMSE as its scorer metric.

Ridge Regression

- Important Note!
  - This allows for uniformity across **all** scorer metrics, even across different tasks types.
  - The same idea of uniformity across model classes applies to referring to the penalty strength parameter as **alpha**.

03-Regularization-Ridge-Lasso-ElasticNet[LECs8-9].ipynb