Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 08

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic disadvantage

Topic Why LR → overfitting, Unstable model

Topic Why we need Regularization

Topic

Topic

Topics to be Covered

Topic Ridge Regression

Topic H.w

Topic

Topic

Topic

Be the change that you wish to see in the world.

— MAHATMA GHANDI

BRIAN TRACY INTERNATIONAL

REVISION

Problems in LR ...

→ overfit

→ unstable

→ multicollinearity shd not be present in data.

REVISION

Problems in LR ...

done

Linear Regression

Space and Time Complexity of Linear Regression

Training $\Rightarrow$ $O(K^3 + NK^2)$

N datapoint

$D+1 = K$

Testing $O(K)$.

Space complexity $\Rightarrow K$

Question 12: In simple linear regression, which variable is considered the independent variable?

A. The variable being predicted y ⇒ dependent

B. The response variable ⇒ y

C. The predictor variable ⇒ x

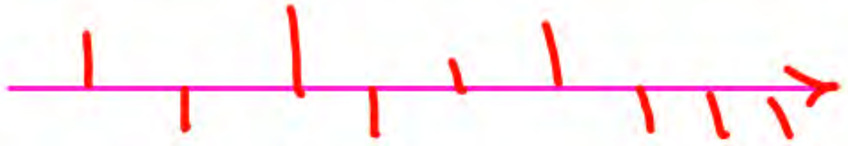
D. There is no independent variable in simple linear regression

Question 19: If the R-squared value in simple linear regression is 0.75, what does it indicate?

done

A. A strong linear relationship between the variables
B. A weak linear relationship between the variables
C. No linear relationship between the variables
D. The model is overfitting

Question 20: Which of the following statements is true regarding the residual plot in simple linear regression?



A. Residuals should exhibit a clear linear pattern.
B. Residuals should be randomly scattered around the horizontal line.
C. Residuals should be negatively correlated with the predictor variable.
D. Residuals should have a positive correlation with the dependent variable.

5.FOr a give N independent input variables (X1,X2... Xn) and dependent (target) variable Y a linear regression is fitted for the best fit line using least square error on this data. The correlation coefficient for one of it's variable(Say X1) with Y is -0.97. Which of the following is true for X1? ?

$\rightarrow \rho_{X_1 Y} = r = -.97$

- A) Relation between the X1 and Y is weak
- B) Relation between the X1 and Y is strong
- C) Relation between the X1 and Y is neutral
- D) Correlation does not imply relationship

6.Given below characteristics which of the following option is the correct for Pearson correlation between V1 and V2? If you are given the two variables V1 and V2 and they are following below two characteristics. 1. If V1 increases then V2 also increases 2. If V1 decreases then V2 behavior is unknown ?

Highly Correlated

- A) Pearson correlation will be close to 1
- B) Pearson correlation will be close to -1
- C) Pearson correlation will be close to 0
- D) None of these

1) A regression analysis is inappropriate when;

a) you have two variables that are measured on an interval or ratio scale.

b) you want to make predictions for one variable based on information about another variable. → True

c) the pattern of data points forms a reasonably straight line. → True

d) **there is heteroscedasticity in the scatter plot.**

2) In regression analysis, the variable that is being predicted is;

a) the independent variable

b) **the dependent variable** y

c) usually denoted by x

d) usually denoted by r

3) In the regression equation $y = b_o + b_1 x$, b_o is the;

a) slope of the line

b) independent variable

c) **y intercept**

d) coefficient of determination

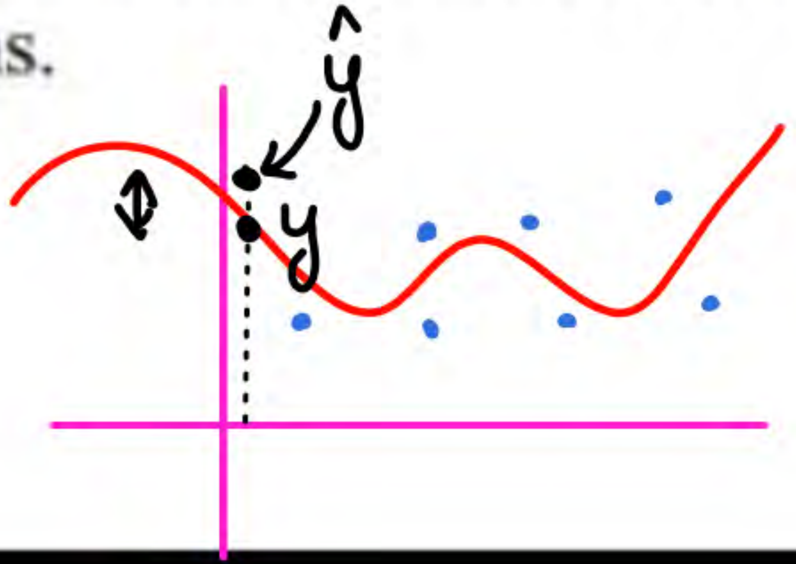
6) Least square method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the Vertical deviations.

a) **Vertical**

b) Horizontal

c) Both of these

d) None of these



7) Which one is the least square method formula;

a) $\min \sum(y_i - \hat{y}_i)^2$

b) $\min \sum(\hat{y}_i - y_i)$

c) $\boldsymbol{\min \sum(y_i - \hat{y}_i)^2}$

d) $\min \sum(y_i - \hat{y}_i)$

13) Below you are given a summary of the output from a simple linear regression analysis from a sample of size 15, SSR=100, SST = 152. The coefficient of determination is;

a) 0.5200

b) ~~0.6579~~ .342 ✓

c) 0.8111

d) 1.52

$$R^2 = 1 - \frac{SSR}{SST} \Rightarrow 1 - \frac{100}{152} \Rightarrow .342$$

10) **A residual is defined as**

a) The difference between the actual Y values and the mean of Y.

b) **The difference between the actual Y values and the predicted Y values.**

c) The predicted value of Y for the average X value.

d) The square root of the slope.

11) If the regression equation is equal to y=23.6−54.2x, then 23.6 is the _____ while -54.2 is the ____ of the regression line.

a) Slope, intercept

b) Slope, regression coefficient

c) **Intercept, slope**

d) Radius, intercept

Q8. Suppose we have N independent variables (X1, X2... Xn) and Y's dependent variable.

Now Imagine that you are applying linear regression by fitting the best-fit line using the least square error on this data. You found that the correlation coefficient for one of its variables (Say X1) with Y is -0.95.

done

Which of the following is true for X1?

A) Relation between the X1 and Y is weak

B) Relation between the X1 and Y is strong

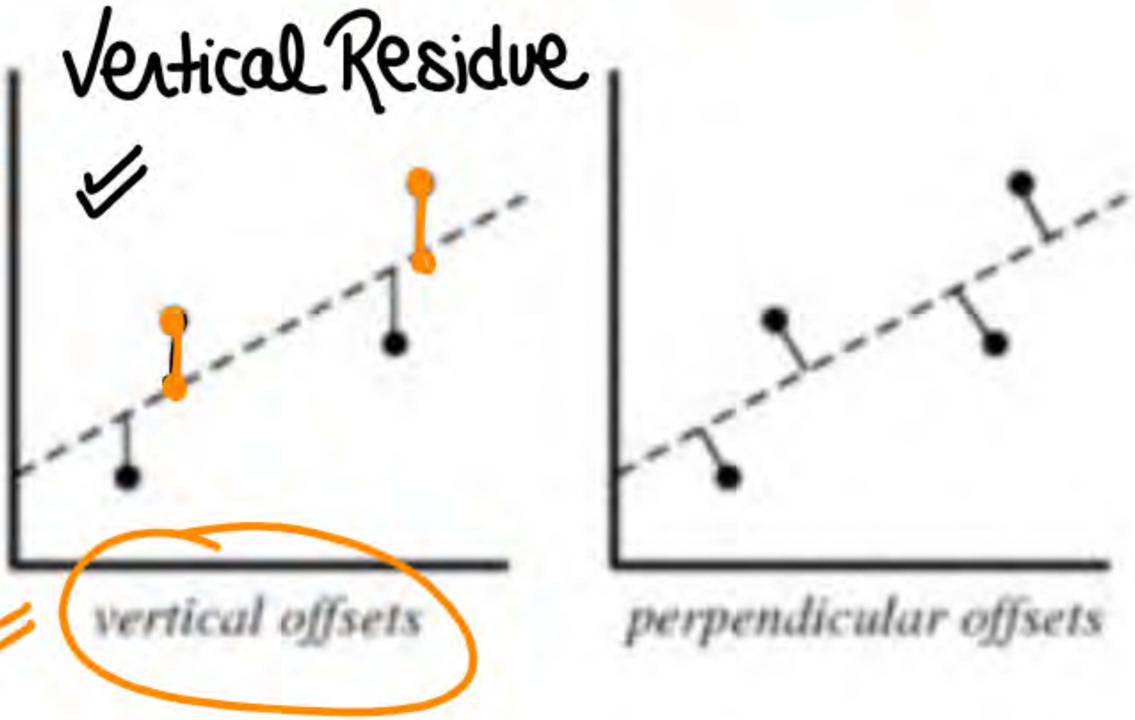
C) Relation between the X1 and Y is neutral

D) Correlation can't judge the relationship

Solution: (B)

Q11. Suppose the horizontal axis is an independent variable and the vertical axis is a dependent variable. Which of the following offsets do we use in linear regression's least square line fit?

a)



~~B)~~ Perpendicular offset

C) Both, depending on the situation

D) None of above

Q12. True- False: Overfitting is more likely when you have a huge amount of data to train.

Overfitting depend on algo.

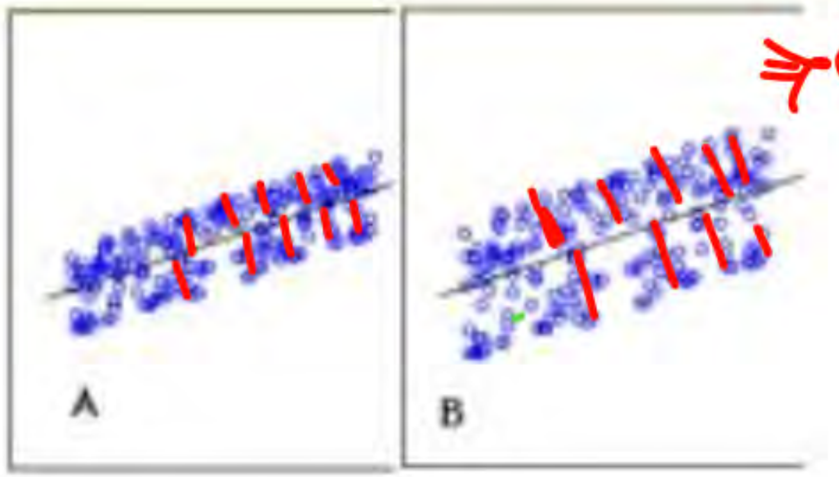
A) TRUE

B) FALSE

Solution: (B)

Q14. Which of the following statement is true about the sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases, A and B.



$\Rightarrow$ Sum = 0

But Sum of $(\text{Residue})^2 \Rightarrow$ in Case A < in Case B.

A) A has a higher sum of residuals than B

B) A has a lower sum of residual than B

C) Both have the same sum of residuals

D) None of these

C ✓

Q18. Which of the following statement is true about outliers in Linear regression?

A) Linear regression is sensitive to outliers

B) Linear regression is not sensitive to outliers

C) Can't say

D) None of these

Q19. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and found a relationship between them. Which of the following conclusion do you make about this situation?

Residue and predicted value are linked.

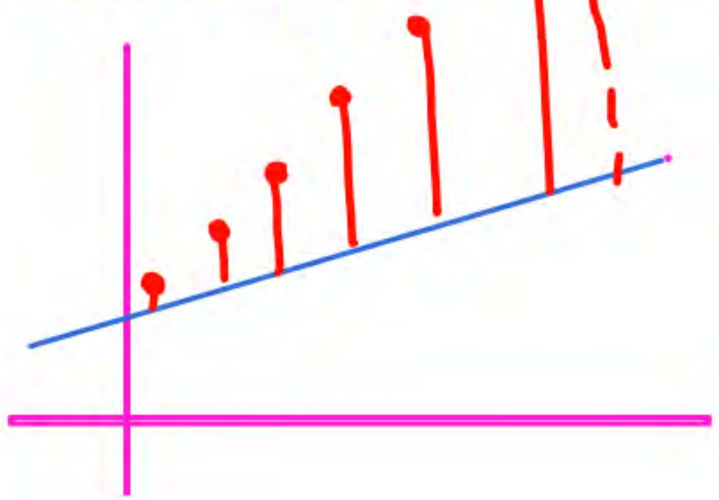
Residue shd be Random.

A) Since there is a relationship means our model is not good

B) Since there is a relationship means our model is good

C) Can't say

D) None of these



$\Rightarrow \left(y = ax^3 + bx^2 + cx + d\right) \Leftarrow$ model

Suppose that you have a dataset D1 and you design a ~~linear~~ model of degree 3 polynomial and find that the training and testing error is “0” or, in other words, it perfectly fits the data.

* Training error = 0   So degree 3 poly ⇒ overfit

Q20. What will happen when you fit a degree 4 polynomial in ~~linear~~ regression?

So if we inc the degree to 4 ⇒ Complexity of polynomial inc

A) There is a high chance that degree 4 polynomial will overfit the data

B) There is a high chance that degree 4 polynomial will underfit the data

C) Can’t say

D) None of these

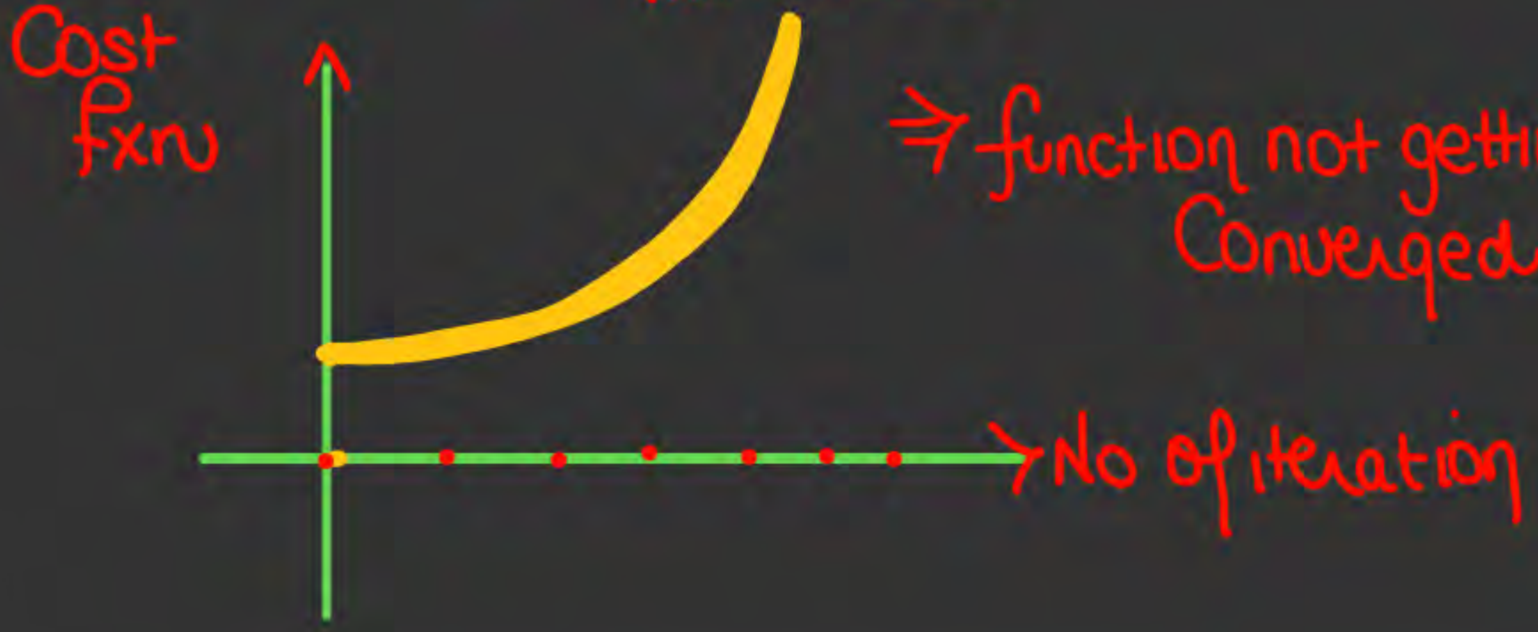
degree 3 polynomial give zero training error. overfit

then degree 2 4

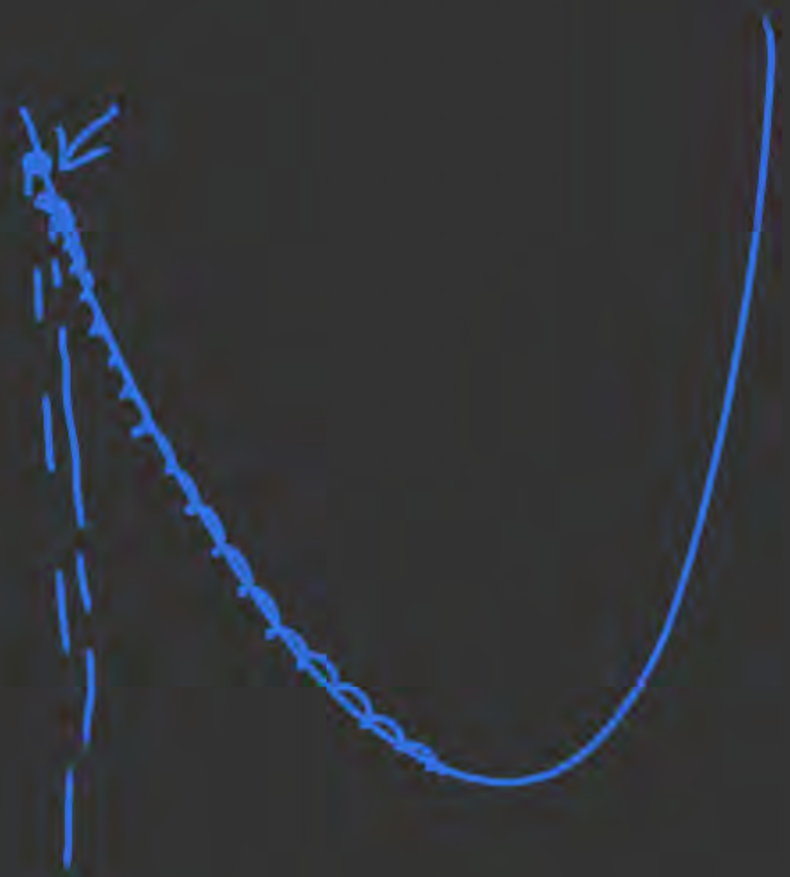
Complexity of model reduces

high chances that this model donot overfit.

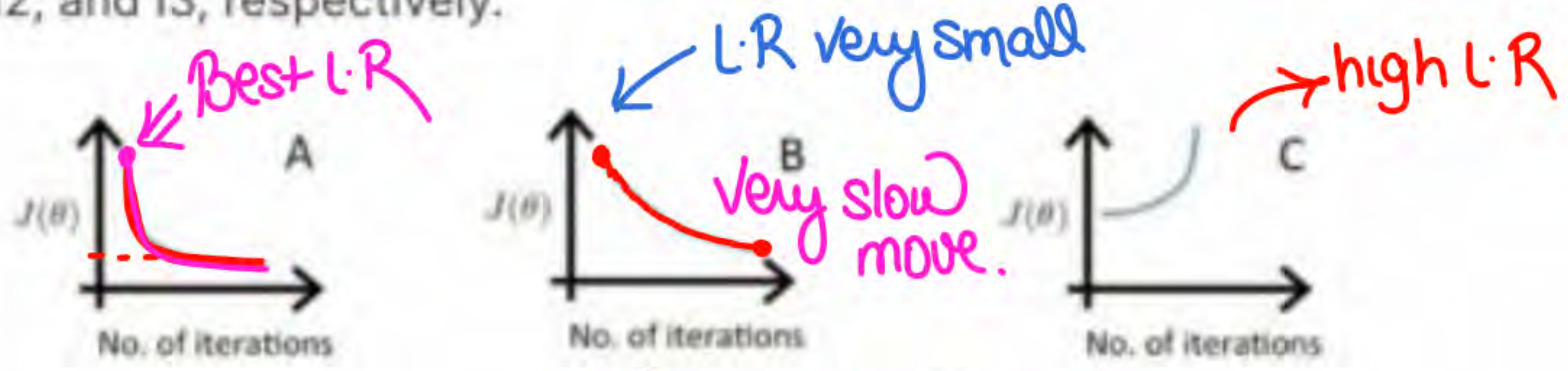
* Cost function ⇒ Function that is to be minimized



⇒ function not getting Converged



Below are three graphs, A, B, and C, between the cost function and the number of iterations, l1, l2, and l3, respectively.



Gradient descent.

Q23. Suppose l1, l2, and l3 are the three learning rates for A, B, and C, respectively. Which of the following is true about l1,l2, and l3?

A) l2 < l1 < l3

B) l1 > l2 > l3

C) l1 = l2 = l3

D) None of these

QUESTION 1

How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

2 β_0, β_1

- 1
- 2
- 3
- 4

QUESTION 2

In a linear regression model, which technique can find the coefficients?

- Ordinary Least Squares $\beta = (X^T X)^{-1} (X^T Y)$
- Gradient Descent
- Regularization
- All of the above

Which one is the disadvantage of Linear Regression?

- The assumption of linearity between the dependent variable and the independent variables. In the real world, the data is not always linearly separable.
- Linear regression is very sensitive to outliers
- Before applying Linear regression, multicollinearity should be removed because it assumes that there is no relationship among independent variables.
- All of the above

QUESTION 4

Which parameter determines the size of the improvement step to take on each iteration of Gradient Descent?

- [x] learning rate
- [ ] epoch
- [ ] batch size
- [ ] regularization parameter

QUESTION 5

5 marks

For a linear regression model, start with random values for each coefficient. The sum of the squared errors is calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible. This technique is called _______?

- Gradient Descent
- Ordinary Least Squares
- Homoscedasticity
- Regularization

QUESTION 6

In a linear regression model, which technique cannot find the coefficients?

Can be used

- Ordinary Least Squares
- Gradient Descent

Cannot be used.

- Regularization
- Normalization

QUESTION 8

What is predicting y for a value of x that is within the interval of points that we saw in the original data called?

- Regression
- Extrapolation
- Intrapolation
- Polation

QUESTION 9

5 marks

The correlation coefficient between the age of a person and their IQ test score is found to be -1.0087. What can you conclude from this?

- Age is not a good predictor of IQ.
- Age is a good predictor of IQ.
- None of the above

So hypothesis testing in L.R ⇒

H0: Null hypothesis ⇒ all β's are zero

$\beta_0 = 0$
$\beta_1 = 0$
$\beta_2 = 0$
⋮

H1: $\beta_0 \neq 0$
$\beta_1 \neq 0$
$\beta_2 \neq 0$
⋮

So we calculate β values from LR

Now we find Zscore for each β

Zscore ⇒ $\dfrac{\beta_{i_{H_1}} - \underbrace{\beta_{i_{H_0}}}_{\to 0}}{(SE_{\beta_i})}$

OR α

5% ⇒ P ⇒ .05

(H_0, H_1) Zscore Pvalue

QUESTION 10

5 marks

In order to determine whether the coefficient in a simple linear regression model is significant or not, which Null Hypothesis do we propose?

- ○ $\beta 0 \neq 0$
- ○ $\beta 1 = 0$
- ○ $\beta 0 = 0$
- ○ $\beta 1 \neq 1$

QUESTION 4

A term used to describe the case when the independent variables in a multiple regression model are correlated is

- regression
- correlation
- multicollinearity
- none of the above

QUESTION 5

A multiple regression model has the form: $y = 2 + 3x1 + 4x2$. As x1 increases by 1 unit (holding x2 constant), y will

- increase by 3 units
- decrease by 3 units
- increase by 4 units
- decrease by 4 units

QUESTION 6

5 marks

The adjusted multiple coefficient of determination accounts for

- $R^2 \Rightarrow$ if the No of features inc then R^2 also inc
i.e even if irrelevant feature inc then R^2 inc.

- the number of dependent variables in the model
- the number of independent variables in the model
- unusually large predictors
- none of the above

QUESTION 7

A multiple regression model has

- only one independent variable
- more than one dependent variable
- more than one independent variable
- none of the above

Questions

26. In a linear regression model, if the sum of squared residuals (SSE) is 100 and the total sum of squares (SST) is 200, what is the coefficient of determination (R-squared)?

a) 0.5
b) 1
c) 0
d) -1

$R^2 = 1 - \frac{100}{200}$

$= .5$

Lets Practise...

Questions

29. You are using the mean squared error (MSE) as an evaluation metric for a regression model. The predicted values are [3, 4, 5, 6], and the actual values are [2, 3, 4, 7]. What is the MSE?

$SE \Rightarrow (3-2)^2 + (4-3)^2 + (5-4)^2 + (6-7)^2$

$MSE \Rightarrow \frac{1}{4} SE \Rightarrow 1$

a) 0.5
b) 1.0
c) 1.5
d) 2.0

Lets Practise...

Questions

32. You are performing linear regression with the following data points:

X: [1, 2, 3, 4] $\Rightarrow \bar{x} \Rightarrow 10/4 = 2.5$

Y: [4, 3, 6, 5] $\Rightarrow \bar{y} \Rightarrow 4.5$

$a = \frac{Cov(X, Y)}{Var(X)} \Rightarrow 3/5$

$b = \bar{y} - a\bar{x}$

So $b = 4.5 - \frac{3}{5} \times \frac{5}{2} \Rightarrow 4.5 - 3/2 \Rightarrow 3$

What is the intercept (b) of the regression line, assuming a simple linear model Y = aX + b?

$Cov = \frac{1}{4-1}\left[\sum_{i=1}^{4}(x_i - \bar{x})(y_i - \bar{y})\right] = 1$

$Var(x) = \frac{1}{4-1}\left[\sum_{i=1}^{4}(x_i - \bar{x})^2\right] = 5/3$

a) 1.5
b) 2
c) 2.5
d) 3 ✓

Shrinkage Methods : Ridge Regression

- Ridge regression is a regularisation techniques...

$$\mathcal{L} = \min \left(\frac{1}{2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \underbrace{\frac{\lambda}{2} \sum_{i=1}^{D} \beta_i^2}_{} \right)$$

→ In Ridge Reg. we add Regularization term i.e β^2

→ Ridge Reg has L2 Regularisation

⇒ we have not included β_0 in this eq

Shrinkage Methods : Ridge Regression

coefficient of

- "In regularization technique, we reduce the magnitude of the features by keeping the same number of features.

- This helps in ....

we try to Reduce β's → unwanted features β ≈ 0, Solve multicollinearity
→ Solve problem of large β ⇒ unstable model
→ overfitting in LR Solved

Ridge Regression

Shrinkage Methods : Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size. ⇒ Penalty term in loss fxn.
- The ridge coefficients minimize a penalized residual sum of squares of the weights.

The loss function are updated

Shrinkage Methods : Ridge Regression

Loss fxnw

The loss function are updated

Shrinkage Methods : Ridge Regression

The main reason for not regularizing the intercept term is that it represents the mean value of the target variable when all the features are zero. Regularizing the intercept can lead to shifting this mean value away from its natural value, which might not be desirable in many cases.

Why the bias term is not included in regularisation ..

The Bias term has a very Imp → Role →

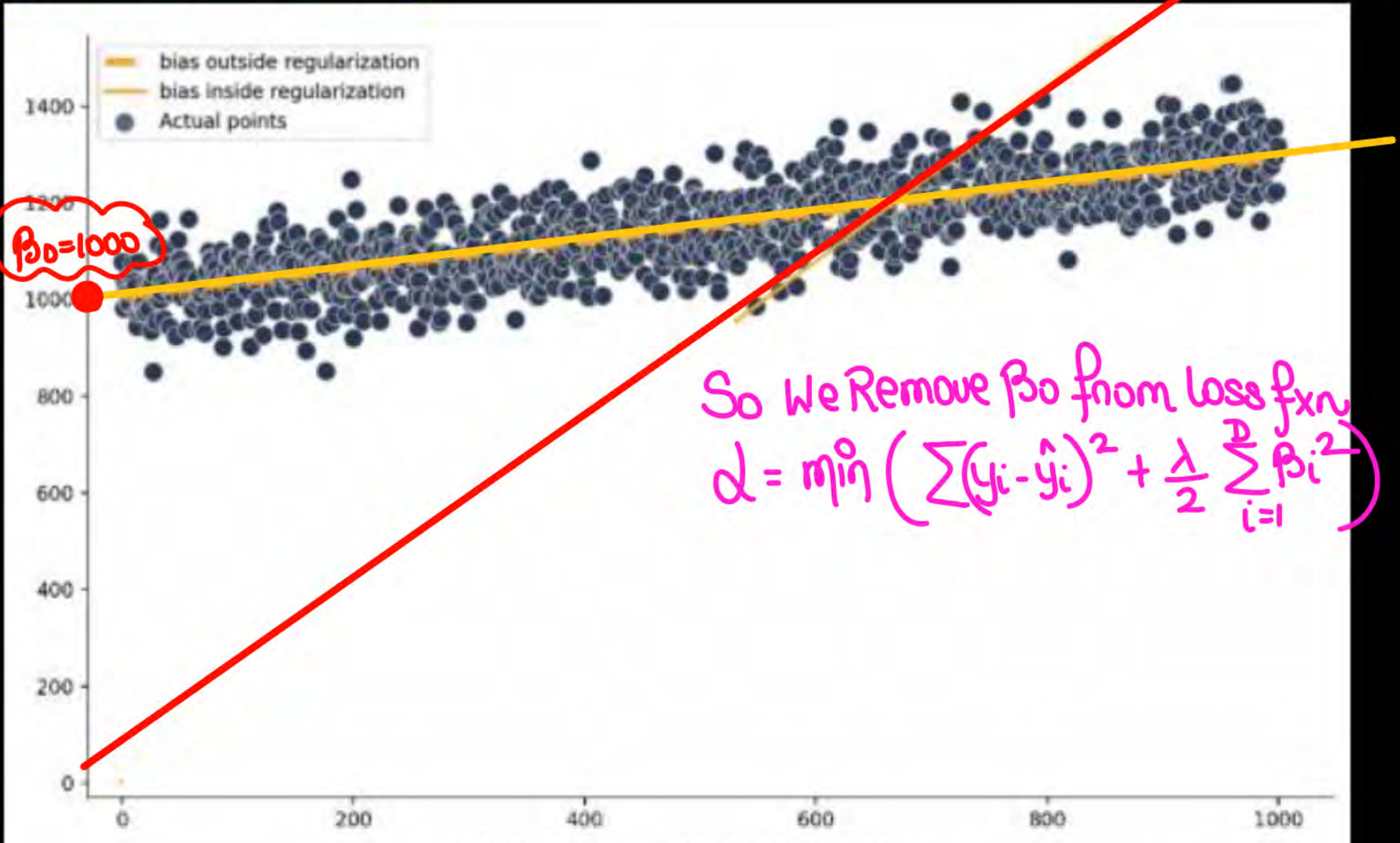
$$\left(y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \dots \right)$$

- So when all the x values are zero then $(y = \beta_0)$

and $\left(\beta_0 = \bar{y} - \beta_1 \overline{x^1} - \beta_2 \overline{x^2} \dots \right)$

↳ β_0 is directly related to avg values of $y, x^1, x^2, \dots$

Ridge Regression



So We Remove β_0 from loss fxn

$$\mathcal{L} = \min\left(\sum (y_i-\hat{y}_i)^2 + \frac{\lambda}{2}\sum_{i=1}^{D}\beta_i^2\right)$$

This GIF has been sourced from the author's website

Shrinkage Methods : Ridge Regression

❖ Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage:

- $\lambda \Rightarrow$ hyperparameter $\Rightarrow$ Const.
- Since β_0 is linked with data's $\bar{y}, \overline{x^1}, \overline{x^2}$ ---- $\Big\}\ \beta_0 = \bar{y} - \beta_1 \overline{x^1} - \beta_2 \overline{x^2}$ ---

  So β_0 shd not be included in loss fxn, because if β_0 is effected then intercept is changed that can ruin whole model (Prev. example)

So we Centre the data $\Rightarrow$

$$x^1 \Rightarrow x^1 - \overline{x^1}$$
$$x^2 \Rightarrow x^2 - \overline{x^2}$$
$$\vdots$$
$$y = y - \overline{y}$$

So we got a new data

$$X = \begin{bmatrix} x_1^1 & x_1^2 & x_1^3 & \cdots & x_1^D \\ x_2^1 & - & - & - & x_2^D \\ x_N^1 & - & - & - & x_N^D \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_D \end{bmatrix}$$

- after centering of data we get a new data, β's which we get here are applicable to the original data Also.

Shrinkage Methods : Ridge Regression

- Here λ is very important control parameter:

Now $\mathcal{L} = \min\left(\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{\lambda}{2}\sum_{i=1}^{D}\beta_i^2\right)$

- 2D and 3 datapoints
- data is centered
- $\hat{y}_i = \beta_1 x_i^1 + \beta_2 x_i^2$

$$\mathcal{L} = \min \frac{1}{2}\sum_{i=1}^{3}\left(y_i - \beta_1 x_i^1 - \beta_2 x_i^2\right)^2 + \frac{\lambda}{2}\left(\beta_1^2 + \beta_2^2\right)$$

$$\frac{\partial L}{\partial \beta_1} \Rightarrow -\left[\sum_{i=1}^{3} y_i x_i^1 - \beta_1 \sum_{i=1}^{3}(x_i^1)^2 - \beta_2 \sum_{i=1}^{3} x_i^2 x_i^1\right] + \lambda\beta_1 = 0$$

$$\frac{\partial L}{\partial \beta_2} \Rightarrow -\left[\sum_{i=1}^{3} y_i x_i^2 - \beta_1 \sum_{i=1}^{3} x_i^1 \cdot x_i^2 - \beta_2 \sum_{i=1}^{3}(x_i^2)^2\right] + \lambda\beta_2 = 0$$

$$\frac{\partial L}{\partial \beta_1} \Rightarrow -\sum_{i=1}^{3} y_i x_i^1 + \beta_1 \sum_{i=1}^{3} (x_i^1)^2 + \beta_2 \sum_{i=1}^{3} x_i^1 x_i^2 + \lambda \beta_1 = 0$$

$$\frac{\partial L}{\partial \beta_2} \Rightarrow -\sum_{i=1}^{3} y_i x_i^2 + \beta_1 \sum_{i=1}^{3} (x_i^1) x_i^2 + \beta_2 \sum (x_i^2)^2 + \lambda \beta_2 = 0$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\frac{\partial L}{\partial \beta} \Rightarrow -X^T Y + (X^T X)\beta + \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \beta = 0$$

$$\Rightarrow -X^T Y + (X^T X)\beta + \lambda I \beta = 0$$

$$\Rightarrow -X^T Y + (X^T X + \lambda I)\beta = 0$$

**

$$\boxed{\text{So } \beta = (X^T X + \lambda I)^{-1} (X^T Y)}$$

$X \Rightarrow \begin{bmatrix} \text{datapoint } 1 \\ \text{" \quad " \quad } 2 \\ \\ \\ \end{bmatrix}$

No Column of '1'

Now from above eq we get β_1, β_2 - - -

$\rightarrow \left(\beta_0 = \bar{y} - \beta_1 \overline{x^1} - \beta_2 \overline{x^2} \cdot - - - \right)$

Ridge Reg $\Rightarrow$ $L = \min\left(\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{\lambda}{2}\sum_{i=1}^{D}\beta_i^2\right)$

$\Rightarrow$ if $\lambda = 0$ $\longrightarrow$ Normal LR $\Rightarrow$ overfitting $\Rightarrow$ Train error = 0
Testing error $\Rightarrow$ high.

if λ = v.v. large $\longrightarrow$ So model try to give β's = 0

↳ So model donot understand data
↳ underfitting
Training error & Testing error high

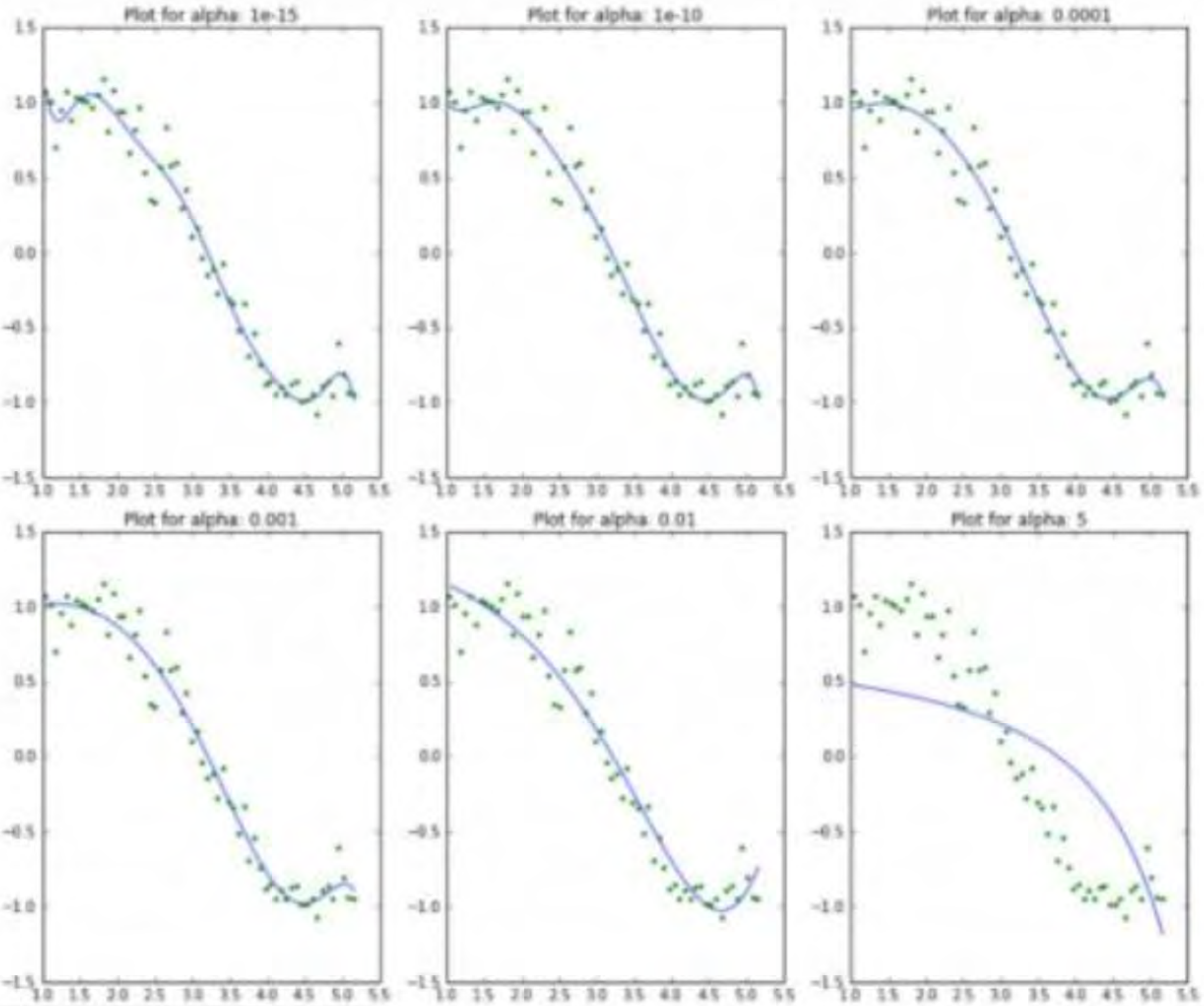
So we have to find best λ by validation.

$\min \ 2f(x)$

$\min \ f(x)$

$\min \ \frac{1}{2}f(x)$

$\Rightarrow$ (The solution for min loc will be same)


Plot for alpha: 1e-15
Plot for alpha: 1e-10
Plot for alpha: 0.0001
Plot for alpha: 0.001
Plot for alpha: 0.01
Plot for alpha: 5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5

Shrinkage Methods : Ridge Regression

- Lets find the solution to this ridge regression problem

Shrinkage Methods : Ridge Regression

- How to find λ (can this be negative?)

Ridge Regression – lets practise

Ridge Regression is a regularization technique used in linear regression to:

A) Increase model complexity.
B) Reduce model complexity and prevent overfitting.
C) Make the model fit the training data perfectly.
D) Enhance the interpretability of the model.

Ridge Regression – lets practise

In Ridge Regression, the penalty term added to the cost function is based on:

A) The absolute values of the regression coefficients.
B) The square of the regression coefficients.
C) The number of features.
D) The dependent variable.

Ridge Regression – lets practise

What happens to the magnitude of regression coefficients in Ridge Regression compared to ordinary linear regression?

A) They become larger.
B) They become smaller.
C) They stay the same.
D) It depends on the dataset.

2 mins Summary

Topic

Topic

Topic

Topic

Topic

THANK - YOU