BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session – 11

Correlation & Regression

IMP Note to Self

Contact Session 11

Contact Session 11: Module 5: Prediction & Forecasting

Contact Session	List of Topic Title	Reference
CS - 11	Correlation & regression	T1& T2
HW	Problems on correlation and regression	T1& T2
Lab	Regression – Analysis of model summary	Lab 5

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?

If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of what degree and in which direction?

To find solutions to the above questions, two approaches are known.

Correlation Analysis

Regression Analysis

Correlation Analysis

Example: Wage Data

A large data regarding the wages for a group of employees from the western region of a country is given.

In particular, we wish to understand the following relationships:

- ***Employee's age and wage:*** **How wages vary with ages?**
- ***Calendar year and wage:*** **How wages vary with time?**
- ***Employee's age and education:*** **Whether wages are anyway related with employees' education levels?**

- Case I. Wage versus Age

- *Employee's age and wage:* How wages vary with ages?

Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

➢ Case II. Wage versus Calendar

➢ ***Wage and calendar year:*** **How wages vary with years?**

Interpretation: ?

➢ Case III. Wage versus Age

Wage and education level: **Whether wages vary with employees' education levels?**

Interpretation: ?

Covariance

The statistical technique used for studying the existence and extent of relationship among 2 or more variables is through Covariance

$$\text{Covariance}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- So in bivariate and multivariate case, the statistical technique used for studying the existence of relationship between all variables is through **Covariance.**
- Covariance signifies the direction of the linear relationship between the two variables.
- By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative effect on the value of the other variable).

- The values of covariance can be any number between the two opposite infinities. Also, it's important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.
- The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables. Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude.

If X_1 and X_2 are two variables, the covariance between X_1 and X_2 is defined as

$$\mathrm{Cov}(X_1, X_2) = \frac{\sum_{i=1}^{n} (X_1 - \overline{X}_1)(X_2 - \overline{X}_2)}{n-1}$$

If $X_1, X_2, \ldots, X_n$ are n variables, then the Covariance between these n variables is the Variance – Covariance matrix given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & . & \sigma_{12} \\ \sigma_{21} & o_{22} & . & . & . & \sigma_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{n1} & o_{n2} & . & . & . & o_{n1} \end{bmatrix}$$

where

$$\sigma_{ij} = \begin{cases} \text{Variance, for } i = j \\ \text{Covariance, for } i \neq j \end{cases}$$

All diagonal elements in the variance-covariance matrix are variances and off diagonal elements are covariances.

Σ is a symmetric matrix.

- Even though the covariance explains the relationship between two or more variables, the units the measurement of variable is attached with the value of the covariance. If they have different unit of measurements, based on the sign it is possible to say whether they are related or not, but the degree (strength) of relationship is not possible to decide based on the value of covariance due to different units of measurement.

Example 1 :

Find the covariance between age (months) and weight (kgs) for the following data

Age (years)	7	6	8	5	6	9
Weight (kgs)	12	8	12	10	11	13

Solution -1

Age (yrs) (X)	Weight (kgs) (Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
5	20	-2	-3	6
4	18	-3	-5	15
7	23	0	0	0
10	32	3	9	27
8	26	1	3	3
8	25	1	2	2
$\sum X = 42$	$\sum X = 144$			$\sum (X-\bar{X})(Y-\bar{Y}) = 53$

$$\mathrm{Cov}(X,Y) = \frac{53}{5} = 10.6 \text{ years.kgs} \rightarrow \text{Interpret this answer !}$$

Covariance based on Probability distribution.

If X and Y are two random variables with probability mass function p(x)/ probability density function f(x), then the covariance between X and Y

$$\mathbf{Cov\ (X, Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - E(X)E(Y)}$$

where $E(X) = \mu_X$ and $E(Y) = \mu_Y$ are the means of X and Y.

- $\mathbf{Covariance(x, x) = Var(x)}$

- $\mathbf{Covariance(x, \boldsymbol{y}) = Covariance(y, x)}$

Application of Covariance.

1. Multivariate linear regression:

- If there a dependent variable and n independent variables, to find out the relationship between these (n+1) variables, a variance – covariance matrix will be helpful.

2. Time series:

- In the given series of data over the time, it helps in finding the autocorrelation at different time lags. For example, if X_{t1} , X_{t2} , ... , X_{tn} denote the time series data over a given period, then the autocovariance

Application of Covariance. at

lag h is given by

$$\gamma_X(t+h, t) = \mathrm{Cov}\,(X_{t+h}, X_t)$$

For example, if $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ denote the time series data over a given period, then the auto-covariance.

The auto-covariance at lag h is then given by

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(X_{t+|h|} - \bar{X})(X_t - \bar{X}), -n < h < n.$$

Application of Covariance.

With the help of auto-covariance, the autocorrelation is given by

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

Karl Pearson's Correlation coefficient

- To overcome this, Karl Pearson's suggestion was to divide the covariance of the variables by their respective standard deviations. The ratio of covariance to the product of standard deviations is called Karl Pearson's product moment correlation, which is free of unit of measurement.
- Karl Pearson's product moment correlation returns a numerical value, popularly known as **Coefficient of Correlation**, denoted by **'r'**
- Correlation coefficient is a statistic that assesses the **strength** and **direction of relationship** of two continuous variables

Formula for computing Karl Pearson's correlation coefficient (r) is standardized covariance (unit less)

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(Y)}}$$

$$r = \frac{\dfrac{\sum_{i=1}^{n}(X-\overline{X})(Y-\overline{Y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(X-\overline{X})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(Y-\overline{Y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y-\overline{Y})^2}}$$

Karl Pearson's Correlation coefficient

Or alternatively we can use

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}, \; x = (X - \overline{X}), \; y = (Y - \overline{Y})$$

Simple Correlation coefficient

- If the sign is +ve this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).
- If the sign is -ve this means an inverse or indirect relationship (an increase in one variable is associated with a decrease in the other).
- The value of r ranges between (-1) and (+1), i.e., $-1 \leq r \leq +1$
- The value of r denotes the strength of the association as illustrated by the following diagram.

Strong
Intermediate
Weak
Weak
Intermediate
Strong
- 1
- 0.75
- 0.25
0
0.25
0.75
+ 1
indirect
Direct
Perfect - ve correlation
No relation
Perfect +ve correlation

Simple Correlation coefficient

- r = - 1 ⇒ Perfect negative correlation
- - 0.99 to - 0.76 ⇒ Strong negative correlation
- - 0.75 to - 0.26 ⇒ Intermediate (Moderate) negative correlation
- -0.25 to 0 ⇒ Weak negative correlation
- r = 0 ⇒ Zero correlation
- 0 to 0.25 ⇒ Weak positive correlation
- 0.26 to 0.75 ⇒ Intermediate (Moderate) positive correlation
- 0.76 to 0.99 ⇒ Strong positive correlation
- r = = 1 ⇒ Perfect positive correlation

Example

Serial No	Age (yrs)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Solution

Sl no.	Age (yrs) (X)	Wt (Kg) (Y)	XY	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	$\sum X=41$	$\sum Y=66$	$\sum XY= 461$	$\sum X^2= 291$	$\sum Y^2= 742$

Karl Pearson's Correlation coefficient

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^{!} - (\sum X)^{!}}\sqrt{n\sum Y^{!} - (\sum Y)^{!}}}$$

$$r = \frac{6*461 - 41*66}{\sqrt{6*291 - (41)^{!}}\sqrt{6*742 - (66)^{!}}}$$

r = 0.759 (Positive (Direct) strong correlation)

Karl Pearson's Correlation coefficient

Example: Relationship between anxiety and test Scores

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY = 129$

Correlational Analysis ➡ Karl Pearson's Correlation coefficient

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^{!} - (\sum X)^{!}}\sqrt{n\sum Y^{!} - (\sum Y)^{!}}}$$

$$r = \frac{6*129 - 32*32}{\sqrt{\left(6*230 - (32)^2\right)\left(6*204 - (32)^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -0.94$$

r = - 0.94 (Negative (Indirect) strong correlation)

Example : A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Solution :

In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus *Y* = birth weight and *X* = gestational age.

For the given data, it can be shown the following

$$\bar{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

$$s_x^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\Sigma(Y-\bar{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

$$r^* = \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{s_x.s_y} = 0.82$$

Conclusion: The sample's correlation coefficient indicates a **strong positive correlation** between Gestational Age and Birth Weight.

Homework Problems

Q1) Consider the following data.

Expenditure (in 1000 Rs)	30	70	40	20	10	40	10	20
Actual Sales(in 1000 Rs)	10	16	19	14	17	16	13	18

For the above data,

i) Find coefficient of correlation between expenditure and actual sales. **[Ans = 0.8108]**

ii) Find rank correlation between expenditure and actual sales. **[Ans = - 0.2424]**

Q2) Find coefficient of correlation between X & Y.

X	79	29	45	61	24	38	33	52	65	63	82	50
Y	15	12	12	6	6	7	3	10	12	13	13	14

Ans = r= 0.59

Q3) Check whether X and Y are having any relation.

X	12	14	15	17	17	16	16	15	14	14
Y	52	52	57	62	67	67	67	62	62	52

Ans = r = 0.7866

Regression

x	1	2	3	4	5	6
y	1	4	9	16	25	36

Regression

x Income (Lakhs)	25	30	20	15	18	22
y Exp (lakhs)	6	8	5	7	4	5

when $x = 32$, then

$y = ?$

100
80
60
40
30
20
10
1
2
3
4
5
6
Linear

Non-linear

CORRELATION	REGRESSION
Measuring strength or degree of the relationship between two variables	Having an algebraic equation between two variables
No estimation	Estimation
Both variables are independent	One is dependent variable and the other is independent variable

Method of Least squares

y : Dependent Variable

x : Independent Variable

predictor variable

response Variable

$$y_i = \beta_0 + \beta_1 x_i$$

Method of Least squares

Method of Least squares

$$S(\beta_0, \beta_1) = \underbrace{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$$

we need to choose β_0 and β_1 which minimizes the error.

Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)(-1)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)(2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression –

$$y = \beta_0 + \beta_1 x$$

$$\left.\begin{aligned} \Sigma y &= \beta_0 n + \beta_1 \Sigma x \\ \Sigma xy &= \beta_0 \Sigma x + \beta_1 \Sigma x^2 \end{aligned}\right\}$$

Normal equations.

$$Y = b_0 + b_1 X$$

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

Example :-

company	Advt Expt	Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$$y = a + bx$$

$$\Sigma y = an + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Example :-

$y = \beta_0 + \beta_1 x$

Sales Revenue y	Advt. expt. x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
$\Sigma 40$	$\Sigma 56$	$\Sigma 524$	$\Sigma 373$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$\Rightarrow 40 = 8\beta_0 + 56\beta_1$

$373 = 56\beta_0 + 524\beta_1$

on solving

$\beta_0 = 0.072$

$\beta_1 = 0.704$

$\therefore y = (0.072) + (0.704)x$

ie $y = (0.072) + (0.704)x$

when $x = 0.075$, then

$$y = (0.072) + (0.704)(0.075)$$

$$= 0.1248 \simeq 12.48\%$$

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma\ 3.5$	$\Sigma\ 10.5$	$\Sigma\ 21$

$$y = \beta_0 + \beta_1 x$$

$$\Sigma y = n\beta_0 + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$$

$$3.5 = 4\beta_0 + \beta_1 (7)$$

$$10.5 = 7\beta_0 + \beta_1 (21)$$

on solving these

$$\beta_0 = 0$$

$$\beta_1 = 0.5$$

ie $y = 0 + (0.5)x$

$$\boxed{\text{when } x = 5, \quad y = (0.5)5 = 0.25}$$

Assumptions of the Linear Regression

Assumptions about the Error

- $E(\varepsilon_i) = 0$ for $i = 1, 2, \ldots, n$.
- $\sigma(\varepsilon_i) = \sigma_\varepsilon$ where σ_ε is unknown.
- The errors are independent, that is, the error in the *i*th observation is independent of the error observed in the *j*th observation.
- The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).

$$y_i = \hat{y}_i + (y_i - \hat{y}_i)$$

observed — Fit — Deviation from fit

subtracting $\bar{y}$ from both sides

$$(y_i - \bar{y}) = (\hat{y} - \bar{y}) + (y_i - \hat{y}_i)$$

Deviation from mean — Deviation due to fit — Residual

RSS $\longrightarrow$ Residual sum of squares

$$= \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i) \right)^2$$

TSS $\longrightarrow$ $\sum_{i=1}^{n} (y_i - \bar{y})^2$ mean of respective variables

$$R^2 = 1 - \frac{RSS}{TSS}$$

Overfitting

Which of these two models would be a better fit to the data?

Multicollinearity

- It refers to the phenomenon of having related predictor variables in the input dataset.
- In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. You drop some of these related independent variables as a way of dealing with multicollinearity.
- **Multicollinearity affects:**

Interpretation:

- Does “change in Y, when all others are held constant” apply?

Inference:

- Coefficients swing wildly, signs can invert
- p-values are, therefore, not reliable

Dealing with Multicollinearity

- Two basic ways of dealing with multicollinearity
- Looking at **pairwise correlations**
  - Looking at the correlation between different pairs of independent variables

Checking the **Variance Inflation Factor** (VIF)

- Sometimes pairwise correlations aren't enough
- Instead of just one variable, the independent variable might depend upon a combination of other variables
- VIF calculates how well one independent variable is explained by all the other independent variables combined

Variance Inflation Factor (VIF)

- The VIF is given by:

$$VIF_i = \frac{1}{1-R_i^2}$$

where *'i'* refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

- The common heuristic we follow for the VIF values is:

> 10: Definitely high VIF value and the variable should be eliminated.

> 5: Can be okay, but it is worth inspecting.

< 5: Good VIF value. No need to eliminate this variable.

Example:-

x_0	Size	No of rooms	No of floors	Age of home	price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
	x_1	x_2	x_3	x_4	y

Multiple Linear Regression

The data consists of n observations on a dependent or response variable Y and p predictor or explanatory variables

$X_1, X_2, \ldots, X_p$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

β_i's are regression coefficients.

normal equations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

Multiple Linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2 + \beta_3 \sum x_3 + \beta_4 \sum x_4$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 + \beta_3 \sum x_1 x_3 + \beta_4 \sum x_1 x_4$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 + \beta_3 \sum x_2 x_3 + \beta_4 \sum x_2 x_4$$

$$\sum x_3 y = \beta_0 \sum x_3 + \beta_1 \sum x_1 x_3 + \beta_2 \sum x_2 x_3 + \beta_3 \sum x_3^2 + \beta_4 \sum x_3 x_4$$

$$\sum x_4 y = \beta_0 \sum x_4 + \beta_1 \sum x_1 x_4 + \beta_2 \sum x_2 x_4 + \beta_3 \sum x_3 x_4 + \beta_4 \sum x_4^2$$

solve $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

normal equations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

Non - Linear Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\sum y = \beta_0 n + \beta_1 \sum x + \beta_2 \sum x^2$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2 + \beta_2 \sum x^3$$

$$\sum x^2 y = \beta_0 \sum x^2 + \beta_1 \sum x^3 + \beta_2 \sum x^4$$

1) $Y = \alpha x^{\beta}$

2) $Y = \alpha e^{\beta x}$

3) $Y = \alpha + \beta \log x$

4) $Y = \dfrac{\alpha}{\alpha x - \beta}$

Other regressions

just a look

Suppose $y = ae^{bx}$

$$\underbrace{\log y}_{Y} = \underbrace{\log a}_{A} + \underbrace{b \log x}_{X}$$

ie $Y = A + bX$

$$\Sigma Y = An + b\Sigma X \longrightarrow (1)$$

$$\Sigma XY = A\Sigma X + b\Sigma X^2 \longrightarrow (2)$$

Hence, we get

$$y = ae^{bx}$$

R – Squared vs Adjusted R - Squared

- In multiple regression, adjusted R – squared is better metric than R – squared asses the goodness of fit of the model

- R – squared always increases if additional variables are added into model , even if they are not related to the dependent variable

Lasso & Ridge

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- OLS estimation: $$\min SSE = \sum \left(Y - \hat{Y} \right)^2$$
- LASSO estimation: $$\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$
- Ridge regression estimation: $$\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^{\,2} \right|$$

IMP Note to Self

Thanks