Regression

Ansar Shahzadi
School of Electrical Engineering & Computer Science
National University of Science and Technology(NUST)

Regression

The regression model is a statistical procedure that allows a researcher to estimate the linear, or straight line, relationship that relates two or more variables. This linear relationship summarizes the amount of change in one variable that is associated with change in another variable or variables.

Regression

- Regression: technique concerned with predicting some variables by knowing others
- The process of predicting variable Y using variable X
- The dependent variable assumed to be random variable whereas the independent Variables are assumed to have fixed values.

A probabilistic model for linearly related data

We observe paired data points (x_1, y_1),(x_2, y_2), . . . ,(x_n, y_n), where we assume that as a function of x_i , each y_i is generated by using some true underlying line

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

or in terms of sample data

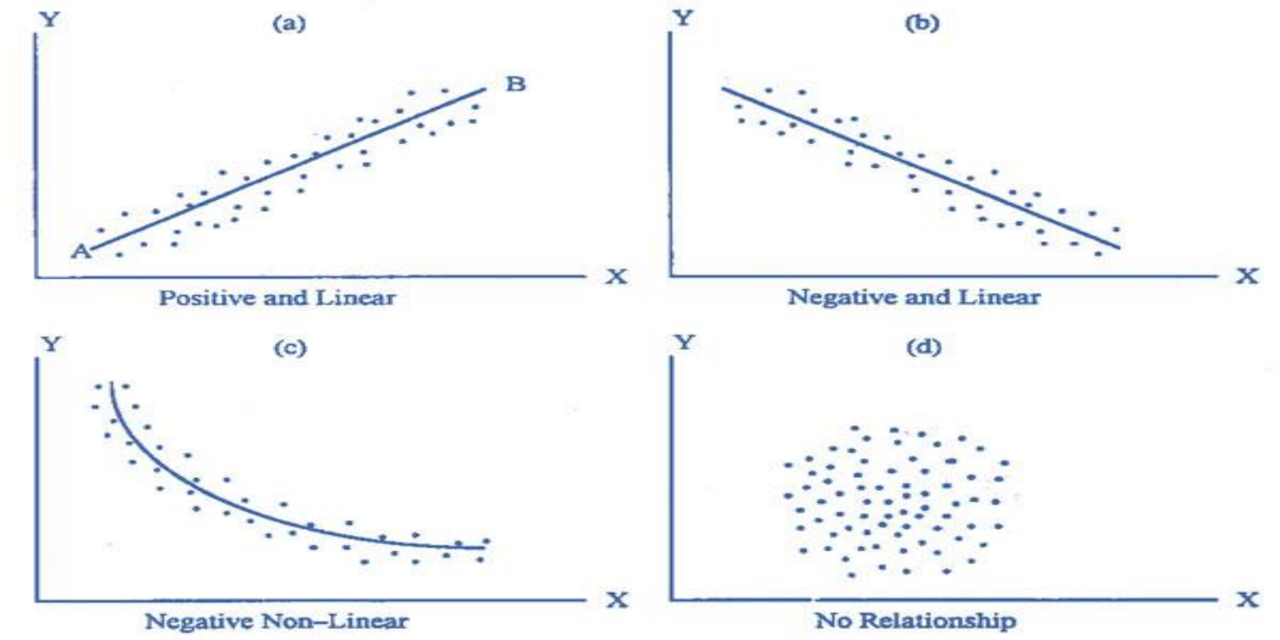
$$Y_i = a + bX_i + e_i$$

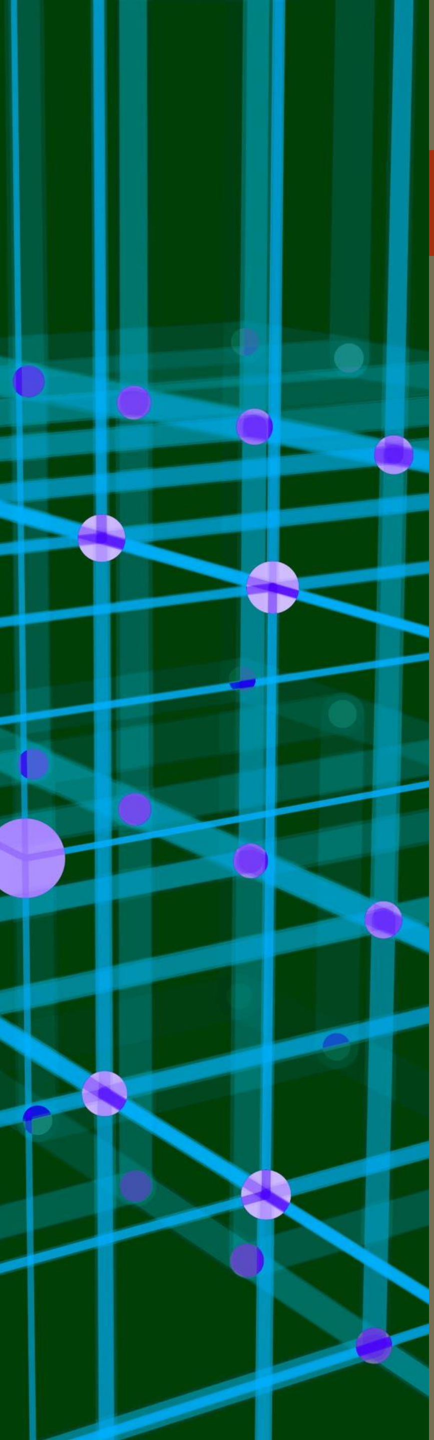
Where

- a is intercept
- b is slope
- e_i are the unknown random error.

Scatter Diagram

Scatter Plots (also called scatter diagrams) are used to investigate the possible relationship between two variables that both relate to the same "event." A straight line of best fit (using the least squares method) is often included.





Least Square Method

By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of: $\hat{Y} = a + bX$

Here $b = \frac{\sum\ (\boldsymbol{x}-\overline{\boldsymbol{x}})(y-\bar{y})}{\sum(x-\bar{x})^2}$ or $b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$

And $a = \bar{Y} - b\bar{X}$

We also use normal equations to find the values a and b

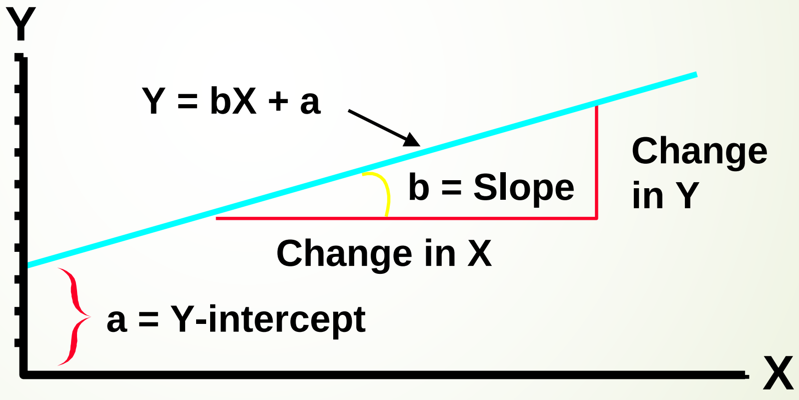
$\sum y$=na+b$\sum x$,

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

Straight Line Equation

Regression equation describes the regression line mathematically

- Intercept
- Slope



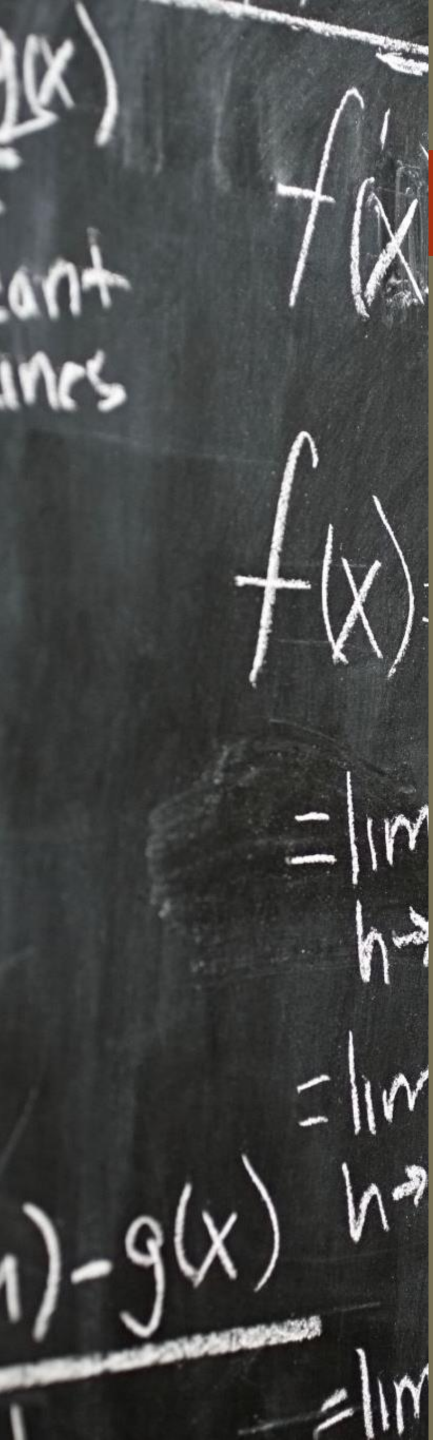
Properties of the Least Square Regression Line

- The least regression line always goes through the point $(\bar{X}, \bar{Y})$, the mean of the data.
- The sum of deviations of the observed values of Y_i from the least squares regression line is always equal to zero, i.e.$\sum(Y_i - \hat{Y})$=0
- The sum of squared deviations of the observed values from the least squares regression is a minimum, i.e.$\sum(Y_i - \hat{Y})^2$=minimum
- The least squared regression line obtained from a random sample is the line of best fit because a and b are two unbiased estimates of the parameter a and β.

Example 1

Find the regression equation for the following data and also predict value of Y at 14.

X	5	6	8	10	12	13	15	16	17
Y	16	19	23	28	36	41	44	45	50



Multiple Regression

Multiple Regression Analysis refers to a set of techniques for studying the straight-line relationships among two or more variables. Multiple regression estimates the β_i's in the equation

$Y = a + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \beta_k X_{ki} + \varepsilon_i$

The X's are the independent variables . Y is the dependent variable. The a and β_i's are the unknown regression coefficients. Their estimates are represented by b_i's. Each β represents the original unknown (population) parameter, while b is an estimates of population parameters. The corresponding regression equation estimated from the sample data is given below

$Y = a + b_1 X_{1i} + b_2 X_{2i} + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_k X_{ki}$

Multiple Regression Examples

- **Predicting a systolic blood pressure** of a person based on their age, height, weight, etc
- Y: **crop yield**

X: rainfall, temperature, humidity,, etc.

- Y: **income**

X: age, education, occupation, etc.

- Y: **selling price of homes**

X: size, location, quality, etc.

- Y: **test scores**

X: teaching method, ability, time of study etc.

Understanding the relationship between the predictors and the response.

Multiple Regression With two Regressors

The estimated multiple regression based on sample data is $\hat{Y}$ = a+ b_1 X_{1i} + b_2 X_{2i}

For a set of n observations.

By using the least squares method, we get the following three normal equations

$\sum Y$ = na+ $b_1 \sum X_1 + b_2 \sum X_2$

$\sum X_1 Y = a \sum X_1$+ $b_1 \sum X_1^2$+ $b_2 \sum X_1 X_2$

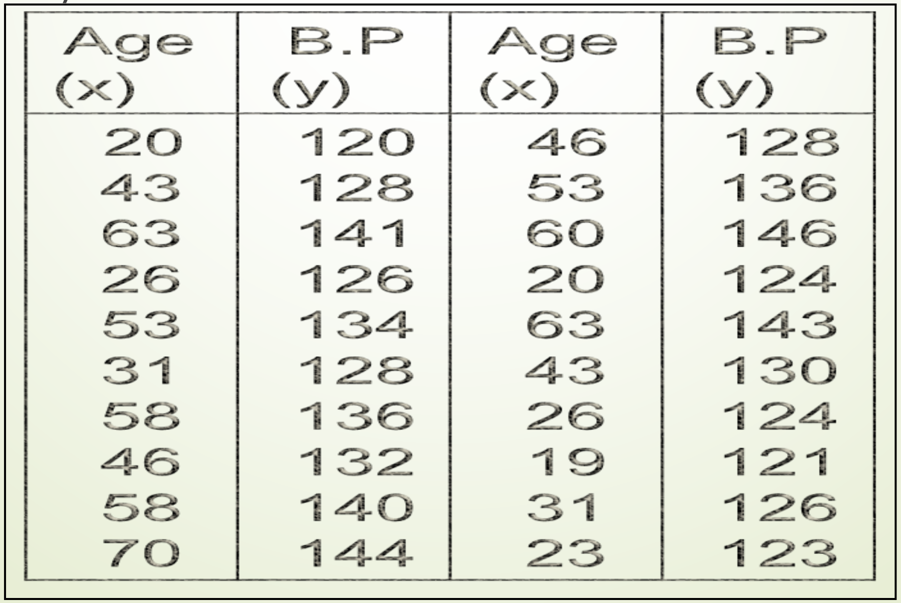
$\sum X_2 Y = a \sum X_2$+ $b_1 \sum X_1 X_2$ + $b_2 \sum X_2^2$

The value of a, b_1 and b_2 are determinad by solving these three normal equation simultaneously.

Question#1

The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

- Find the regression equation?
- What is the predicted blood pressure for a man aging 25 years?



Age (x)	B.P (y)	Age (x)	B.P (y)
20	120	46	128
43	128	53	136
63	141	60	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123

Question#2

The data set in table concerns the relationship between temperature and resistance of vacuum transducer bobbins, which are used in automobile industry.

- Fit a linear regression model with resistance as the dependent variable and temperature as the explanatory variable.
- What is the predicted resistance when the temperature is 69 F?
- Plot scatter diagram. Comment about the relationship?

T(degree F)	60	61	63	63	65	65	65	66	68
R(ohms)	59.47	63.44	61.59	62.44	65.84	64.21	66.87	64.82	68.93
T(degree F)	69	70	70	72	72	72	74	75	77
R(ohms)	69.80	67.95	71.14	68.71	71.64	73.04	70.73	74.75	70.86