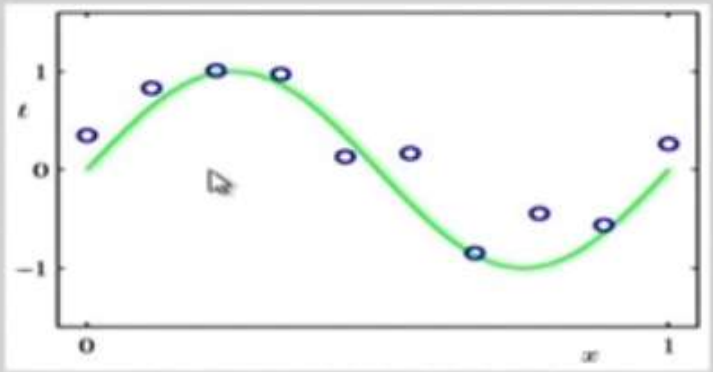
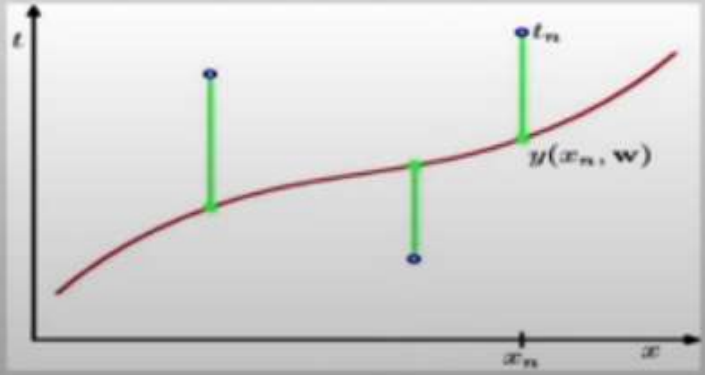
A Simple Example: Fitting a Polynomial

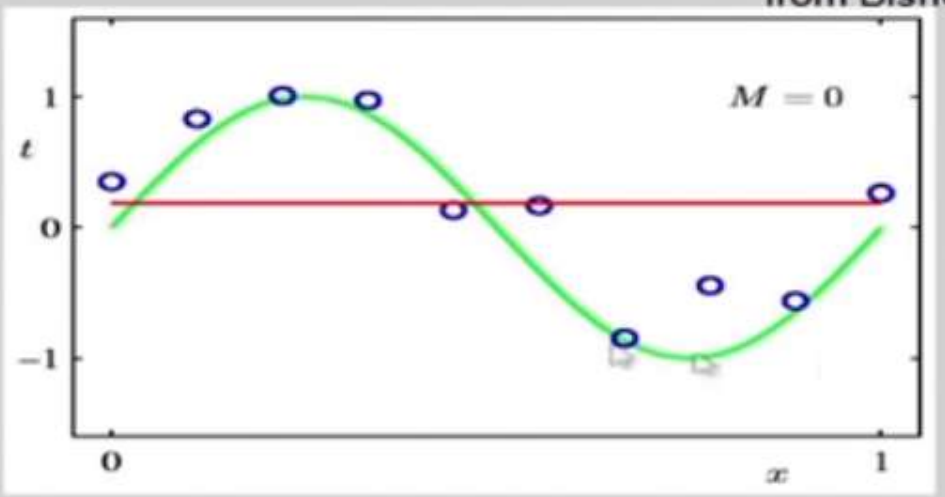
- The green curve is the true function (which is not a polynomial)
- We may use a loss function that measures the squared error in the prediction of y(x) from x.

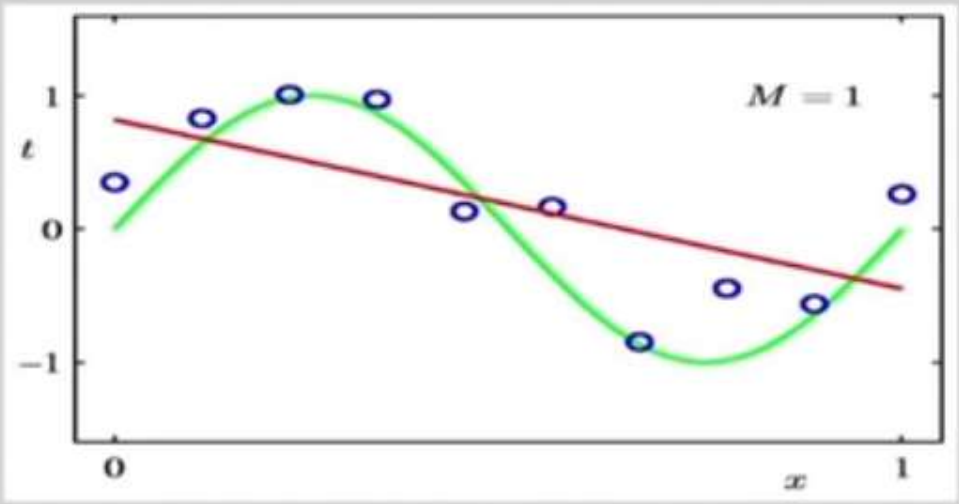


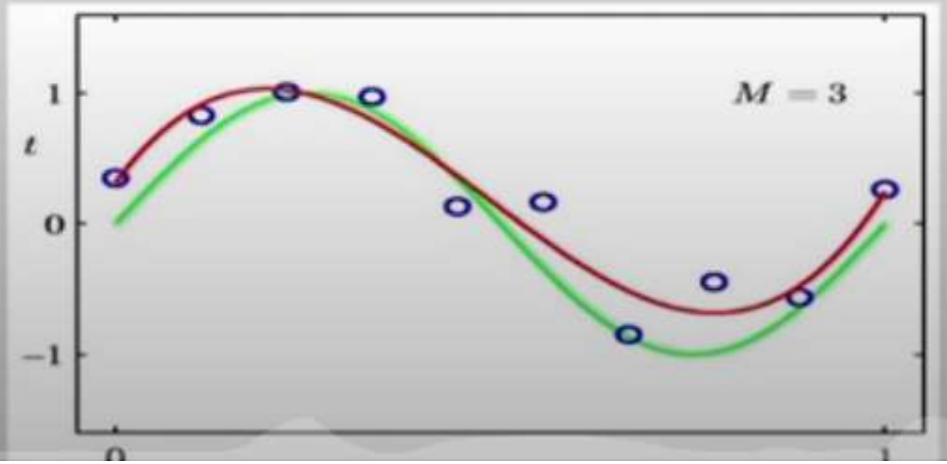


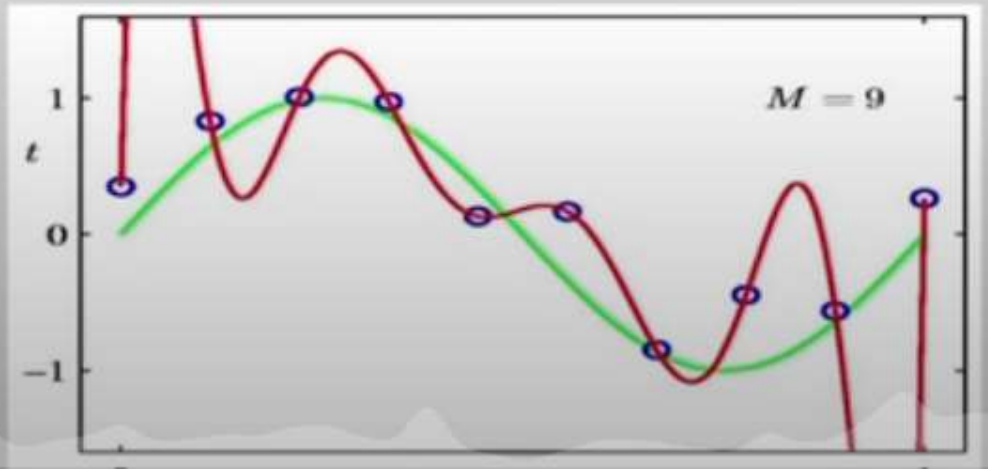
Some fits to the data: which is best?

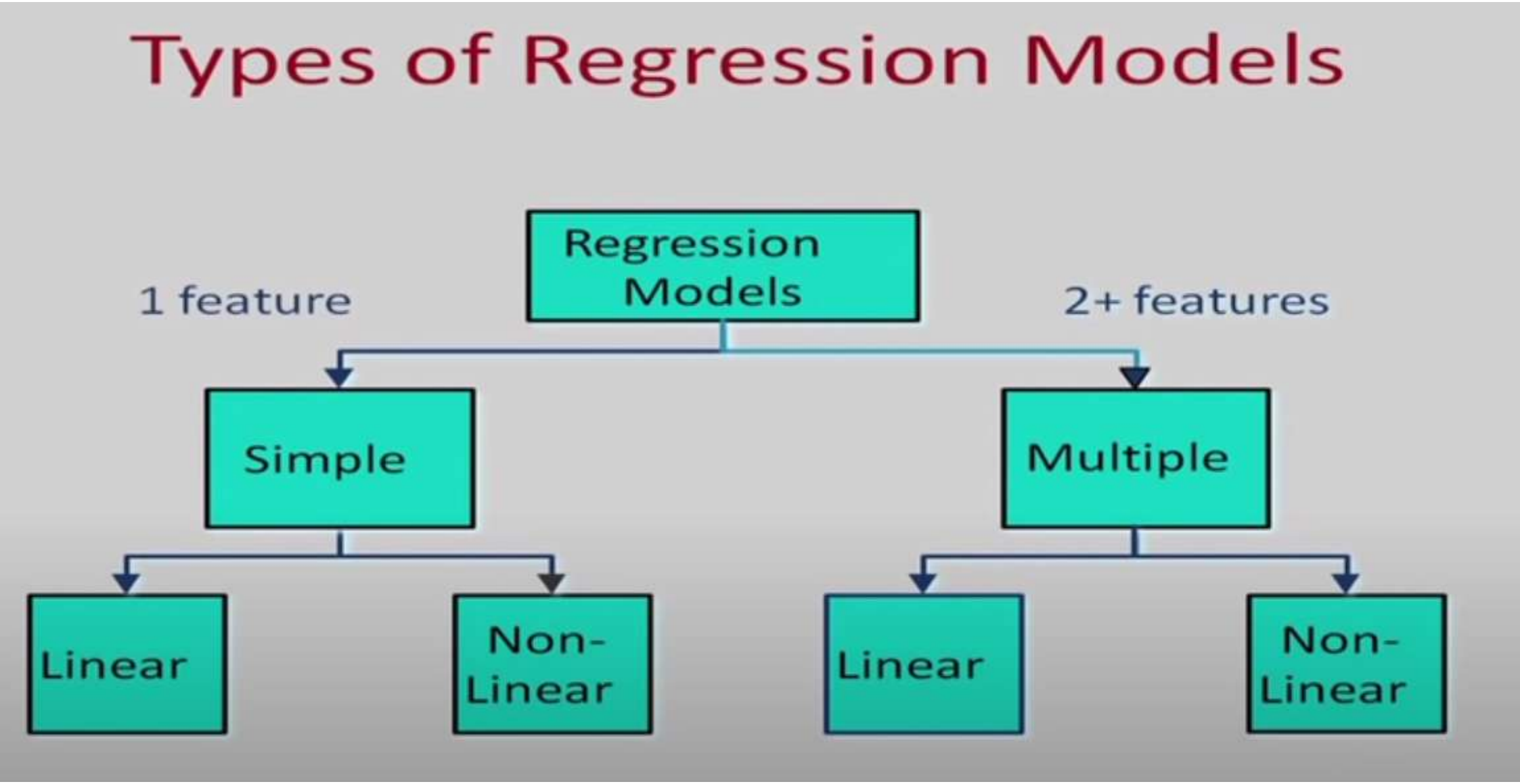
from Bishop









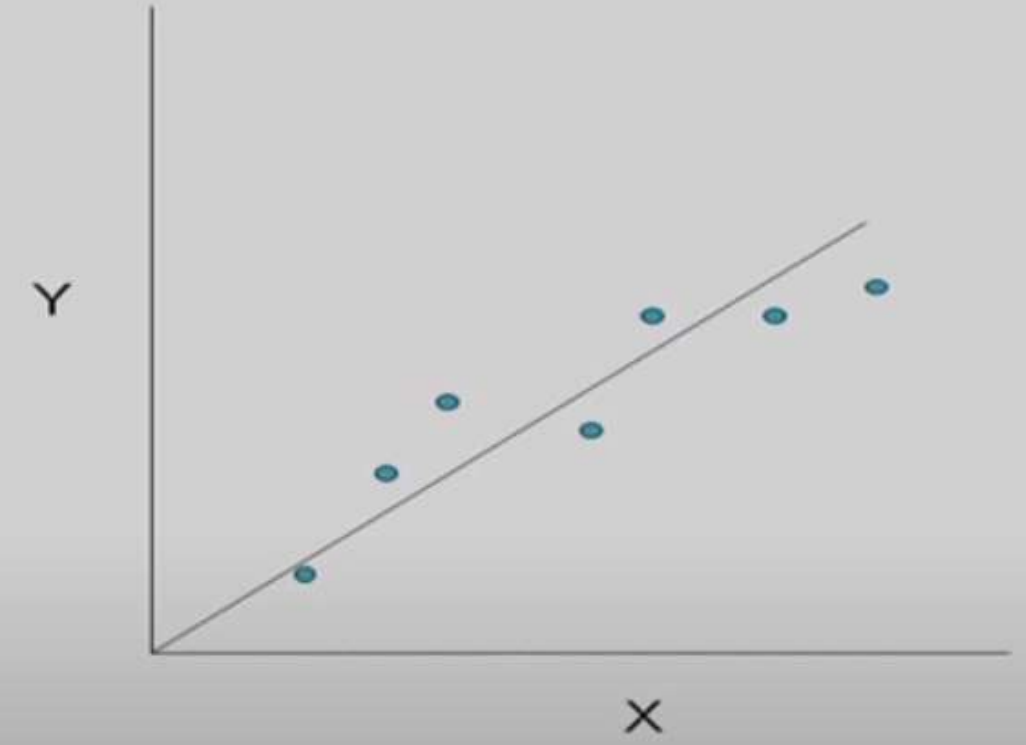

Types of Regression Models
Regression Models
1 feature
2+ features
Simple
Multiple
Linear
Non-Linear
Linear
Non-Linear

Linear regression

Given an input x compute an output y

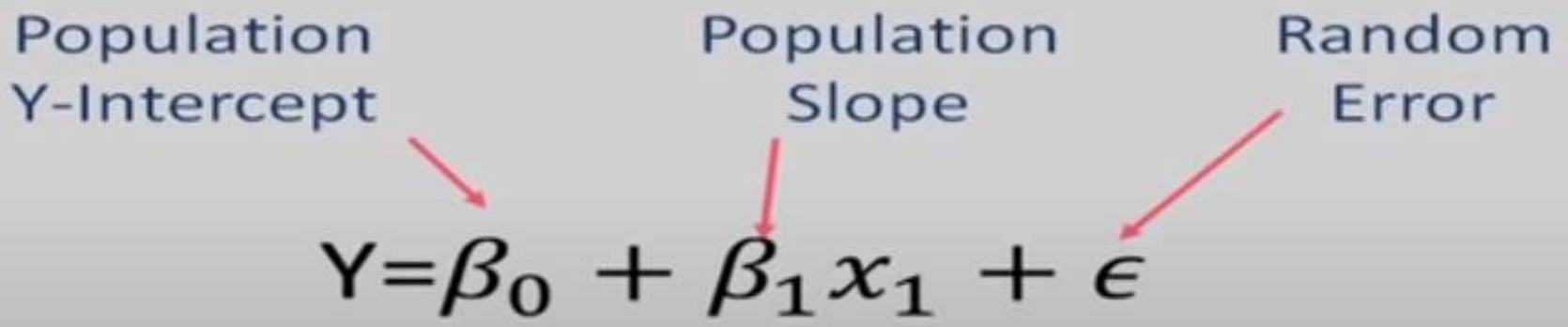
For example:

- Predict height from age
- Predict house price from house area
- Predict distance from wall from sensors



Linear Regression Model

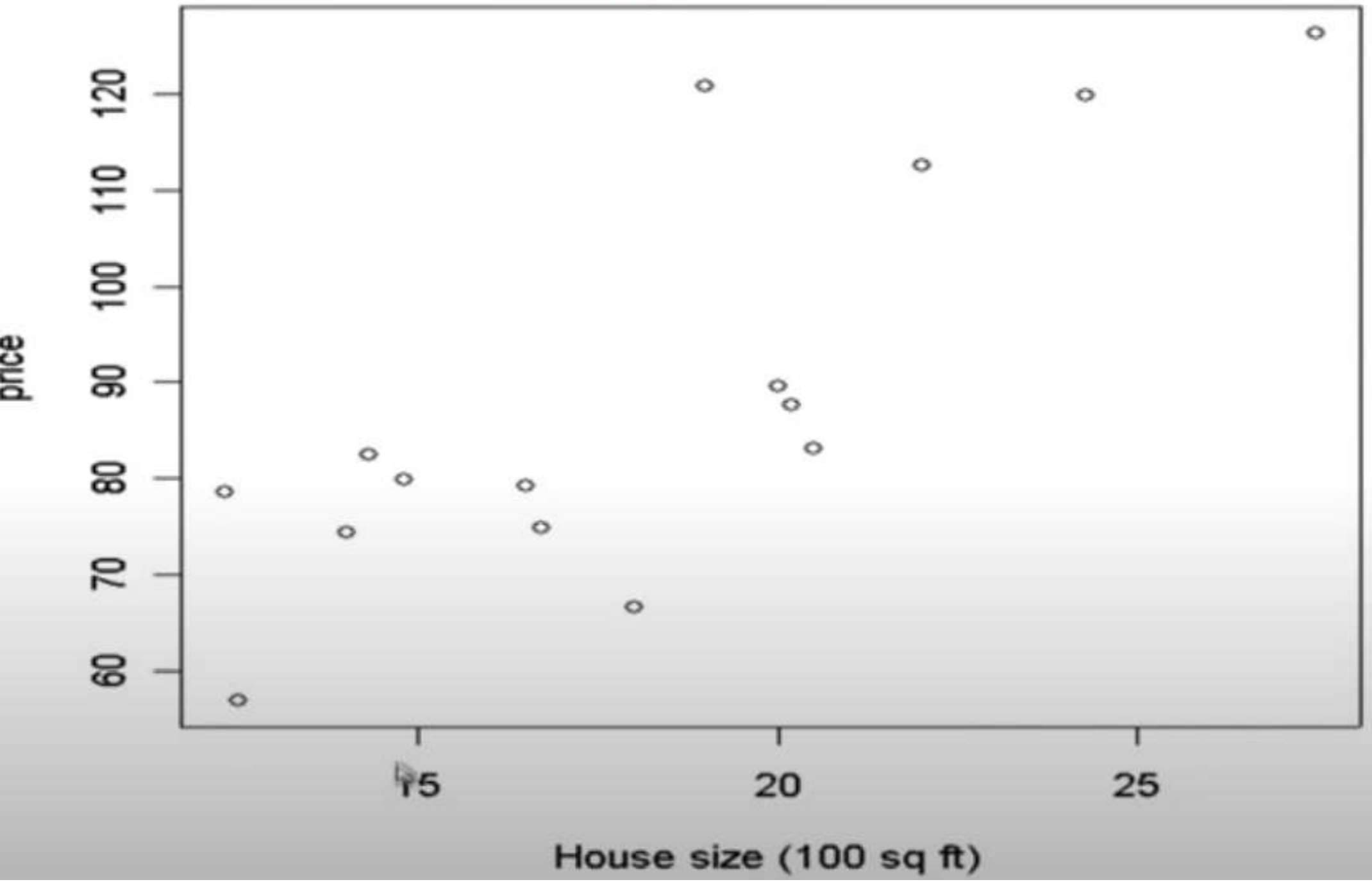
- Relationship Between Variables Is a Linear Function



House Number	*Y*: Actual Selling Price	*X*: House Size (100s ft2)
1	89.5	20.0
2	79.9	14.8
3	83.1	20.5
4	56.9	12.5
5	66.6	18.0
6	82.5	14.3
7	126.3	27.5
8	79.3	16.5
9	119.9	24.3
10	87.6	20.2
11	112.6	22.0
12	120.8	.019
13	78.5	12.3
14	74.3	14.0
15	74.8	16.7
Averages	88.84	18.17

Sample 15 houses from the region.

House price vs size



Assumption

- The data may not form a perfect line.
- When we actually take a measurement (i.e., observe the data), we observe:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where ε_i is the random error associated with the *i*th observation.

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.
- The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).

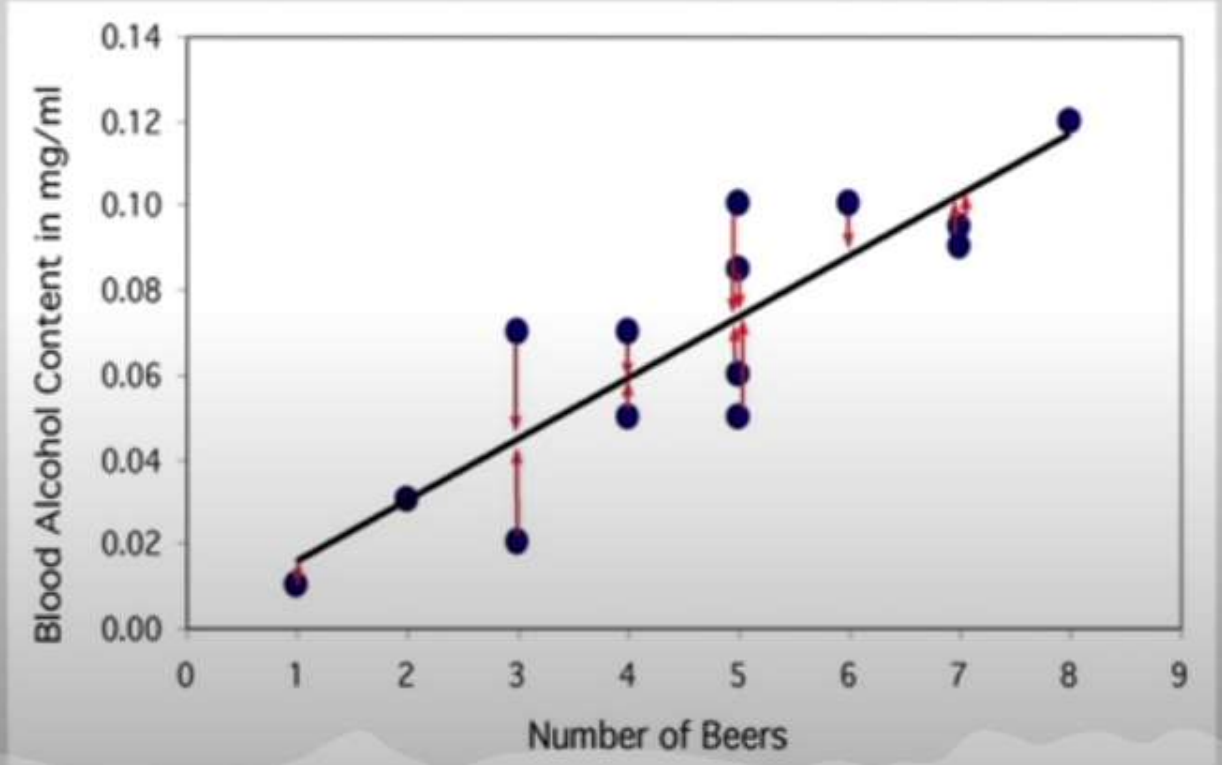
Linear Regression – Multiple Variables

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- β_0 is the intercept (i.e. the average value for Y if all the X's are zero), β_j is the slope for the *j*th variable X_j

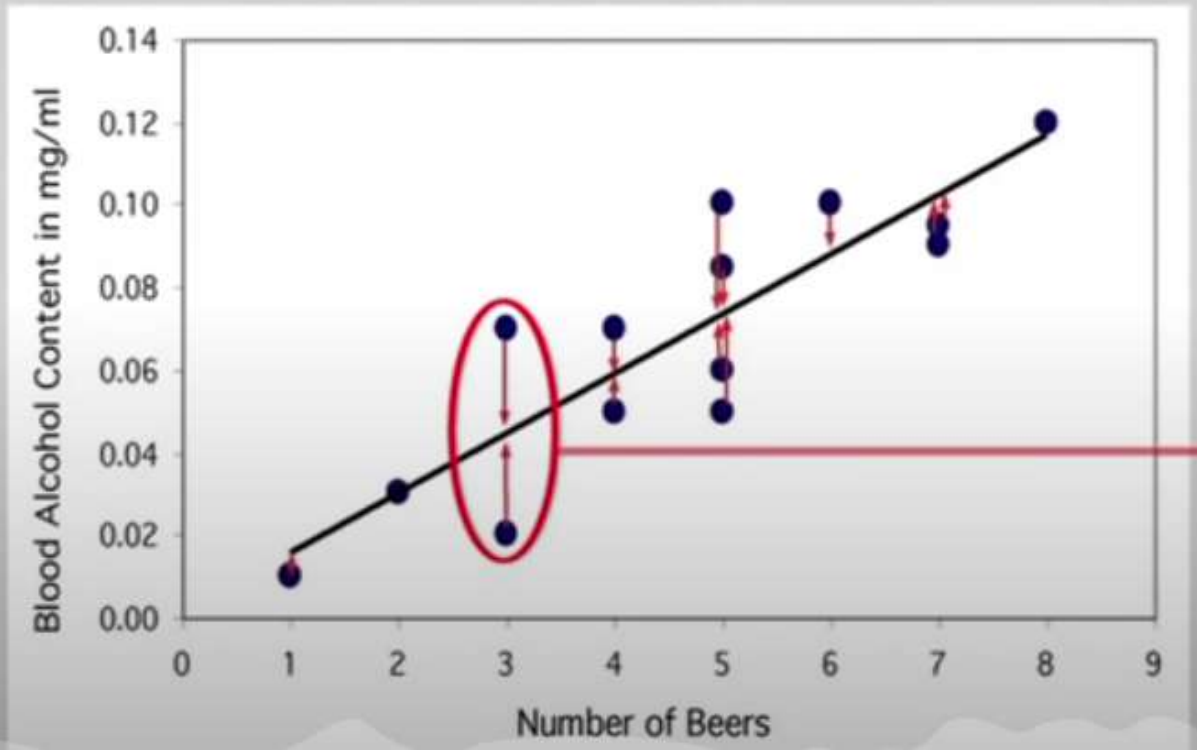
The regression line

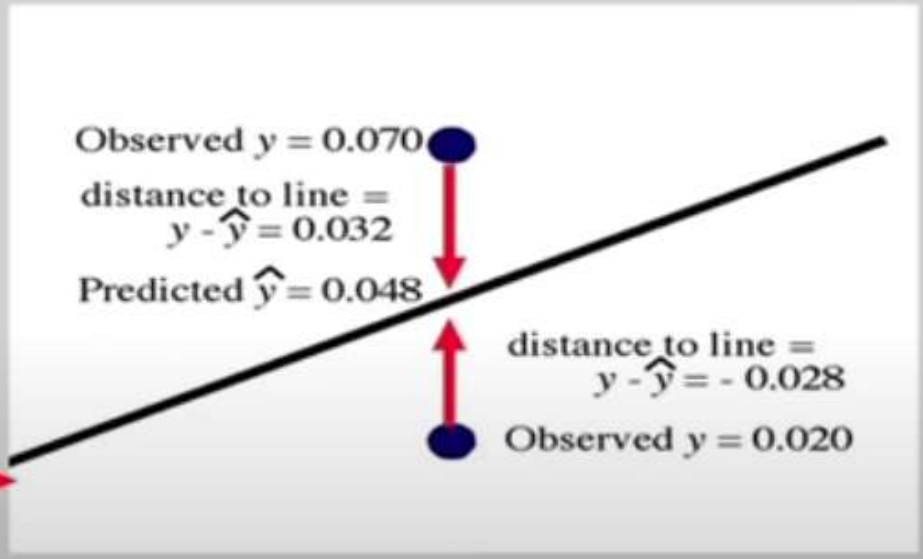
The least-squares regression line is the unique line such that the sum of the squared vertical (y) distances between the data points and the line is the smallest possible.



The regression line

The least-squares regression line is the unique line such that the sum of the squared vertical (y) distances between the data points and the line is the smallest possible.





Criterion for choosing what line to draw: method of least squares

- The method of least squares chooses the line ($\widehat{\beta_0}$ and $\widehat{\beta_1}$) that makes the sum of squares of the residuals $\sum \varepsilon_i^{\,2}$ as small as possible
- Minimizes

$$\sum_{i=1}^{n}[y_i-(b_0+b_1x_i)]^2$$

for the given observations (x_i, y_i)

How do we "learn" parameters

- For the 2-*d* problem

$$Y = \beta_0 + \beta_1 X$$

- To find the values for the coefficients which minimize the objective function we take the partial derivates of the objective function (SSE) with respect to the coefficients. Set these to 0, and solve.

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

$$h(x) = \sum_{i=0}^{n} \beta_i x_i$$

- There is a closed form which requires matrix inversion, etc.
- There are iterative techniques to find weights
  - delta rule (also called LMS method) which will update towards the objective of minimizing the SSE.

Linear Regression

$$h(x) = \sum_{i=0}^{n} \beta_i x_i$$

To learn the parameters θ (β_i) ?

- Make h(**x**) close to *y*, for the available *training examples*.
- Define a cost function $J(\theta)$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h(x)^{(i)} - (y)^{(i)})^2$$

- Find θ that minimizes J(θ).

LMS Algorithm

- Start a search algorithm (e.g. gradient descent algorithm,) with initial guess of θ.
- Repeatedly update θ to make J(θ) smaller, until it converges to minima.

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\theta)$$

- *J* is a convex quadratic function, so has a single global minima. gradient descent eventually converges at the global minima.
- At each iteration this algorithm takes a step in the direction of steepest descent(-ve direction of gradient).

Linear Regression Models -

a. The term "regression" generally refers to predicting a real number. However, it can also be used for classification (predicting a category or class.)

b. The term "linear" in the name “linear regression” refers to the fact that the method models data with linear combination of the explanatory variables.

c. A linear combination is an expression where one or more variables are scaled by a constant factor and added together.

d. In the case of linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

$$\text{response} = \text{intercept} + \text{constant} * \text{explanatory}$$

e. In its most basic form fits a straight line to the response variable. The model is designed to fit a line that minimizes the squared differences (also called errors or residuals.).

Linear Regression Models -

a. Before we generate a model, we need to understand the degree of relationship between the attributes Y and X

b. Mathematically correlation between two variables indicates how closely their relationship follows a straight line. By default we use Pearson's correlation which ranges between -1 and +1.

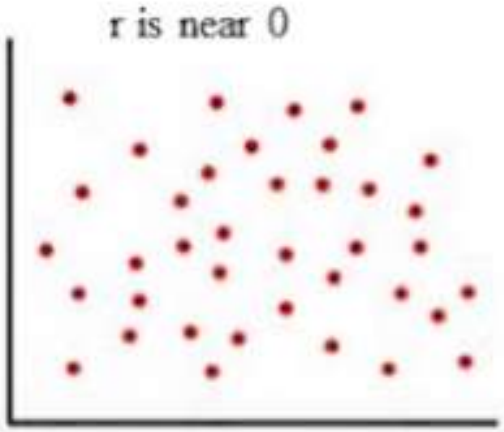
c. Correlation of extreme possible values of -1 and +1 indicate a perfectly linear relationship between X and Y whereas a correlation of 0 indicates absence of linear relationship

   I. When r value is small, one needs to test whether it is statistically significant or not to believe that there is correlation or not

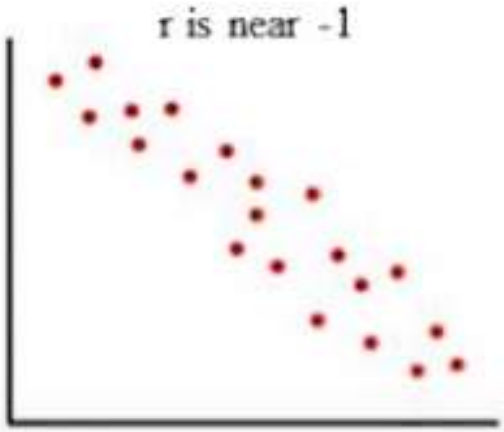
Linear Regression Models (Recap) -

d. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

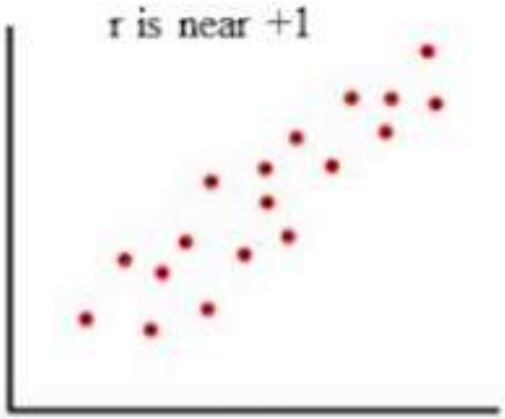
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



No Correlation



Negative



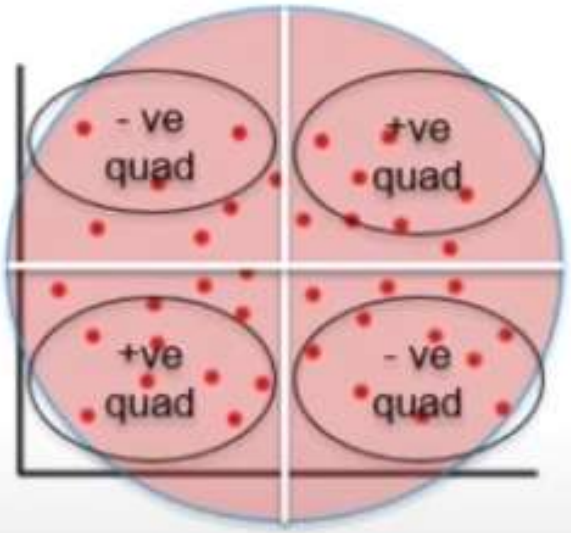
Positive

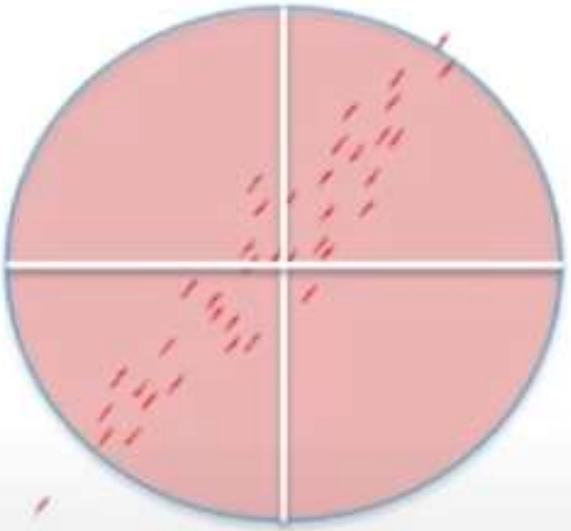
e. Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! Nonlinear models may be better in such cases

Linear Regression Models (Recap) -

f. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$





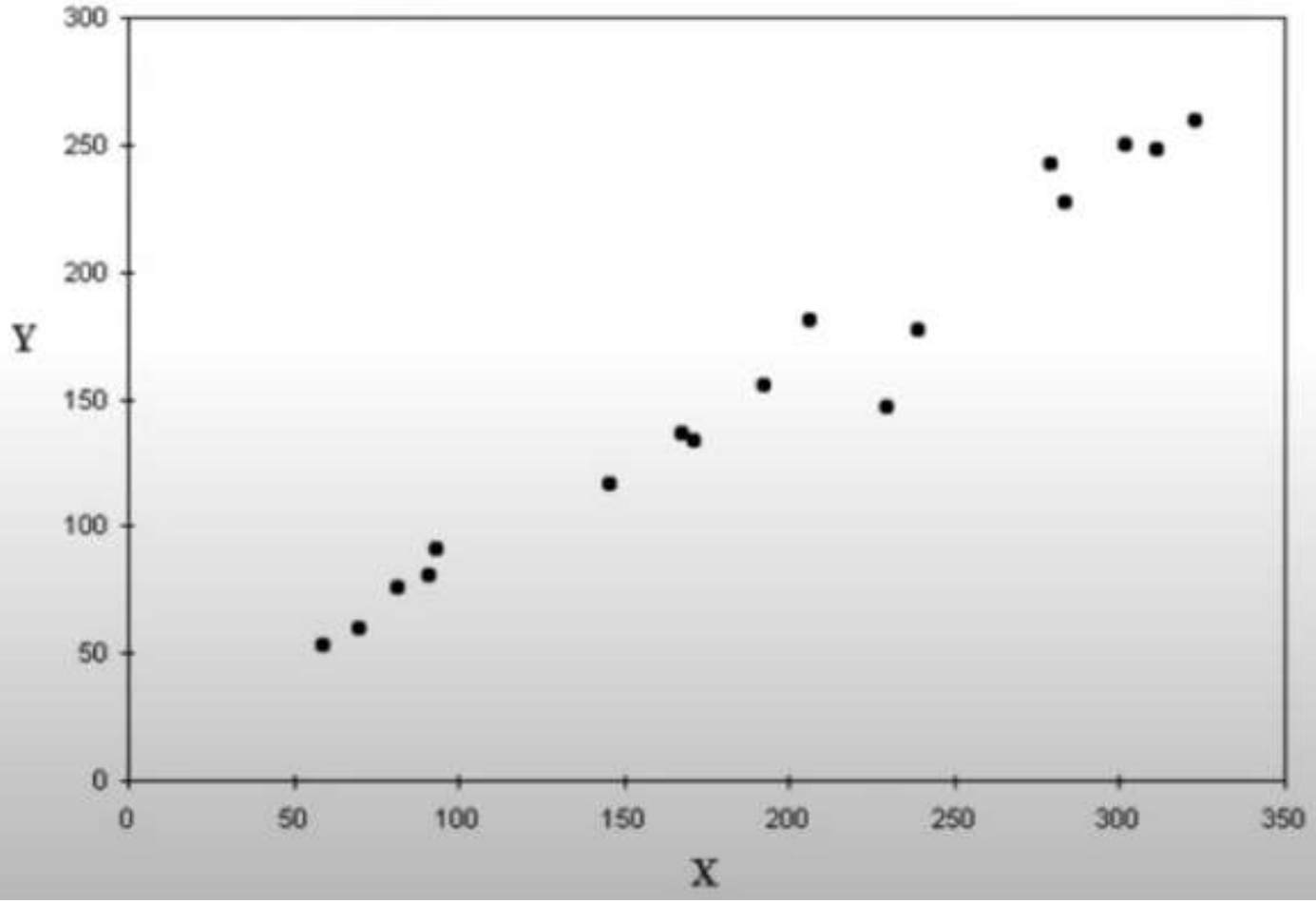
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = 0$$

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} > 0$$

Linear Regression Models -

g. Given Y = f(x) and the scatter plot shows apparent correlation between X and Y Let's fit a line into the scatter which shall be our model

h. But there are infinite number of lines that can be fit in the scatter. Which one should we consider as the model?



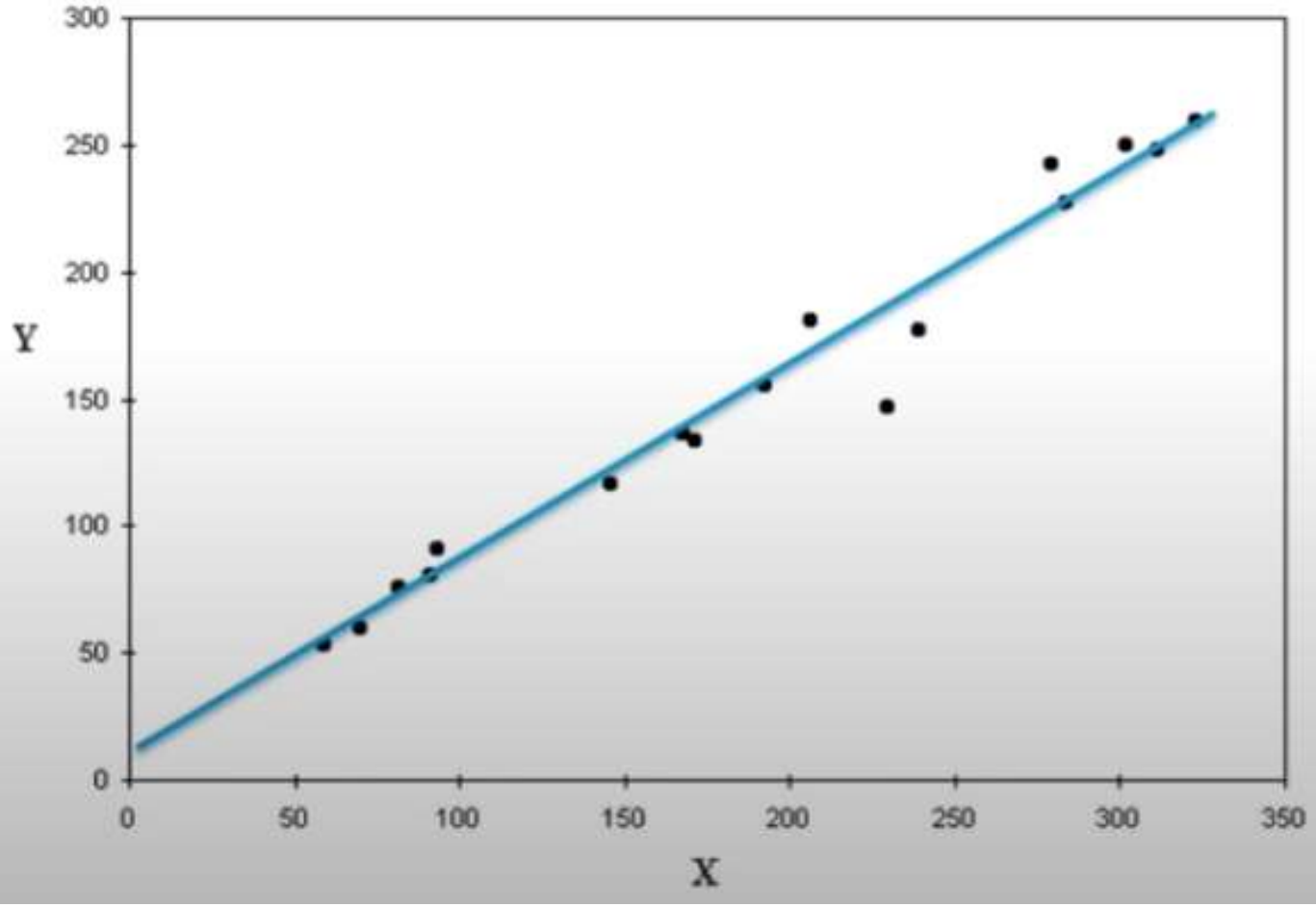
i. This and many other algorithms use gradient descent or variants of gradient descent metho for finding the best model

j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

Linear Regression Models -

g. Given Y = f(x) and the scatter plot shows apparent correlation between X and Y Let's fit a line into the scatter which shall be our model

h. But there are infinite number of lines that can be fit in the scatter. Which one should we consider as the model?



i. This and many other algorithms use gradient descent or variants of gradient descent method for finding the best model

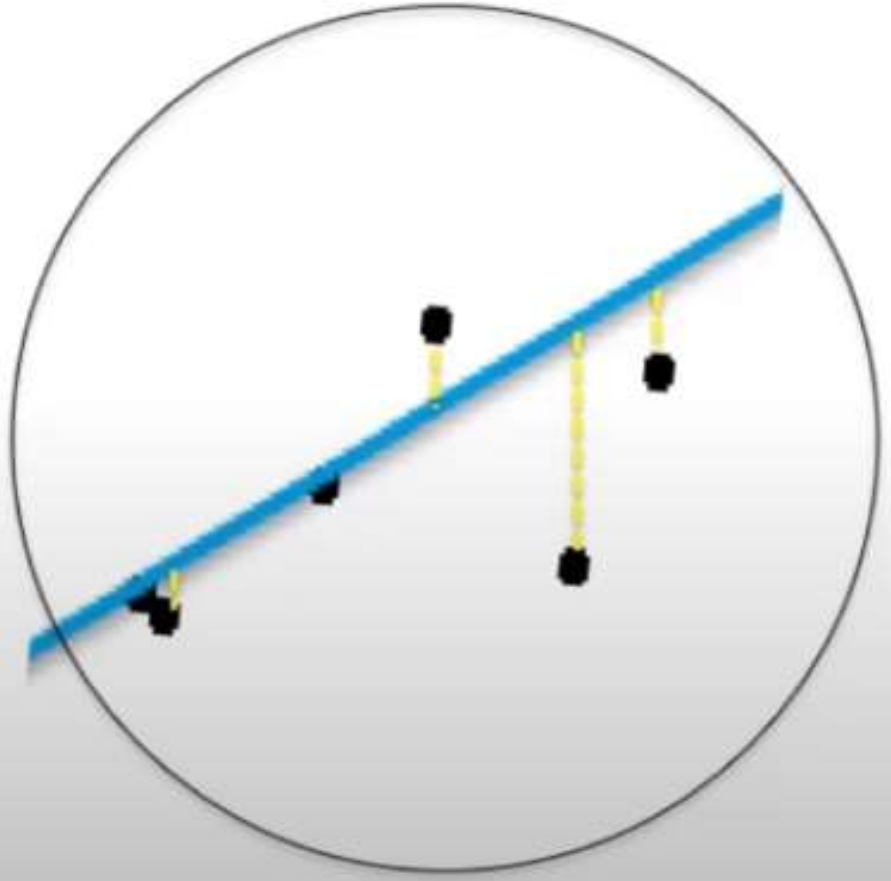
j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

Linear Regression Models (Recap) -

k. Whichever line we consider as the model, it will not pass through all the points.

l. The distance between a point and the line (drop a line vertically (shown in yellow)) is the error in prediction

m. That line which gives least sum of squared errors is considered as the best line



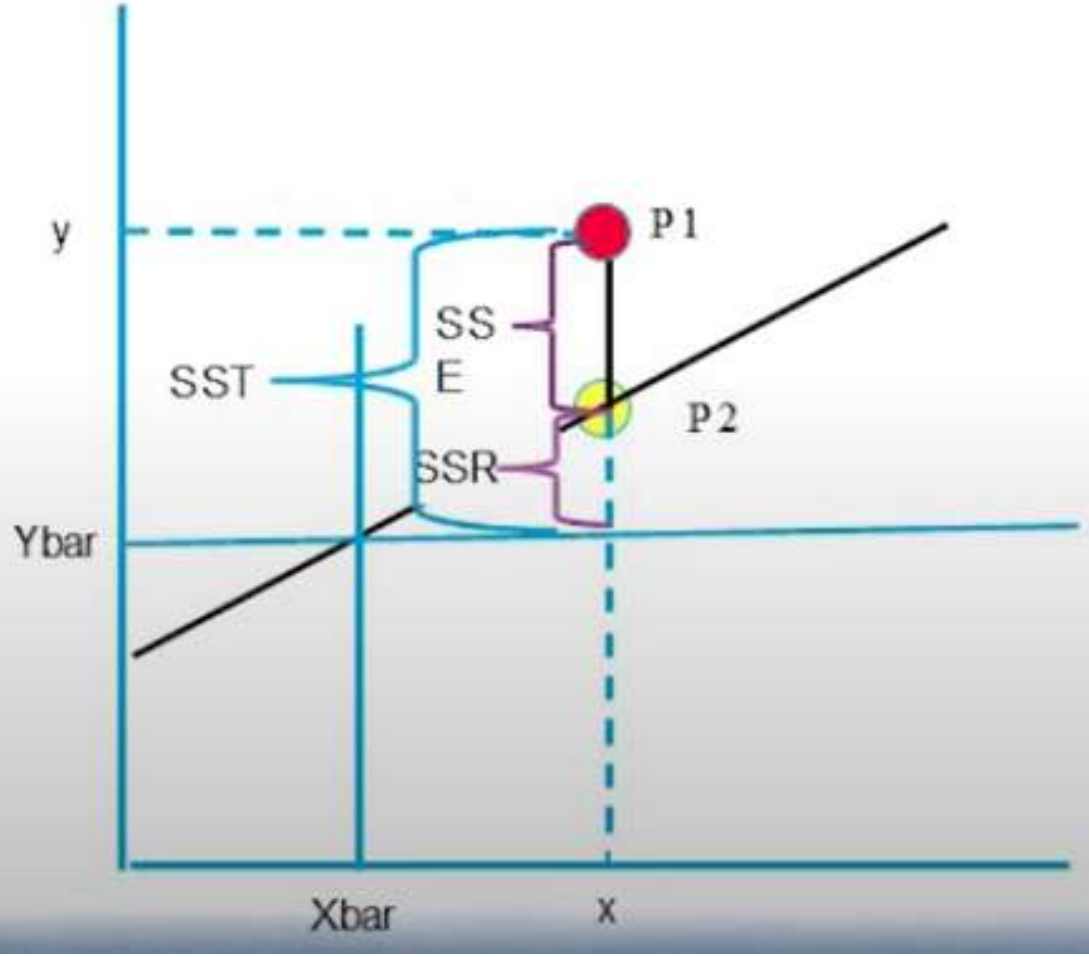
Linear Regression Models -

o. Coefficient of determinant (Contd…)

I. There are a variety of errors for all those points that don't fall exactly on the line.

II. It is important to understand these errors to judge the goodness of fit of the model i.e. How representative the model is likely to be in general

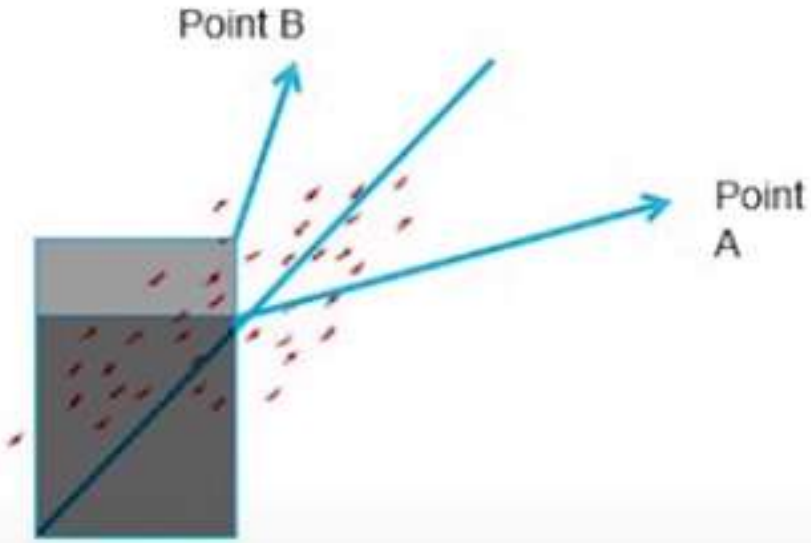
III. Let us look at point P1 which is one of the given data points and associated errors due to the model

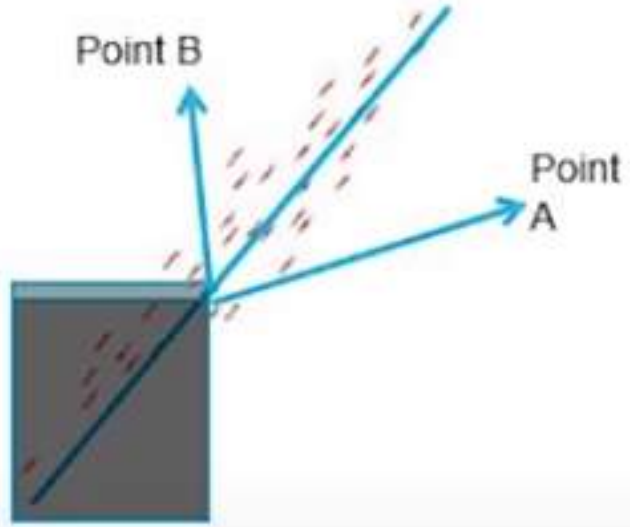


1. P1 – Original y data point for given x
2. P2 - Estimated y value for given x
3. Ybar – Average of all Y values in data set
4. SST – Sum of Square error Total (SST) Variance of P1 from Ybar (Y – Ybar)^2
5. SSR - Regression error (p2 – ybar)^2 (portion SST captured by regression model)
6. SSE - Residual error (p1 – p2)^2

Linear Regression Models (Recap) -

q. Coefficient of determinant (Contd…) -





In case of point "A", the line explains the variance of the point

Whereas point "B" the is a small area (light grey) which the line does not represent.

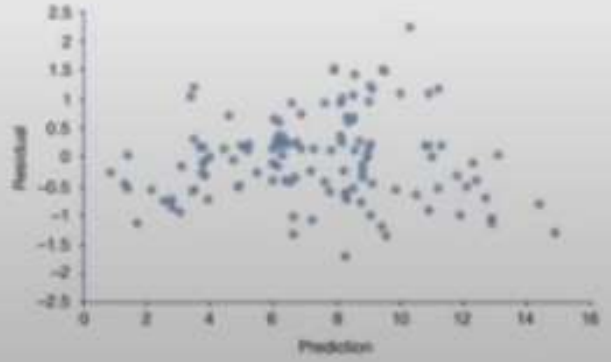
%age of total variance that is represented by the line is coeff of determinant

Linear Regression Assumptions

Linear regression model is based on a set of assumptions. If the underlying dataset does not meet these assumptions, then data may have to be transformed or linear model may not be good fit

1. Assumption of linearity. assumes a linear relation between the dependent / target variable and the independent / predictor variables.

2. Assumption of normality of the error distribution.
   a. The errors should be normally distributed across the model.
   b. This assumption can be tested using a frequency histogram, skew and kurtosis of a normal plot. If the distribution does not approximate normal distribution, data transformation may be necessary
   c. A scatter plot between the actual values and the predicted values should show the data distributed equally across the model.
   d. Another way of doing this is to plot residual values against the predicted values. We should not see any trends



Linear Regression Model -

Advantages –

1. Simple to implement and easier to interpret the outputs coefficients

Disadvantages -

1. Assumes a linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them
2. Outliers can have huge effects on the regression
3. Linear regression assume independence between attributes
4. Linear regression looks at a relationship between the mean of the dependent variable and the independent variables.
5. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables
6. Boundaries are linear