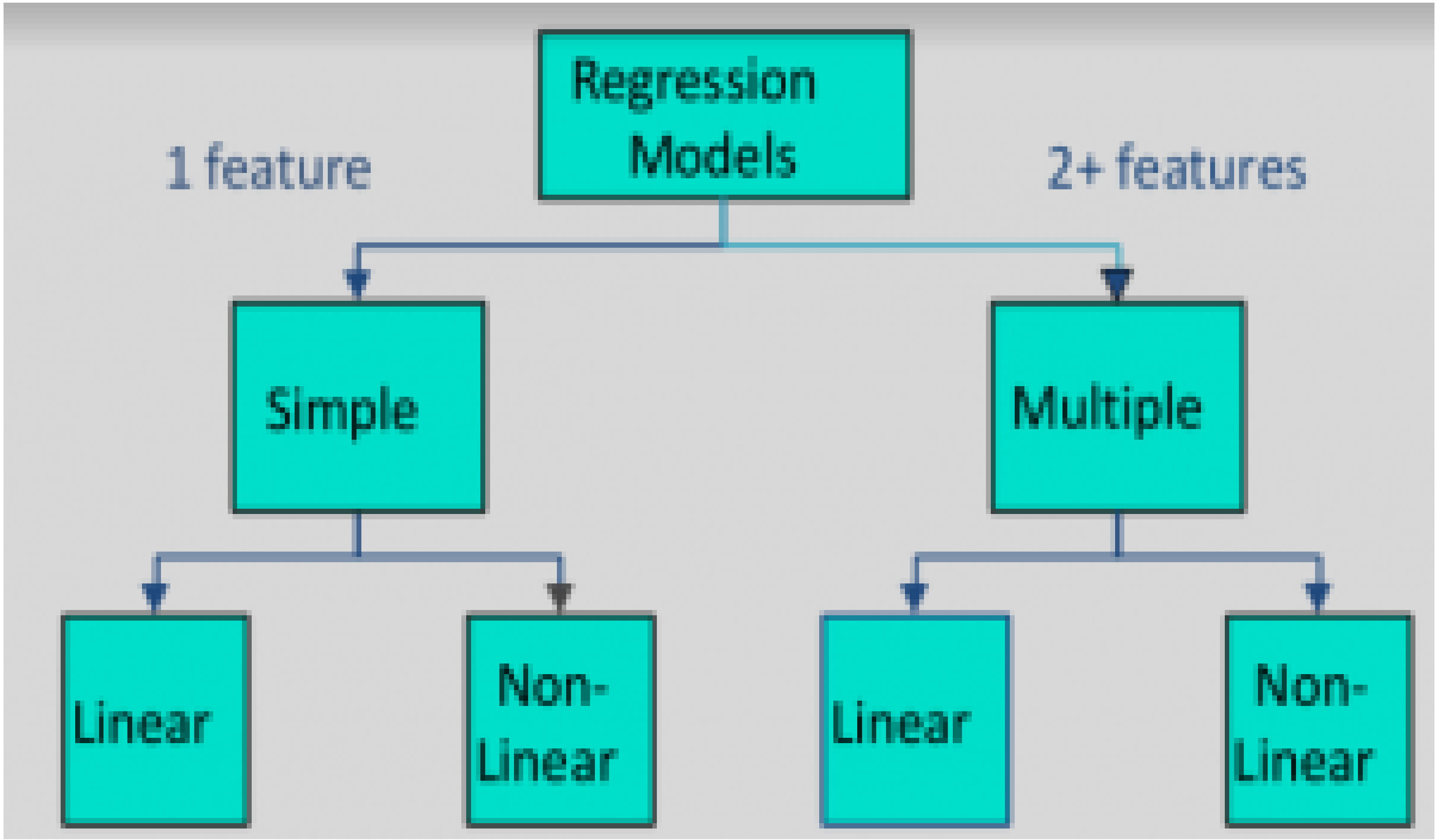
Regression


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

Examples

- Dependent variable is employment income – independent variables might be hours of work, education, occupation, sex, age, region, years of experience, unionization status, etc.

- Price of a product and quantity produced or sold:
  - Quantity sold affected by price. Dependent variable is quantity of product sold – independent variable is price.
  - Price affected by quantity offered for sale. Dependent variable is price – independent variable is quantity sold.

Univariate and multivariate models

Simple regression model

(Education) $x \longrightarrow y$ (Income)

Multivariate or multiple regression model

(Education) $x_1 \rightarrow y$

(Sex) $x_2 \rightarrow y$ (Income)

(Experience) $x_3 \rightarrow y$

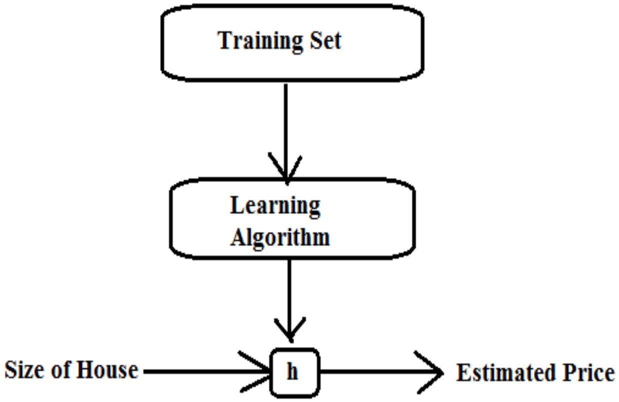
(Age) $x_4 \rightarrow y$

Simple Regression

- Is example of Supervised Learning
- One input variable (feature) ,also called univariate regression
- Univariate linear regression is used when you want to predict a **single output** value y from a **single input** value x
- Input variable/feature values: x_i *(Height)*
- Output variable/target variable values : y_i (Weight)
- A pair (x_i, y_i) is $\boldsymbol{ith}$ training example
- Observation pairs: (x_i, y_i); $i = 1,2, \ldots, m$
- A list of training examples $\{(x_i, y_i) |\ i = 1,2, \ldots, m\}$ is called **training set**
- X = space of input values and Y = space of output values , (X= Y = $\mathcal{R}$)
- $m =$ number of observations (number of training examples)

Simple Regression Statement

- Given correct values of output variable y_i for the training data set having m values of input variable $x_i \; \forall \; i = 1,2, \ldots , m$
- Regression problem is to predict real valued output for a given value of feature/input
- Given a training set, to learn a function h : X → Y so that h(x) is a "good" predictor for the corresponding value of y. For historical reasons, this function h is called a hypothesis



Training Set

Sr No	Height (cms)	Weight (Kg)
1	135	57
2	165	70
3	155	63
4	160	65
5	150	62….

- **Quiz**: Suppose there are **35** pair of (height, weight) = (x_i, y_i)
- $m =?$
- What is x_2?
- What is target variable?
- What is input variable?
- What is feature?
- What is output variable?
- What is y_3?
- What is (x_5, y_5)?
- What is training example for $i = 4$?
- What is training set?
- Why is it supervised learning?
- Does target variable take continuous or discrete values?
- Describe hypothesis h function

Hypothesis function Linear Regression

Sr No	Height (cms) = x	Weight (Kg)= y
1	135	57
2	165	70
3	155	63
4	160	65
5	150	62….

- $y = h(x) = a_0 + a_1 x$
- $a_0, a_1 \, are \; parameters$
- How to choose the parameters? (Many concepts)
- Just say $h(x) = 2 + 0.5\,x$
- Weight = 0.5 * height + 2

Quiz: Predict weight for height as 162 cm

Predict weight for height as 160 cm

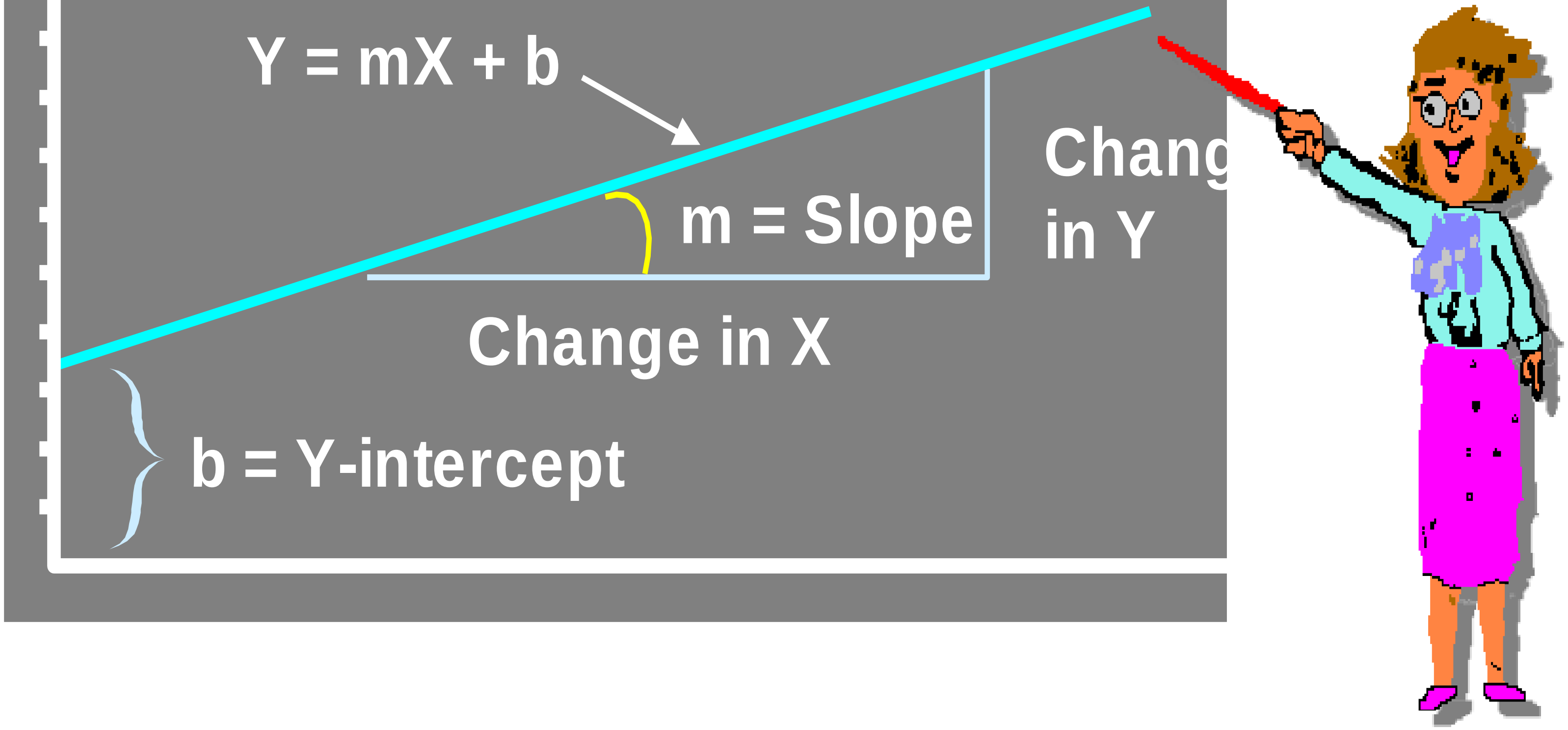
Same for height 165 cm , 150 cm?

What are the errors in above two predictions?

Draw scatter diagram, and h(x)

Linear Regression

Linear Equations



Univariate or simple linear regression

- *x* is the independent variable
- *y* is the dependent variable
- The regression model is

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

- The model has two variables, the independent or explanatory variable, *x*, and the dependent variable *y*, the variable whose variation is to be explained.
- The relationship between *x* and *y* is a linear or straight line relationship.
- Two parameters to estimate – the slope of the line β_1 and the *y*-intercept β_0 (where the line crosses the vertical axis).
- ε is the unexplained, random, or error component. Much more on this later.

Linear Regression Model

Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

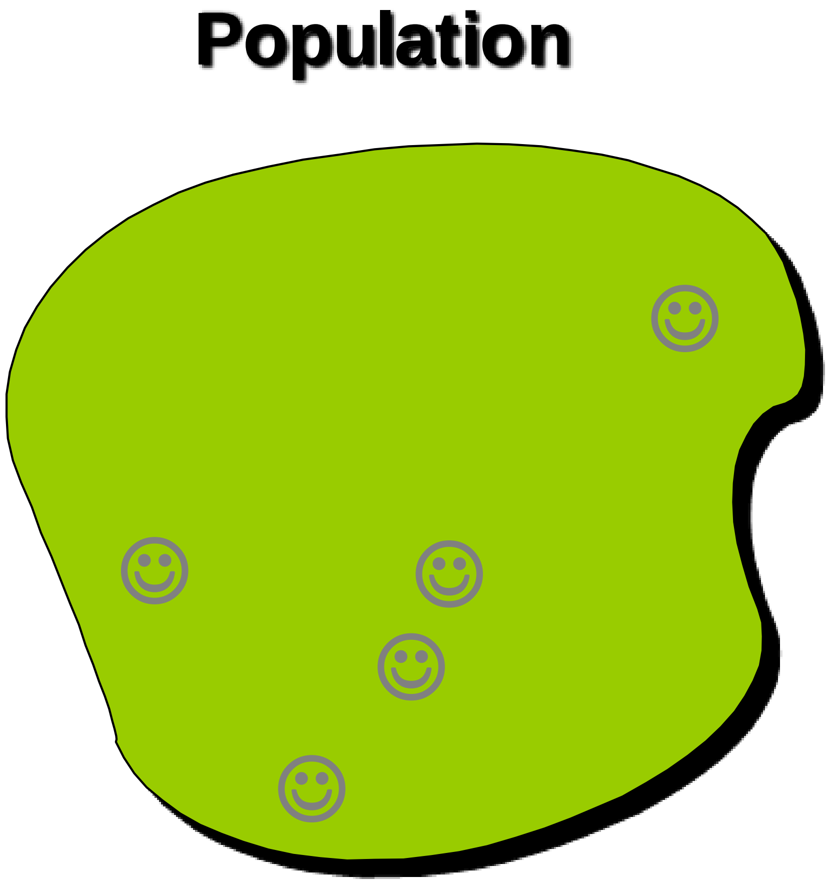
Random Error

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

Dependent (Target/Output) Variable
(e.g.Weight)

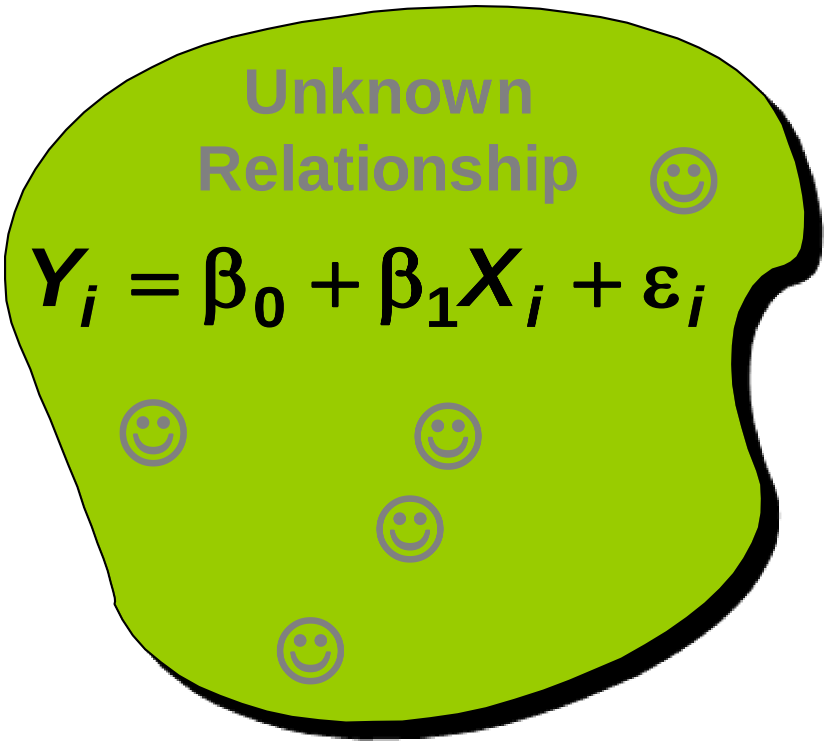
Independent (Explanatory/Input/Feature) Variable
(e.g. Height)

Population & Sample Regression Models



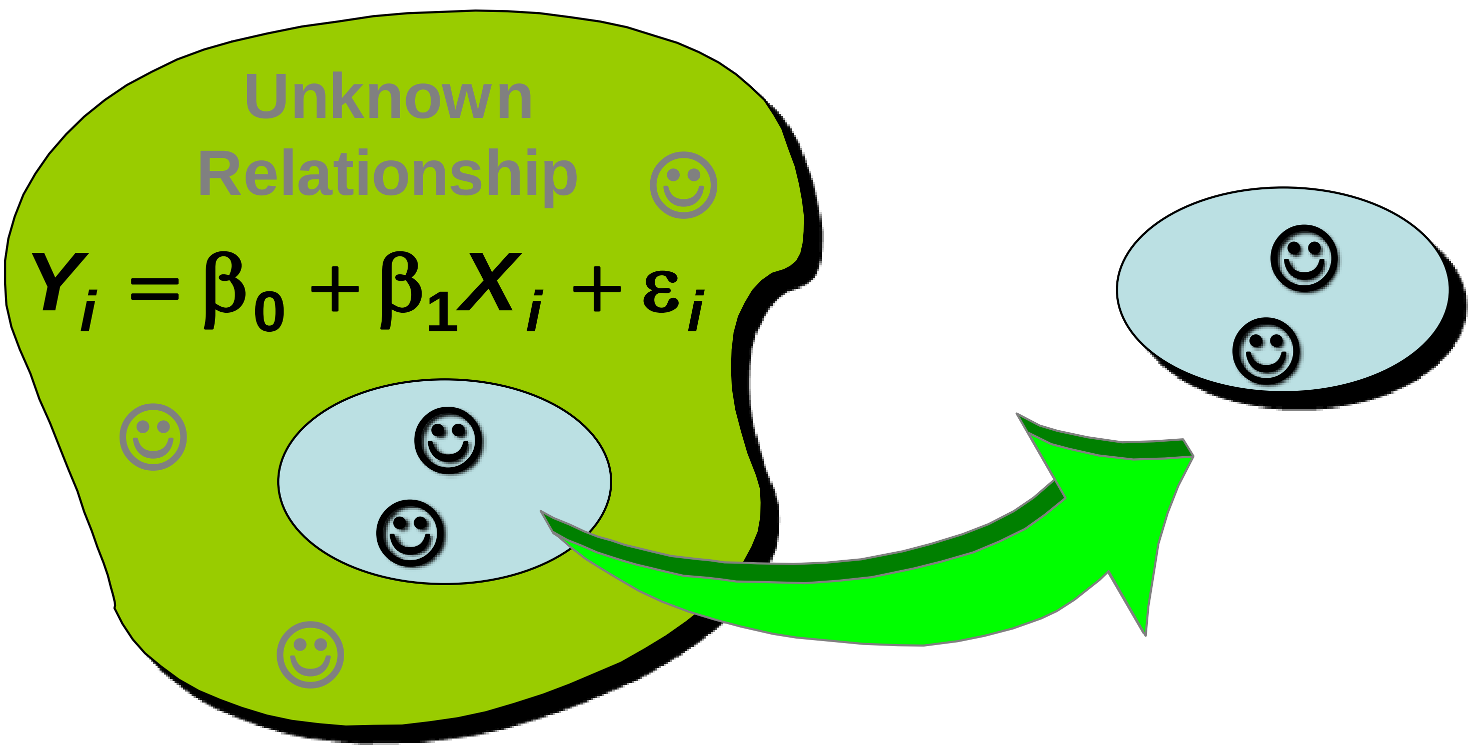
Population & Sample Regression Models

Population

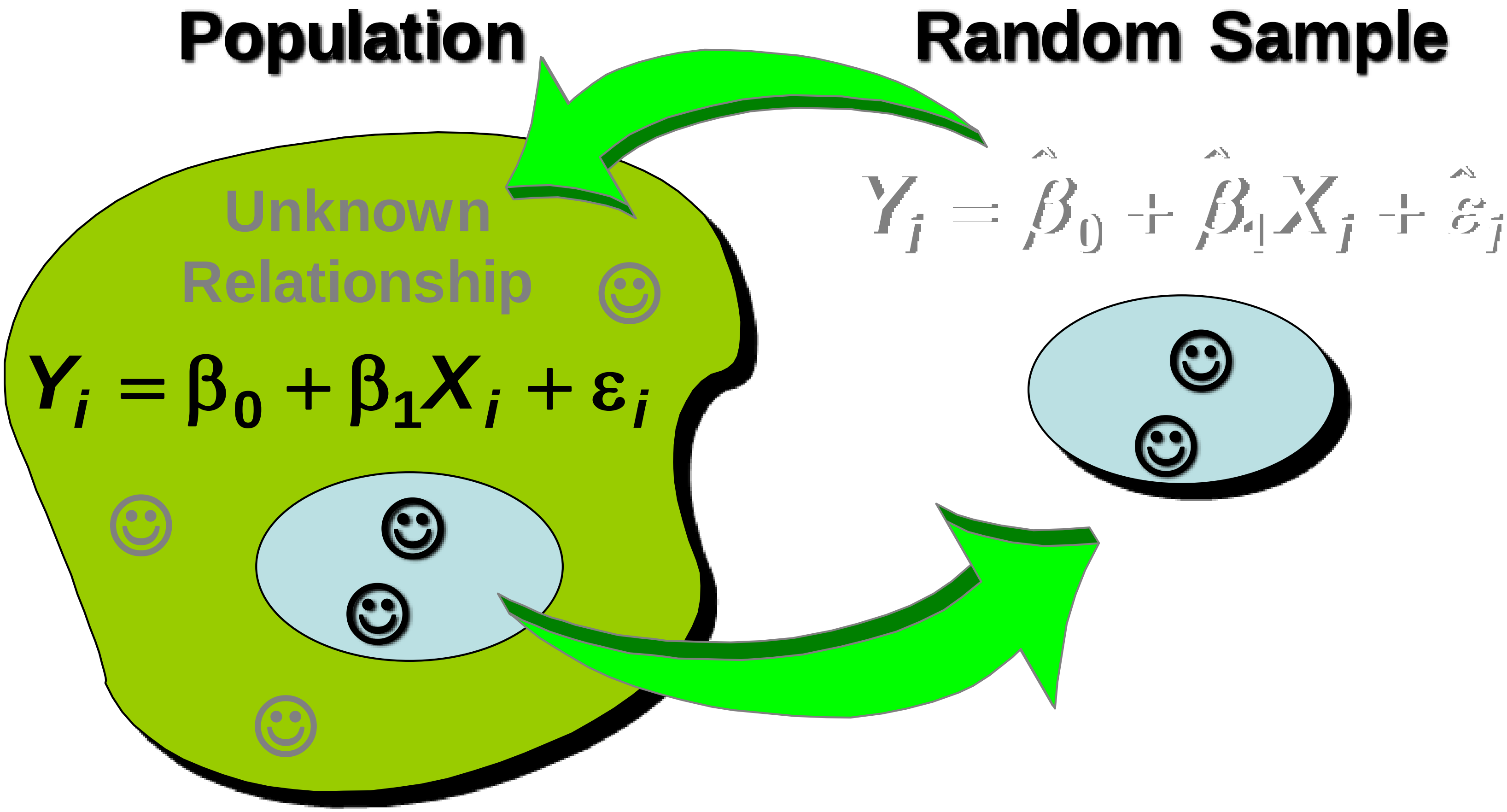


Population & Sample Regression Models

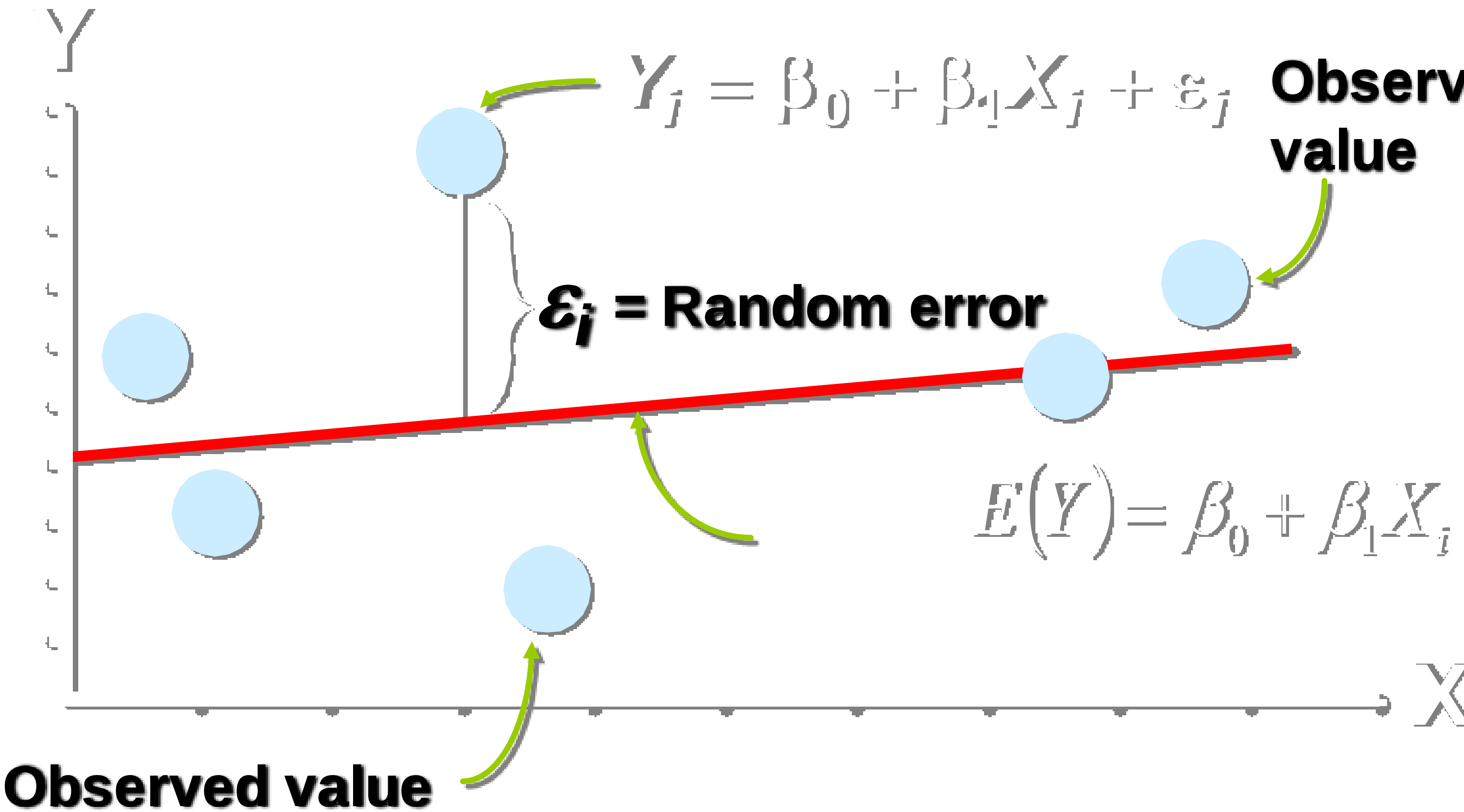
Population **Random Sample**



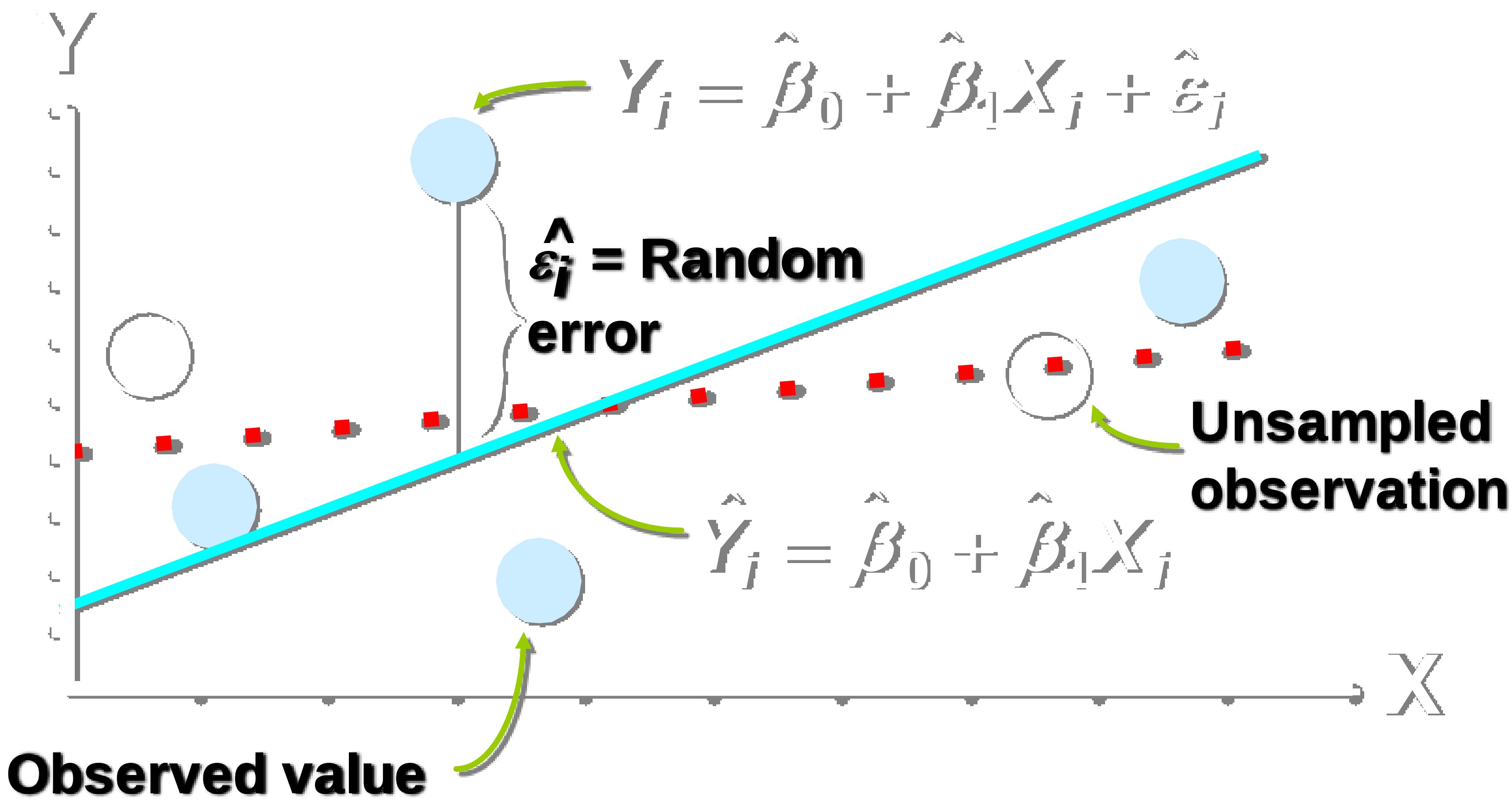
Population & Sample Regression Models



Population Linear Regression Model



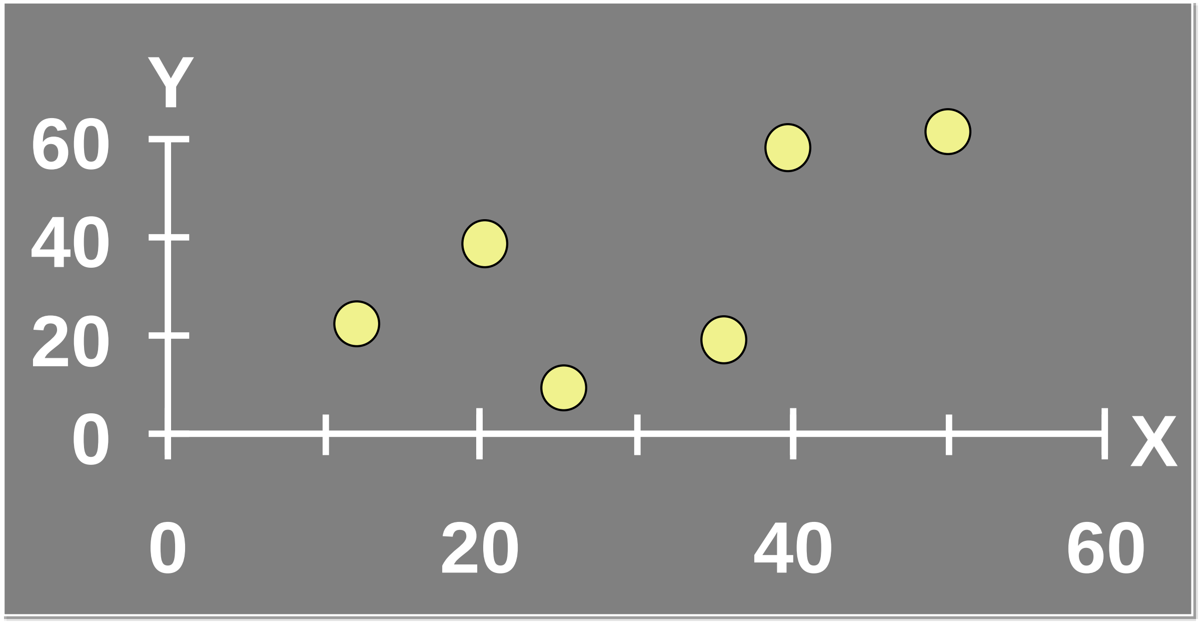
Sample Linear Regression Model



Estimating Parameters: Least Squares Method

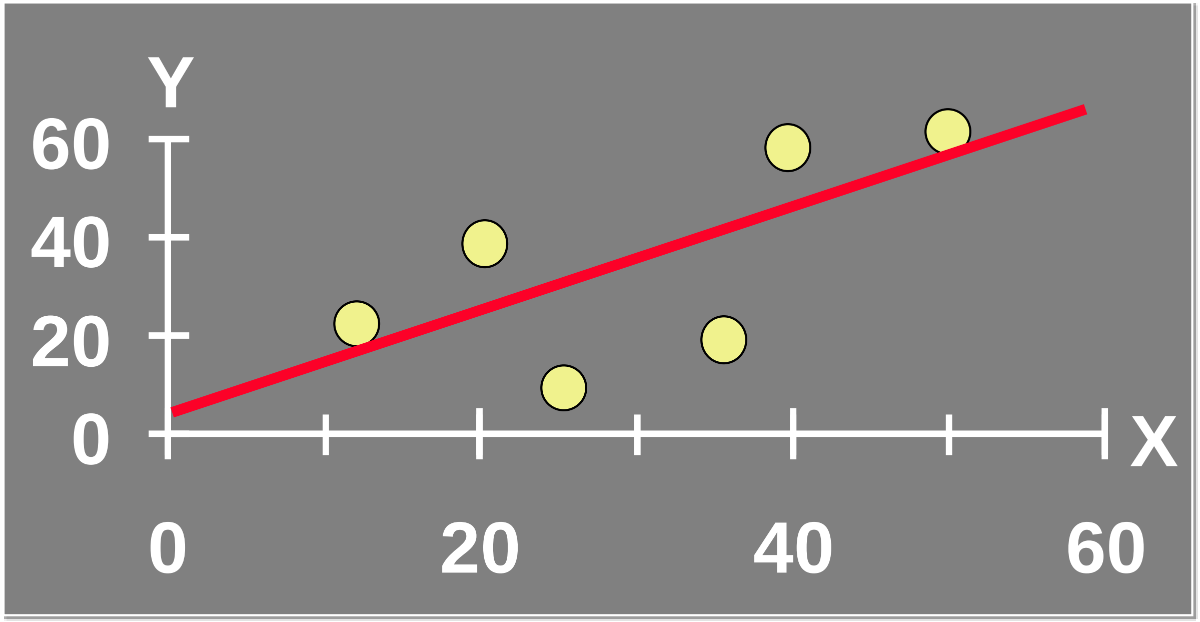
Scatter plot

- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit



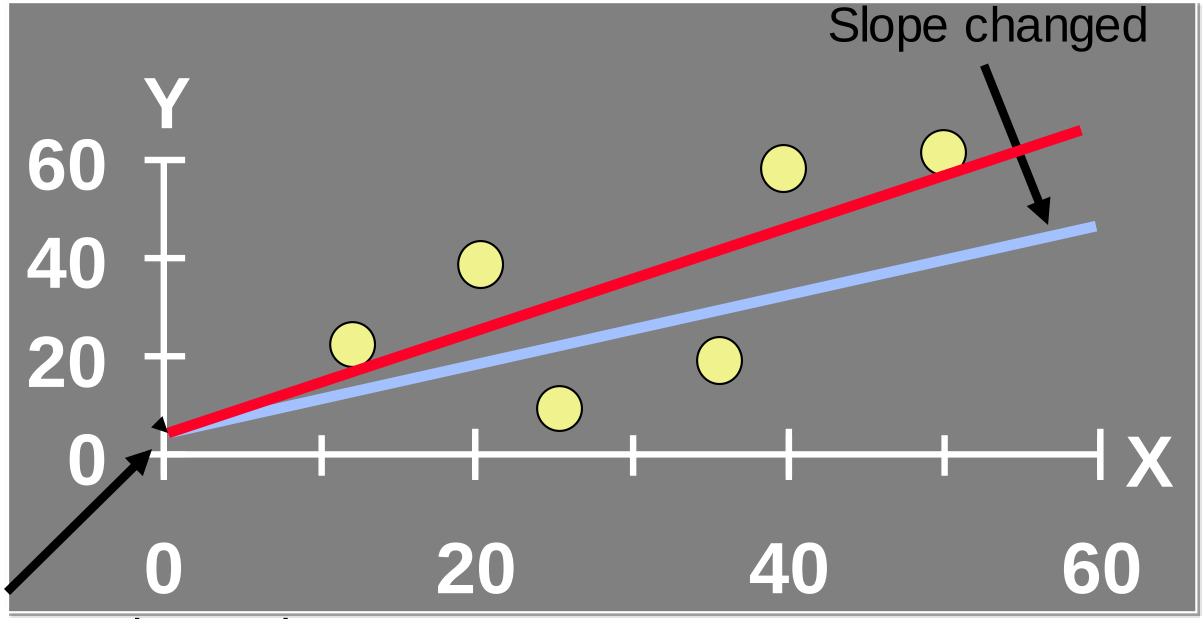
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Thinking Challenge

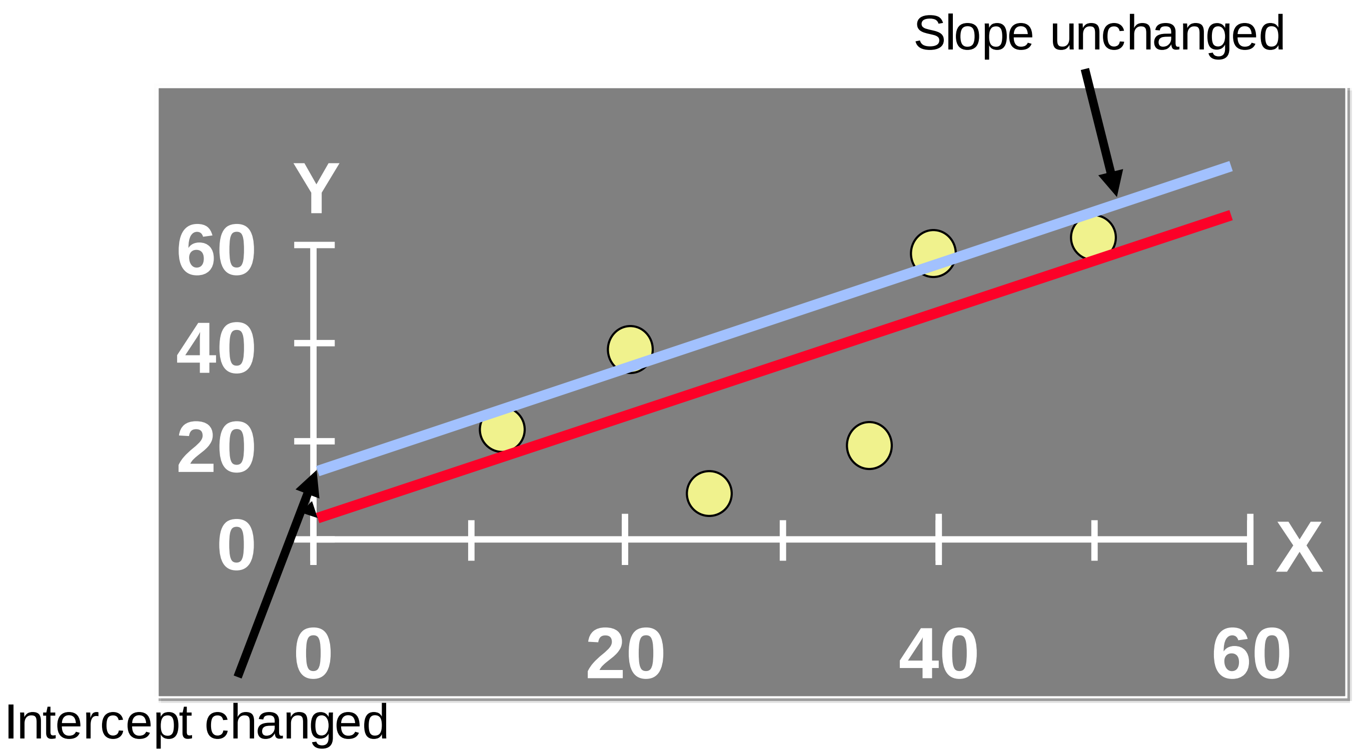
How would you draw a line through the points? How do you determine which line 'fits best'?



Intercept unchanged

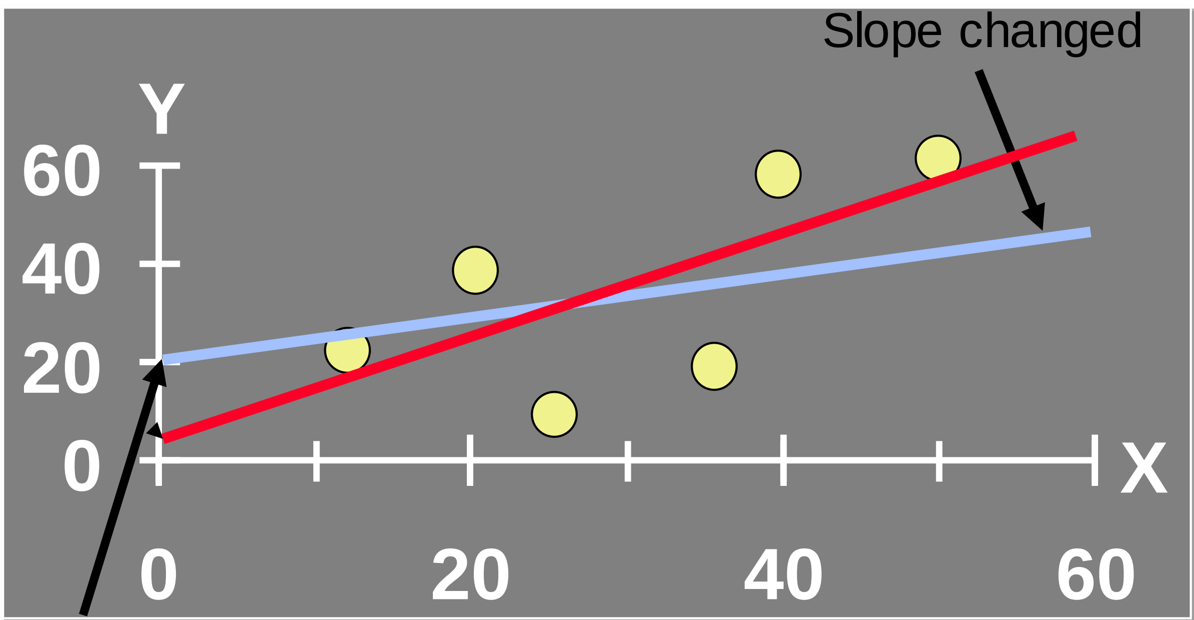
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Intercept changed

Regression line

- The regression model is $y = \beta_0 + \beta_1 x + \varepsilon$
- Data about *x* and *y* are obtained from a sample.
- From the sample of values of *x* and *y*, estimates b_0 of β_0 and b_1 of β_1 are obtained using the least squares or another method.
- The resulting estimate of the model is

$$\hat{y} = b_0 + b_1 x$$

- The symbol $\hat{y}$ is termed "*y* hat" and refers to the predicted values of the dependent variable *y* that are associated with values of *x*, given the linear model.

Least Squares

- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative ones

Least Squares

- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Least Squares

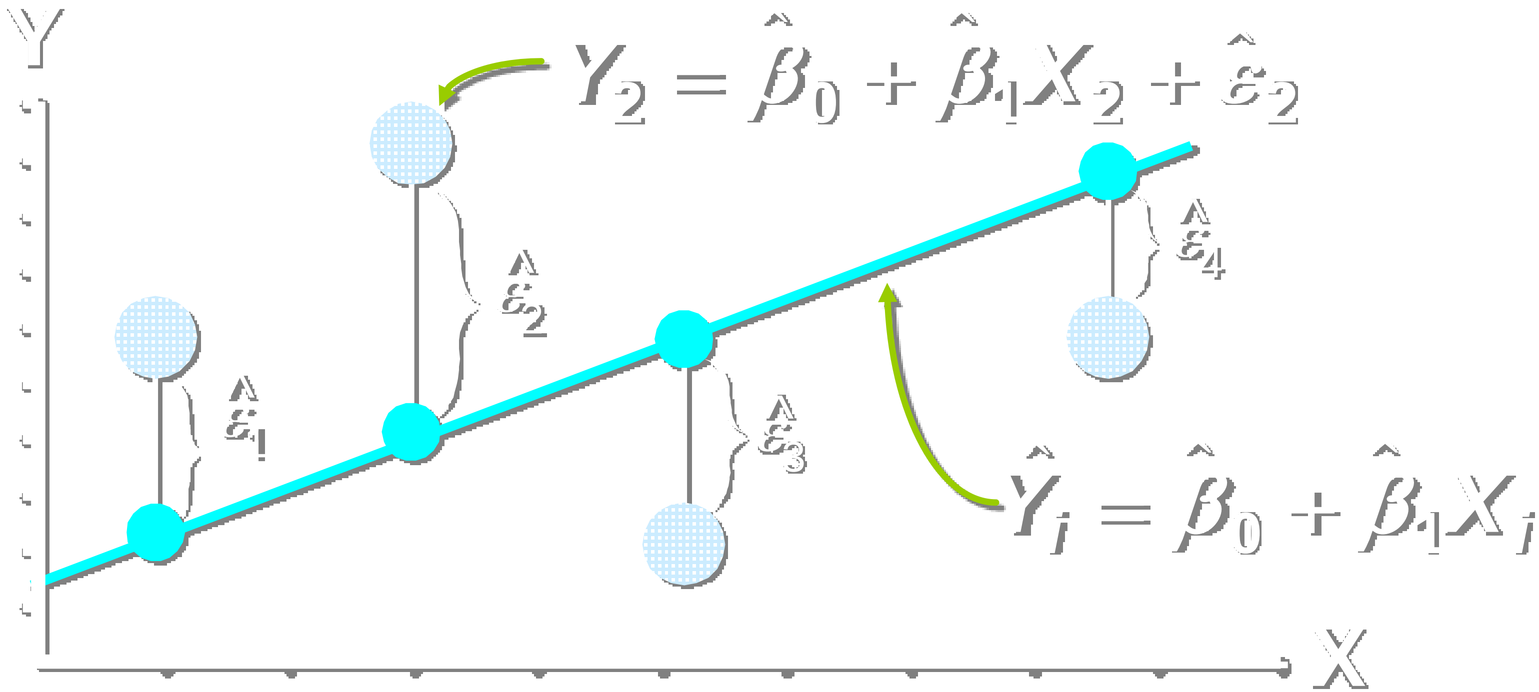
- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Objective Function/Cost/Loss Function

- Our objective is to get the best possible line. **The best possible line will be such so that the total/average squared vertical distances of the scattered points from the line will be the least**.
- Ideally, the line should pass through all the points of our training data set. (Zero Error)
- Consider that the data points (training set) $\{(x_i, y_i); i = 1,2, \ldots, m\}$; are given. We would like to fit a straight line $h(x) = a_0 + a_1 x$ to this data.
- For that we have to find values of $a_0 \; and \; a_1$ which minimize the total square error:

Cost Function/Sum of square of errors

Hence, the sum of the squares of the errors,

$$S = \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)]^2.$$

For S to be minimum, we have

$$\frac{\partial S}{\partial a_0} = -2 \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)]$$

and

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{m} x_i [y_i - (a_0 + a_1 x_i)].$$

Normal Equations

The above equations are simplified to

$$ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$$

and

$$ma_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i.$$

Since the x_i and y_i are known quantities, the above two equations (called the **normal equations**), can be solved for the two unknown a_0 and a_1. Differentiating $\frac{\partial S}{\partial a_0}$ and $\frac{\partial S}{\partial a_1}$ with respect to a_0 to a_1 respectively, we find

$$\frac{\partial^2 S}{\partial a_0^2} \quad \text{and} \quad \frac{\partial^2 S}{\partial a_1^2}$$

and both will be positive at the points. Hence these values provide a minimum of S.

Computation Table

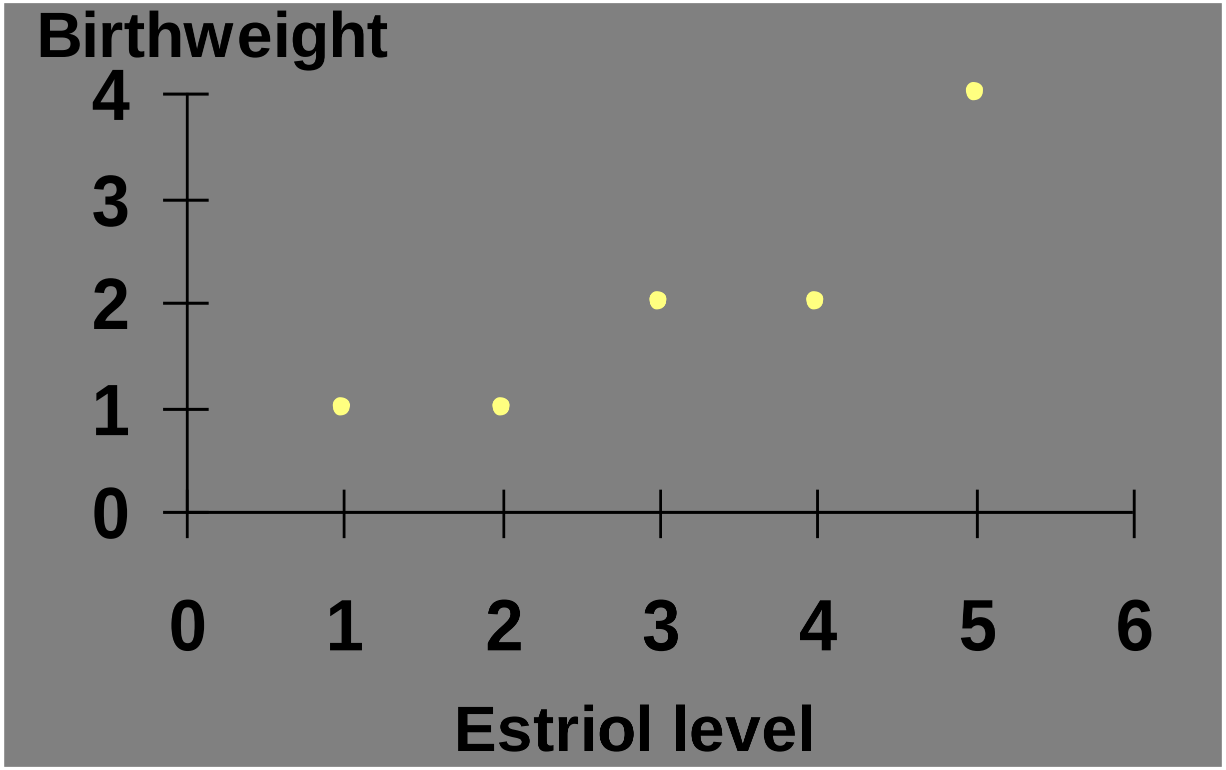
X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Parameter Estimation Example

- Obstetrics: What is the **relationship** between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	**Birthweight** (g/1000)
1	1
2	1
3	2
4	2
5	4

Scatterplot
Birthweight vs. Estriol level



Parameter Estimation Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	**10**	**55**	**26**	**37**

Sr	Height	Weight	xy	
iii	(cms) = x	(Kg)= y		
1	135	57	7695	18225
2	165	70	11550	27225
3	155	63	9765	24025
4	160	65	10400	25600
5	150	62	9300	22500
m= 5	$\sum x_i$ 765	317	48710	117575