Simple Linear Regression and Correlation

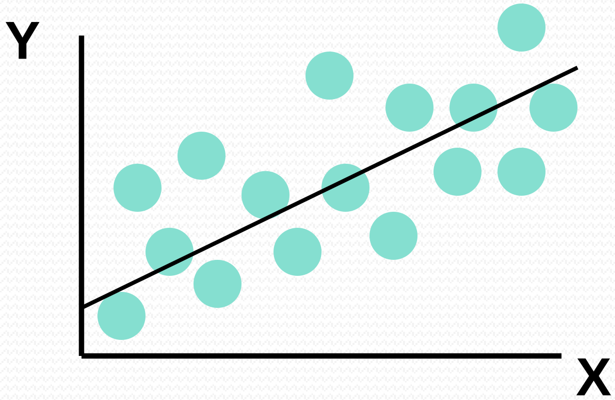
Why?

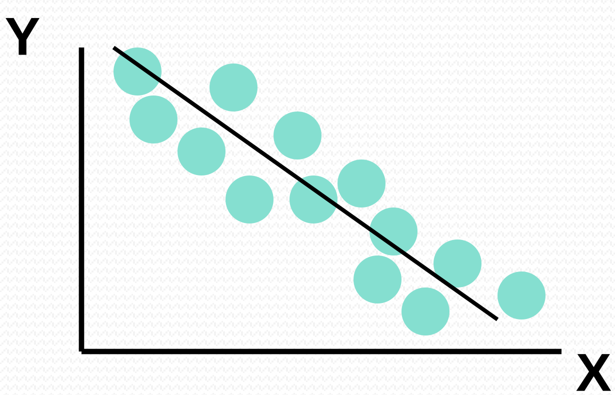
- Regression Analysis is Used Primarily to Model Causality and Provide Prediction
  - Predict the values of a dependent (response) variable based on values of at least one independent (explanatory) variable
  - Explain the effect of the independent variables on the dependent variable

Scatter Plot

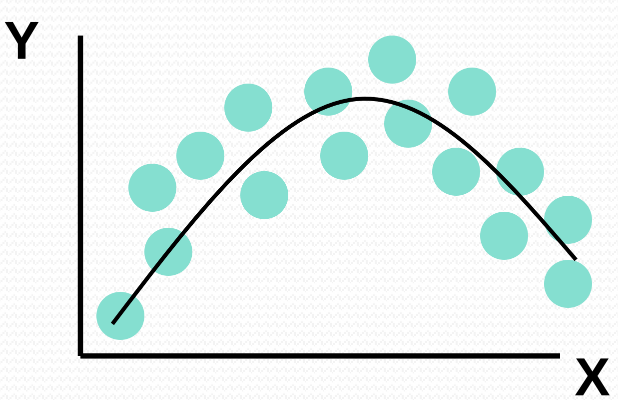
Types of Relationships

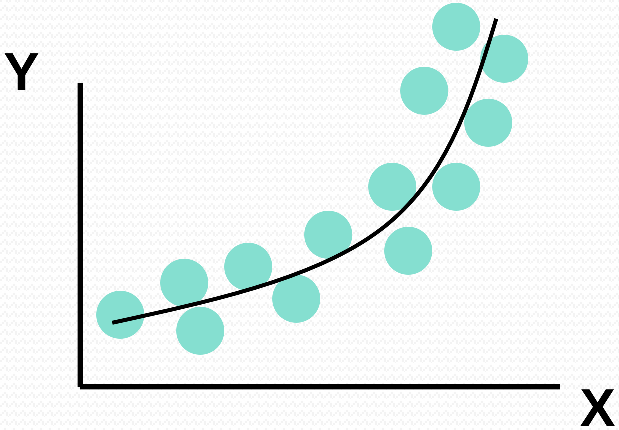
Linear relationships





Curvilinear relationships

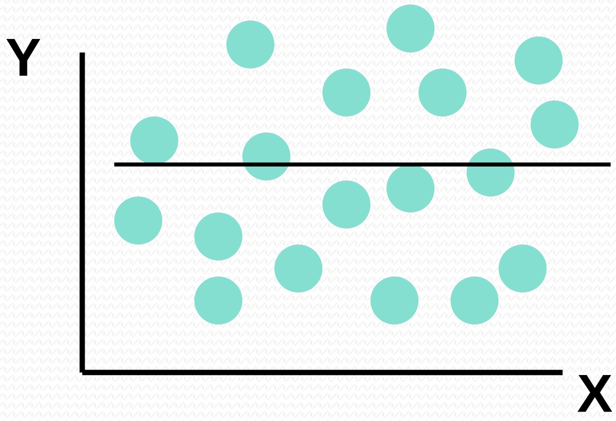


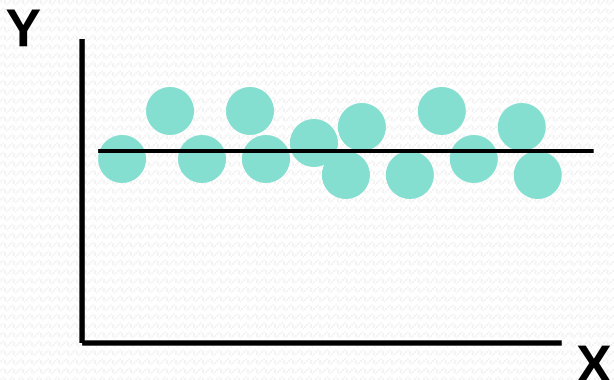


Types of Relationships

(continued)

No relationship

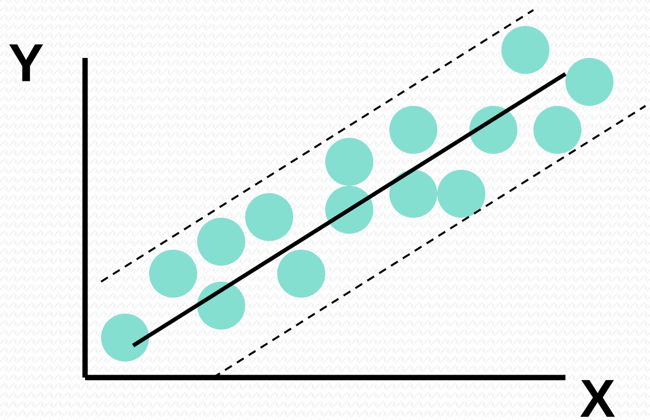


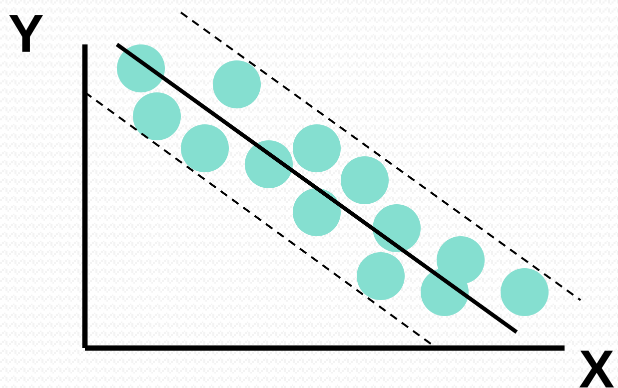


Types of Relationships

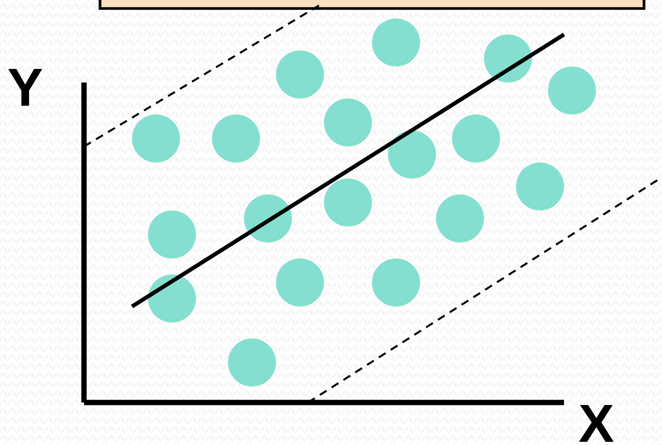
(continued)

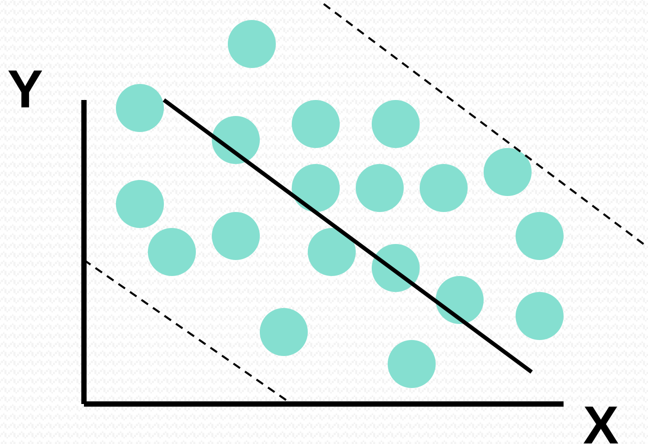
Strong relationships





Weak relationships





Simple Linear Regression Model

Population
Y Intercept

Population
Slope
Coefficient

Random
Error

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\mu_{Y|X}} + \varepsilon_i$$

Dependent
(Response)
Variable

Population
Regression
Line
(Conditional Mean)

Independent
(Explanatory)
Variable

Meaning of Parameters

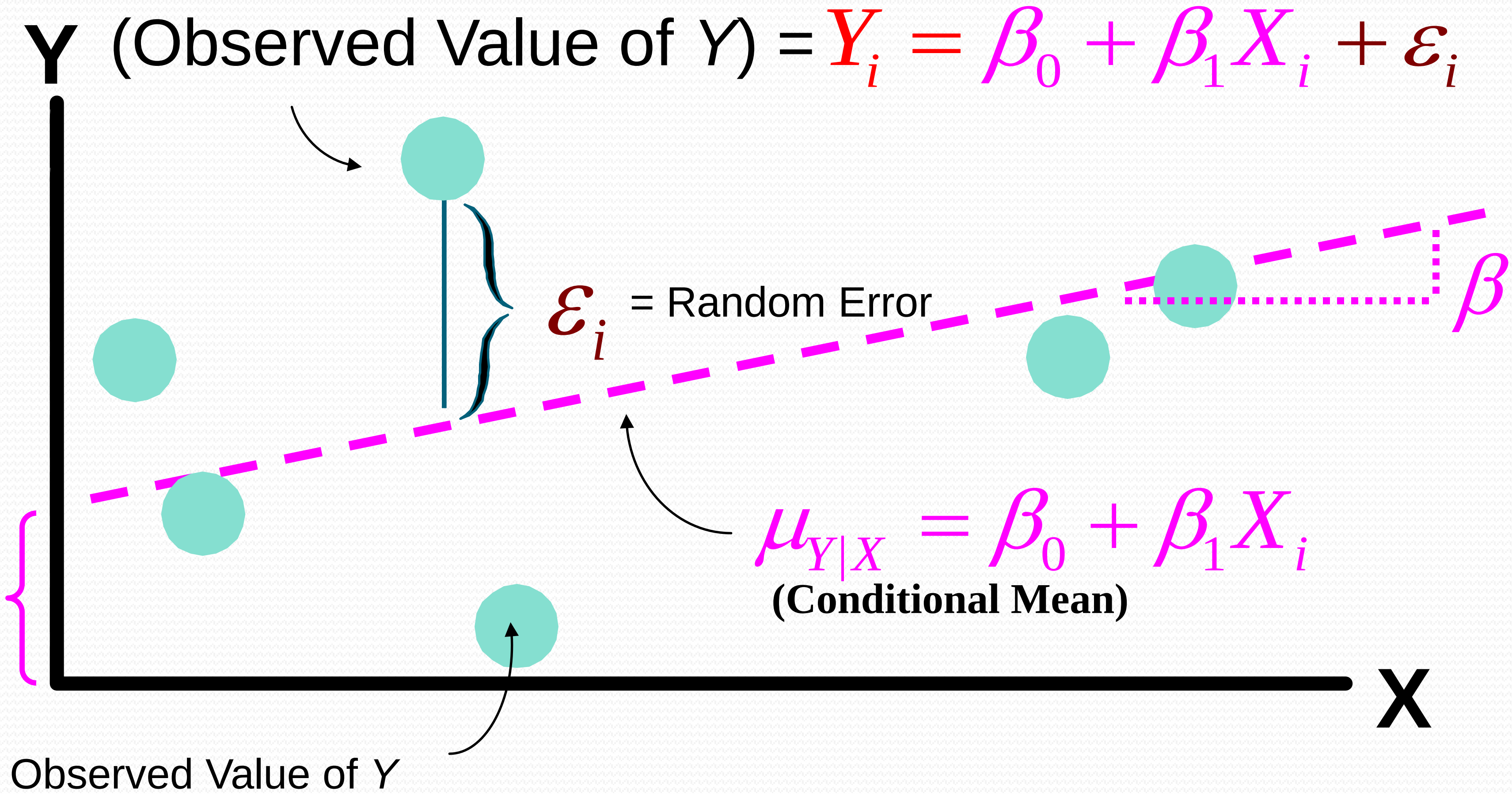
$$\beta_0 = \mu_{Y|X=0}$$

$$\beta_1 = \frac{\text{change in } \mu_{Y|X}}{\text{change in } X}, \text{ or}$$

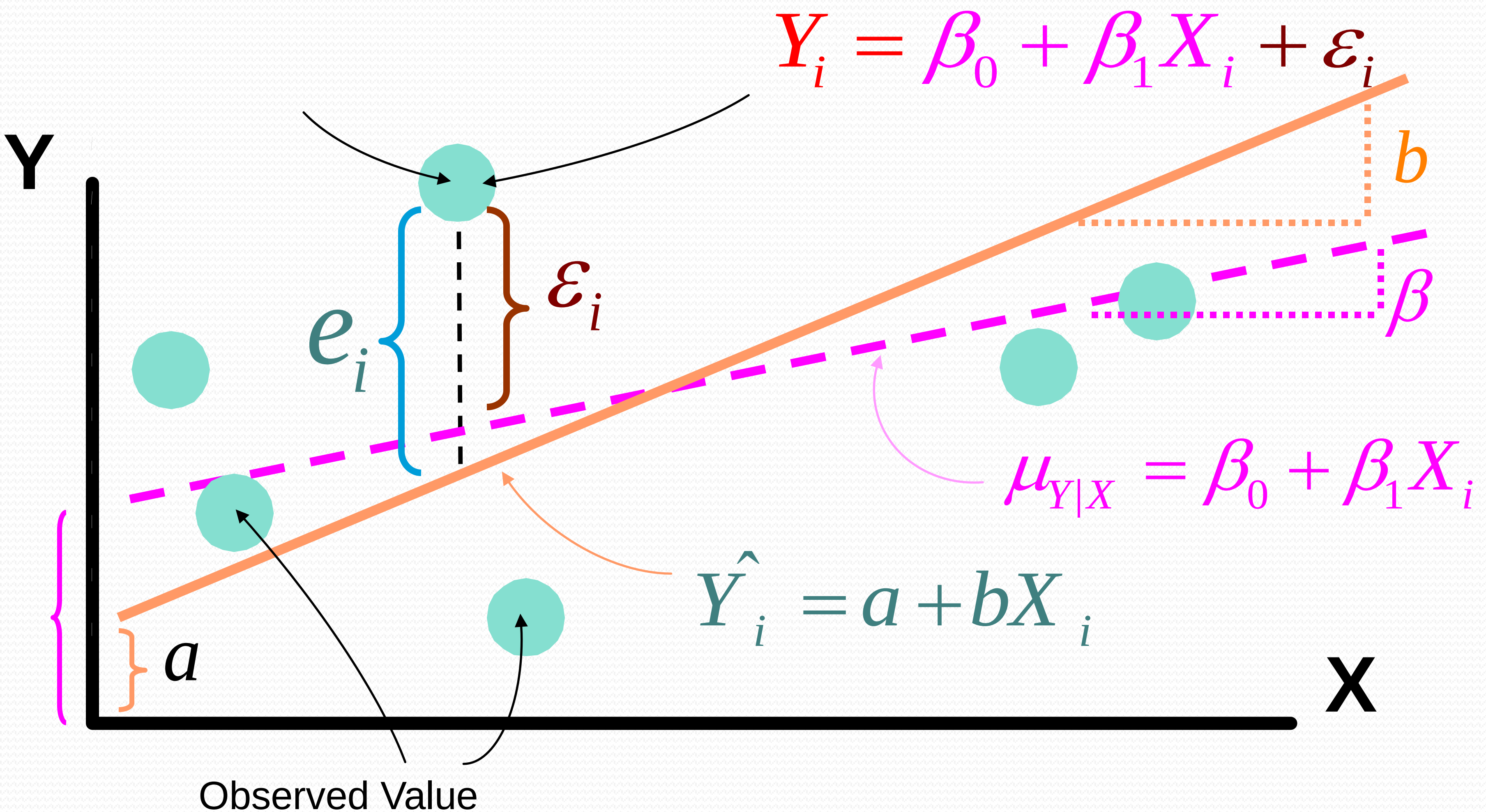
the change in the mean $\mu_{Y|X}$ for one unit change in X

Why the error term?

- Measurement Error
- Uncontrollable Factors
- Factors not included in the regression



Y (Observed Value of *Y*) = $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
ε_i = Random Error
β
$\mu_{Y|X} = \beta_0 + \beta_1 X_i$
(Conditional Mean)
X
Observed Value of *Y*



$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
X
b
β
e_i
ε_i
$\mu_{Y|X} = \beta_0 + \beta_1 X_i$
$\hat{Y}_i = a + bX_i$
a
Observed Value

- Estimated Intercept

Estimated Slope

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{} + e_i$$

Residual

$$\hat{Y} = \beta_0 + \beta_1 X =$$

Simple Regression Equation

(Fitted Regression Line, Predicted Value)

Parameter Estimation

- Least Squares Estimation

Minimize:

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

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

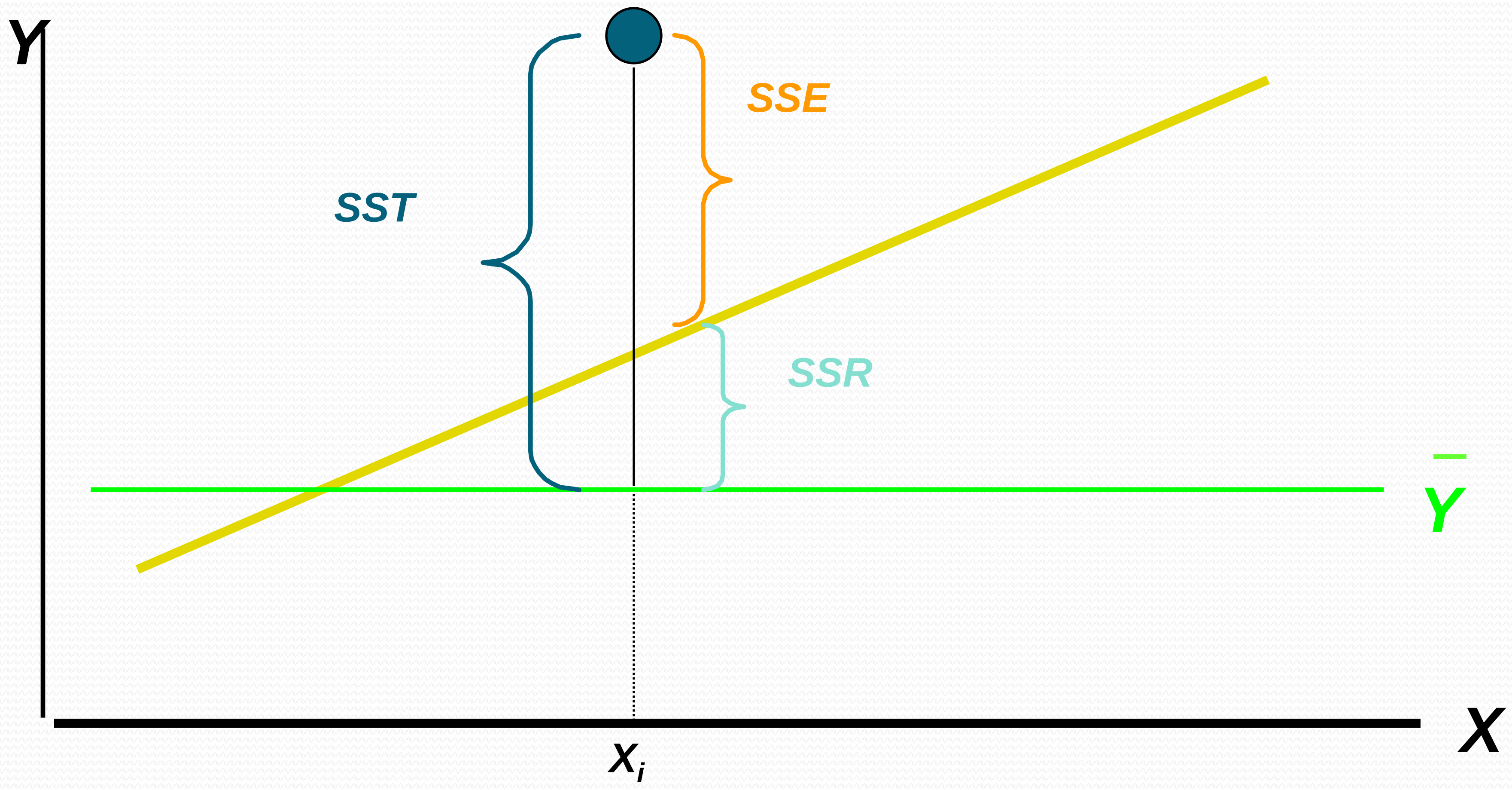
$$\beta_1 = \frac{S_{xy}}{S_{xx}}$$

$$S_{xy} = \sum (x_i - \bar{X})(y_i - \bar{Y}) = \sum x_i y_i - n\bar{X}\bar{Y}$$

$$S_{xx} = \sum (x_i - \bar{X})^2 = \sum x_i^2 - n\bar{X}^2$$

$$S_{yy} = \sum (y_i - \bar{Y})^2 = \sum y_i^2 - n\bar{Y}^2$$

Partition of the Sum of Squares



SST	=	***SSR***	+	***SSE***
Total Variability	=	**Explained Variability**	+	**Unexplained Variability**

- SST = Total Sum of Squares
  - Measures the variation of the Y_i values around their mean,
- SSR = Regression Sum of Squares
  - Explained variation attributable to the relationship between X and Y
- SSE = Error Sum of Squares
  - Variation attributable to factors other than the relationship between X and Y

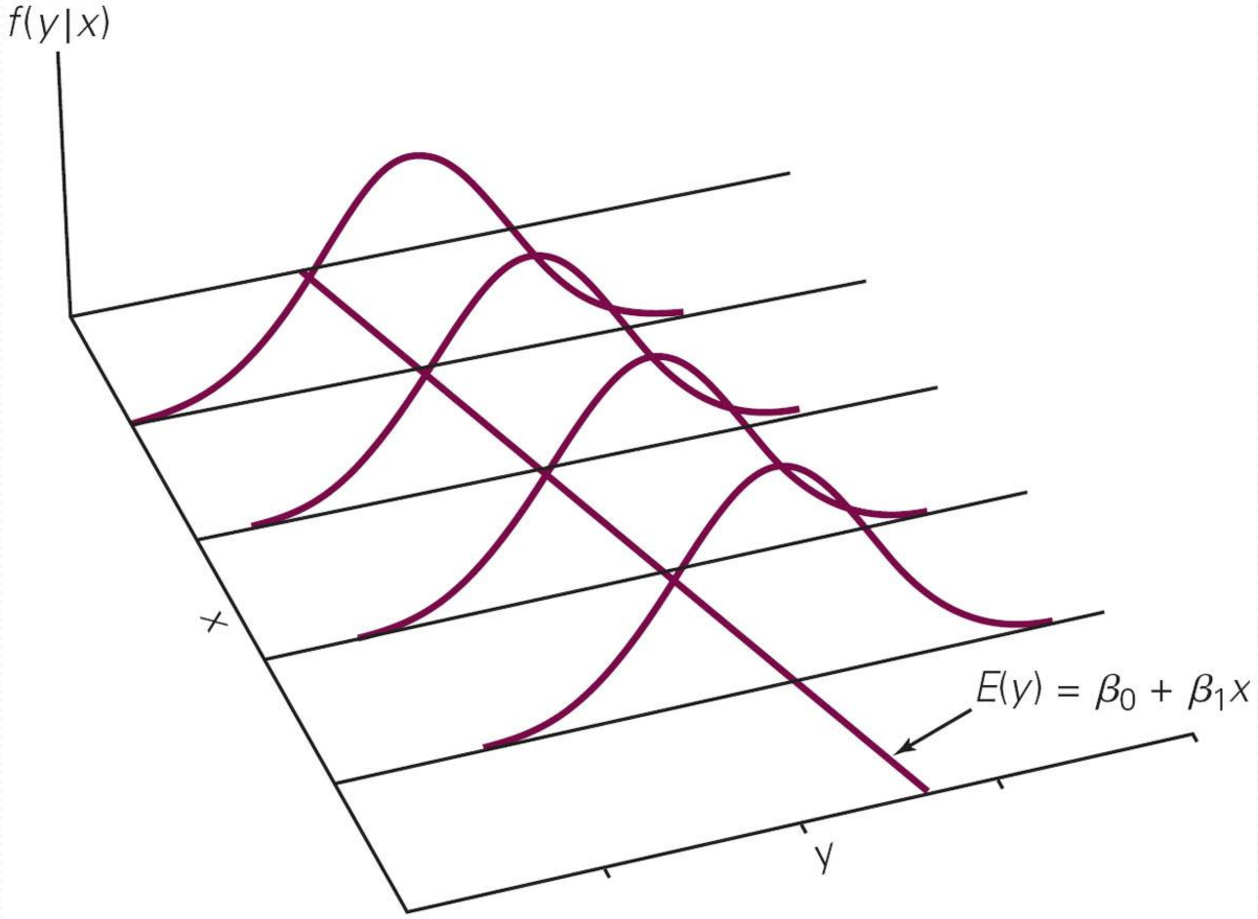
Assumptions

- The errors are independent
- They are normally distributed with mean zero and constant variance
- i.e

$$\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$$

- In other words,

$$y_i \sim N(\beta_0 + \beta_1 x, \sigma^2), i = 1, \ldots, n$$



$f(y|x)$
x
y
$E(y) = \beta_0 + \beta_1 x$

What about σ^2

$$\sigma^2 = \frac{SSE}{n-2} = MSE = \frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2}$$

Properties of the Estimators

$$\beta_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}}\right)\right)$$

$$\beta_0 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Inference About the Parameters

β_0

- A 100(1-α)% Confidence Interval

$$\beta_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}}\right)}$$

Inference About the Parameters

β_1

- A 100(1-α)% Confidence Interval

$$\beta_1 \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}}$$

$$H_0: \ \beta_1 = \beta_{1,0}$$

- Test Statistic:

$$t = \frac{\beta_1 - \beta_{1,0}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Decision Rule

- For a two-sided alternative Reject H_o

if $\quad |t| > t_{\frac{\alpha}{2}, n-2}$

Hypothesis of the Significance of the Regression

- H_o : $\beta_1 = 0$ vs H_1 : $\beta_1 \neq 0$
- Test Statistic
- $$t = \frac{\beta_1}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Significance of the Regression Continued

- If the null hypothesis is rejected,
- we conclude that
- the regression is significant.

Inference About the mean

$$\mu_{Y|X=x_0}$$

- A point estimate is

$$y(x_0)$$

Interval Estimate for the Mean

- A 100(1-α)% Confidence Interval

$$y(x_0) \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_0-\bar{X})^2}{S_{xx}}\right)}$$

What About Prediction?

What about Prediction?

$$y(x_{new})$$

- is predicted by

$$y(x_{new})$$

Prediction Interval for $y(x_{new})$

$$y(x_{new}) \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{\left(x_{new}-\bar{X}\right)^2}{S_{xx}}\right)}$$

Coefficient of Determination

$$R^2 = \frac{SSR}{SST}$$

Meaning of R-square

- The proportion of variability explained by the regression.

Sample Correlation Coefficient

- This is a measure of the strength of the linear relationship between x and y

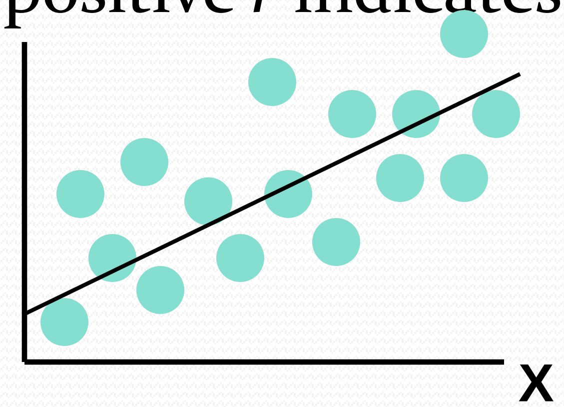
$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$\beta_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

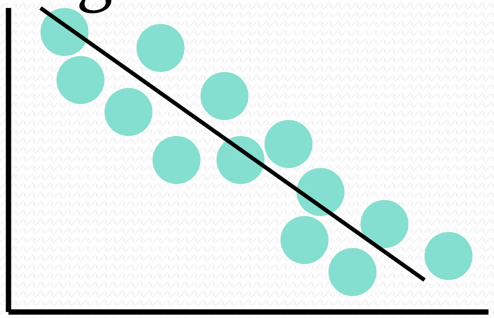
$$|r| \leq 1$$

Sign and Magnitude of *r*

- A positive *r* indicates that *y* increases with *x*



- A negative *r* indicates that *y* decreases with *x*



Sign and Magnitude of *r*

- $r = 0$ indicates no linear relationship between x and y
- While a magnitude close to 1 indicates that the relationship is strong.

- Notice that in Simple Linear Regression
- The coefficient of determination

$$R^2 = r^2$$