



# Topic 3

## Simple Linear Regression



# Learning objectives

- Explain the assumptions of classical linear regression using the Gauss-Markov Theorem (GMT) framework.
- Estimate the parameters of a regression using the ordinary least squares (OLS) method.
- Interpret the coefficients of a regression.
- Understand the steps of hypothesis testing
  - Standard errors
  - Confidence intervals

# Gauss-Markov Theorem:

Under the 5 Gauss-Markov assumptions, the OLS estimator is the best, linear, unbiased estimator of the true parameters ( $\beta$ 's) conditional on the sample values of the explanatory variables. In other words, the OLS estimators is BLUE

# 5 Gauss-Markov Assumptions for Simple Linear Model (Wooldridge, p.65)

1. *Linear in Parameters*

$$y = \beta_0 + \beta_1 x_1 + u$$

2. *Random Sampling of  $n$  observations*

$$(x_i, y_i) : i = 1, 2, \dots, n$$

3. *Sample variation in explanatory variables.  $x_i$ 's are not all the same value*


$$x \neq (x_1 = x_2 = x_3 = \dots = x_n)$$

4. *Zero conditional mean.* The error  $u$  has an expected value of 0, given any values of the explanatory variable

$$E(u|x) = 0$$

5. *Homoskedasticity.* The error has the same variance given any value of the explanatory variable.

$$Var(u|x) = \sigma^2$$



# How Good are the Estimates?

## Properties of Estimators

- *Small Sample Properties*
  - True regardless of how much data we have
  - Most desirable characteristics
- Unbiased
- Efficient
- BLUE (Best Linear Unbiased Estimator)

# “Second Best” Properties of Estimators

- Asymptotic (or large sample) Properties
  - True in hypothetical instance of infinite data
  - In practice applicable if  $N > 50$  or so
- Asymptotically unbiased
- Consistency
- Asymptotic efficiency

# Bias

- A parameter is unbiased if

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, \dots, k$$

- In other words, the average value of the estimator in repeated sampling equals the true parameter.
- Note that whether an estimator is biased or not implies nothing about its dispersion.

# Efficiency

- An estimator is efficient if its variance is less than any other estimator of the parameter.
- This criterion only useful in combination with others. (e.g.  $\hat{\beta}_j = 2$  is low variance, but biased)

$\hat{\beta}_j$  is the “best” Unbiased estimator if

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j)$$

, where  $\tilde{\beta}_j$  is any other unbiased estimator of  $\beta$



$F(\beta x)$

Unbiased and  
efficient estimator  
of  $\beta$

Biased estimator  
of  $\beta$

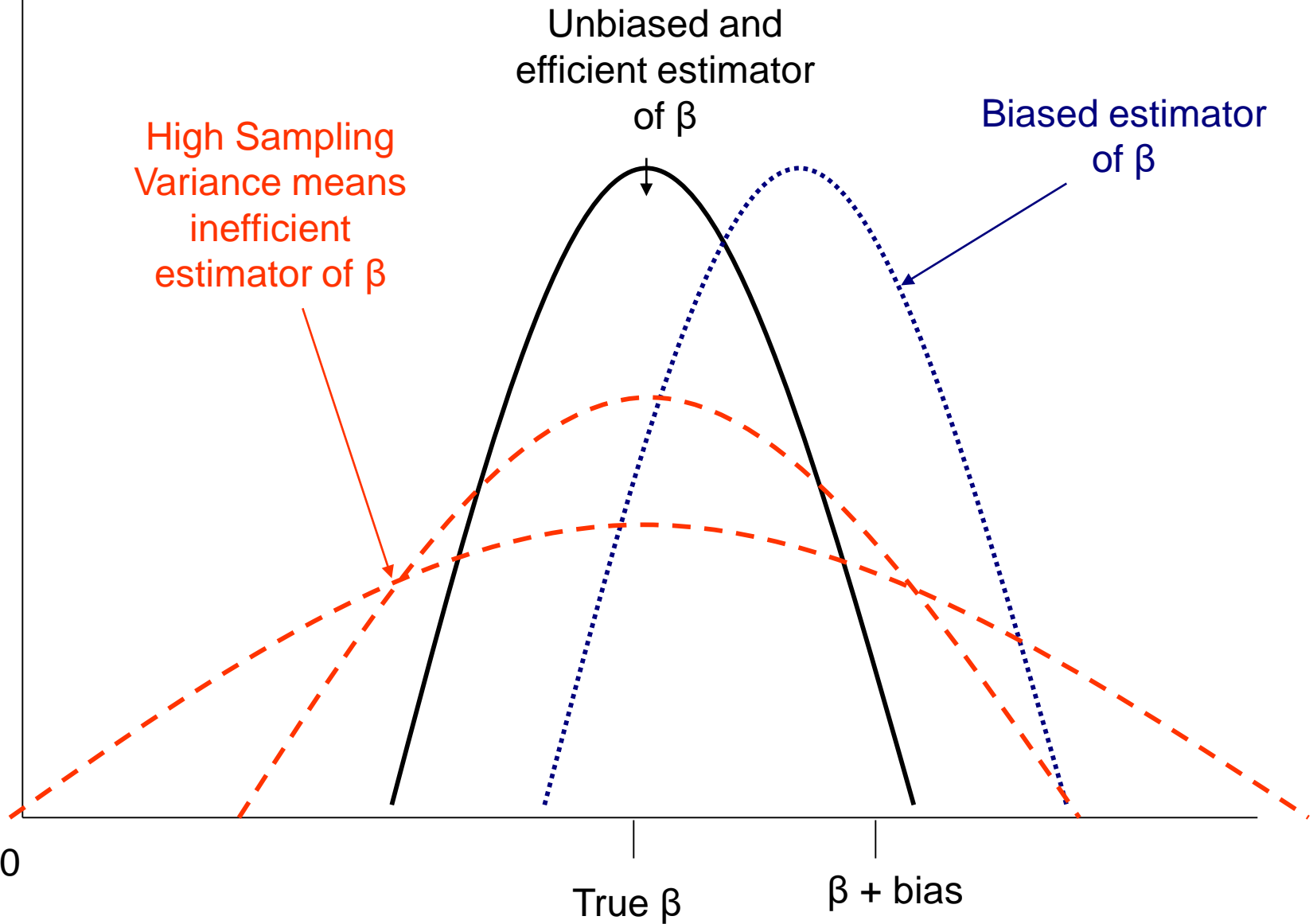
High Sampling  
Variance means  
inefficient  
estimator of  $\beta$

0

True  $\beta$

$\beta + \text{bias}$

9



# BLUE

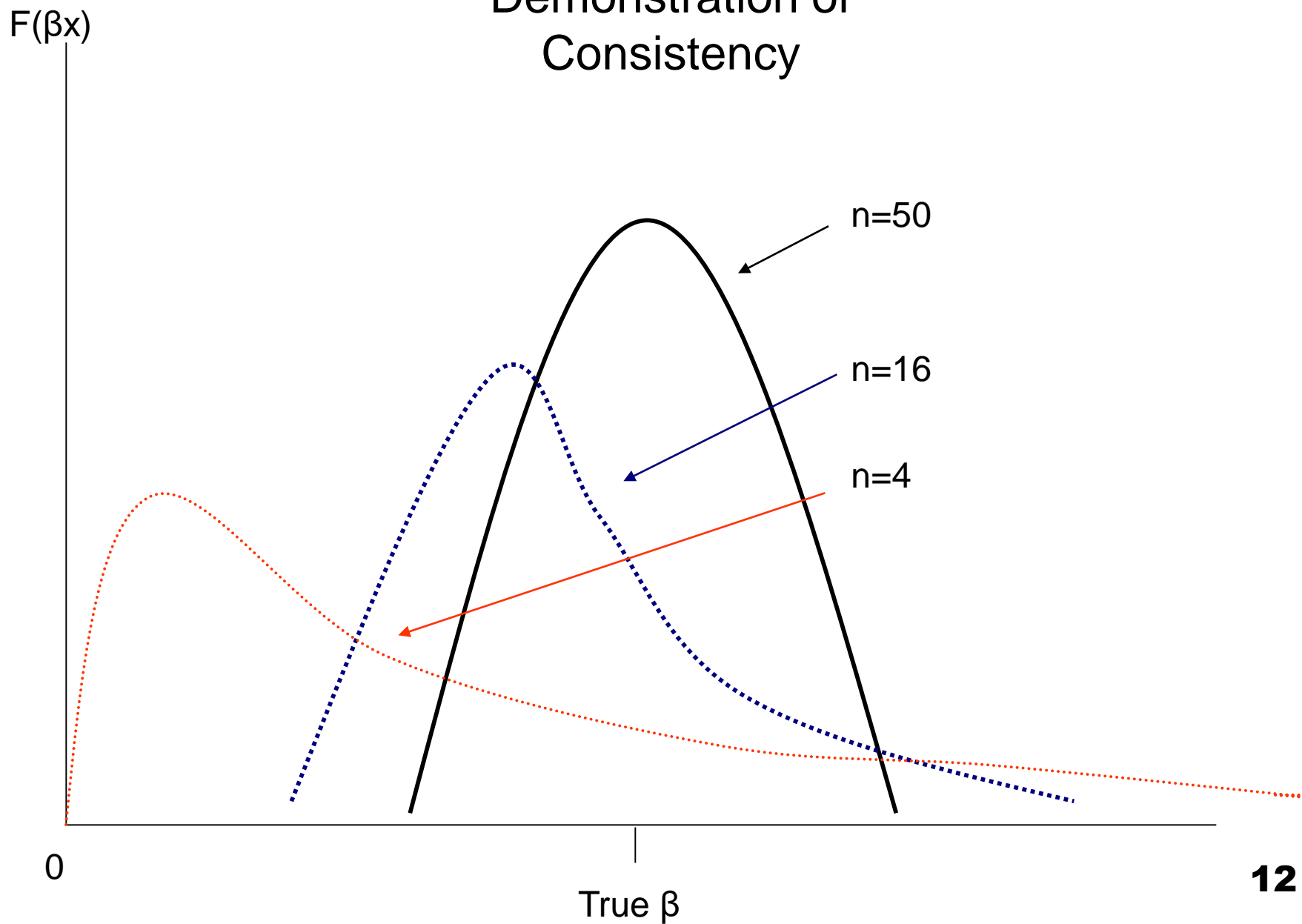
## (Best Linear Unbiased Estimate)

- An Estimator  $\hat{\beta}_j$  is BLUE if:
- $\hat{\beta}_j$  is a linear function
- $\hat{\beta}_j$  is unbiased:  $E(\hat{\beta}_j) = \beta_j, j = 0, 1, \dots, k$
- $\hat{\beta}_j$  is the most efficient:  $Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j)$

# Large Sample Properties

- Asymptotically Unbiased
  - As  $n$  becomes larger  $E(\hat{\beta}_j)$  trends toward  $\beta_j$
- Consistency
  - If the bias and variance both decrease as  $n$  gets larger, the estimator is consistent.
- Asymptotic Efficiency
  - asymptotic distribution with finite mean and variance
  - is consistent
  - no estimator has smaller asymptotic variance

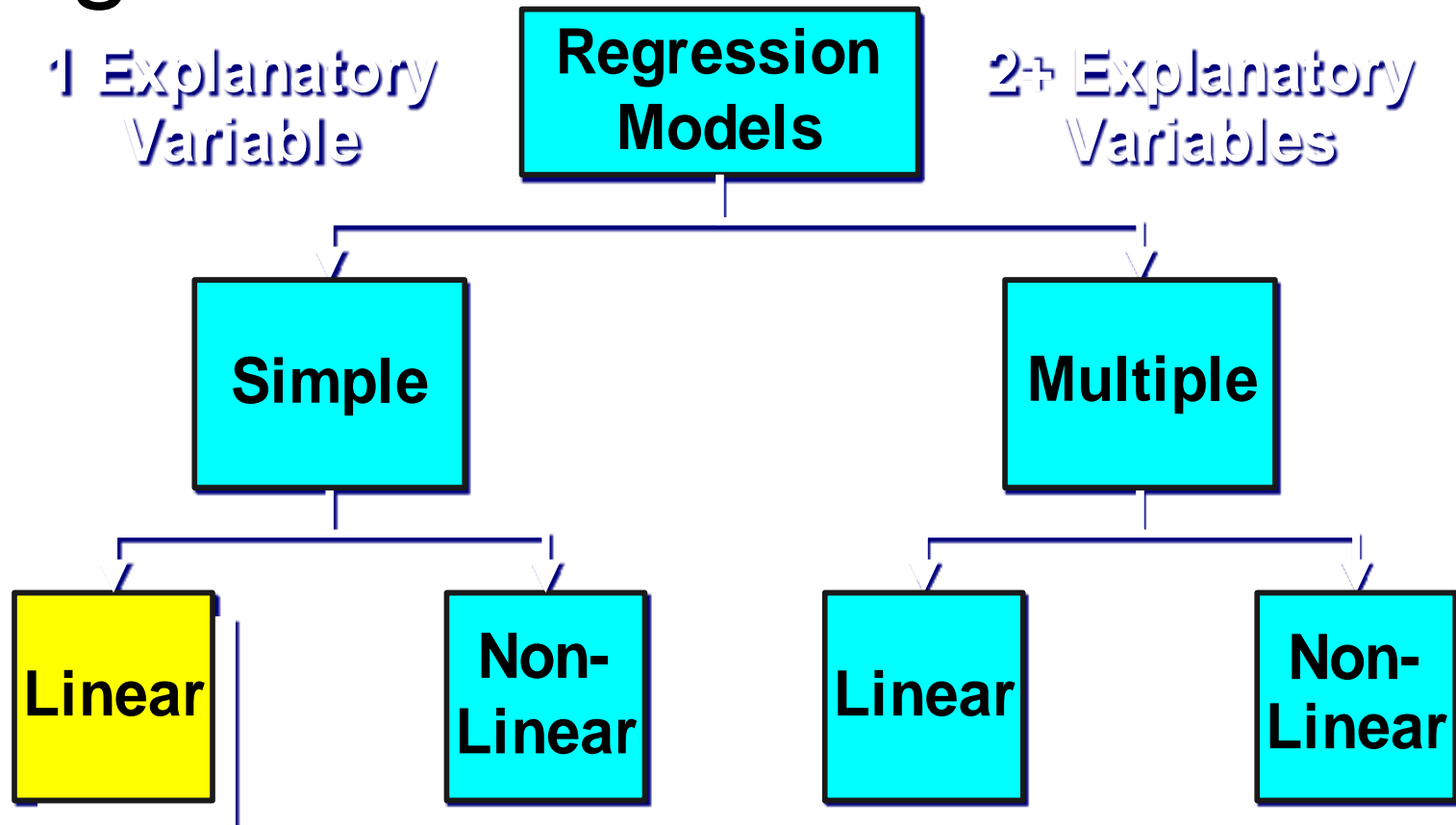
# Demonstration of Consistency



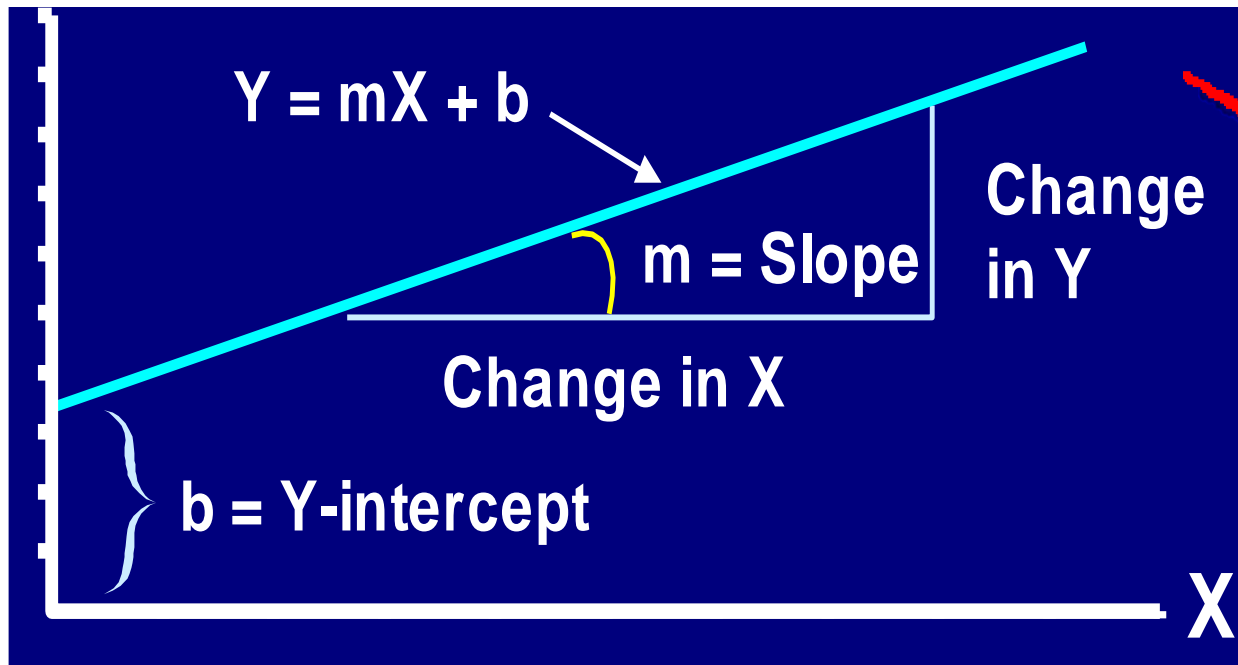


# Linear Regression Model

# Types of Regression Models



# Linear Equations



# Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

The diagram illustrates the Linear Regression Model equation,  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ , with labels and arrows pointing to each term:

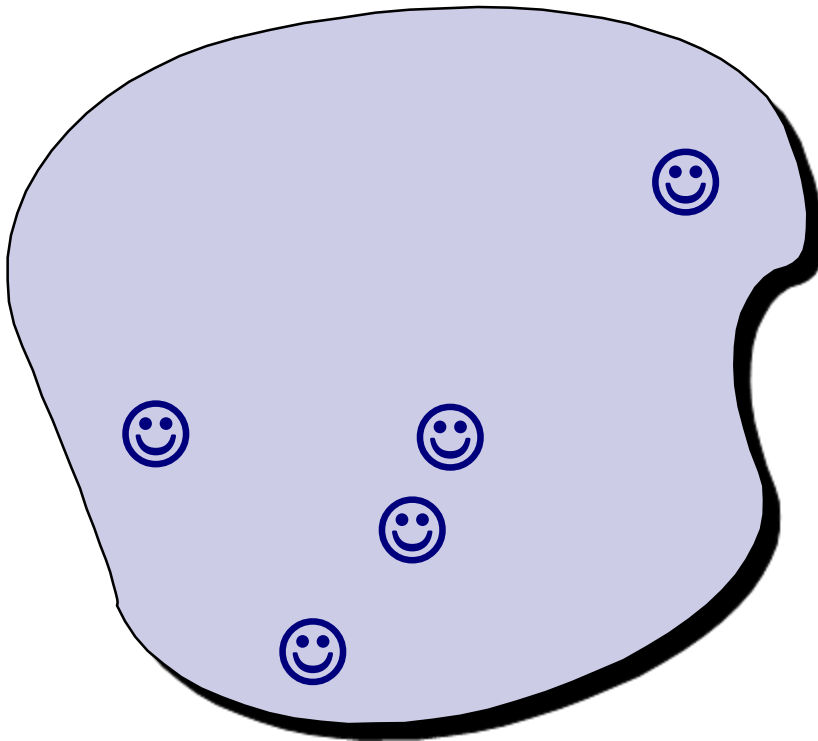
- Population Y-Intercept** points to  $\beta_0$ .
- Population Slope** points to  $\beta_1$ .
- Random Error** points to  $\varepsilon_i$ .
- Dependent (Response) Variable (e.g., CD+ c.)** points to  $Y_i$ .
- Independent (Explanatory) Variable (e.g., Years s. serocon.)** points to  $X_i$ .

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



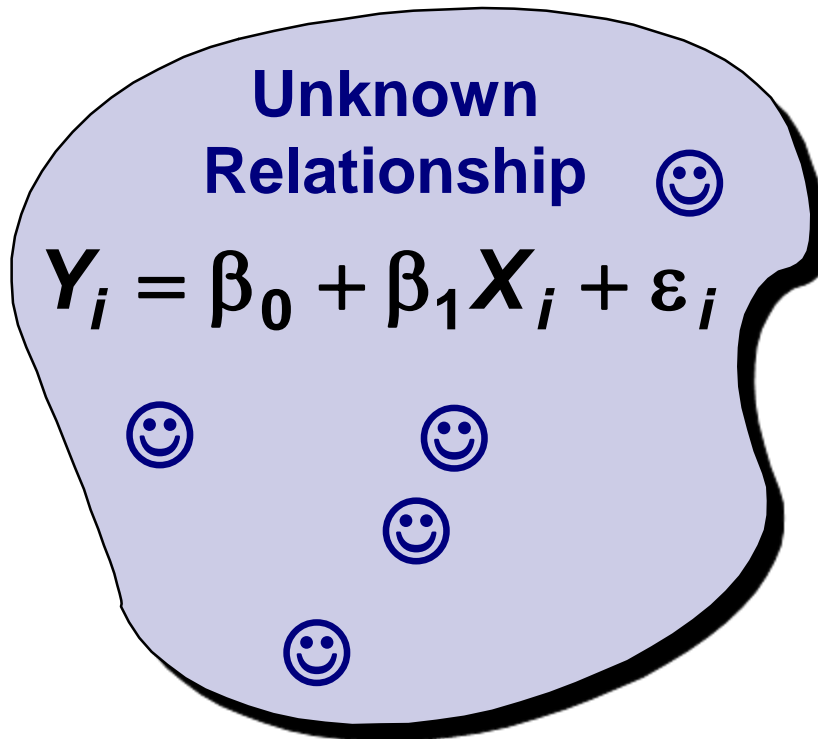
# Population & Sample Regression Models

## **Population**



# Population & Sample Regression Models

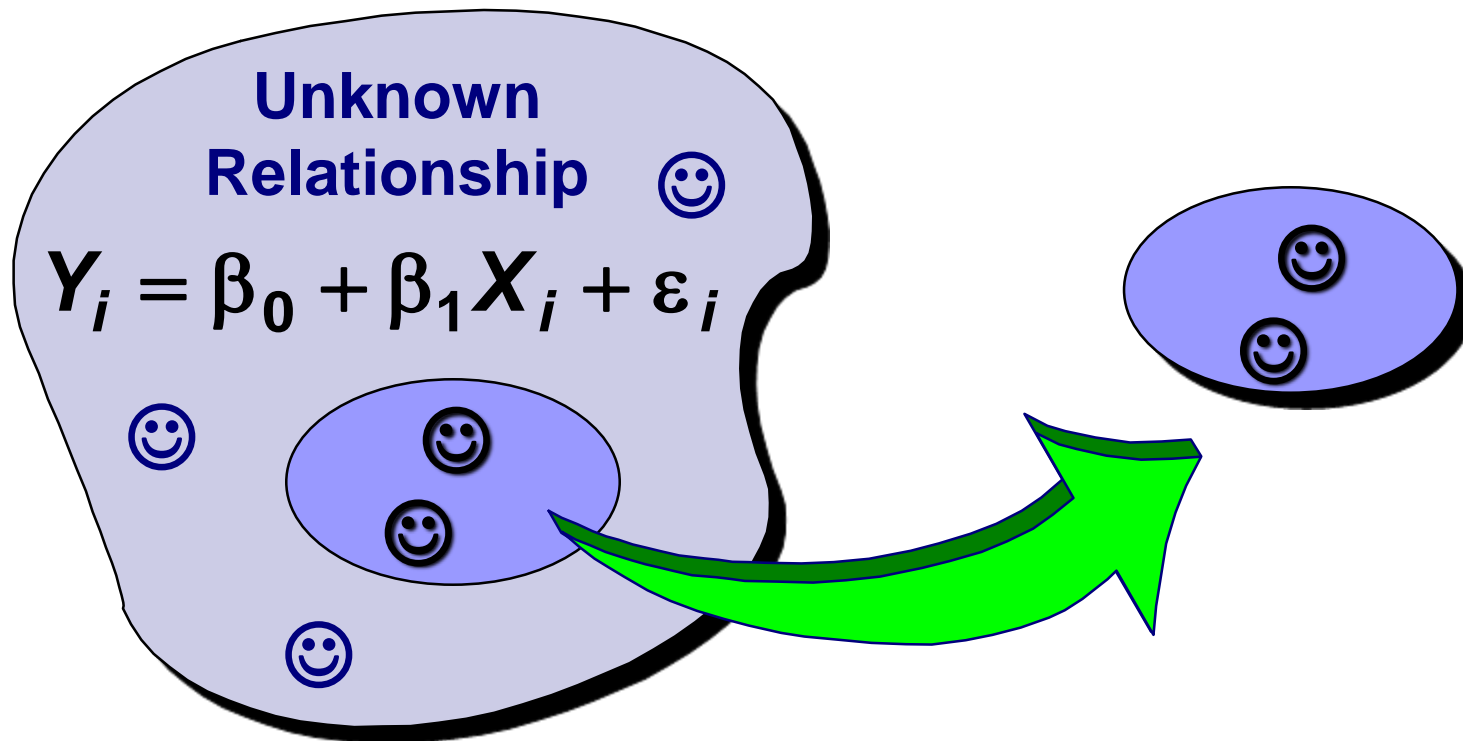
## Population



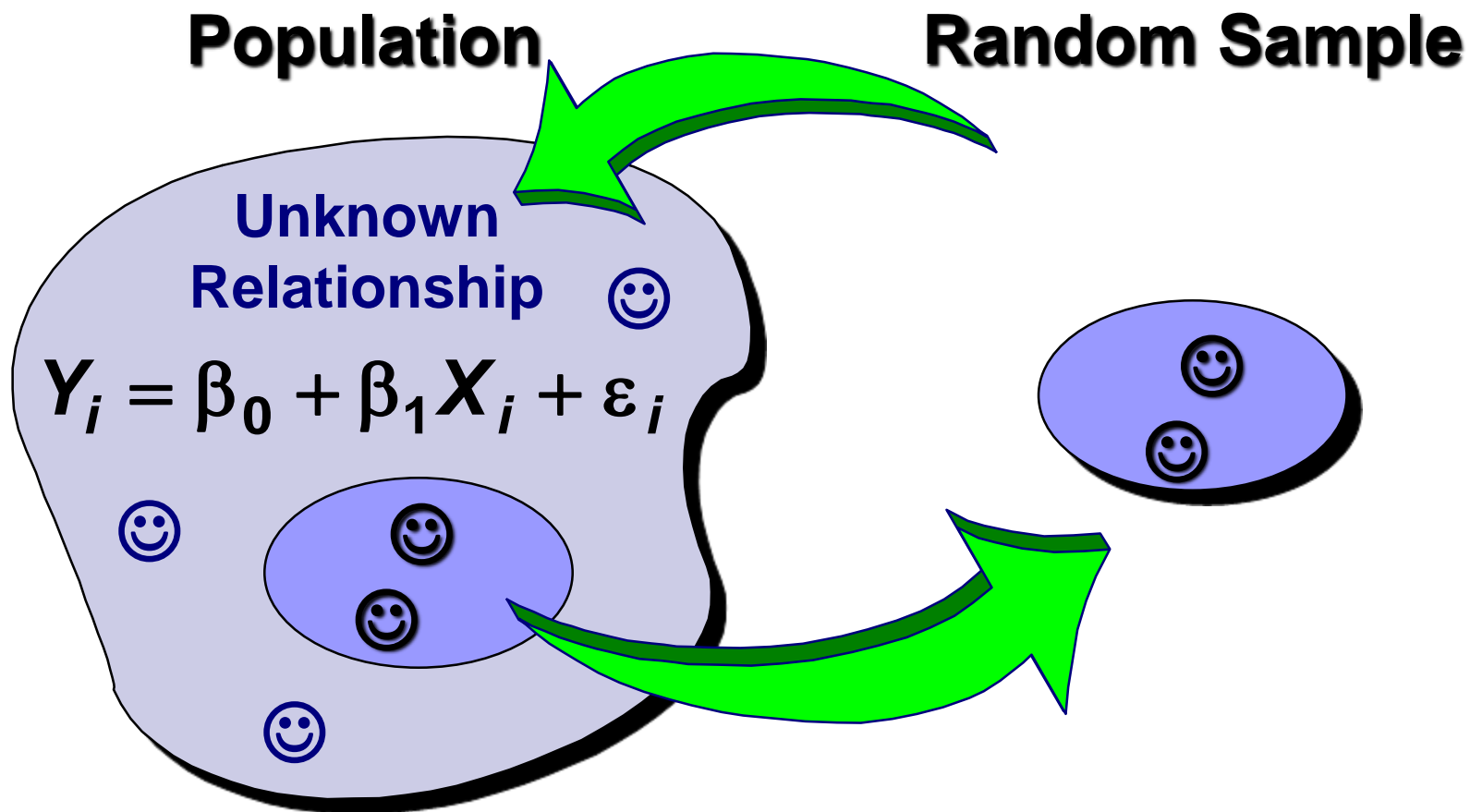
# Population & Sample Regression Models

**Population**

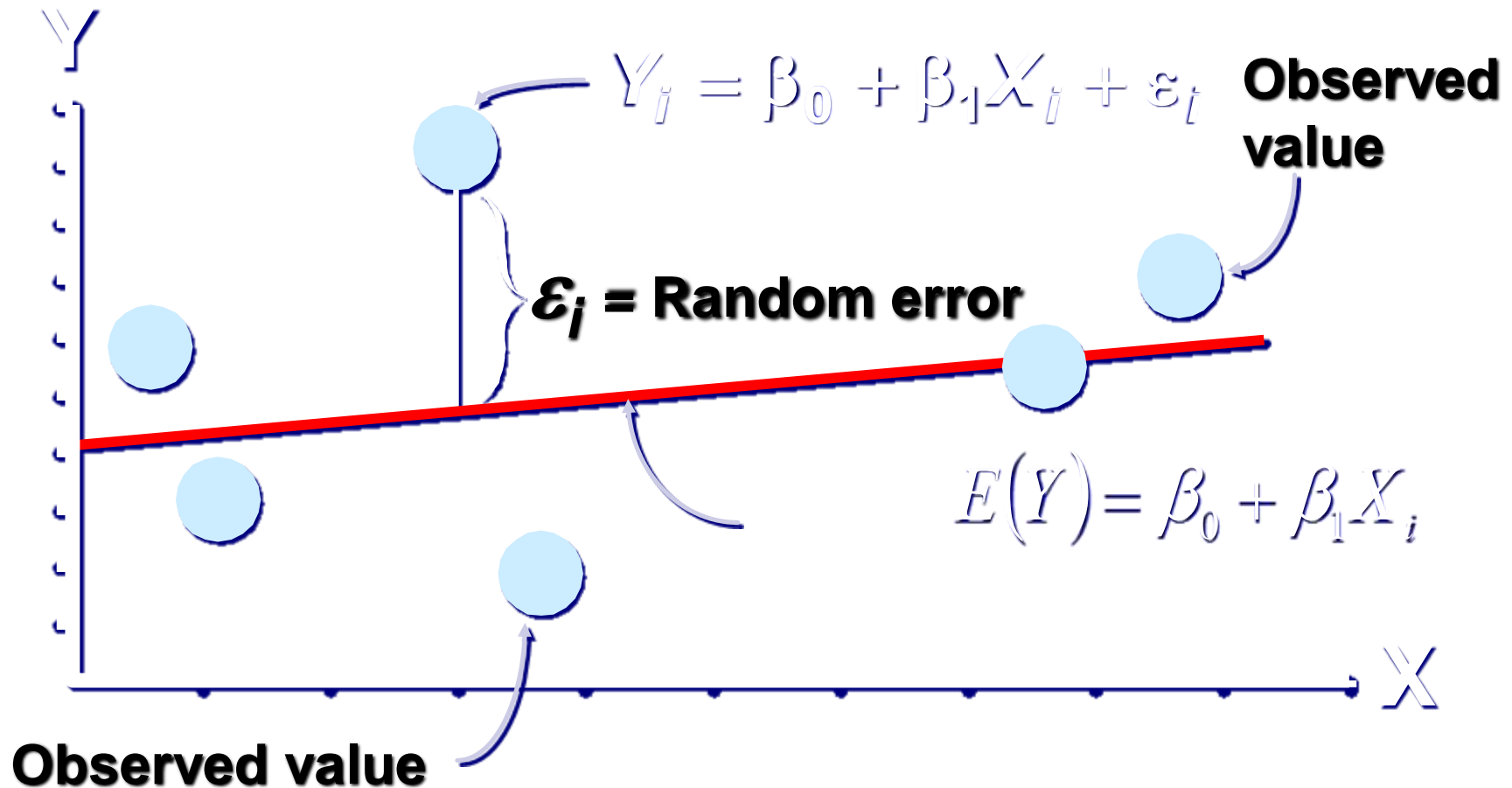
**Random Sample**



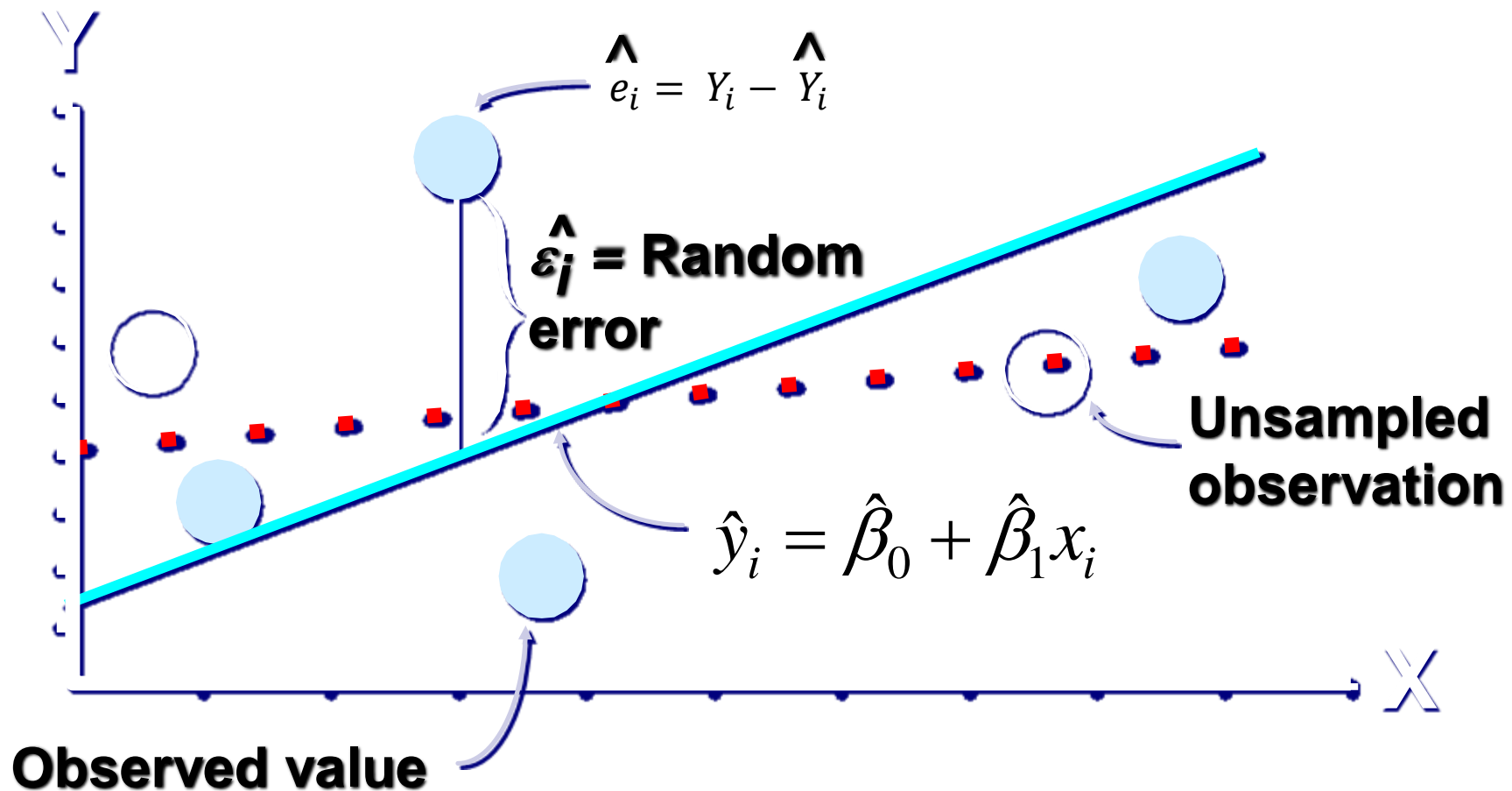
# Population & Sample Regression Models




# Population Linear Regression Model



# Sample Linear Regression Model

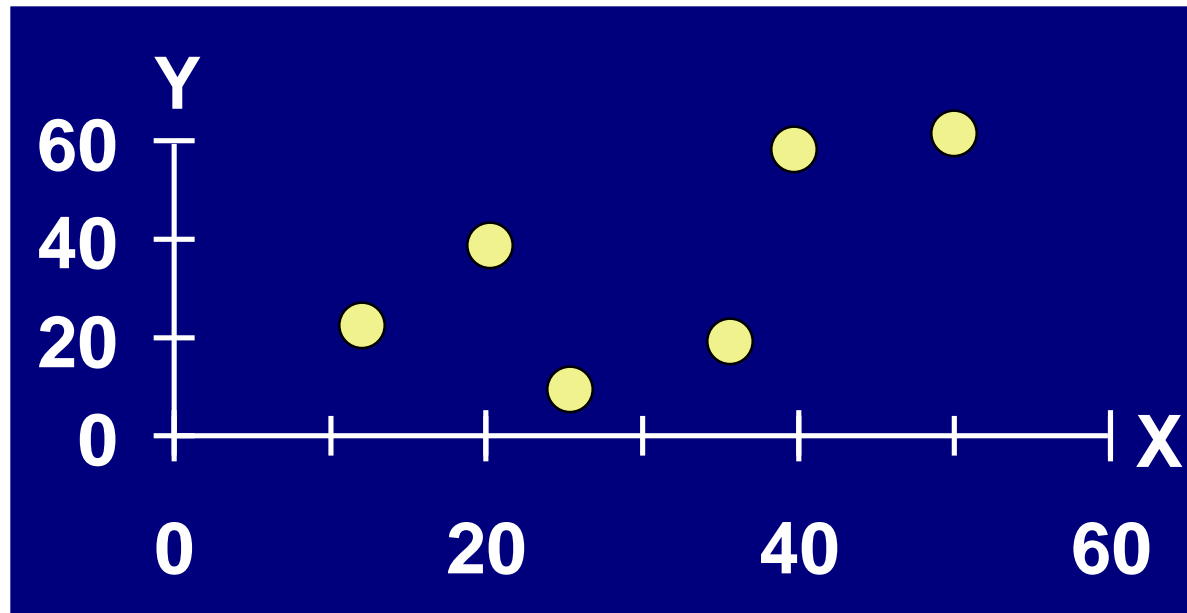




# Estimating Parameters: Ordinary Least Squares Method

# Scatter plot

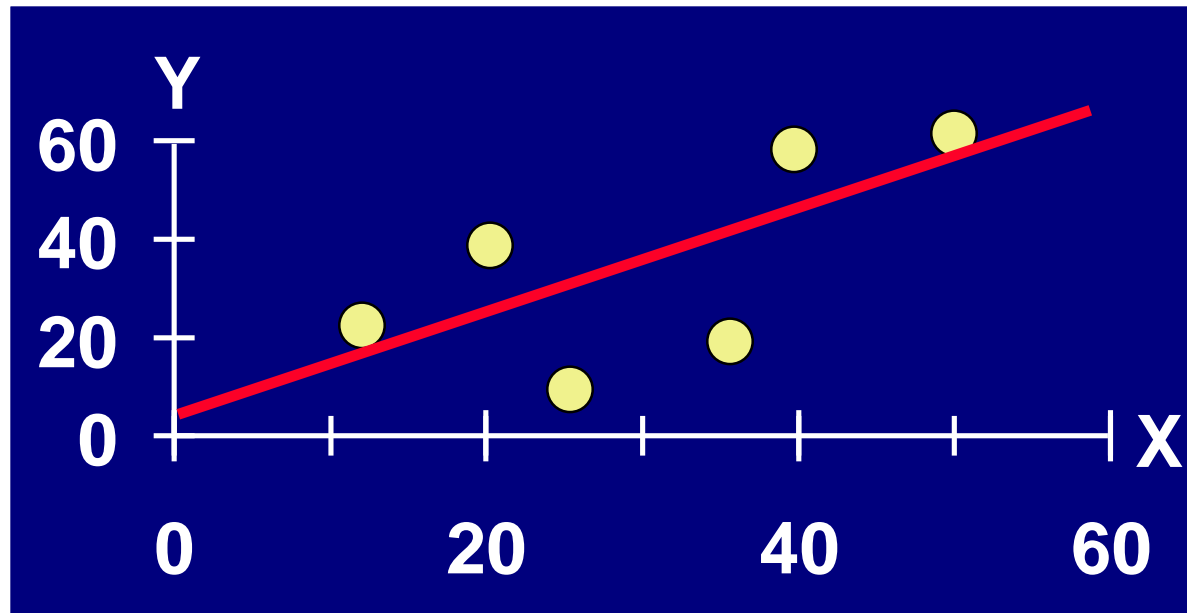
- 1. Plot of All  $(X_i, Y_i)$  Pairs
- 2. Suggests How Well Model Will Fit





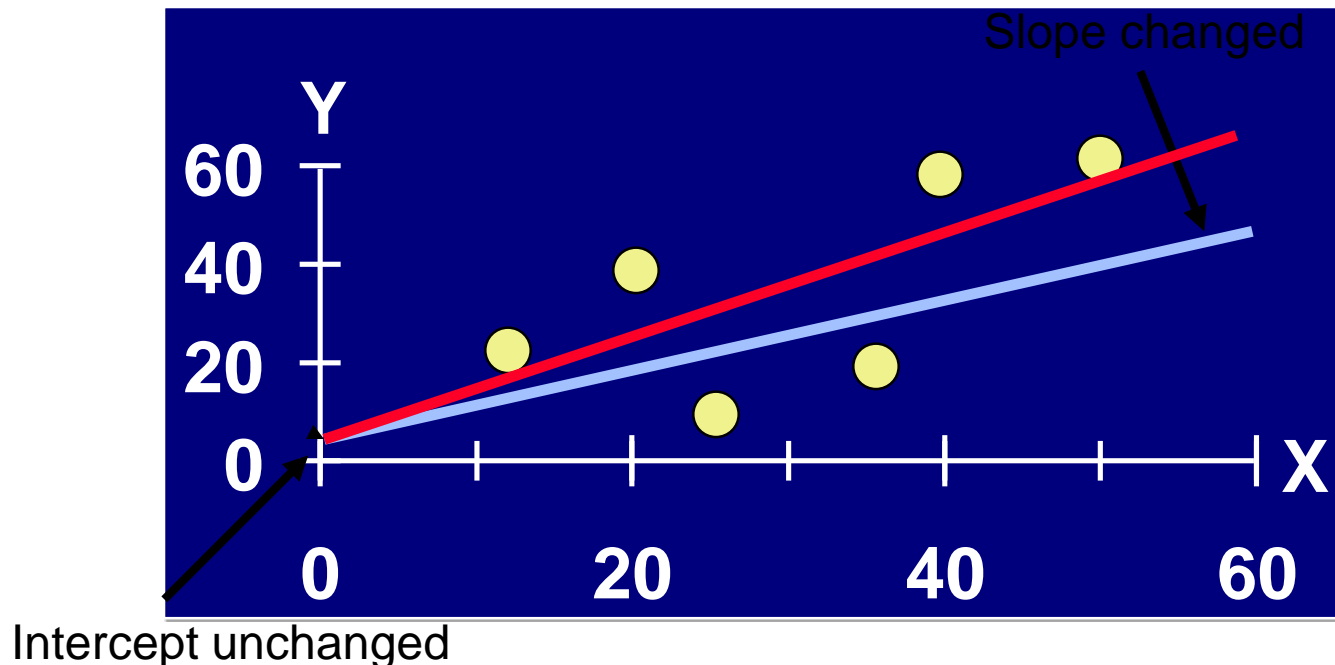
# Thinking Challenge

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



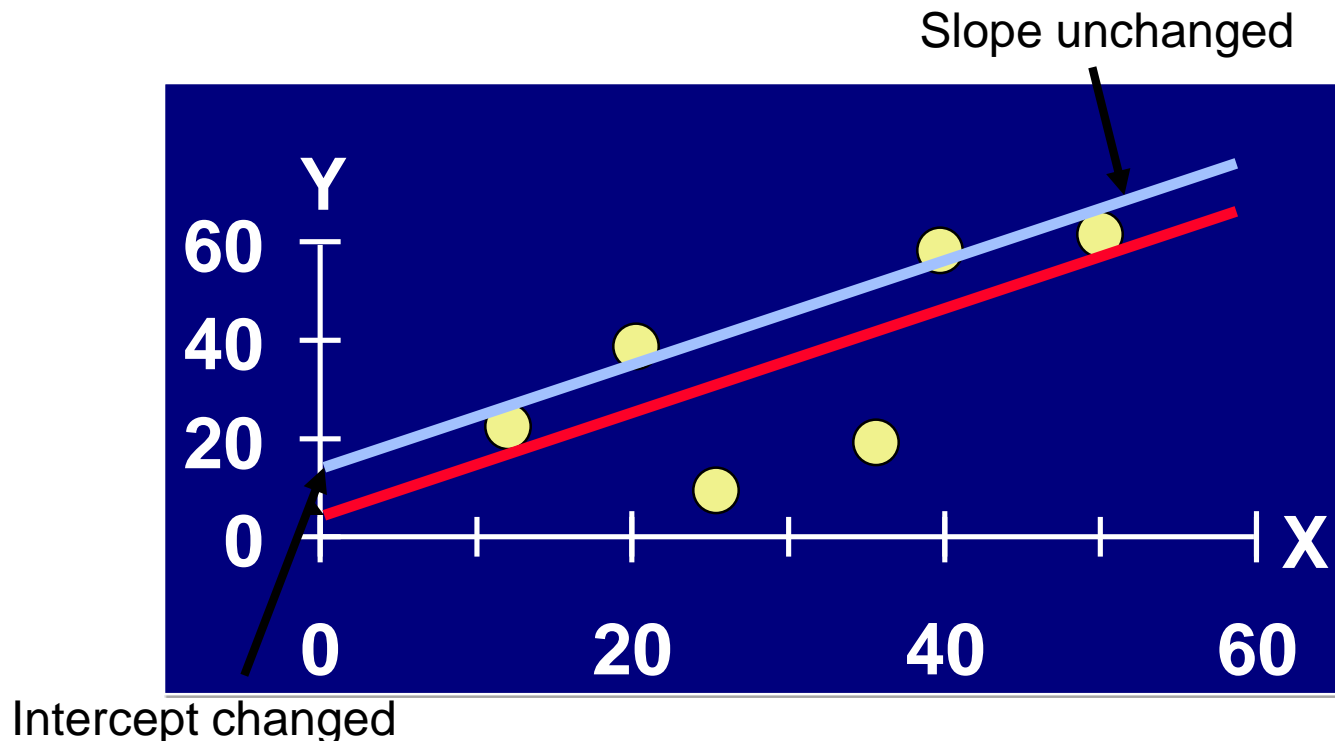
# Thinking Challenge

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



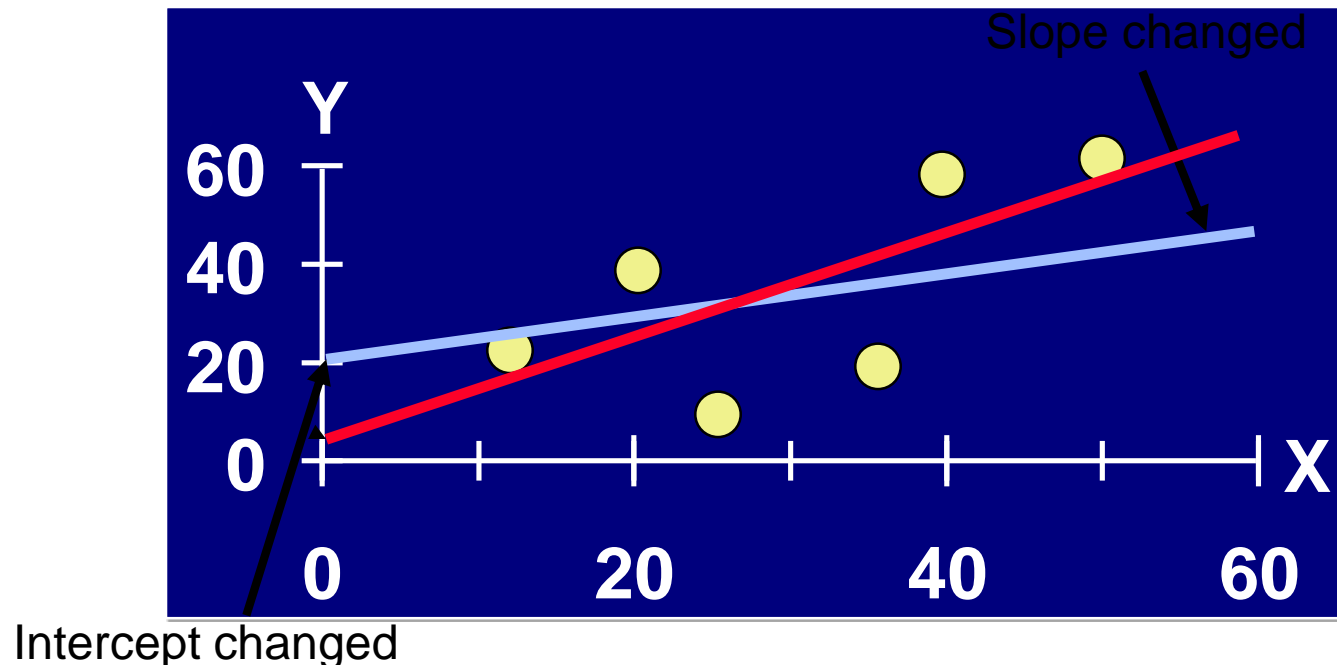
# Thinking Challenge

**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'?**



# Ordinary Least Squares

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

# Ordinary 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!**

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

# Ordinary Least Squares

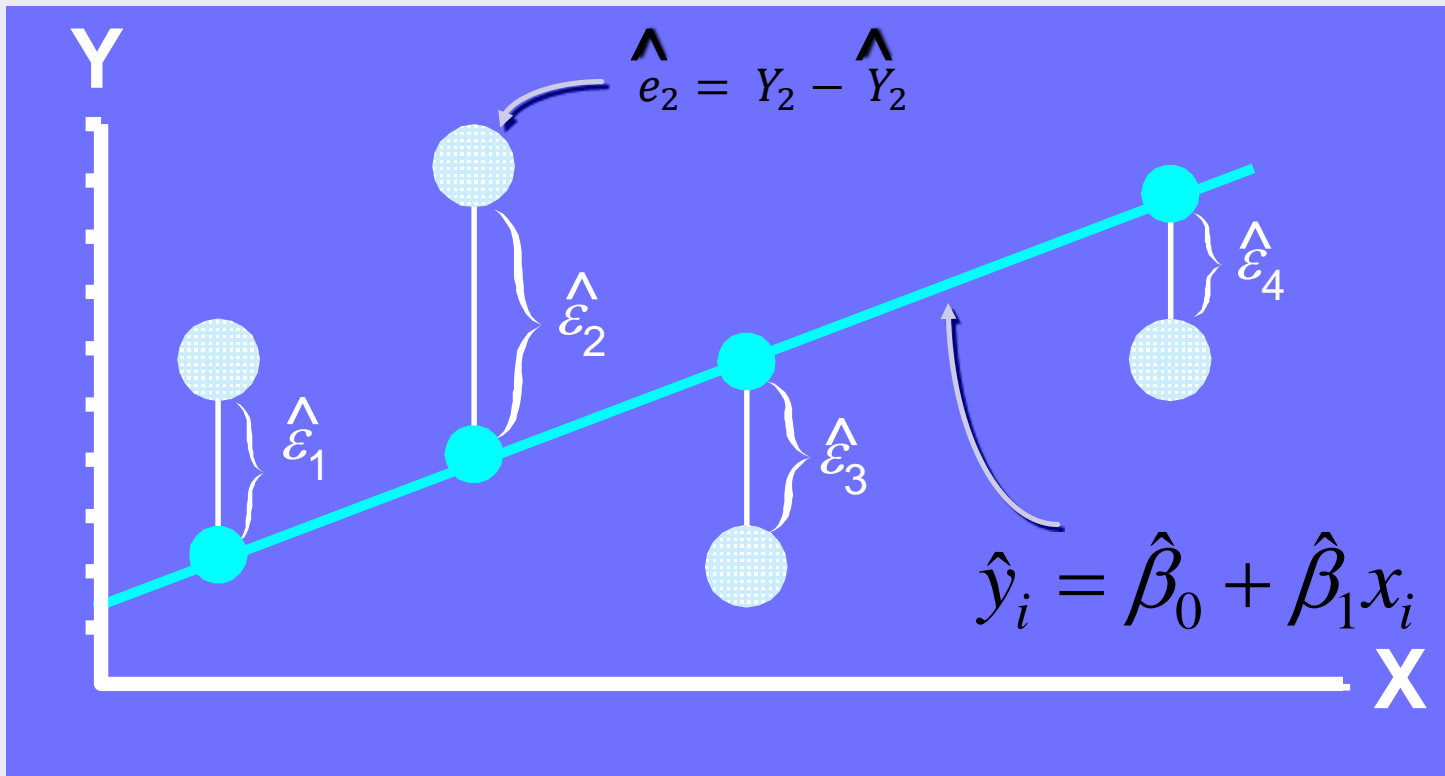
- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\mathcal{E}}_i^2$$

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

# Ordinary Least Squares Graphically

$$OLS \text{ minimizes } \sum_{i=1}^n e_i^2 = \hat{e}_1^2 + \hat{e}_2^2 + \hat{e}_3^2 + \hat{e}_4^2$$





# Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum xy}{\sum x^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# Interpretation of Coefficients

- 1. Slope ( $\hat{\beta}_1$ )

- Estimated  $Y$  Changes by  $\hat{\beta}_1$  for Each 1 Unit Increase in  $X$

- If  $\hat{\beta}_1 = 2$ , then  $Y$  Is Expected to Increase by 2 for Each 1 Unit Increase in  $X$

# Interpretation of Coefficients

## ■ 1. Slope ( $\hat{\beta}_1$ )

- Estimated  $Y$  Changes by  $\hat{\beta}_1$  for Each 1 Unit Increase in  $X$
- *A 1 unit increase in  $X$  leads to a (+/-) unit change in  $Y$* 
  - If  $\hat{\beta}_1 = 2$ , then  $Y$  Is Expected to Increase by 2 for Each 1 Unit Increase in  $X$

## ■ 2. Y-Intercept ( $\hat{\beta}_0$ )

- Average Value of  $Y$  When  $X = 0$ 
  - If  $\hat{\beta}_0 = 4$ , then Average  $Y$  Is Expected to Be 4 When  $X$  Is 0

# Parameter Estimation Example

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

<u>Estriol</u>	<u>Birthweight</u>
(mg/24h)	(g/1000)
1	1
2	1
3	2
4	2
5	4



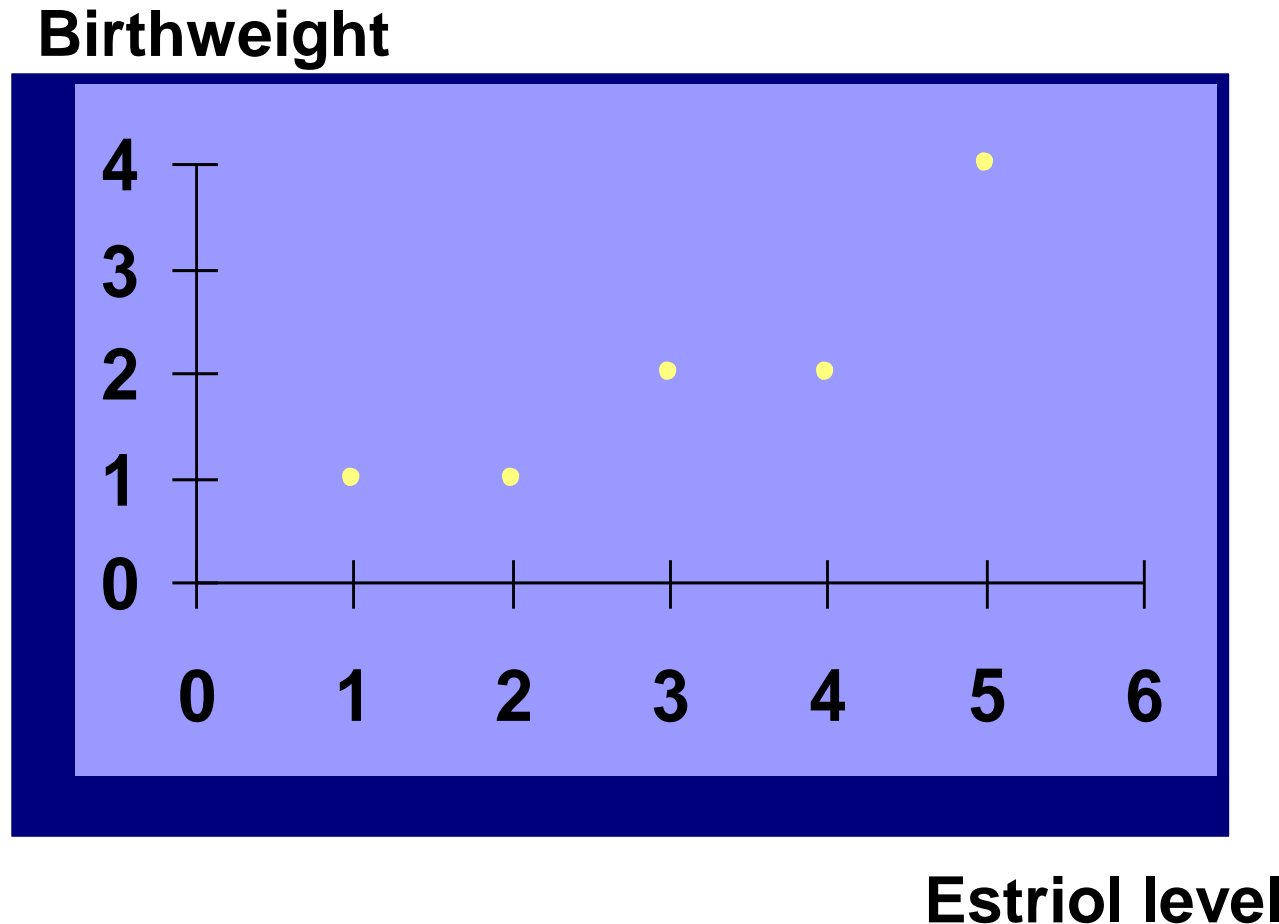
# Exercise

- Plot a scatter diagram
- Estimate the linear regression
- Interpret your results based on economic theory
- Show that  $\sum_{i=1}^n e_i = 0$
- Show that the SSE  $\approx 0$

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0$$

# Scatterplot

## Birthweight vs. Estriol level



# Parameter Estimation Solution Table

$X_i$	$Y_i$	$X_i^2$	$Y_i^2$	$X_i Y_i$
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

# Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = 0.70$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2 - (0.70)(3) = -0.10$$



# Coefficient Interpretation

## Solution

- 1. Slope ( $\hat{\beta}_1$ )
  - A 1 unit Increase in Estriol (X) leads to a 0.7 unit increase in birthweight (Y)

# Coefficient Interpretation Solution

## ■ 1. Slope ( $\hat{\beta}_1$ )

- Birthweight ( $Y$ ) Is Expected to Increase by .7 Units for Each 1 unit Increase in Estriol ( $X$ )

## ■ 2. Intercept ( $\hat{\beta}_0$ )

- Average Birthweight ( $Y$ ) Is -.10 Units When Estriol level ( $X$ ) Is 0
  - Difficult to explain
  - The birthweight (or any weight for that matter) should always be positive

rename var12 Birthweight

. regress Birthweigh Estriol

Source	SS	df	MS	Number of obs	=	5
Model	4.9	1	4.9	F(1, 3)	=	13.36
Residual	1.1	3	.366666667	Prob > F	=	0.0354
				R-squared	=	0.8167
				Adj R-squared	=	0.7556
Total	6	4	1.5	Root MSE	=	.60553

Birthweight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Estriol	.7	.1914854	3.66	0.035	.0906079	1.309392
_cons	-.1	.6350853	-0.16	0.885	-2.121125	1.921125

.

Command



# Limitations of simple linear regression

- Only considers one independent variable.
- The dependent variable must be continuous.
- Cannot show causation.
- Sensitive to outliers.
- Can only describe linear relationships.
- Only looks at the mean of the dependent variable.



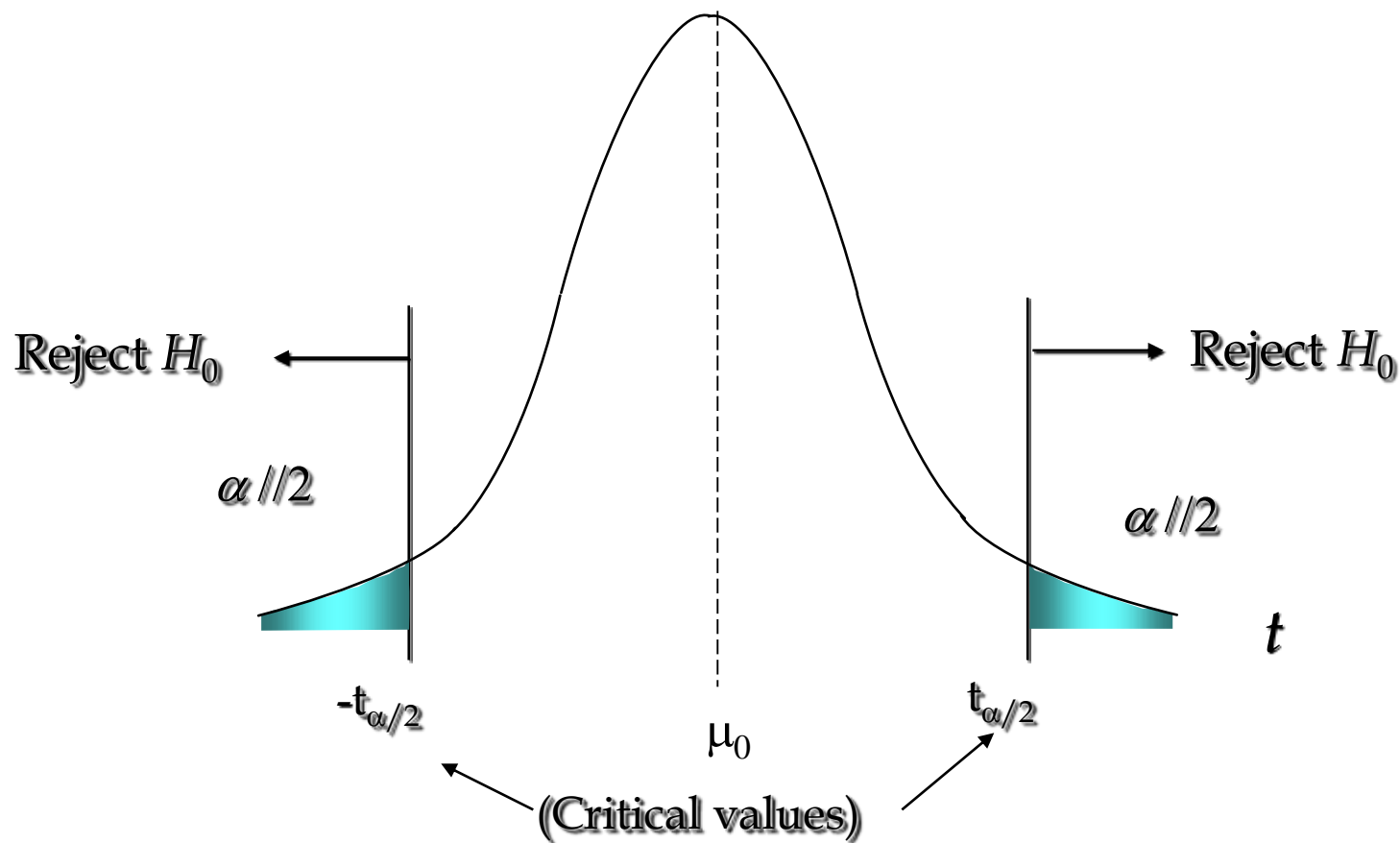
# Statistical Inference

Hypothesis testing

Confidence intervals

# Hypothesis Testing

Two-Tailed Test about a Population Mean: Small  $n$



# Student's $t$ -test

- The  $t$ -test is used to test hypotheses about means when the population variance is unknown (the usual case). Closely related to  $z$ , the unit normal.
- Remember: If the sample is small ( $n < 30$ ) and the population variance  $\sigma$  is unknown, then we use the  $t$ -test and not the  $z$ -test.

# Steps of Hypothesis Testing

1. Determine the null and alternative hypotheses.
2. Specify the level of significance  $\alpha$ .
3. Select and calculate the test statistic that will be used to test the hypothesis.

## Using the Test Statistic

4. Use  $\alpha$  to determine the critical value for the test statistic. The critical value comes from the Student's t-distribution table.
5. State the rejection rule for  $H_0$ . Use the value of the test statistic and the rejection rule to determine whether to reject  $H_0$ .
6. Make a conclusion on the statistical significance of the coefficient.





# How do we compute the test statistic?

$$t^* \text{ or } t_{\text{calculated}} = \frac{\hat{\beta}_i - \text{null mean}}{\sigma_{\beta_i}}$$

For our cases *null mean*=0

# How do we get the t-critical?

- From the Student's t-distribution tables

$$t_{n-k-1, \alpha/2}$$

Recall that this is a 2-tailed test, so check  $\alpha = 0.05$  from tables

# Finding the Standard Errors

$$\sigma_{\alpha}^2 = \text{Var}(\hat{\alpha}) = \frac{\delta^2 \sum X_i^2}{n \sum x_i^2} \text{ and } \sigma_{\alpha} = \text{SE}(\hat{\alpha}) = \sqrt{\frac{\delta^2 \sum X_i^2}{n \sum x_i^2}}$$

$$\sigma_{\beta_i}^2 = \text{Var}(\hat{\beta}) = \frac{\delta^2}{\sum x_i^2} \text{ and } \sigma_{\beta_i} = \text{SE}(\hat{\beta}) = \sqrt{\frac{\delta^2}{\sum x_i^2}}$$

Where:

$$\delta^2 = \frac{\sum_{i=1}^n e_i^2}{n-k-1}$$

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

# How do we find the critical values? t distribution values

With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.28
.90	1.812	1.725	1.697	1.64
.95	2.228	2.086	2.042	1.96
.99	3.169	2.845	2.750	2.58

Note:  $t \rightarrow Z$  as  $n$  increases



# Confidence Intervals

**Confidence Interval:** An interval of values computed from the sample, that is almost sure to cover the true population value.

We make confidence intervals using values computed from the sample, not the known values from the population.

The confidence level is the probability that we do not find a statistically significant effect of the effect of an independent variable is zero.

It is related to the significance level and it is defined as  $1 - \alpha$



# Confidence Intervals

Interpretation: In 95% of the samples we take, the true population proportion (or mean) will be in the interval.

We are 95% confident that  $\beta_i$  lies between the lower limit and the upper limit


This is also the same as saying we are 95% confident that the true population proportion (or mean) will be in the interval



How do we compute the intervals?

Single population mean (small n, normally distributed)

$$CI = \hat{\beta}_i \pm t_{n-k-1, \alpha/2} * \sigma_{\beta_i}$$



How do we compute the intervals?  
Single population mean (small n, normally distributed)

$$P \left[ \hat{\beta}_i - t_{\frac{\alpha}{2}, n-k-1} * \sigma_{\hat{\beta}_i} \leq \hat{\beta}_i \leq \hat{\beta}_i + t_{\frac{\alpha}{2}, n-k-1} * \sigma_{\hat{\beta}_i} \right] = 95\%$$



# Hypothesis Testing Example

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

<u>Estriol</u>	<u>Birthweight</u>
(mg/24h)	(g/1000)
1	1
2	1
3	2
4	2
5	4



# Exercise 2

- Compute the standard errors for  $\hat{\alpha}$  and  $\hat{\beta}$
- Test the statistical significance of the slope at 5% level ( $\alpha = 5\%$ )
- Compute the confidence intervals for  $\hat{\alpha}$  and  $\hat{\beta}$
- Write out the compact form of the regression equation:

$$\hat{Y}_i = -0.1 + 0.7X + e$$

$$(SE(\hat{\alpha})) \quad (SE(\hat{\beta}))$$

$$R^2 = ?$$

$$n = ?$$

# Exercise 3

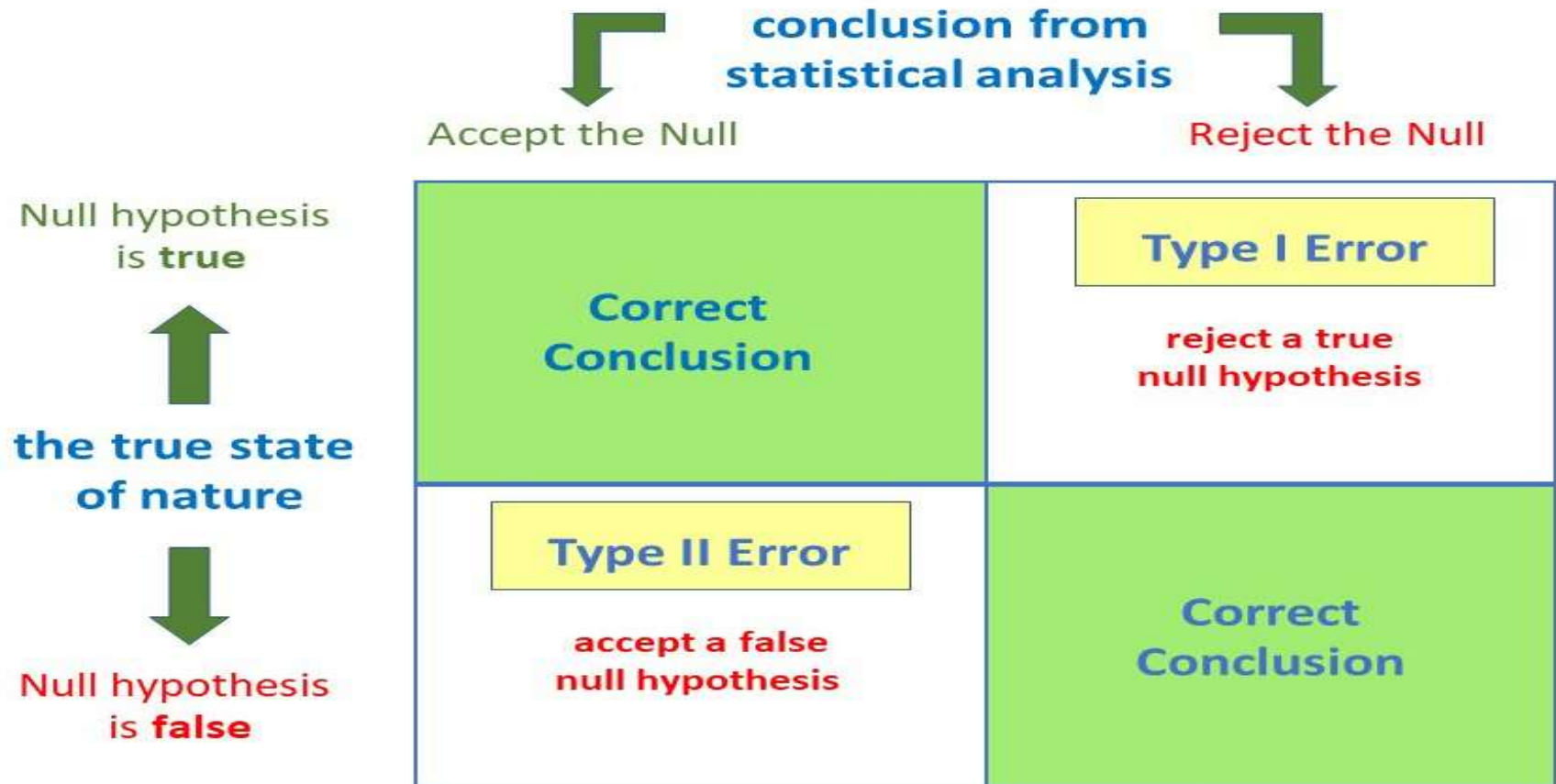
- The following data relates to the quantity demanded and price of a commodity collected from five markets.

<b>Price</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>Quantity demanded</b>	<b>15</b>	<b>10</b>	<b>14</b>	<b>8</b>	<b>3</b>

# Exercise 3

- Plot a scatter diagram
- Estimate the linear regression
- Interpret your results based on economic theory
- Show that  $\sum_{i=1}^n e_i = 0$
- Show that the SSE  $\approx 0$
- Compute the standard error for  $\hat{\beta}$
- Test the statistical significance of the slope at 5% level ( $\alpha = 0.05$ )
- Write out the compact form of the regression equation
- Compute the confidence intervals for  $\hat{\beta}$

# Conclusion from Statistical Analysis



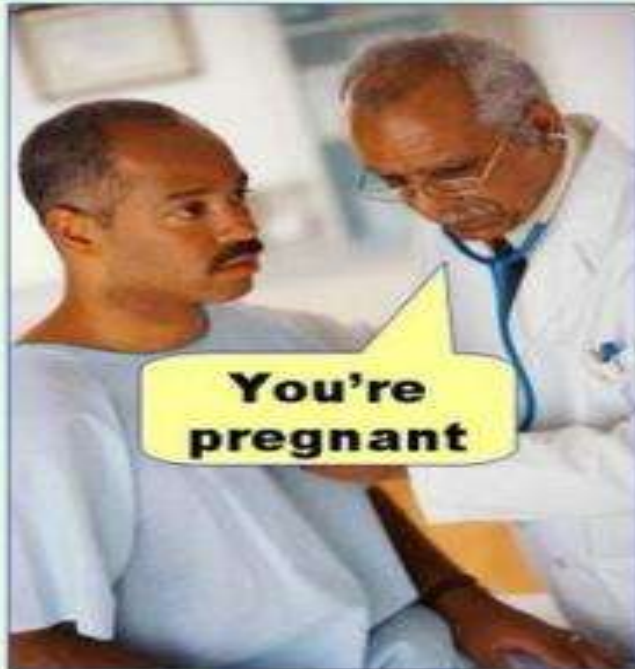
# Types of Statistical Errors

Table 1. Types of Statistical Errors

	$H_0$ is actually:	
	True	False
Reject $H_0$	Type I error	Correct
Accept $H_0$	Correct	Type II error

# Type I and Type II Error

**Type I error**  
(false positive)



**Type II error**  
(false negative)





# Type I and Type II Error

- **False Positive: (Type 1 Error)**

- Interpretation: You predicted positive and it's false.
- You predicted that a man is pregnant but he actually is not.

- **False Negative: (Type 2 Error)**

- Interpretation: You predicted negative and it's false.
- You predicted that a woman is not pregnant but she actually is.