Chapter 8 Simple Linear Regression Model

L1 – Simple Linear Regression by Least Squares Method

- Methods to Assess the Model

Learning Outcomes

At the end of the lesson, the student should be able to

- **Use the least squares method (LSM) to estimate a linear model**
- **Assessing the model to determine whether the model obtained is an adequate fit to the data**
- Construct confidence intervals on regression parameters
- Use the regression model to make prediction of a future observation and construct appropriate prediction interval on the future observation

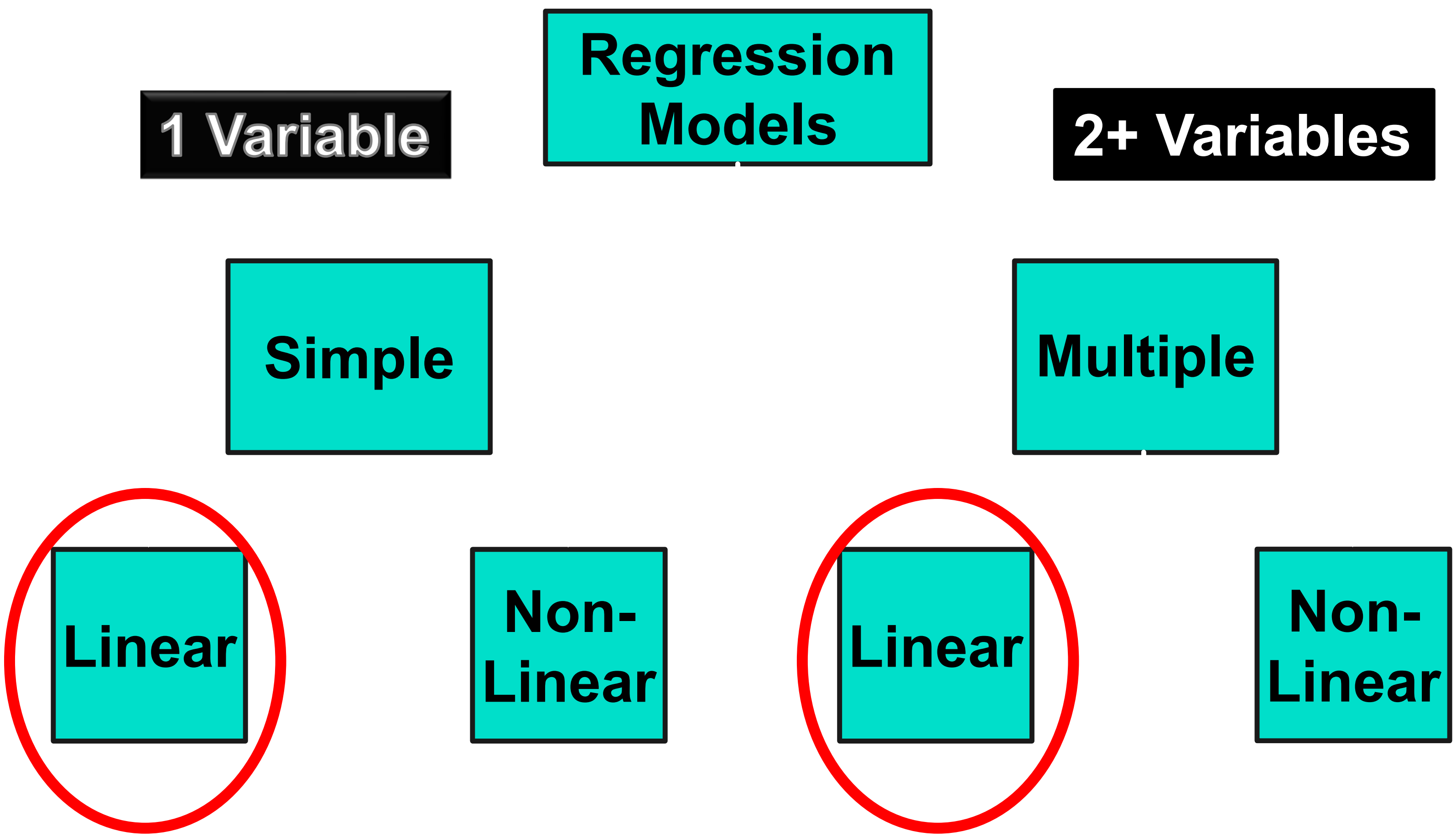
Introduction - Linear Regression

- Sometime we wish to investigate the result of statistical enquiry or experiment by **comparing two set of data**, *x* and *y*,
- Example
  - The age of a plant VS the quantity of fruit produced by a plant.
  - Pupil's mark in ODE VS Statistics

Regression Function

- Look for a relationship *y=f(x)*, where the function *f* is to be determined.
  - *i.e* given the point only (scatter digram)we have to ***"work back-ward"*** or ***"regress"*** the function.
- We only consider the simple type of function, *y=f(x)* which is a ***straight line.***
- We try to ***estimate*** fairly accurately the position of the line – ***a regression line.***

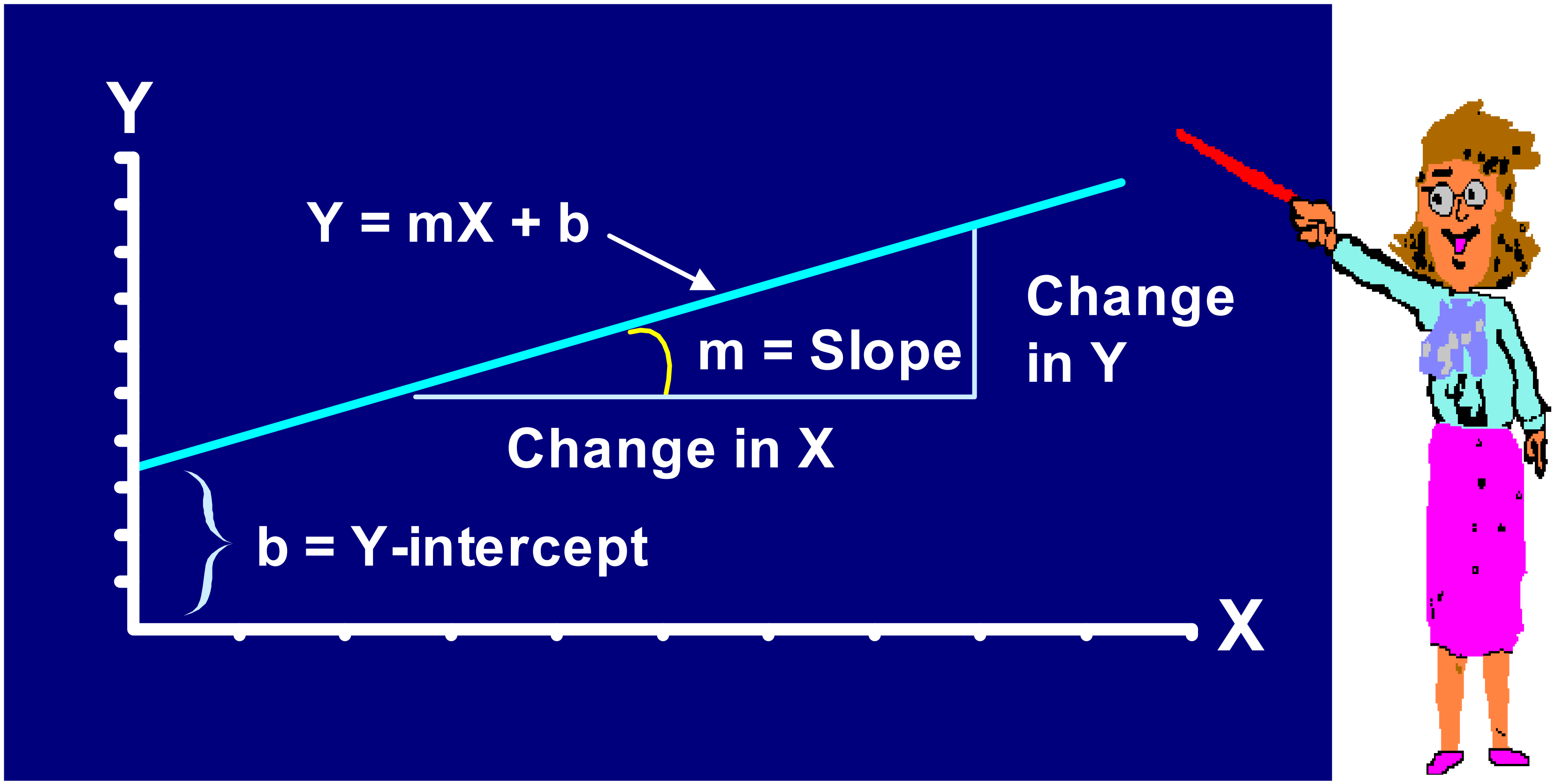
Types of Regression Models



Simple Linear Regression Model

- Determine the random relationship between
  - Y (dependent variable) and X (independent variables) on the base of n observations $(x_1, y_1),\ldots,(x_n, y_n)$
- The Model Parameters are estimated by Least Squares Method (LSM).
- Make predictions for Y from the model

Linear Equations…. You know...



Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

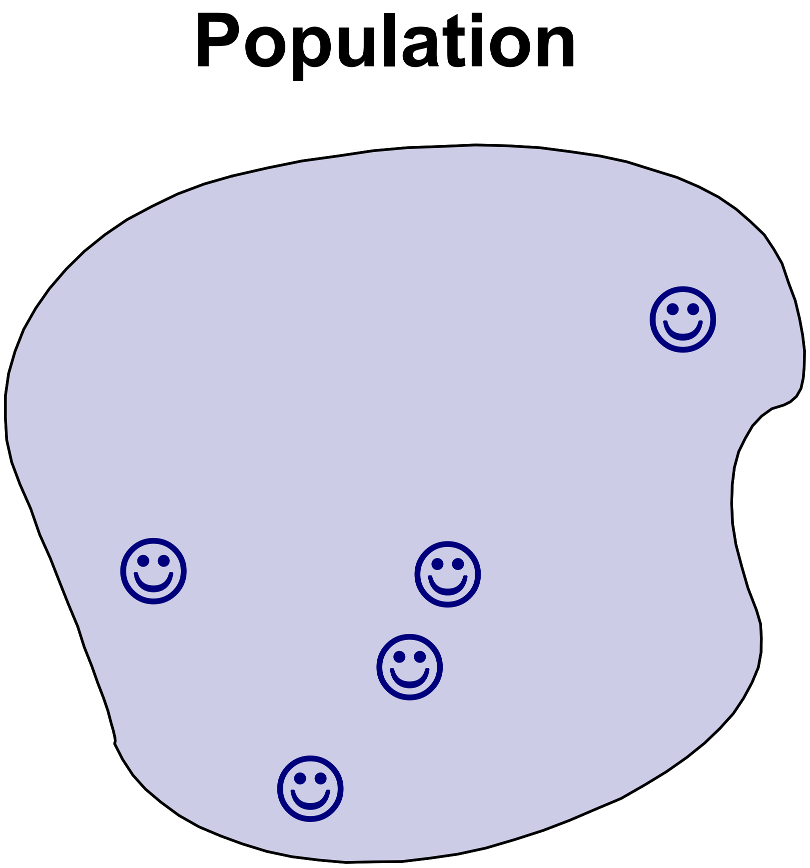
Random Error

$$Y_i \quad \beta_0 + \beta_1 X_i + \varepsilon$$

Dependent Variable

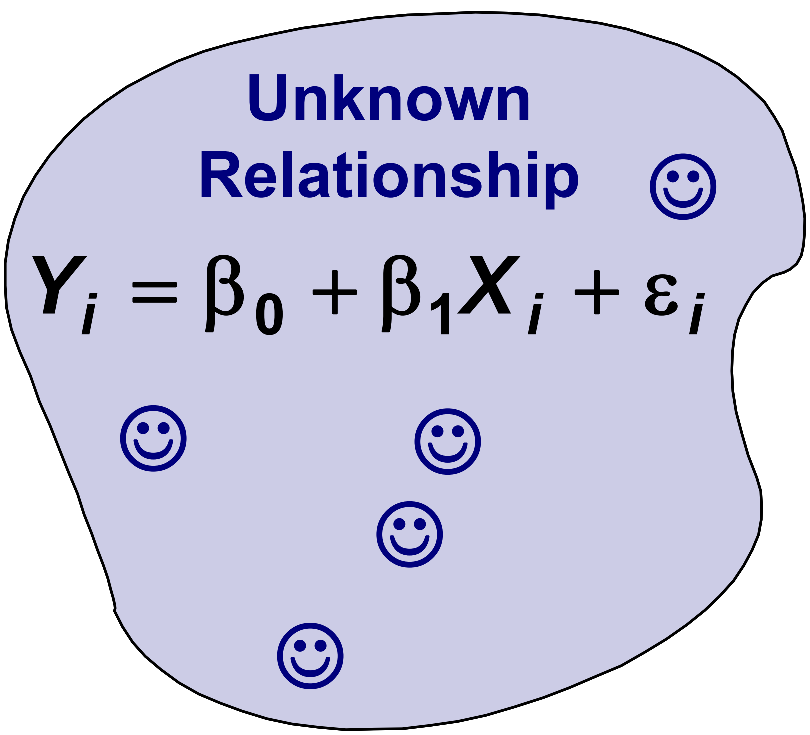
Independent Variable

Population & Sample Regression Models

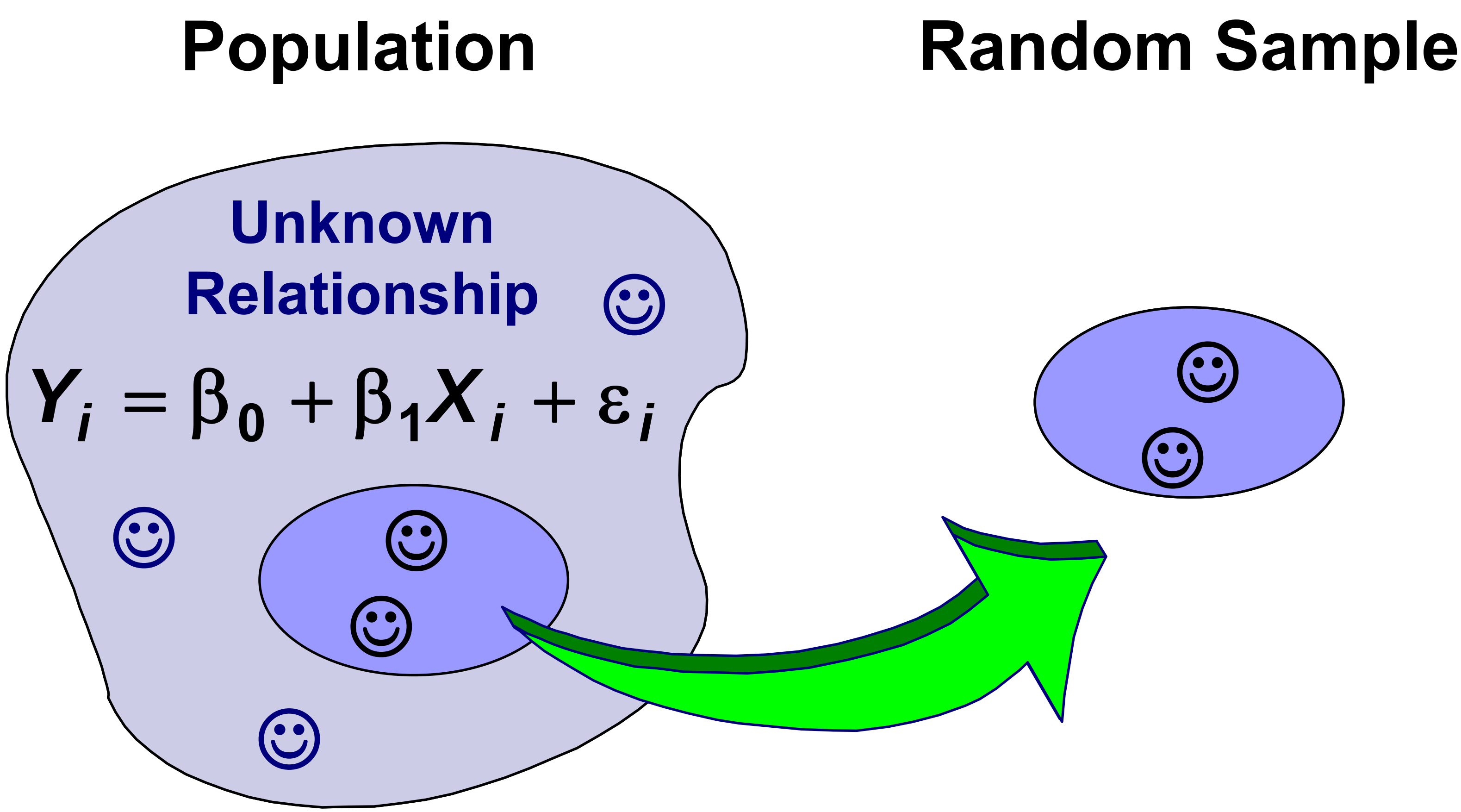


Population & Sample Regression Models

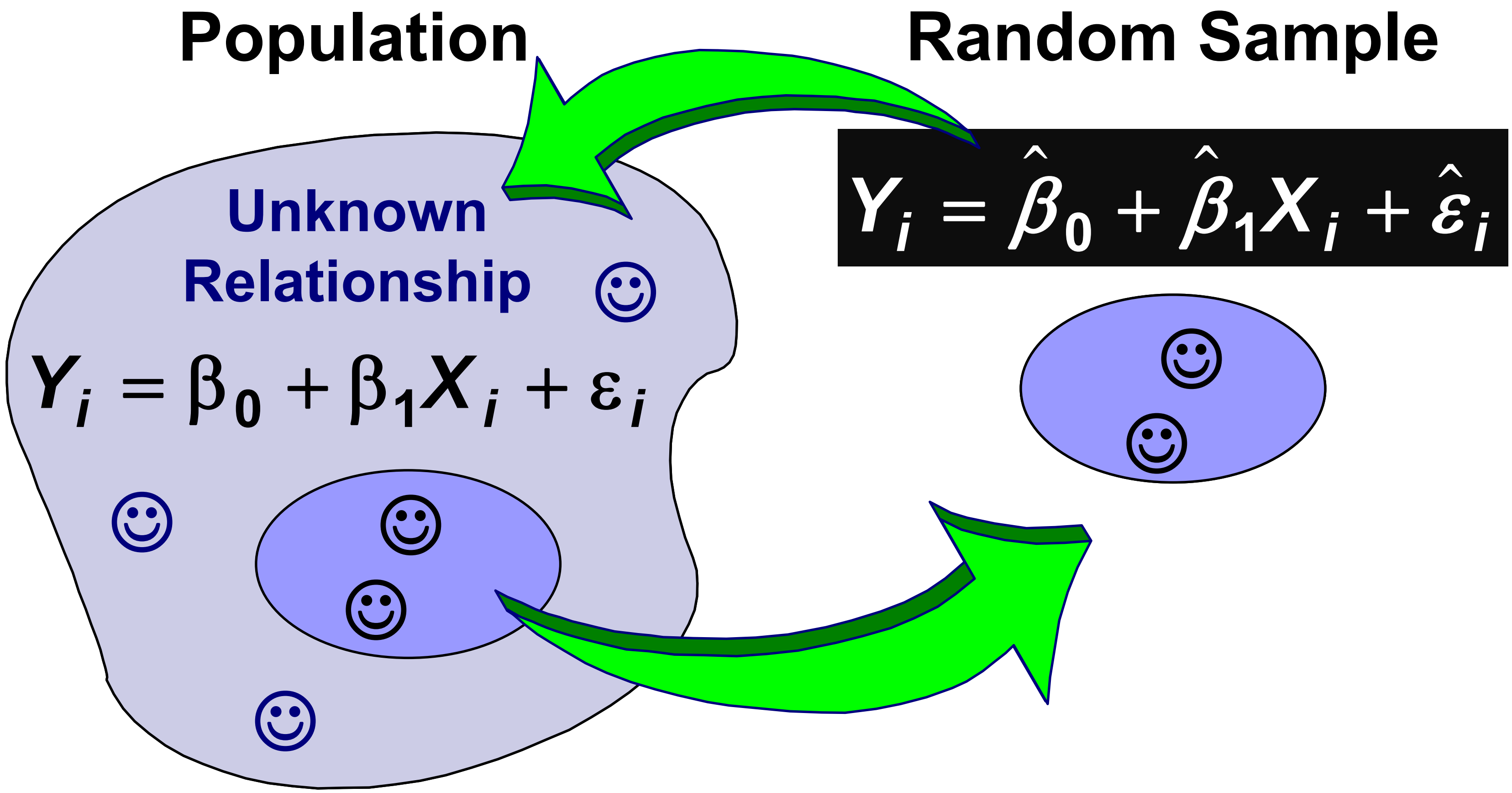
Population



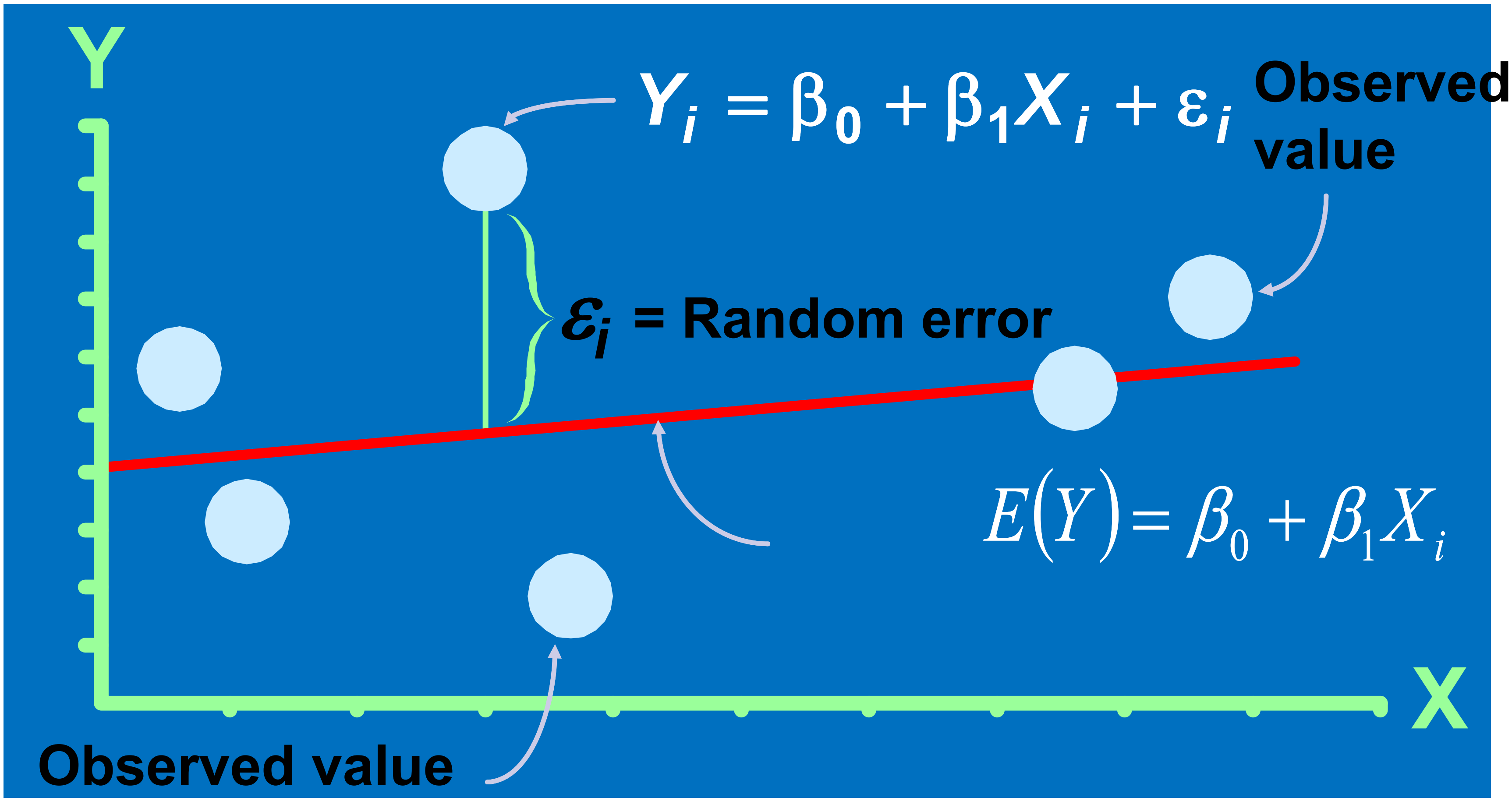
Population & Sample Regression Models



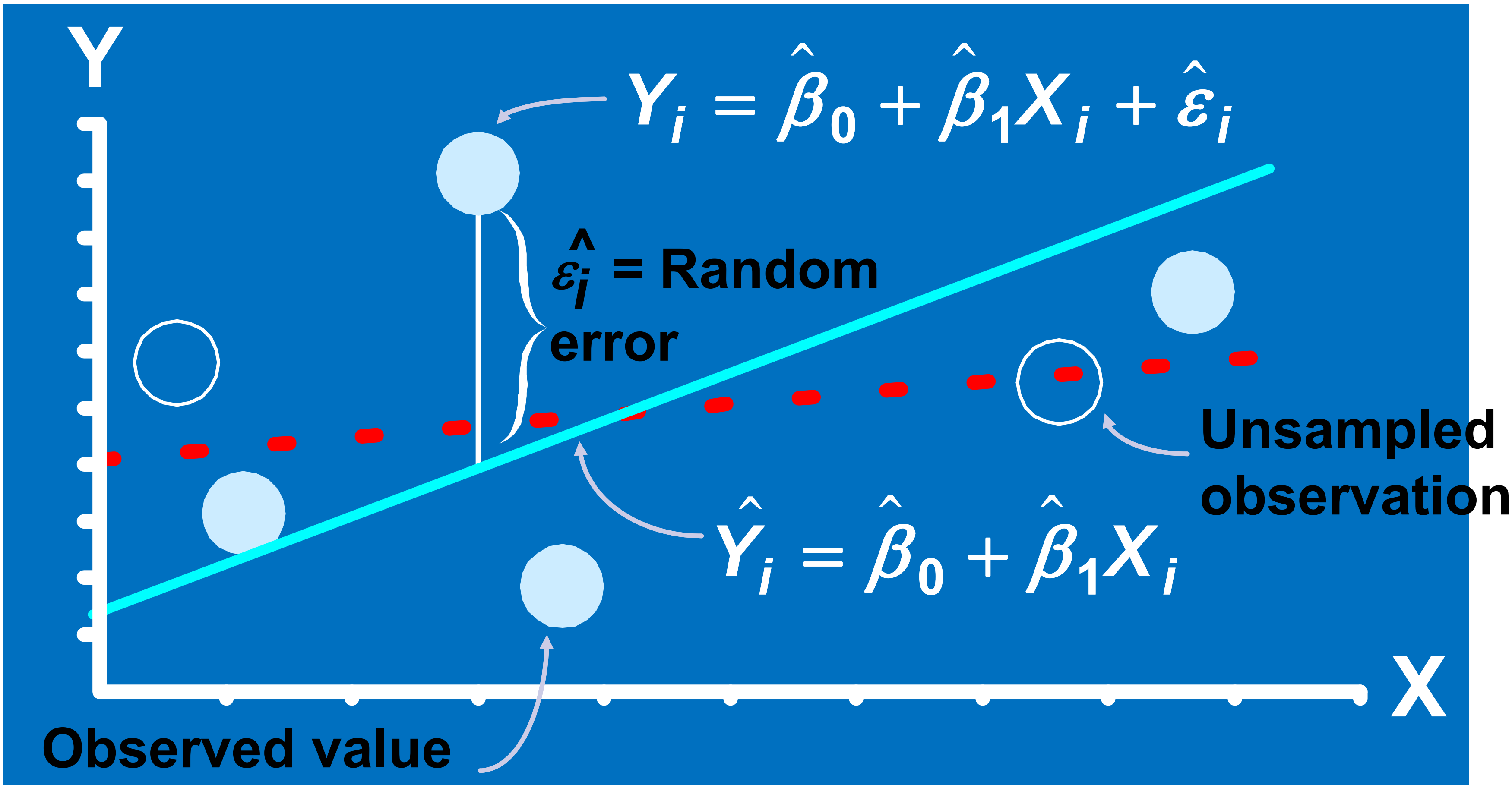
Population & Sample Regression Models



Population Linear Regression Model



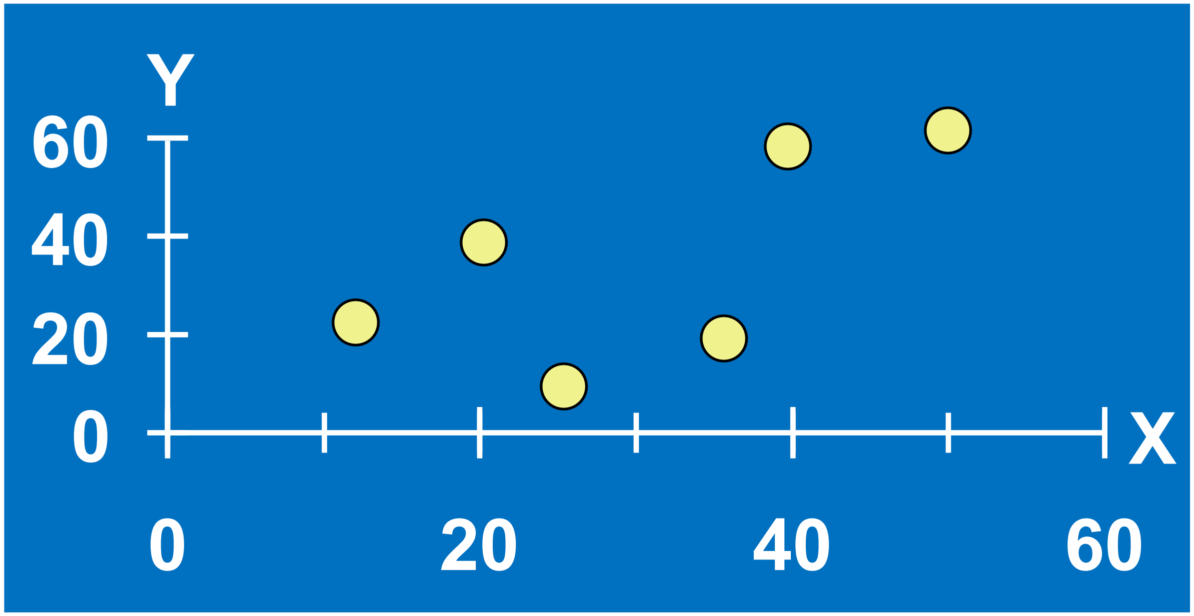
Sample Linear Regression Model



- Estimating Parameters by Least Squares Method

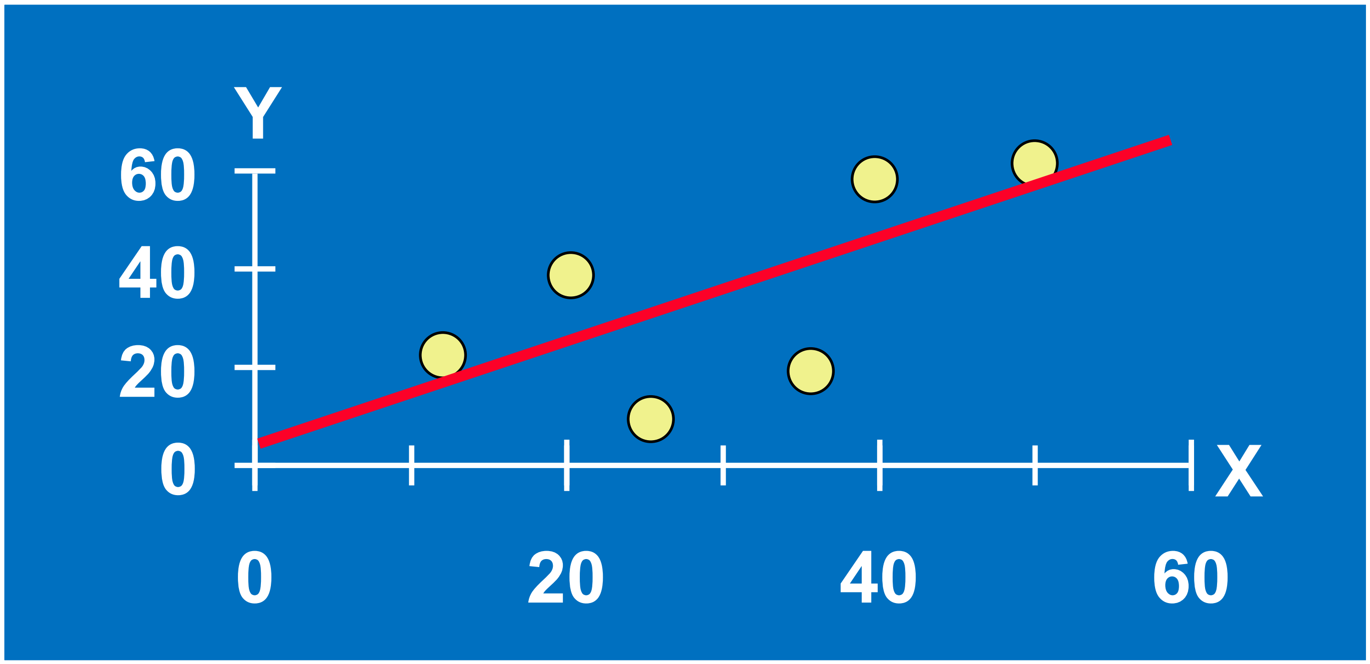
Scatter plot

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



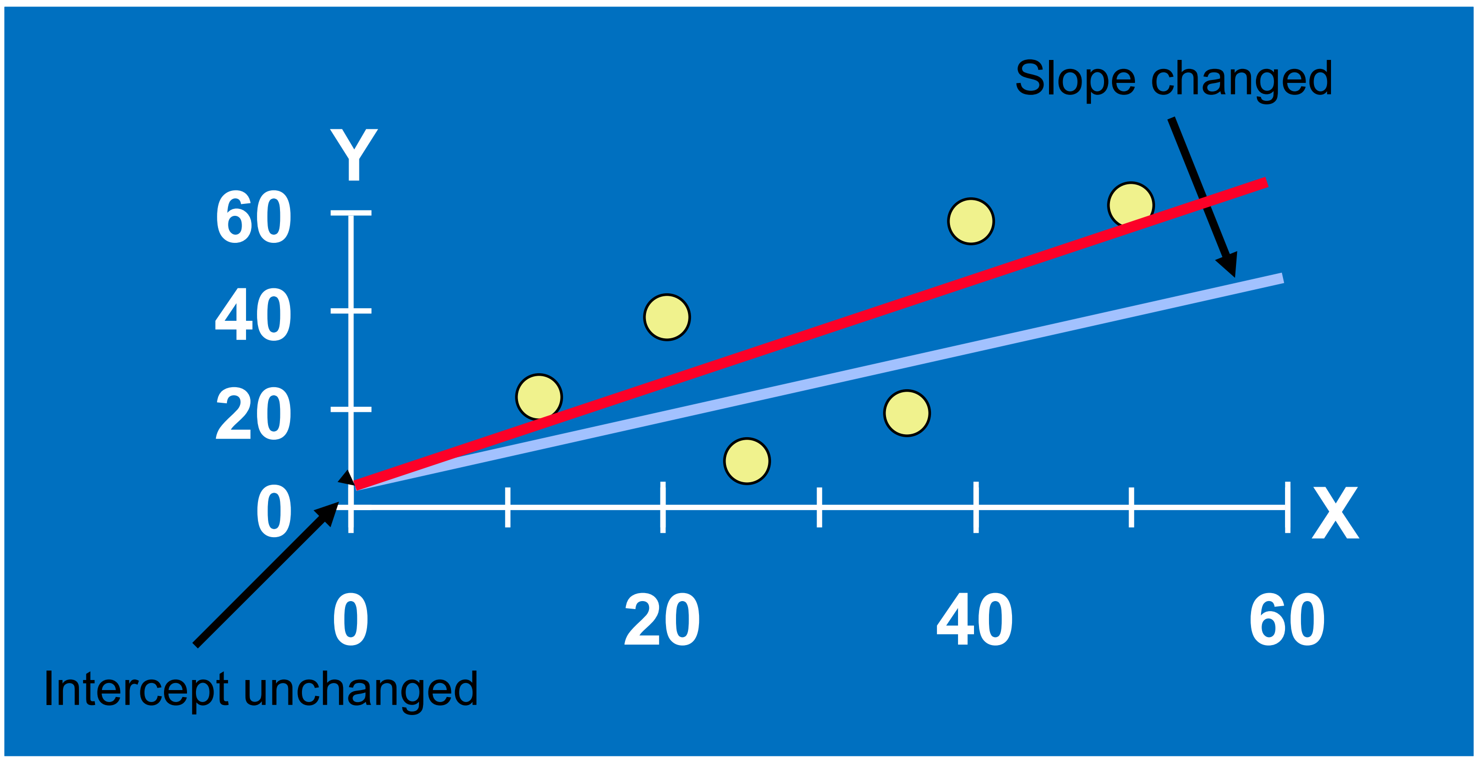
Thinking Challenge

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



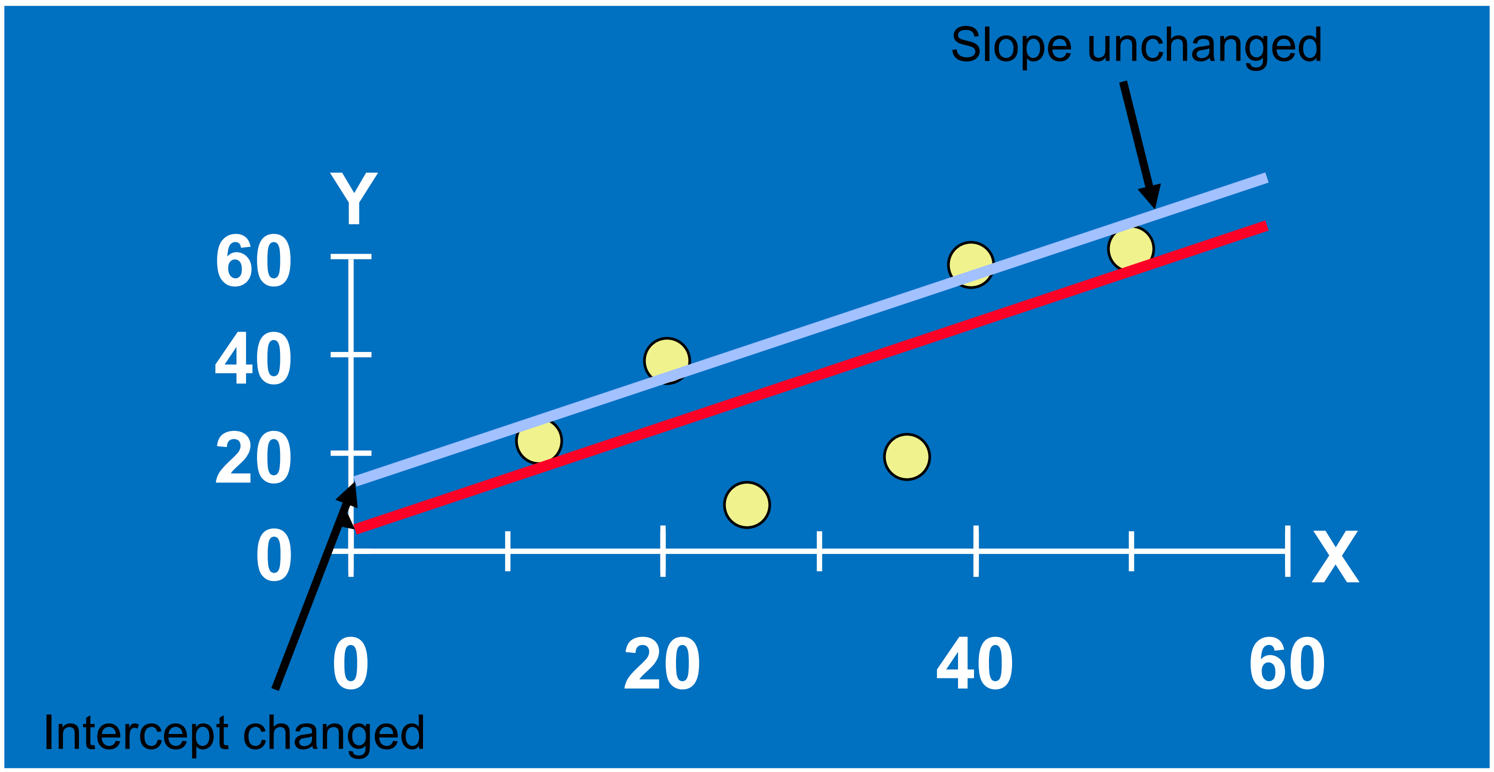
Thinking Challenge

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



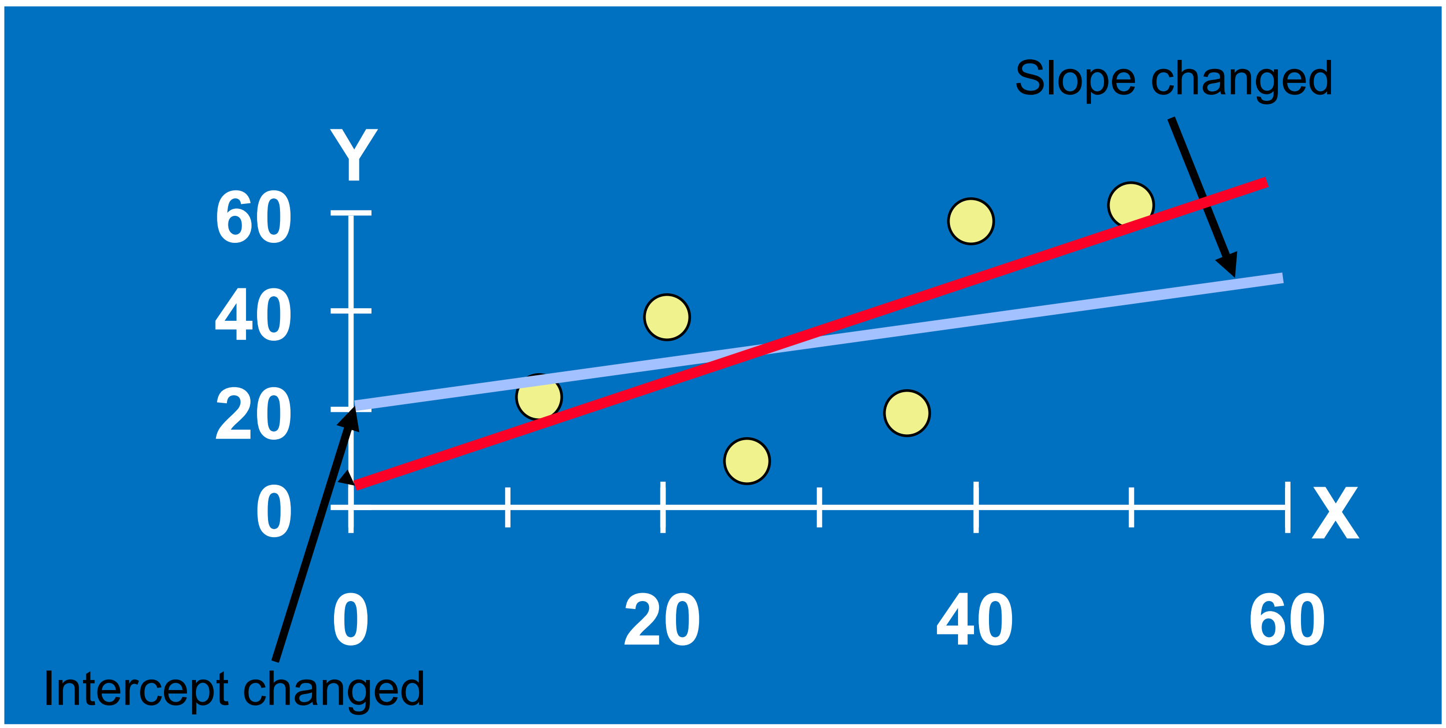
Thinking Challenge

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'?



How to draw the best fit line???.... Answer→By Least Squares Method

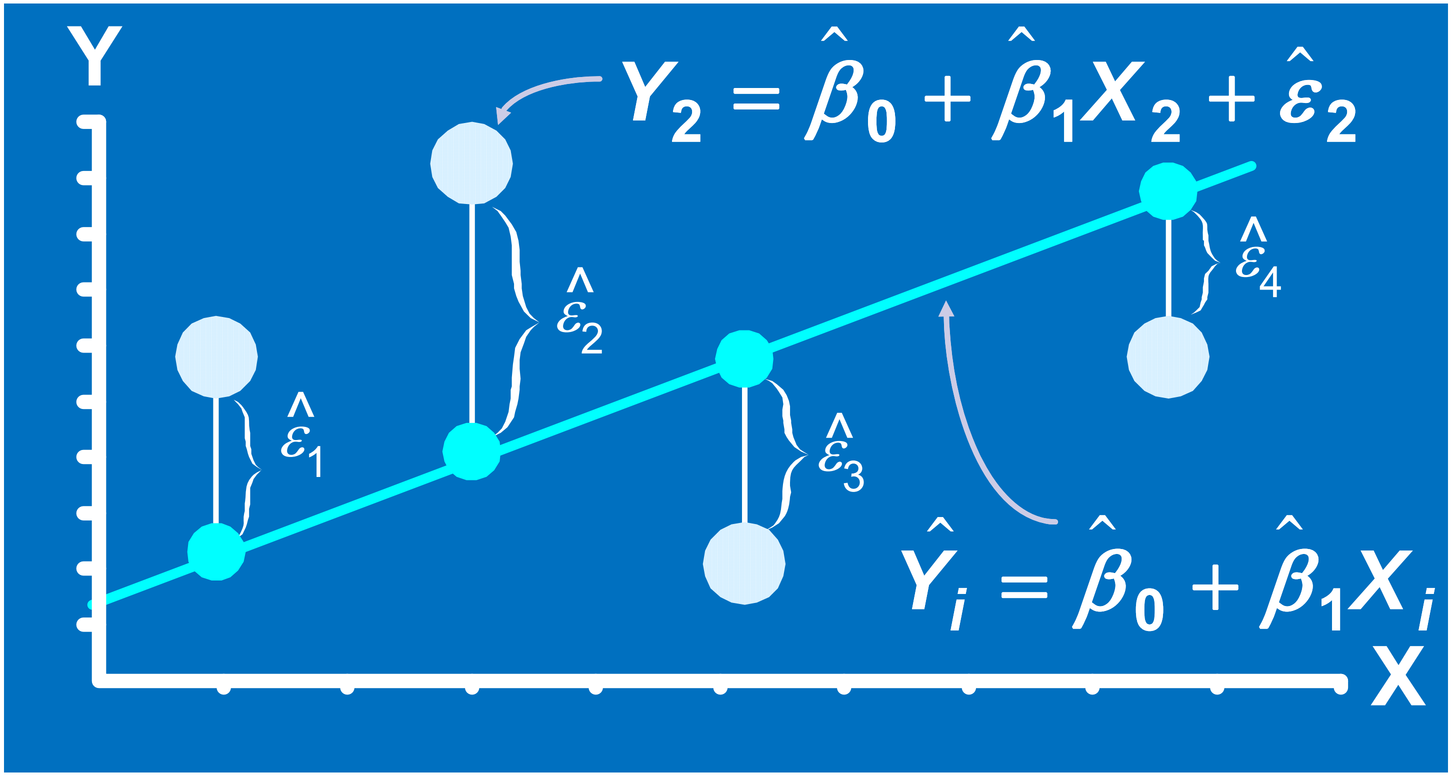
- 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum.
- But Positive Differences Off-Set Negative ones. So square errors!

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

- 2. LS methods minimizes the ***Sum of the Squared Differences (errors) (SSE)***

Least Squares Method Graphically

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



Coefficient Equations

- **Predicted/Estimated equation** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- **Sample slope**

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

$$\boxed{\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ S_{xy} &= \sum_{i=1}^{n} (x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right), \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{aligned}}$$

- **Sample Y - intercept** $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad (12.2)$$

Computing formulas for the numerator and denominator of $\hat{\beta}_1$ are

$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n \qquad S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2/n$$

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (12.3)$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0}$$

$$= -2\left(n\overline{y} - n\beta_0 - n\beta_1 \overline{x} \right)$$

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

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$
\begin{aligned}
0 &= \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i\right)^2}{\partial \beta_1} \\
&= -2 \sum x_i \left(y_i - \beta_0 - \beta_1 x_i\right) \\
&= -2 \sum x_i \left(y_i - \overline{y} + \beta_1 \overline{x} - \beta_1 x_i\right)
\end{aligned}
$$

$$
\begin{aligned}
&\beta_1 \sum x_i \left(x_i - \overline{x}\right) = \sum x_i \left(y_i - \overline{y}\right) \\
&\beta_1 \sum \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right) = \sum \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) \\
&\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}
\end{aligned}
$$

Example 1:

The manager of a car plant wishes to investigate how the plant's electricity usage depends upon the plant production. The data is given below **estimate the linear regression equation**

Production ($million) (x)	**4.51**	**3.58**	**4.31**	**5.06**	**5.64**	**4.99**	**5.29**	**5.83**	**4.7**	**5.61**	**4.9**	**4.2**
Electricity Usage (y)	**2.48**	**2.26**	**2.47**	**2.77**	**2.99**	**3.05**	**3.18**	**3.46**	**3.03**	**3.26**	**2.67**	**2.53**

Solution : You need to write the following equation

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

x	4.51	3.58	4.31	5.06	5.64	4.99	5.29	5.83	4.7	5.61	4.9	4.2	$\sum x$ =58.62
y	2.48	2.26	2.47	2.77	2.99	3.05	3.18	3.46	3.03	3.26	2.67	2.53	$\sum y$ =34.15
xy	11.18	8.09	10.65	14.02	16.86	15.22	16.82	20.17	14.24	18.29	13.08	10.63	$\sum xy$ =169.25
x^2	20.34	12.82	18.58	25.60	31.81	24.90	27.98	33.99	22.09	31.47	24.01	17.64	$\sum x^2$ =291.23

$$S_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$= 169.25 - (58.62)(34.15)/12 = 2.43045$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$= 291.23 - (58.62)^2/12 = 4.8723$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = 2.43045 / 4.8723 = 0.4988$$

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

$$= (34.15/12) - (0.4988)(58.62/12) = 0.4091$$

Estimated Regression Line $\hat{y} = 0.4091 + 0.4988x$

The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine. Determination of this number for a biodiesel fuel is expensive and time-consuming. The article "Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (*J. of Automobile Engr.*, 2009: 565–583) included the following data on x = iodine value (g) and y = cetane number for a sample of 14 biofuels. The iodine value is the amount of iodine necessary to saturate a sample of 100 g of oil. The article's authors fit the simple linear regression model to this data, so let's follow their lead.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

The necessary summary quantities for hand calculation can be obtained by placing the x values in a column and the y values in another column and then creating columns for x^2, xy, and y^2 (these latter values are not needed at the moment but will be used shortly). Calculating the column sums gives $\Sigma x_i = 1307.5, \Sigma y_i = 779.2, \Sigma x_i^2 = 128{,}913.93$, $\Sigma x_i y_i = 71{,}347.30, \Sigma y_i^2 = 43{,}745.22$, from which

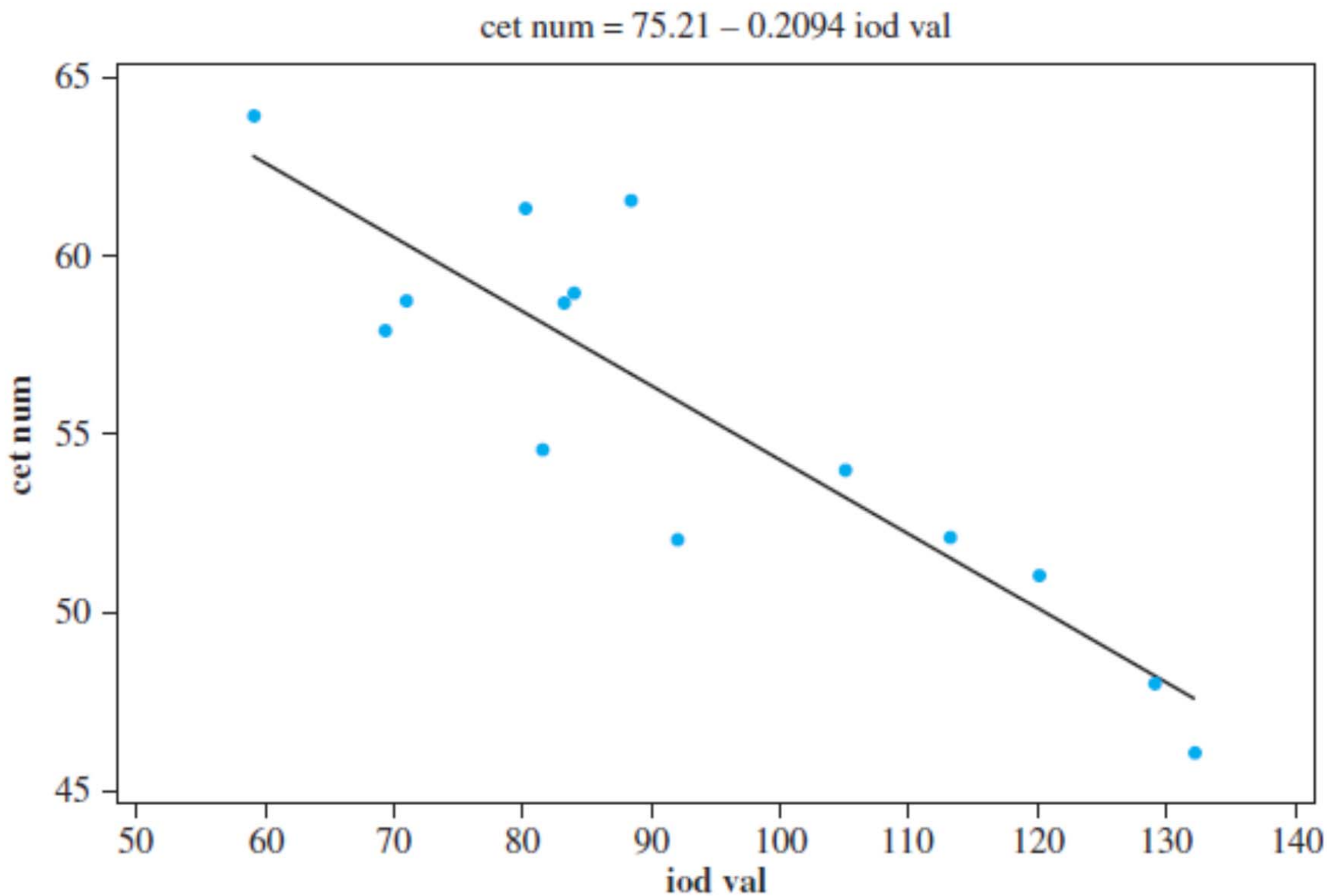
$$S_{xx} = 128{,}913.93 - (1307.5)^2/14 = 6802.7693$$

$$S_{xy} = 71{,}347.30 - (1307.5)(779.2)/14 = -1424.41429$$

The estimated slope of the true regression line (i.e., the slope of the least squares line) is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -.20938742$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 55.657143 - (-.20938742)(93.392857) = 75.212432$$



Assessing or Evaluate the Model

Methods to assess the model **(Based on Sum Squares of Errors – SSE)**

1) Standard error of estimate
2) Coefficient of determination
3) The t-test of the slope

Method 1: Standard Error of Estimate (σ^2)

- **Compute Standard Error of Estimate by** $$\boxed{\hat{\sigma}^2 = \frac{SSE}{n-2}}$$

- **Where Sum of Squares of Errors/Residual (SSE) is**

$$\boxed{SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}$$

where

$$S_{YY} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \qquad S_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \qquad S_{XX} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

- **This is an unbiased estimator for σ^2 (for Population)**

- The standard error of the slope and intercept

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}$$

- The smaller SSE the more successful is the Linear Regression Model in explaining y.

$$\hat{\sigma}^2 = \frac{\mathbf{SSE}}{\mathbf{n-2}}$$

Example 1 (Continued) :

Estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\mathbf{SSE}}{\mathbf{n-2}} = 0.0299$$

There for the Standard Error of Estimate is $\sqrt{0.0299} = 0.173$

Method 2: COEFFICIENT OF DETERMINATION

- Total Sum of Squares (SST):
  - Measure how much variance is in the dependent variable.
  - Made up of the SSE and SSR

$$\boxed{\mathbf{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

$$\text{SST} \quad = \quad \text{SSE} \quad + \quad \text{SSR}$$

SSE – Error or residual sum of squares

- measure how much of variation in dependent variable in our model **unexplained**

$$\boxed{SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

SSR - Regression Sum of Squares

- measure how much of variation in dependent variable in our model **explained**

$$\boxed{SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}$$

Method 2: COEFFICIENT OF DETERMINATION

- Coefficient of determination

$$\boxed{R^2 = SSR/SST = 1 - (SSE/SST)}$$

- proportion of variability in the observed dependent variable that is explained by the linear regression model.

- The coefficient of determination measures the **strength of that linear relationship**, denoted by $\boldsymbol{R^2}$

- The **greater** $\boldsymbol{R^2}$ the more successful is the Linear Model

Example 1 Determine the R square

$$R^2 = SSR/SST = 1 - (SSE/SST)$$

$$SST = S_{YY}$$

$$= \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 98.6967 - \frac{1}{12}(34.15)^2 = 1.51149$$

$$S_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$= 169.2532 - (58.62)(34.15)/12 = 2.43045$$

$SSE = S_{YY} - (S_{XY})^2/S_{XX} = 1.51149 - (2.43045)^2/4.8723 = 0.2991$

$R^2 = 1 - SSE/SST = 1 - (0.2991/1.51149) = 0.802$ **High Value !**

Method 3: Testing the slope

- The T test (method 3) addresses if there is enough evidence to infer linear relationship exists.
- Test the hypothesis

$H_0 : \beta 1 = 0$ **(there is no relationship between *x* and *y*)**

$H_1 : \beta 1 \neq 0$ **(the straight-line model is adequate)**

- Test Statistic: $T -$ distribution: $$T = \frac{\hat{\beta}1 - \beta 1}{\sqrt{\hat{\sigma}^2 / S_{XX}}} = \frac{\hat{\beta}1 - \beta 1}{se(b)}$$

- Critical Region: $|T| > t_{\alpha/2,\, n-2}$.

Example 1 (Cont…) : Applying this method to make inference on the slope

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

Estimated Regression Line , $\hat{\mathbf{y}} = 0.409 + 0.499\mathbf{x}$

$$\alpha = 0.05; \quad t_{\alpha/2,n-2} = t_{0.025,10} = 2.228$$

$$S_{XX} = 4.8723; \quad \hat{\sigma}^2 = 0.0299$$

$$\boxed{T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / S_{XX}}} = \frac{0.499 - 0}{0.073} = 6.37}$$

- Critical Region: $|T| > t_{\alpha/2,\, n-2}$.
- Since 6.37 > 2.228, reject H_0 , thus, the distribution of Electricity usage does depend on level of production

Results of Regression analysis using Minitab

Regression Analysis: electricity usage versus production

The regression equation is Electricity usage = 0.409 + 0.499production

Predictor	Coef	SE Coef	T	P-value
Constant	0.409	0.3860	1.06	0.314
production	0.499	0.07835	6.37	0.000

S = 0.1729 R-Sq = 80.2%

T test results

standard error of estimate

Coefficient of determination

Example 2:

The number of disk I/O's and the processor times of seven programs were measured. The data is given below

Number of disk, x	14	16	27	42	39	50	83
Processor times, y	2	5	7	9	10	13	20

i. Estimate the linear regression equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

ii. Find the standard error of estimate of this regression.

iii. Determine the coefficient of determination of this regression.

iv. Test whether the model is adequate.

Solution:

Number of disk, x	14	16	27	42	39	50	83	$\sum x$ =271
Processor times, y	2	5	7	9	10	13	20	$\sum y$ =66
xy	28	80	189	378	390	650	1660	$\sum xy$ =3375
$x2$	196	256	729	1764	1521	2500	6889	$\sum x^2$ =13855
$y2$								

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

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\sum x}{n}}$$

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$
$$= 13855 - (271)^2/7 = 3363$$

$$S_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$
$$= 3375 - (271)(66)/7 = 820$$

$$\therefore \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{820}{3363} = 0.24$$

$$\therefore \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9.43 - 0.24(38.7)$$
$$= -0.00828$$

So, the estimated regression line is

$$\hat{y} = -0.00828 + 0.24x$$

ii. The standard error of estimate is given by

$$\hat{\sigma}^2 = \frac{\mathbf{SSE}}{\mathbf{n-2}} = \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{y}_i)^2}{n-2} = \frac{\mathbf{S}_{yy} - \frac{(\mathbf{S}_{xy})^2}{\mathbf{S}_{XX}}}{n-2}$$

Where

$$\mathbf{S}_{YY} = \sum_{i=1}^{n}\mathbf{y}_i^2 - \frac{\mathbf{1}}{\mathbf{n}}\left(\sum_{i=1}^{n}\mathbf{y}_i\right)^2$$

$$= 828 - (66)^2/7 = 205.71$$

$$\therefore \quad \hat{\sigma}^2 = \frac{205.71 - (820)^2/3363}{5} = \frac{5.87}{5} = 1.174$$

There for the standard error of estimate is square root of 1.174

iii. The coefficient of determination is given by

$R^2 = SSR/SST = 1 - (SSE/SST)$

$$\boxed{\mathbf{SSE} = 5.87}$$

$$\boxed{\mathbf{SS}T = 205.71}$$

$$\boxed{\therefore \quad R^2 = 1 - \frac{\mathbf{SSE}}{\mathrm{SST}} = 1 - \frac{5.87}{205.71} = 0.9715}$$

Example 14.1

- The number of disk I/O's and processor times of seven programs were measured as: (14, 2), (16, 5), (27, 7), (42, 9), (39, 10), (50, 13), (83, 20)
- For this data: n=7, Σxy=3375, Σx=271, Σx^2=13,855, Σy=66, Σy^2=828, $\bar{x}$= 38.71, $\bar{y}$= 9.43. Therefore,

$$\begin{aligned} b_1 &= \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n(\bar{x})^2} = \frac{3375 - 7 \times 38.71 \times 9.43}{13,855 - 7 \times (38.71)^2} = 0.2438 \\ b_0 &= \bar{y} - b_1\bar{x} = 9.43 - 0.2438 \times 38.71 = -0.0083 \end{aligned}$$

- The desired linear model is:

$$\text{CPU time} = -0.0083 + 0.2438(\text{Number of Disk I/O's})$$

- END

Chapter 8
Simple Linear Regression Analysis

L2 – ANOVA table & Confidence Interval

Learning Outcomes

At the end of the lesson, the student should be able to

- Perform the ANOVA tests - to determine the significance of the model.
- Construct CI on regression parameters.
- Make prediction of a future observation and construct appropriate prediction interval on the future observation base on the regression model.

ANOVA (Analysis of Variance) Approach

- *ANOVA is* another way to test for the significance of regression Model
- Testing for the significance of regression.

Hypotheses: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

Test statistic: $F_0 = \dfrac{MS_R}{MS_E}$ where: $MS_R = \dfrac{SSR}{1} \quad MS_E = \dfrac{SSE}{n-2}$

Rejection criteria: $F_0 = \dfrac{MS_R}{MS_E} > f_{\alpha,1,n-2}$

- The ANOVA test for significance of regression is usually **summarized in table.**

ANOVA Table

Source Of variation	Sum of Squares	Degrees of freedom (df)	Mean Square (Sum of squares / df)	Computed F
Regression	SSR	1	$MS_R = SSR/1$	$F = MS_R/MS_E$
Error	SSE	n – 2	$MS_E = SSE/(n-2)$	
Total	SST	n - 1		

Source of Variance	**Sum of Squares**	**Degrees of freedom**	**Mean Square**	**Computed F**
Regression	**1.2124**	**1**	**1.2124**	**40.53**
Error	**0.2991**	**10**	**0.0299**	
Total	**1.5115**	**11**		

Example 1 (Continued) ANOVA TABLE :

Source of Variance	Sum of Squares	Degrees of freedom	Mean Square	Computed F
Regression	1.2124	1	1.2124	40.53
Error	0.2991	10	0.0299	
Total	1.5115	11		

Testing $H_0 : \beta_1 = 0, \; H_1 : \beta_1 \neq 0$

Taking the level of significance, $\alpha = 0.1$

$$f_{\alpha/2,1,n-2} = f_{0.05,1,10} = 4.96$$

$$F_0 = 40.53 > 4.96$$

Decision : **Reject H_0**. Therefore the linear model is fitted.

ANOVA TABLE (MINITAB)

Regression Analysis: electricity usage versus production

Predictor	Coef	SE Coef	T	P-value
Constant	0.409	0.3860	1.06	0.314
production	0.499	0.07835	6.37	0.000

S = 0.1729 R-Sq = 80.2%

Analysis of Variance

Source	DF	SS	MS	F	P-value
Regression	1	1.2124	1.2124	40.53	0.000
Residual Error	10	0.2991	0.0299		
Total	11	1.5115			

$F = MS_R/MS_E$

Confidence Intervals on the Model Parameters

A 100 (1- a)% confidence level on the slope β_1 in a simple linear Regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} se(\hat{\beta}_1)$$

Similarly, a 100 (1- a)% confidence level on the intercept β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2} se(\hat{\beta}_0)$$

where

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} \qquad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}$$

By using the regression equation

1. **Estimating the expected value of y for a given x**

$E(Y_0) = E(Y \mid x_0) = \beta_0 + \beta_1 x$ can be estimated by: $\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$

That means, need to estimate y_0 for a given x_0

The $(1-\alpha)$ Confidence Interval for the **Expected value** of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\boxed{\begin{gathered}[\hat{\mathbf{y}}_0 - \Delta,\ \hat{\mathbf{y}}_0 + \Delta], \quad \textbf{where} \quad \hat{\mathbf{y}}_0 = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_0, \\ \Delta = \mathbf{t}_{\frac{\alpha}{2}, n-2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(\mathbf{x}_0 - \bar{\mathbf{x}})^2}{S_{xx}}}, \quad \hat{\sigma}^2 = \frac{SSE}{n-2}\end{gathered}}$$

Example 1 (Continued) :

An estimate for the mean electricity usage when x = 5 (M) and α =0.05

Estimated Regression Line , $\hat{\mathbf{y}} = 0.409 + 0.499\mathbf{x}$

$$\hat{\mathbf{y}} = 0.409 + 0.499(5) = 2.904$$

95% Confidence Interval $[2.904 - \boldsymbol{\Delta},\ 2.904 + \boldsymbol{\Delta}]$

$$\Delta = t_{\alpha/2,n-2}(\hat{\sigma})\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}}} = t_{0.025,10}(0.1729)\sqrt{\frac{1}{12} + \frac{(5-4.885)^2}{4.8723}}$$

$$= (2.228)(0.0507) = 0.113$$

$$[2.904 - 0.113,\ 2.904 + 0.113] = [2.791,\ 3.017]$$

$$\hat{\sigma}^2 = \frac{\mathbf{SSE}}{\mathbf{n-2}} = 0.0299$$

<u>Interpretation</u>: with a monthly production of \$5 million, the *Expected* electricity usage is between 2.8 and 3.0kWh

By using the regression equation

2. Predicting the Particular Value of y for a given x

A value of $Y_0 = Y(x_0)$ can be estimated by: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The $(1-\alpha)$ **Confidence Interval for $Y_0 = Y(x_0)$**

$$[\hat{y}_0 - \Delta,\ \hat{y}_0 + \Delta]\,, \quad where \quad \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0,$$

$$\Delta = t_{\frac{\alpha}{2},n-2}.\hat{\sigma}.\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\ ,\ \hat{\sigma}^2 = \frac{SSE}{n-2}$$

Sometimes called prediction interval

Example 1 (Continued) :

Prediction for the electricity usage when x = 5

Estimated Regression Line , $\hat{\mathbf{y}} = 0.409 + 0.499\mathbf{x}$

$$\hat{\mathbf{y}} = 0.409 + 0.499(5) = 2.904$$

95% Confidence Interval $[2.904 - \boldsymbol{\Delta},\ 2.904 + \boldsymbol{\Delta}]$

$$\Delta = t_{\alpha/2,n-2}(\hat{\sigma})\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}} = t_{0.025,10}(0.1729)\sqrt{\frac{13}{12} + \frac{(5-4.885)^2}{4.8723}}$$

$$= (2.228)(0.1799) = 0.401$$

$$[2.904 - 0.401,\ 2.904 + 0.401] = [2.503,\ 3.305]$$

$$\hat{\sigma}^2 = \frac{\mathbf{SSE}}{\mathbf{n-2}} = 0.0299$$

The prediction interval shows that if next month's production target is $5 million, then with 95% confidence next month's electricity usage will be somewhere between 2.5 and 3.3 kWh

x_1

Example 2 :

An engineer at a semiconductor company wants to model the relationship between the device HFE (y) and the parameter Emitter - RS (x_1). Data for Emitter - RS was first collected and a statistical analysis is carried out and the output is displayed in the table given.

a) Estimate HFE when the Emitter - RS is 14.5.

b) Obtain a 95 % confidence interval for the true slope β.

c) Test for significance of regression for $\alpha = 0.05$.

x_1

Regression Analysis: y = 1075.2 – 63.87x

Predictor	Coef	SE Coef	T	P-value
Constant	1075.2	121.1	8.88	0.000
x	-63.87	**8.002**	-7.98	0.000

S = 19.4   R-Sq = 0.78

Analysis of variance

Source	DF	SS	MS	F
Regression	1	23965	23965	63.70
Residual	18	6772	376	
Total	19	30737		

Solution :

a) Estimate HFE when the Emitter - RS is 14.5.

$$\hat{y} = 1075.2 - [(14.5)(63.87)] = 149.085$$

b) Obtain a 95 % confidence interval for the true slope β.

$$t_{0.025,18} = 2.101$$

$$\hat{\beta}_1 - t_{\alpha/2,n-2} se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2} se(\hat{\beta}_1)$$

$$-63.87 - t_{0.025,18}(8.002) \le \beta_1 \le -63.87 + t_{0.025,18}(8.002)$$

$$-80.682 \le \beta_1 \le -47.058$$

c) Test for significance of regression for $\alpha = 0.05$.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

From ANOVA, reject H_0 if $F > f_{\alpha,1,n-2}$

$$\text{Reject } H_0 \ \textit{if} \ F = 63.87 > f_{0.05,1,18} = 4.41$$

Conclusion: since 63.87 > 4.41, we reject H_0 conclude that there is a significance linear relationship.

- END