EE 325

Two-Variable Regression Model: The Problem of Estimation

ee325 2/2012 (Ajarn Kaewkwan Tangtipongkul)

The Two-Variable PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

The Two-Variable SRF:

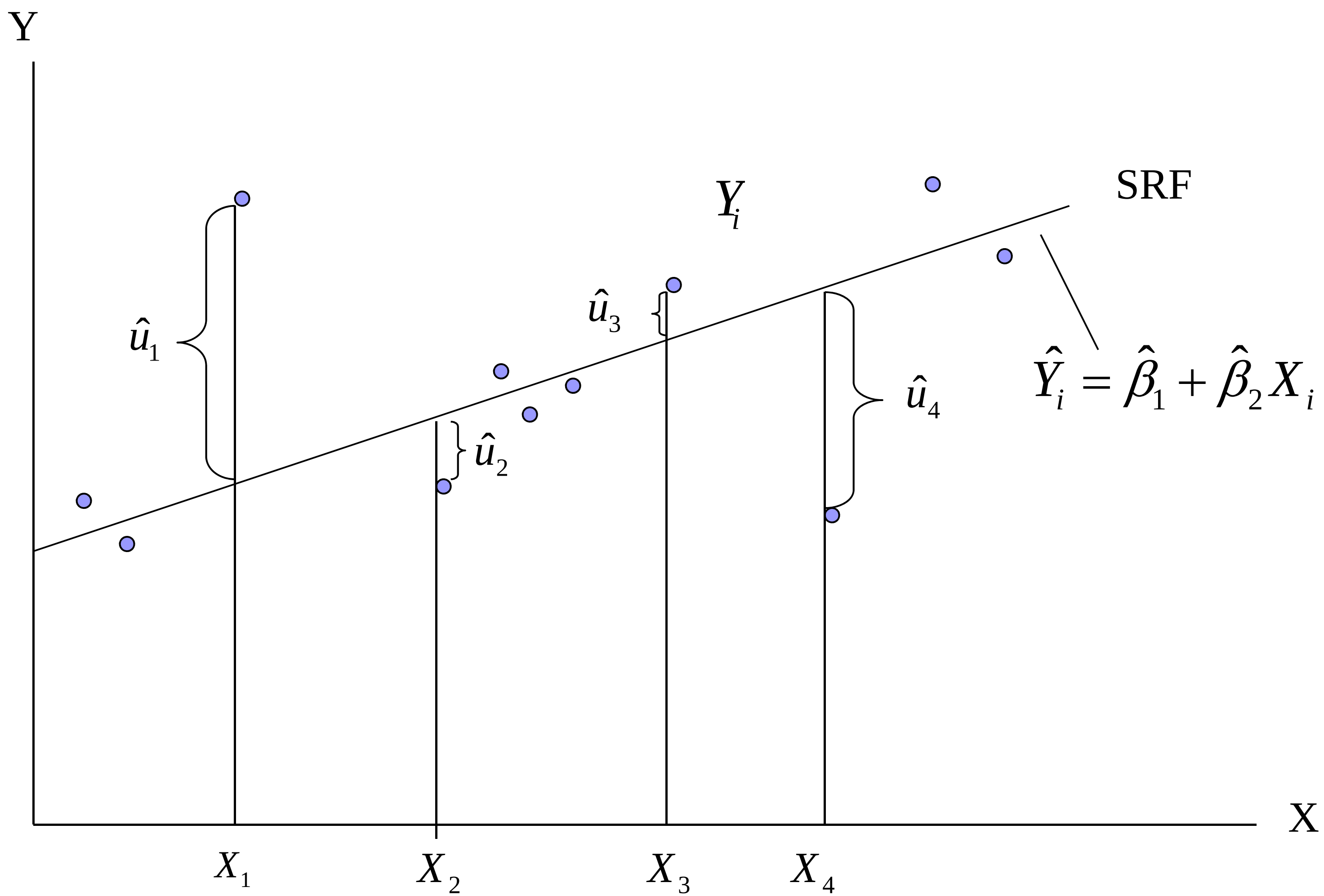
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

$$= \hat{Y}_i + \hat{u}_i$$

$\hat{Y}_i$ is the estimated (conditional mean) value of Y_i

$$\hat{u}_i = Y_i - \hat{Y}_i$$
$$= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

Ordinary Least Squares (OLS)



Ordinary Least Squares (OLS)

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$$
$$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

The principle or the method of least squares chooses $\hat{\beta}_1 and \hat{\beta}_2$ in such a manner that, for a given sample or set of data, $\sum \hat{u}_i^2$ is as small as possible

$$\frac{\partial\left(\sum \hat{u}_i^2\right)}{\partial\hat{\beta}_1} = -2\sum\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = -2\sum \hat{u}_i$$

$$\frac{\partial\left(\sum \hat{u}_i^2\right)}{\partial\hat{\beta}_2} = -2\sum\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) X_i = -2\sum \hat{u}_i X_i$$

Setting these equation to zero

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

Where n is the sample size. These simultaneous equations are known as the **Normal Equation**

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$= \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2}$$

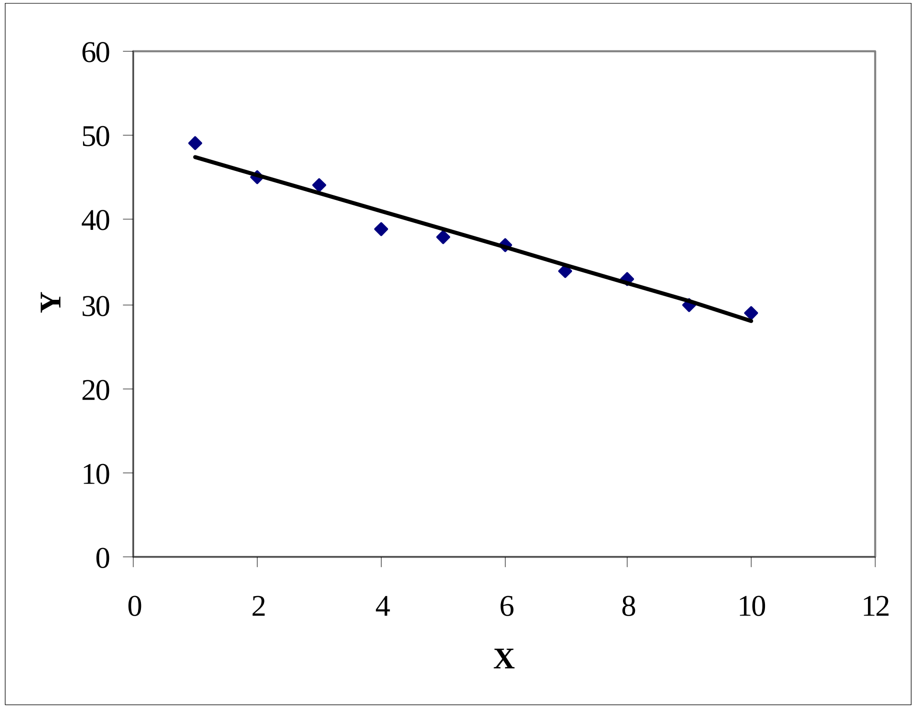
$$= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2}$$

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$= \bar{Y} - \hat{\beta}_2 \bar{X}$$

Example ☺

Y	X
49	1
45	2
44	3
39	4
38	5
37	6
34	7
33	8
30	9



$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{-178}{82.5} \approx -2.1576$$

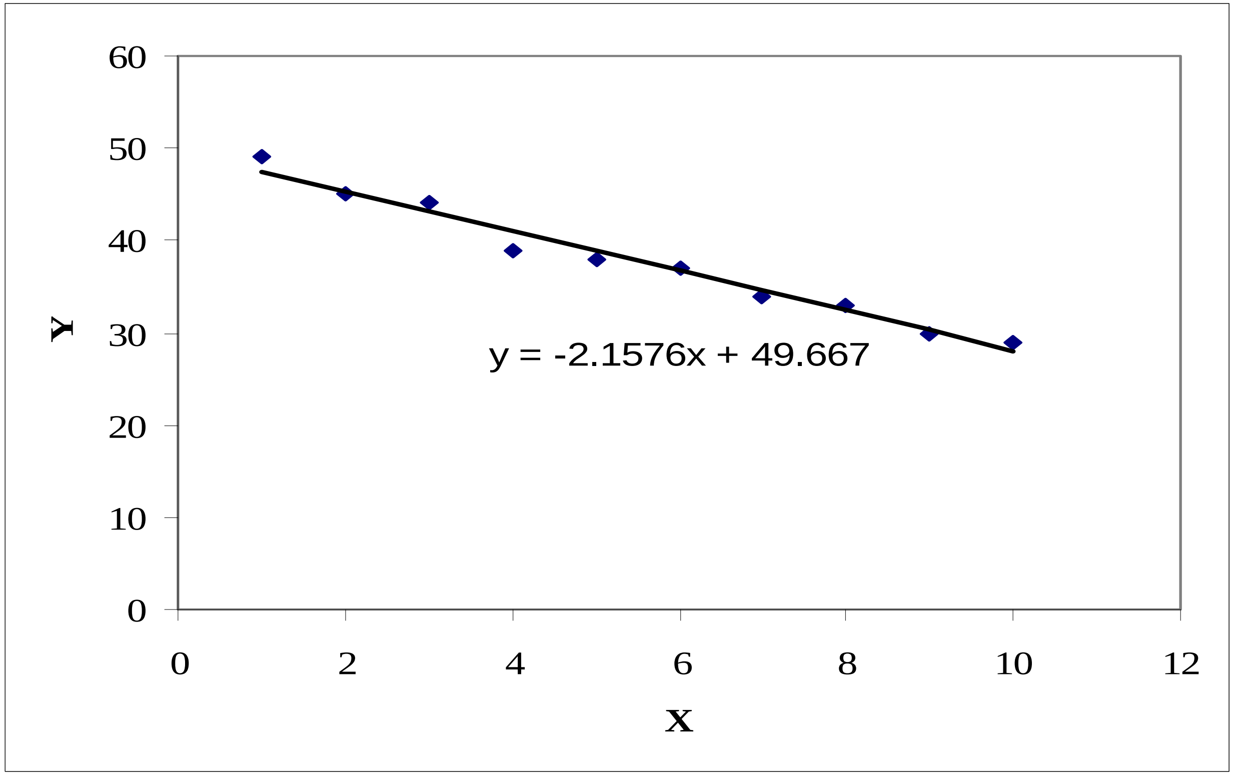
$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 37.8 - (-2.1576)(5.5) = 49.667$$

$$\bar{X} = 5.5$$

$$\bar{Y} = 37.8$$

Y	X	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
49	1				
45	2				
44	3				
39	4				
38	5				
37	6				
34	7				
33	8				
30	9				
29	10				

$$\hat{Y}_i = 49.667 - 2.1576 X_i$$



Practice I

A Random Sample from the Population of Table 2.1

Weekly consumption expenditure $ (Y)	Weekly income $(X)
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

$$\hat{Y} = 0.5091X_i + 24.455$$

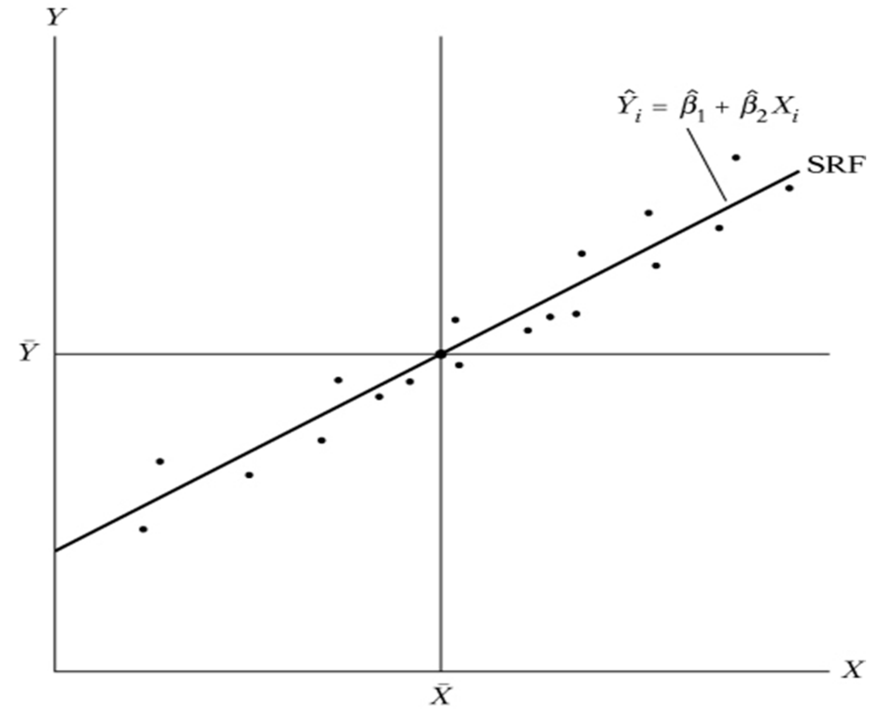
Practice II

Another Random Sample from the Population of Table 2.1	
Weekly consumption expenditure $ (Y)	Weekly income $(X)
Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

$$\hat{Y} = 0.5761 X_i + 17.17$$

The regression line has the following properties:

- Passes through the sample means of Y and X



- The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y for

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

- The mean value of the residuals $\hat{u}_i$ is zero
- The residuals $\hat{u}_i$ are uncorrelated with the predicted $\hat{Y}_i$
- The residuals $\hat{u}_i$ are uncorrelated with X_i

Classical Linear regression model (CLRM)

The assumptions underlying the method of least squares:

1. Linear regression model
2. Fixed X values or X values independent of the error term
3. Zero mean value of disturbance
4. Homoscedasticity or Constant Variance of u_i
5. No autocorrelation between the disturbances u_i
6. The number of observations n must be greater than the number of parameters to be estimated
7. The nature of X variables

Linear regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

- Linear in the parameters
- May or may not be linear in the variables

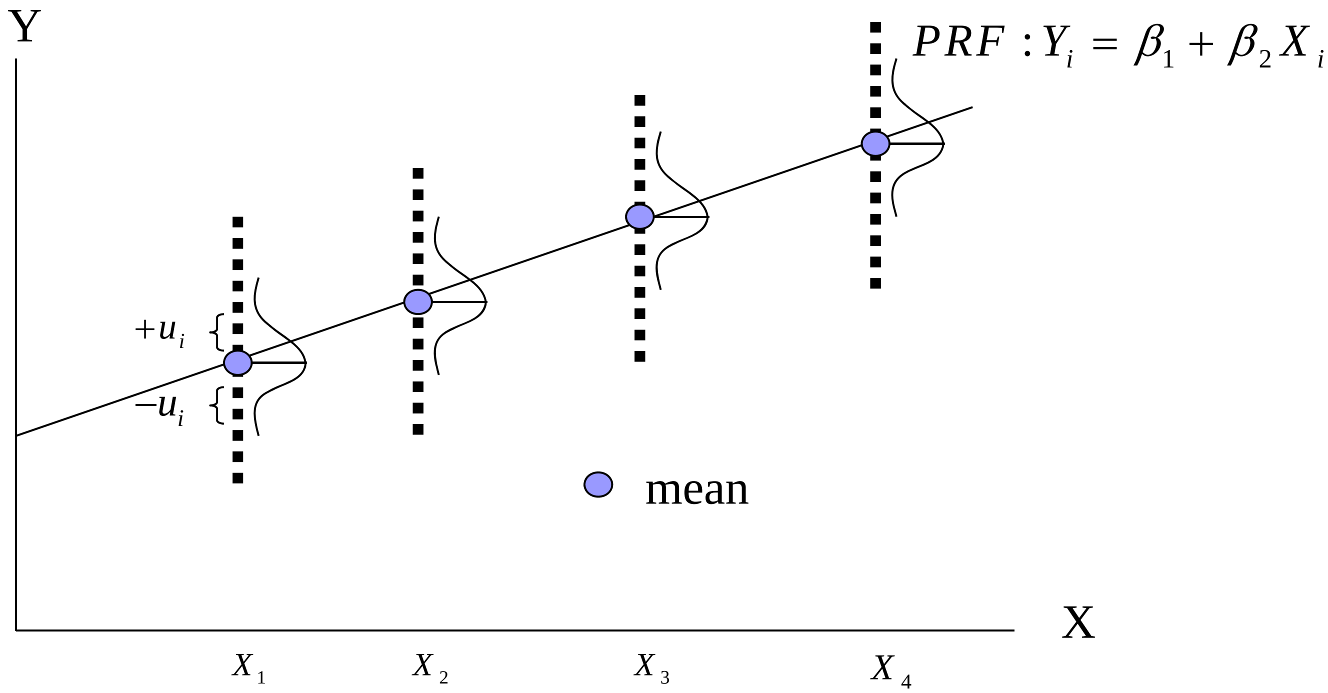
X values independent of the error term

$$Cov(X_i, u_i) = 0$$

Zero mean value of disturbance term

$$E(u_i \mid X_i) = 0$$

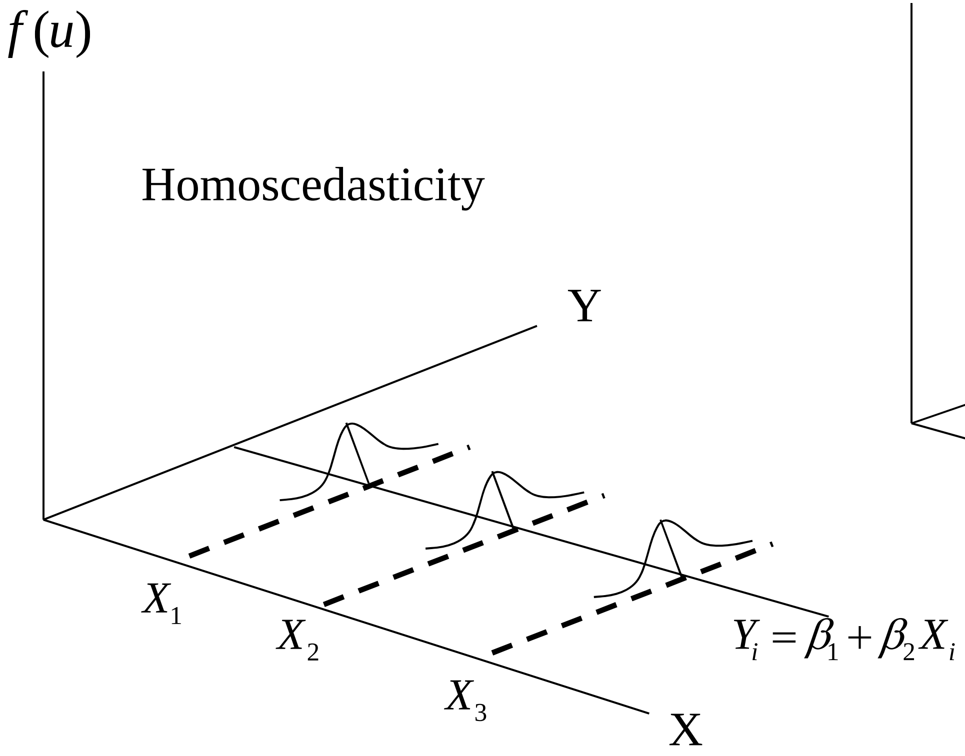
$$E(u_i) = 0$$



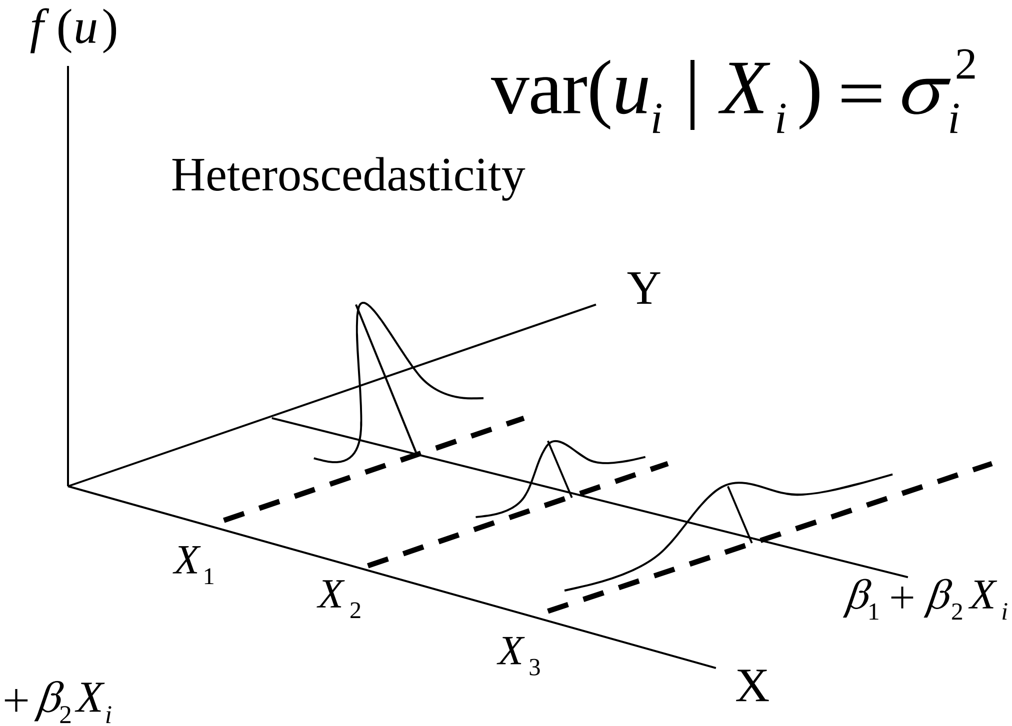
Homoscedasticity

$$
\begin{aligned}
\operatorname{var}(u_i) &= E[u_i - E(u_i \mid X_i)]^2 \\
&= E(u_i^2 \mid X_i), \textit{because of assumption } 3 \\
&= E(u_i^2), \textit{if } X_i \textit{ are nonstochastic} \\
&= \sigma^2
\end{aligned}
$$

Homoscedasticity vs. Heteroscedasticity



$$\text{var}(u_i \mid X_i) = \sigma^2$$



$$\text{var}(u_i \mid X_i) = \sigma_i^2$$

Homoscedasticity v.s. Heteroscedasticity

Homoscedasticity

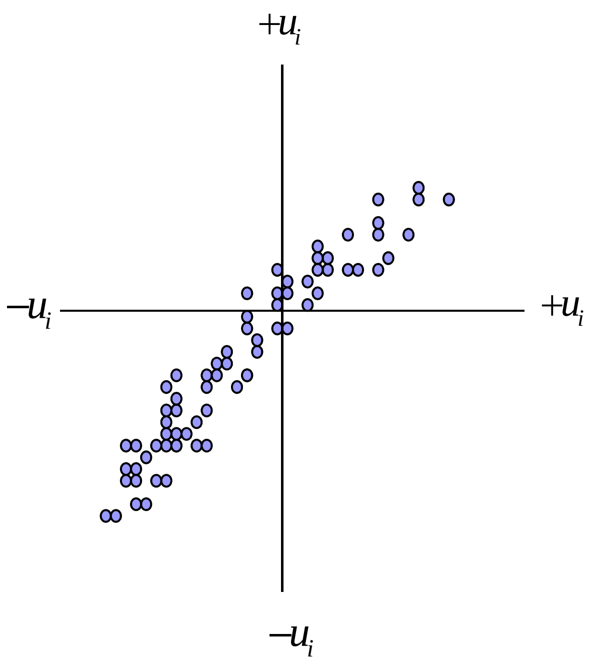
- Equal variance
- The variation around the regression line is the same across the X values

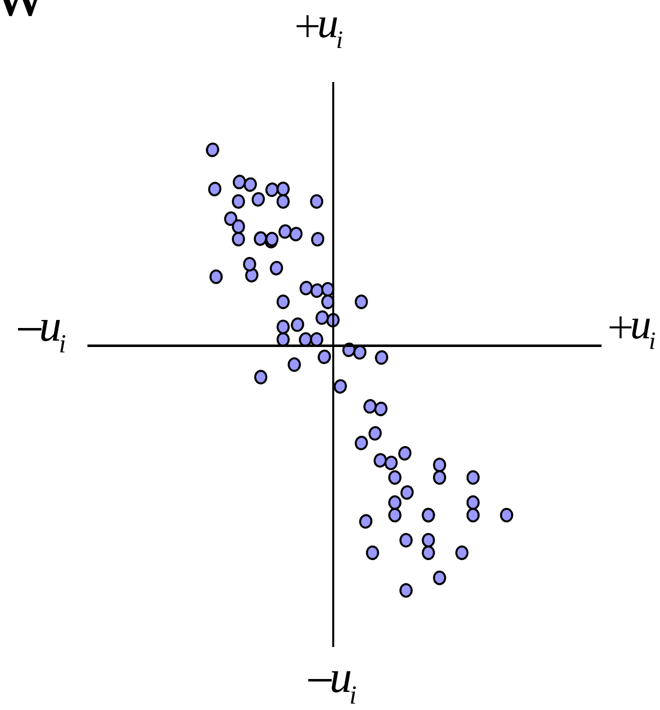
Heteroscedasticity

- Unequal variance

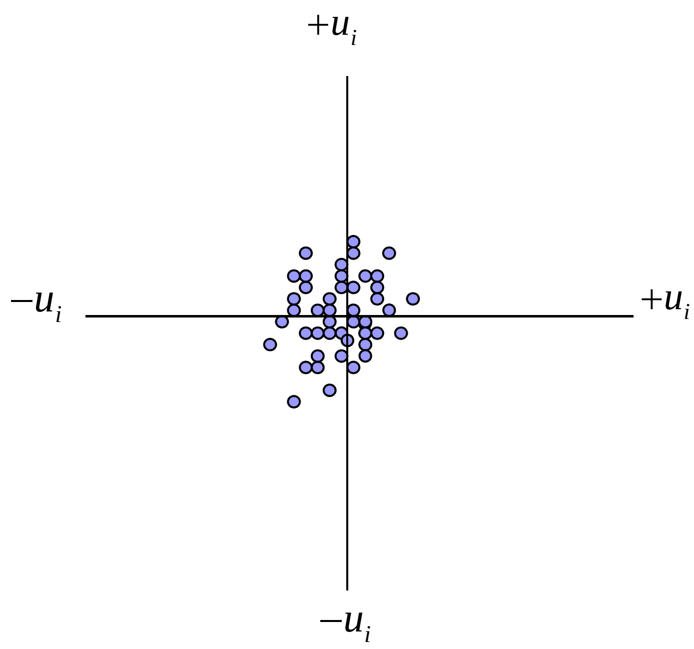
No autocorrelation between the disturbances

Given X_i, the deviations of any two Y values from their mean value do not exhibit patterns such as those shown in figures below





No autocorrelation between the disturbances



No autocorrelation between the disturbances

Given any two X values, X_i and $X_j (i \neq j)$, the correlation between any two u_i and $u_j (i \neq j)$ is zero.

$$\mathrm{cov}(u_i, u_j \mid X_i, X_j) = 0$$

$$\mathrm{cov}(u_i, u_j) = 0, if \quad X \quad is \ nonstochastic$$

Where i and j are two different observation and where cov means covariance

The number of observations n must be greater than the number of parameters to be estimated

- The number of observations must be greater than the number of explanatory variables

The nature of X variables

- The X values in a given sample must not all be the same

- No outliers in the X values

Standard Errors of Least-Squares Estimates

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

$$se(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

$$\operatorname{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\sigma$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

$\hat{\sigma}^2$ is the OLS estimator of the true but unknown σ^2

The expression n-2 is known as the number of degrees of freedom

$\sum \hat{u}_i^2$ is the sum of the residuals squared or residual sum of squares (RSS)

$$COV(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X}\ \mathrm{var}(\hat{\beta}_2)$$

$$= -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right)$$

$$= \frac{-\bar{X}\sigma^2}{\sum (X_i - \bar{X})^2}$$

Example ☺

Y	X
49	1
45	2
44	3
39	4
38	5
37	6
34	7
33	8
30	9

$$\hat{\beta}_1 = 49.667$$

$$\hat{\beta}_2 = -2.1576$$

$$\bar{X} = 5.5$$

Y	X	$\boxed{\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i}$	$\boxed{\hat{u}_i}$	$\boxed{\hat{u}_i^2}$	$\boxed{(X_i - \bar{X})^2}$	$\boxed{X_i^2}$
49	1	47.5094	1.4906	2.2219	20.25	1
45	2	45.3518	-0.352	0.1238	12.25	4
44	3	43.1942	0.8058	0.6493	6.25	9
39	4	41.0366	-2.037	4.1477	2.25	16
38	5	38.879	-0.879	0.7726	0.25	25
37	6	36.7214	0.2786	0.0776	0.25	36
34	7	34.5638	-0.564	0.3179	2.25	49
33	8	32.4062	0.5938	0.3526	6.25	64
30	9	30.2486	-0.249	0.0618	12.25	81
29	10	28.091	0.909	0.8263	20.25	100
			Total	9.5515	82.5	385

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \qquad \frac{9.5515}{10-2} = 1.1939$$

$$\operatorname{var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2} \qquad \frac{1.1939}{82.5} = 0.0145$$

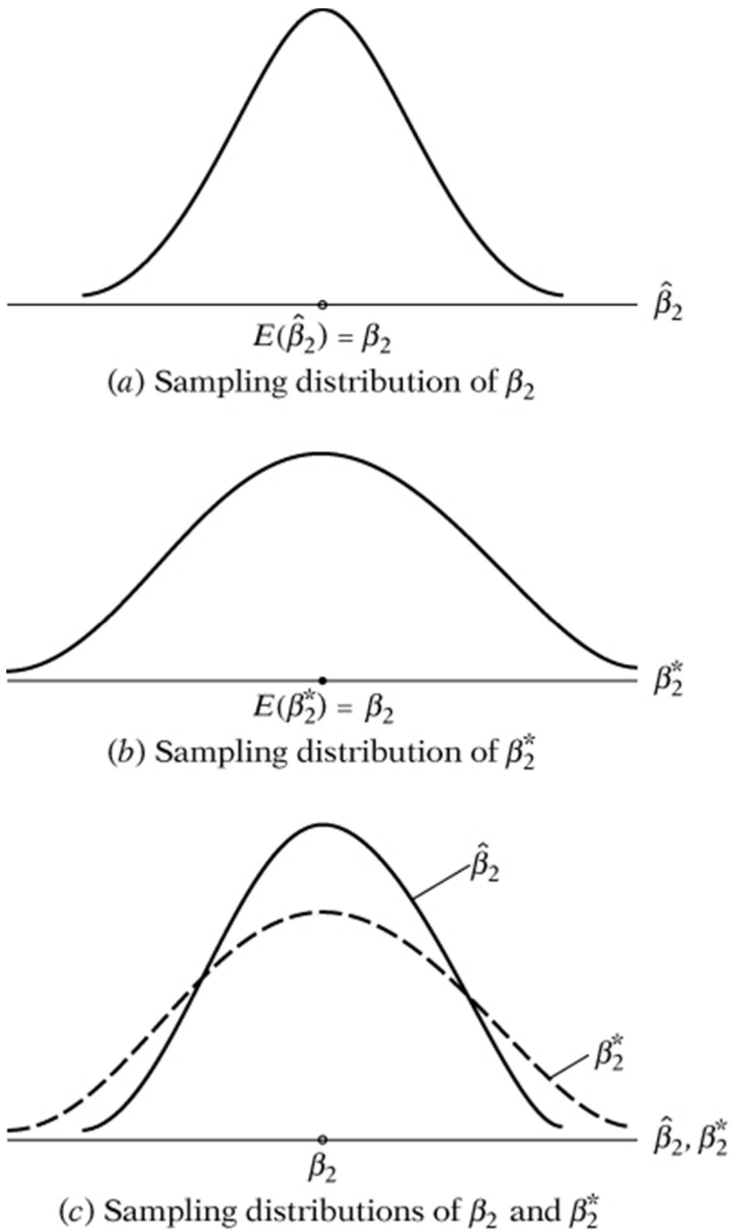
$$\operatorname{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2}\hat{\sigma}^2 = \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}\hat{\sigma}^2 \quad \frac{(1.1939)(385)}{(10)(82.5)} = 0.5572$$

$$COV(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X}\operatorname{var}(\hat{\beta}_2) \quad -(5.5)(0.0145) = -0.07975$$

Properties of Least-Squares Estimators: The Gauss-Markov Theorem

Best Linear Unbiased Estimator (BLUE) of β_2 :

- It is linear, that is, a linear function of a random variable, such as the dependent variable Y in the regression model
- It is unbiased $E(\hat{\beta}_2) = \beta_2$
- Efficient estimator – an unbiased estimator with the least variance



(*a*) Sampling distribution of β_2

(*b*) Sampling distribution of β_2^*

(*c*) Sampling distributions of β_2 and β_2^*

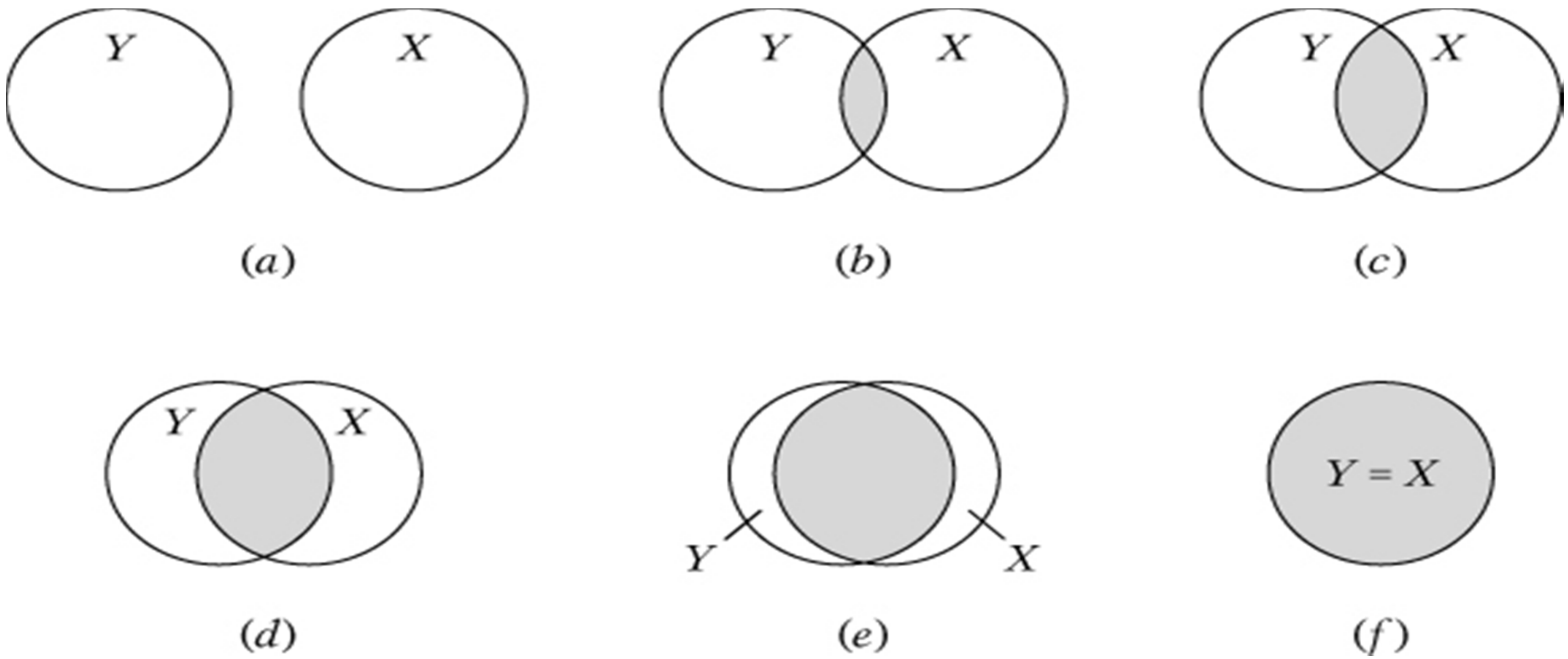
Tangtipongkul)

The Coefficient of Determination r^2

- A measure of goodness of fit

- A summary measure that tells how well the sample regression line fits the data

- Measures the proportion or percentage of the total variation in Y explained by the regression model


Y
X
(a)
Y
X
(b)
Y
X
(c)
Y
X
(d)
Y
X
(e)
Y = X
(f)

To compute this r^2

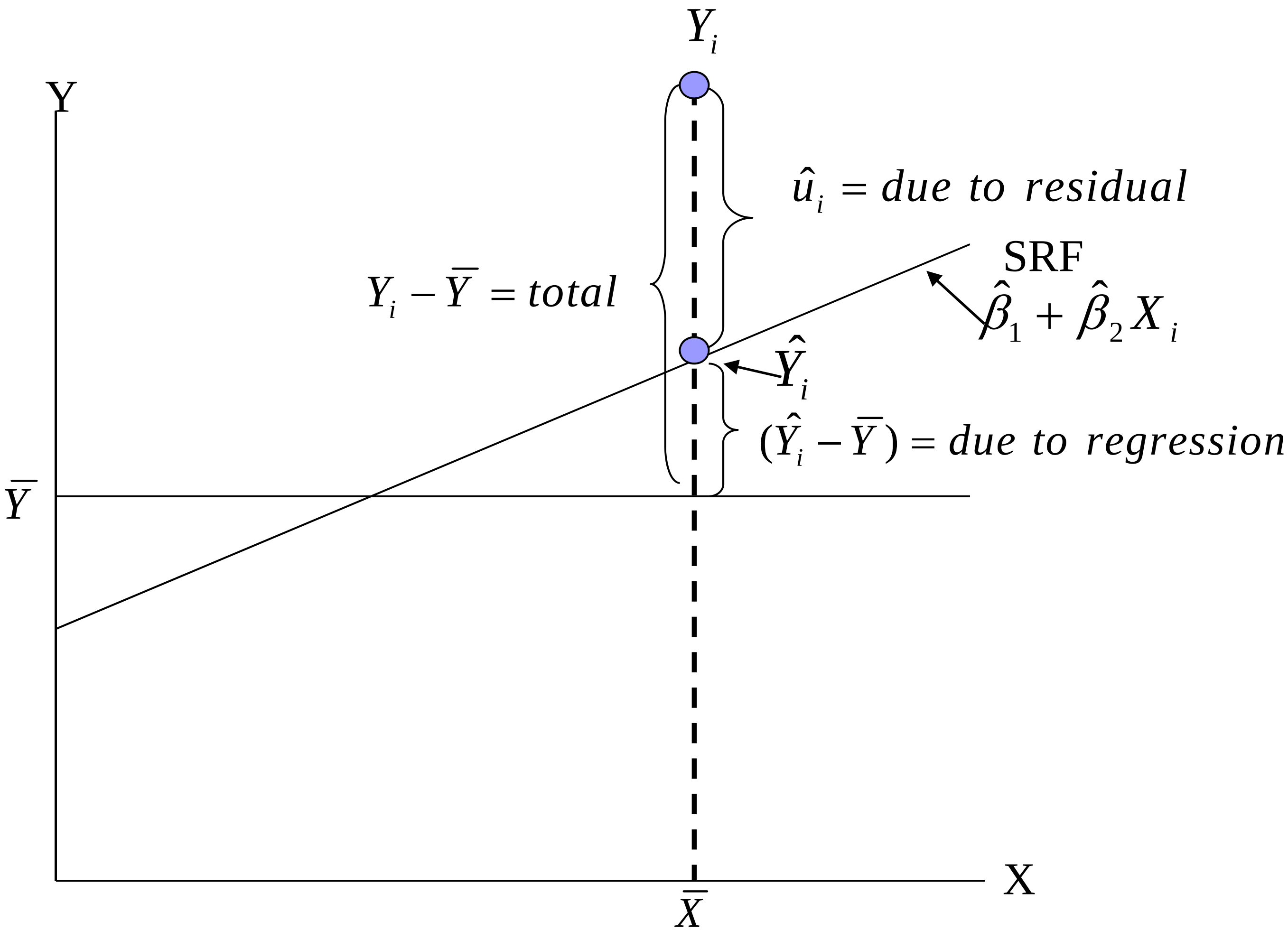
$$Y_i = \hat{Y}_i + \hat{u}_i$$

$$y_i = \hat{y}_i + \hat{u}_i$$

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i$$

$$= \sum \hat{y}_i^2 + \sum \hat{u}_i^2$$

$$= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2$$



Y_i
Y
$\hat{u}_i = due\ to\ residual$
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$Y_i - \bar{Y} = total$
$\hat{Y}_i$
$(\hat{Y}_i - \bar{Y}) = due\ to\ regression$
$\bar{Y}$
X
$\bar{X}$

$$\sum y_i^2 = \sum (Y_i - \bar{Y})^2$$

Total variation of the actual Y values about their sample mean

(Total Sum of Squares, TSS)

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \hat{Y})^2 = \hat{\beta}_2^2 \sum x_i^2$$

Variation of the estimated Y values about their mean (Explained Sum of Squares, ESS)

$$\sum \hat{u}_i^2$$

Residual or unexpected variation of the Y values about the regression line (Residual Sum of Squares, RSS)

$$TSS = ESS + RSS$$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{ESS}{TSS}$$

Example ☺

Y	X
49	1
45	2
44	3
39	4
38	5
37	6
34	7
33	8
30	9

$$\hat{\beta}_1 = 49.667$$

$$\hat{\beta}_2 = -2.1576$$

$$\bar{X} = 5.5$$

$$\bar{Y} = 38$$

Y	X	$\boxed{\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i}$	$\hat{u}_i$	$(Y_i - \overline{Y})^2$	$(\hat{Y}_i - \overline{Y})^2$	$\hat{u}_i^2$
49	1					
45	2					
44	3					
39	4					
38	5					
37	6					
34	7					
33	8					
30	9					
29	10					

Y	X	$\boxed{\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i}$	$\hat{u}_i$	$(Y_i - \overline{Y})^2$	$(\hat{Y}_i - \overline{Y})^2$	$\hat{u}_i^2$
49	1	47.509				
45	2	45.352				
44	3	43.194				
39	4	41.037				
38	5	38.879				
37	6	36.721				
34	7	34.564				
33	8	32.406				
30	9	30.249				
29	10	28.091				

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{ESS}{TSS}$$

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{384.06}{393.6} \approx 0.98$$

Approximately 98 percent of the variation in Y is explained by variation in X.

Two properties of r^2

1. Nonnegative quantity
2. Its limits are $0 \leq r^2 \leq 1$

The coefficient of correlation: r

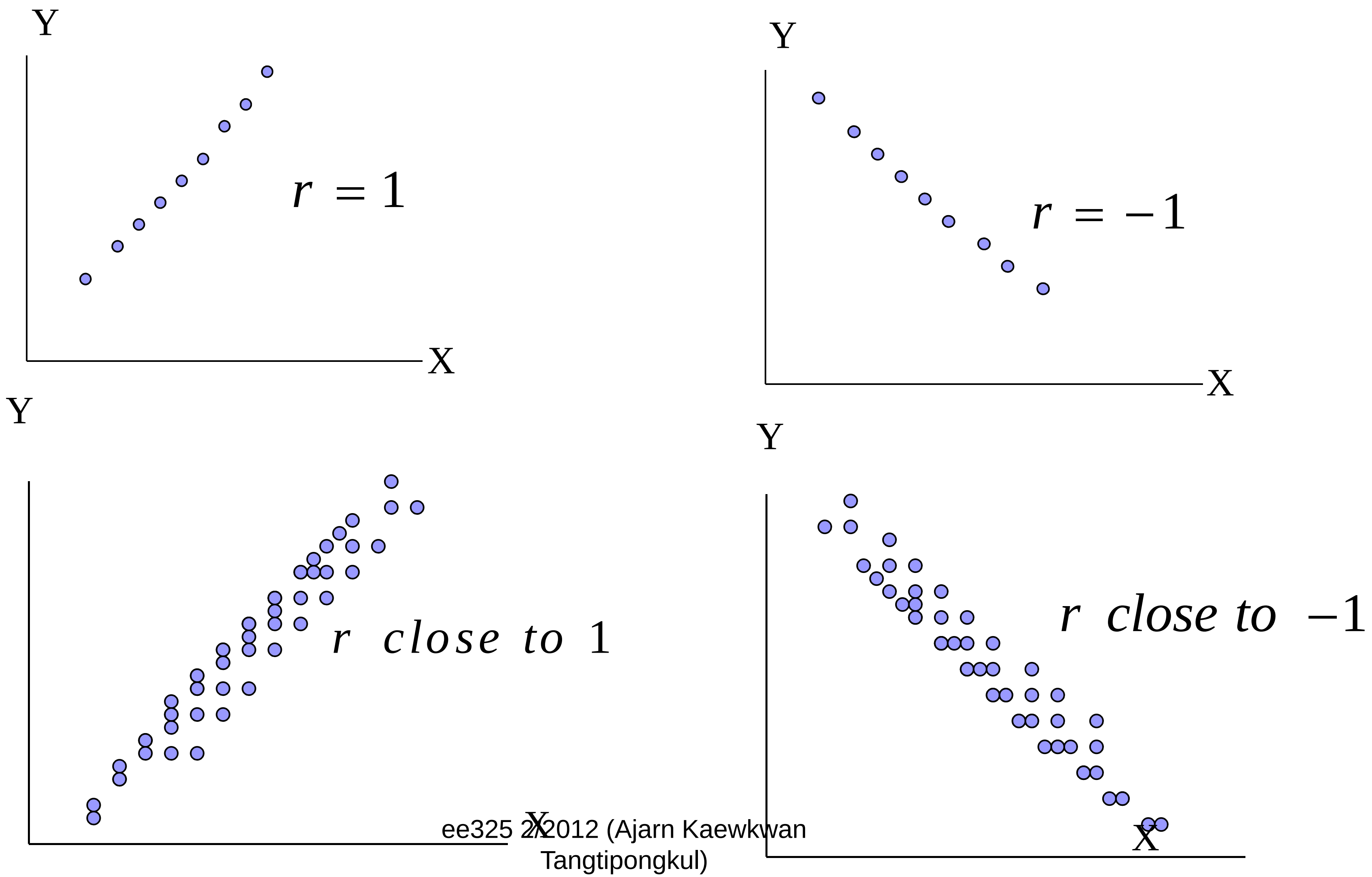
- A measure of the degree of association between two variables

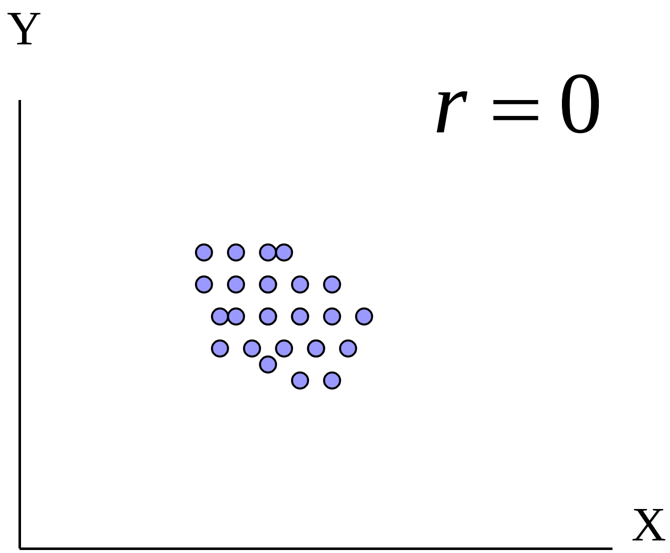
$$r = \pm\sqrt{r^2}$$

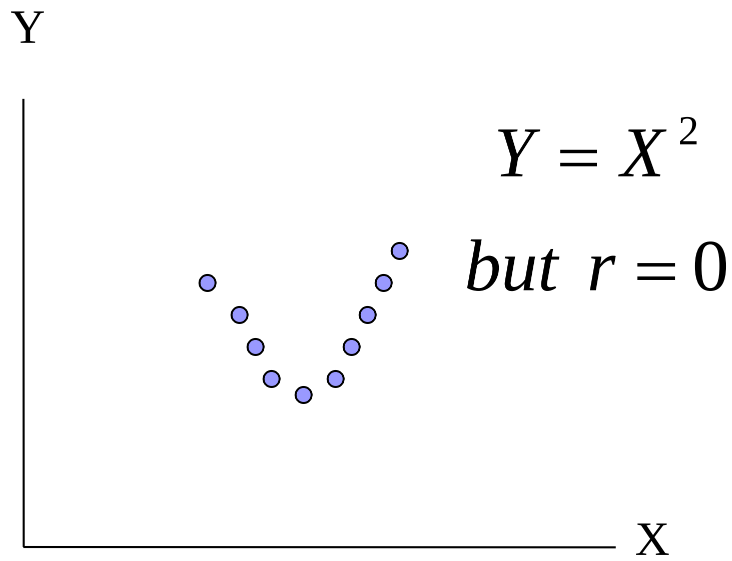
Properties of r

1. Can be positive or negative
2. Lies between the limits of -1 and 1
3. Symmetric in nature
4. Independent or the origin and scale
5. If X and Y are statistically independent, the correlation coefficient between them is zero **but zero correlation does not necessarily imply independence**
6. No meaning for describing nonlinear relations
7. Does not necessarily imply any cause and effect relationship

Correlation patterns




Y
X
$r = 0$


Y
X
$Y = X^2$
$but \ r = 0$

Source

Gujarati, D.N. (2009) Basic Econometrics. 5th ed. Singapore, McGraw-Hill.