GE2262 Business Statistics
Topic 8 Simple Linear Regression

Lecturer: Dr. Iris Yeung

Room : LAU-7239

Tel No.: 34428566

E-mail: msiris@cityu.edu.hk

Outline

- Scatter Plot
- Covariance and the Coefficient of Correlation
- Simple Linear Regression

Reference
Levine, D.M., Krehbiel, T.C. and Berenson, M.L., *Business Statistics: A First Course*, Pearson Education Ltd, Chapter 2 & 3 & 12

Part One

- Scatter Plot
- Covariance and the Coefficient of Correlation
- Simple Linear Regression

Introduction

- This topic studies the relationship among variables which measure different characteristics of items or individuals in a population
  - Example: relationship between property price and floor area, age of building, location, direction, view, floor level etc
- Dependent variable Y: the variable we wish to predict or explain
  - Example: Y=property price
- Independent variable X: the variable used to predict or explain the dependent variable
  - Example: X= floor area, age of building, location, direction, view, floor level etc
- Simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - There is only one independent variable X
  - The relationship between X and Y is described by a linear function
- Multiple linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$
  - There are k independent variables
  - The relationship between X and Y is described by a linear function

Purpose of Regression Analysis

Simple Linear Regression analysis

$$\text{Model: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{Fitted line: } \hat{Y}_i = b_0 + b_1 X_i$$

- **Predict** the value of a quantitative dependent variable based on the value of (at least) one independent variable (quantitative/numerical or qualitative/categorical)
- **Explain** the effect of the independent variable(s) on the dependent variable

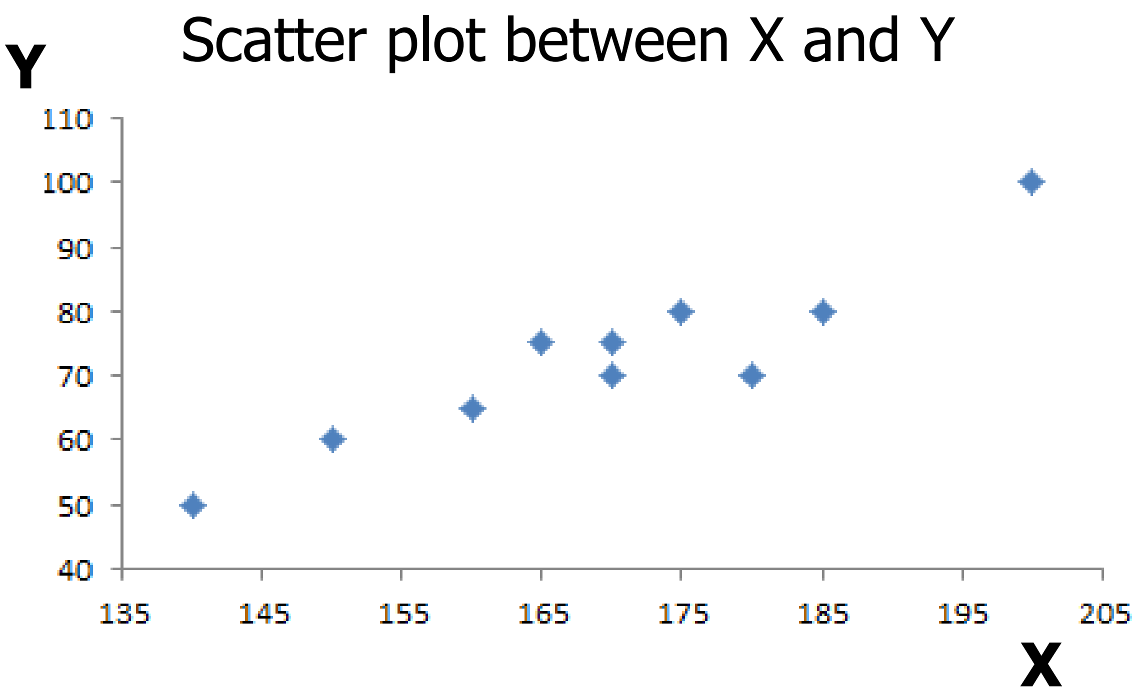
Preliminary Analysis

- Scatter plot
  - To visualize the relationship between X and Y
- Covariance and Coefficient of correlation
  - Both measure the linear relationship between two numerical variables
  - Covariance can determine the direction of the linear relationship between X and Y but cannot determine the strength of the relationship
  - Coefficient of correlation can determine both the strength and direction of the linear relationship between X and Y

Scatter Plot of Sample Data -- Positive Linear Association

Consider the following data for variables X, Y and Z from a sample of 10 observations

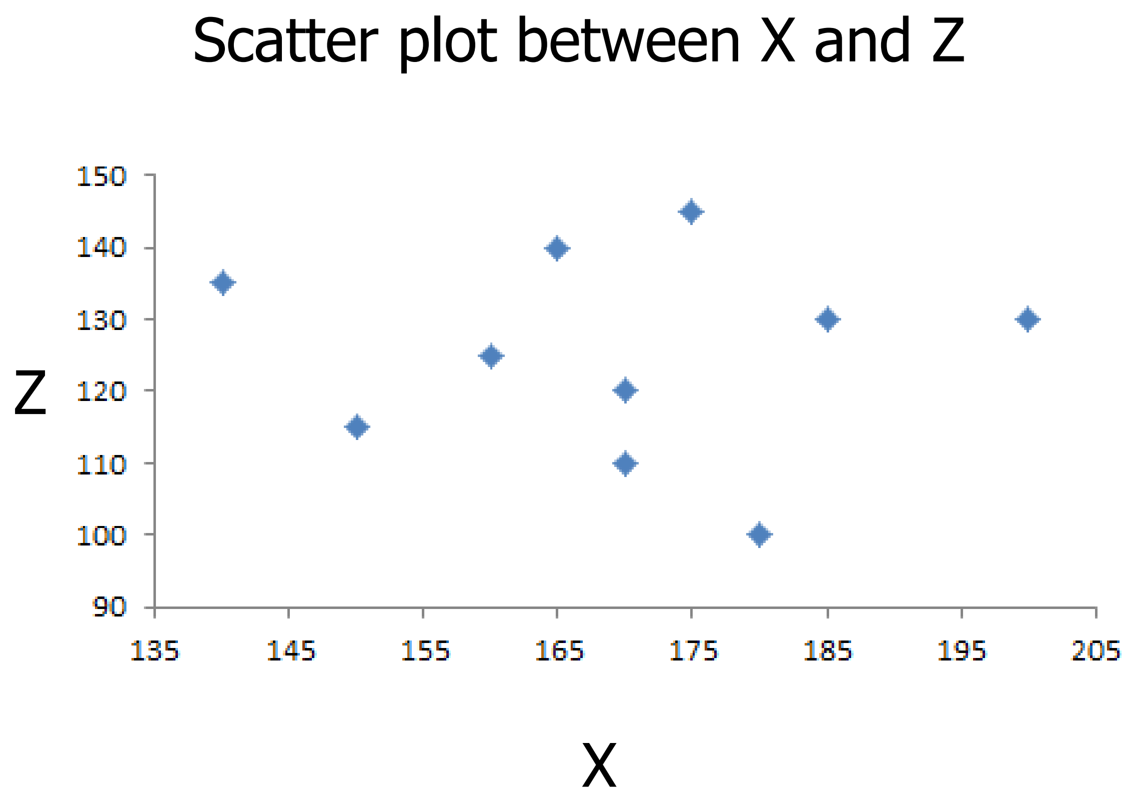
X (height)	*Y* (weight)	Z (IQ)
170	75	120
185	80	130
165	75	140
140	50	135
180	70	100
150	60	115
200	100	130
160	65	125
175	80	145
170	70	110



- The dots on the scatter plot lie "close to" a straight line with a positive slope (X and Y move in the same direction)
- We say that these two variables, X and Y, have a positive linear association

Scatter Plot of Sample Data -- No Linear Association

X (height)	Z (IQ)
170	120
185	130
165	140
140	135
180	100
150	115
200	130
160	125
175	145
170	110

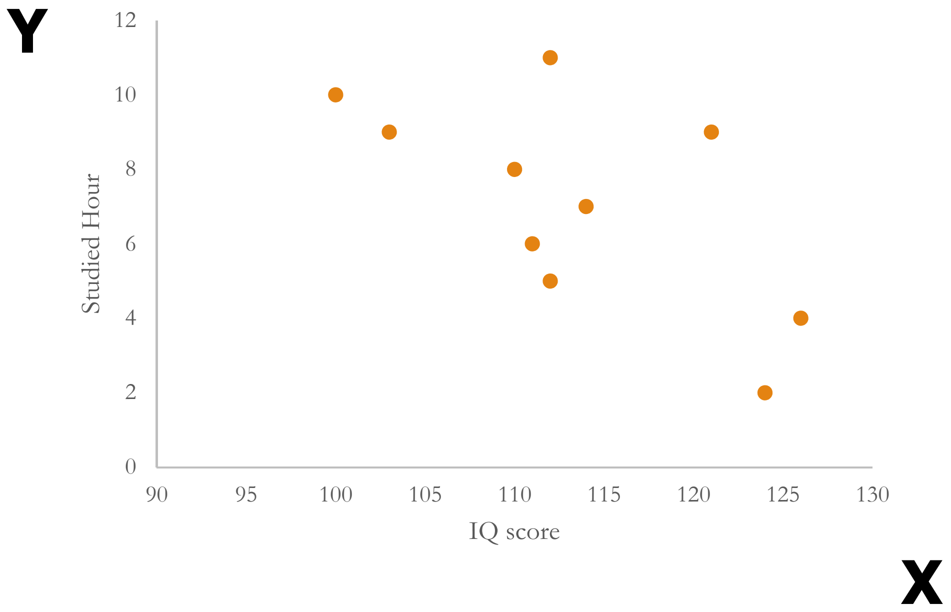


- The diagram indicates no obvious relationship between X and Z

Scatter Plot of Sample Data -- Negative Linear Association

X (IQ Score)	Y (Studied Hour)
112	5
126	4
100	10
114	7
112	11
121	9
110	8
103	9
111	6
124	2

Scatter plot between IQ score and studied hour



- the dots on the scatter plot lie "close to" a straight line with negative slope (X and Y move in opposite direction)
- we say that the variables exhibit a negative linear association

Part Two

- Scatter Plot
- Covariance and the Coefficient of Correlation
- Simple Linear Regression

Preliminary Analysis

- Scatter plot
  - To visualize the relationship between X and Y
- Covariance and Coefficient of correlation
  - Both measure the linear relationship between two numerical variables
  - Covariance can determine the direction of the linear relationship between X and Y but cannot determine the strength of the relationship
  - Coefficient of correlation can determine the strength and direction of the linear relationship between X and Y

Covariance for Positive Linear Association

Population covariance

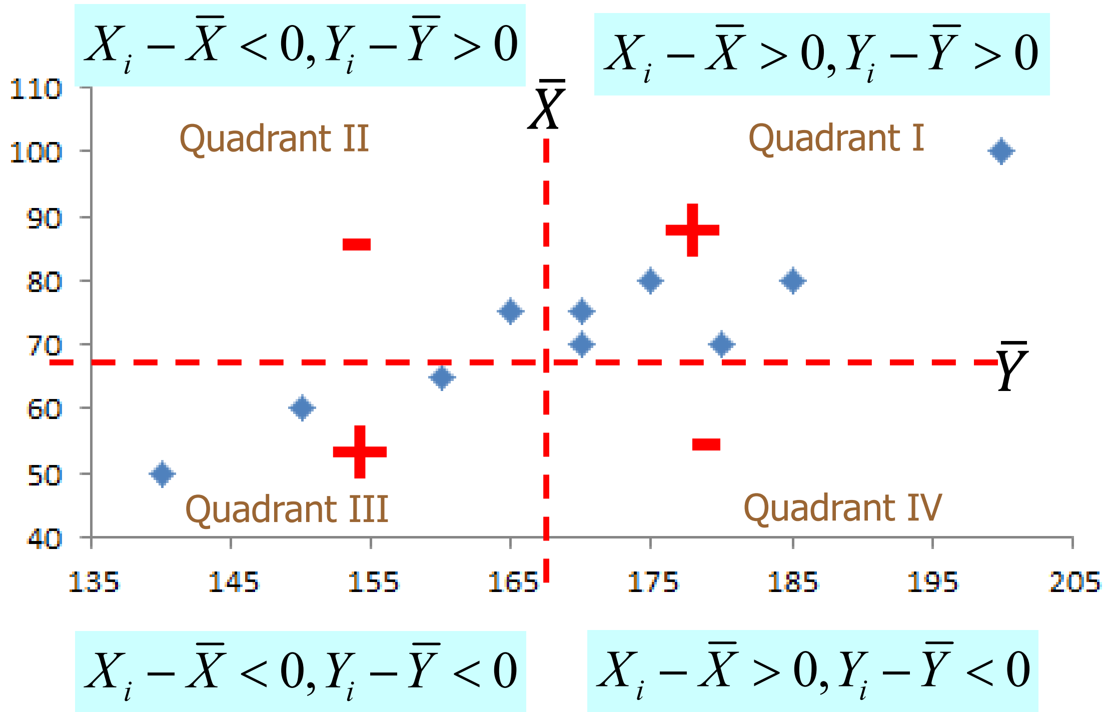
$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Sample covariance

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{n-1}$$

- The cross product term $(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive in quadrants I and III, and negative in quadrants II and IV

- If X and Y have positive linear association, there is a tendency for the dots to lie predominantly in quadrants I and III
- Covariance > 0



Covariance for Negative Linear Association

Population covariance

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

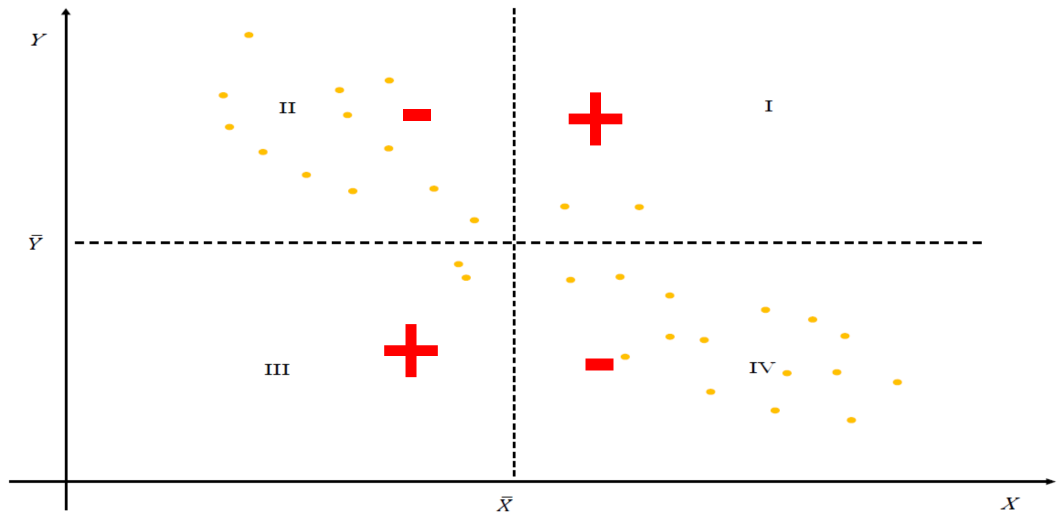
Sample covariance

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{n-1}$$

- The cross product term $(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive in quadrants I and III, and negative in quadrants II and IV

- If X and Y have negative linear association, there is a tendency for the dots to lie predominantly in quadrants II and IV
- Covariance < 0

$X_i - \bar{X} < 0, Y_i - \bar{Y} > 0$ $X_i - \bar{X} > 0, Y_i - \bar{Y} > 0$



$X_i - \bar{X} < 0, Y_i - \bar{Y} < 0$ $X_i - \bar{X} > 0, Y_i - \bar{Y} < 0$

Covariance for No Linear Association

Population covariance

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

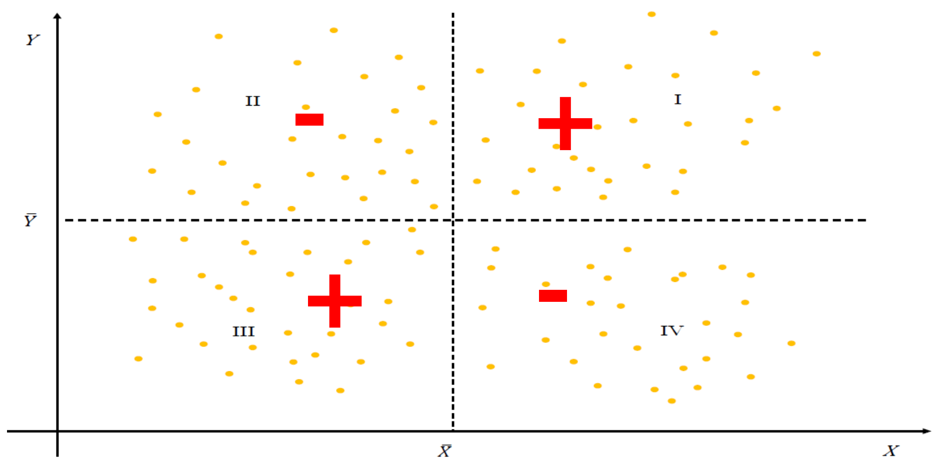
Sample covariance

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{n-1}$$

- The cross product term $(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive in quadrants I and III, and negative in quadrants II and IV

- If X and Y have no or very weak linear association, then there is a tendency for the dots to scatter across all four quadrants
- Covariance close to 0

$X_i - \bar{X} < 0, Y_i - \bar{Y} > 0$ $\quad$ $X_i - \bar{X} > 0, Y_i - \bar{Y} > 0$



$X_i - \bar{X} < 0, Y_i - \bar{Y} < 0$ $\quad$ $X_i - \bar{X} > 0, Y_i - \bar{Y} < 0$

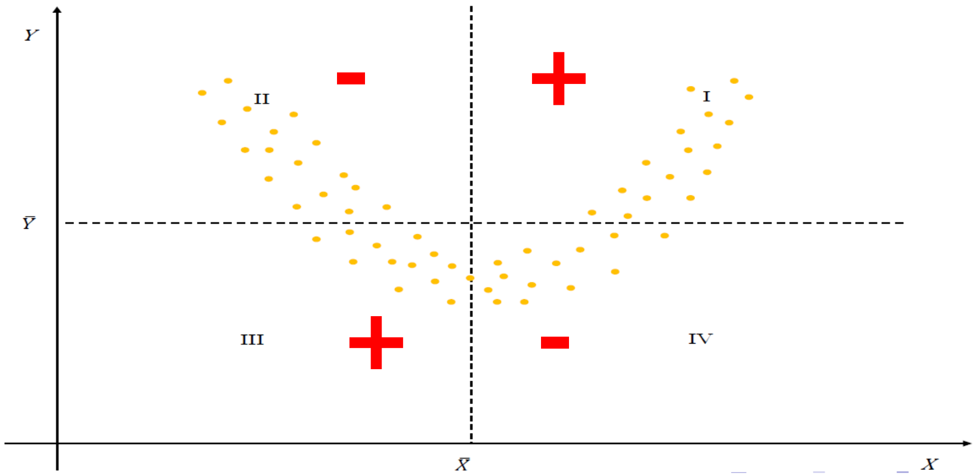
Covariance for Non- Linear Association

Population covariance	Sample covariance
$\sigma_{XY} = \dfrac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$	$S_{XY} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \dfrac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{n-1}$

- The cross product term $(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive in quadrants I and III, and negative in quadrants II and IV

- If X and Y have non-linear association, the dots may also scatter across all four quadrants
- Covariance close to 0
- A covariance of zero does not necessarily imply that X and Y have no association. They may be related in a non-linear way
- Covariance = 0 -> We can only say they have no linear association

$X_i - \bar{X} < 0, Y_i - \bar{Y} > 0$ | $X_i - \bar{X} > 0, Y_i - \bar{Y} > 0$



$X_i - \bar{X} < 0, Y_i - \bar{Y} < 0$ | $X_i - \bar{X} > 0, Y_i - \bar{Y} < 0$

Calculation of Covariance

- Consider the sample data regarding X and Y (n=10)

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	XY
170	75	0.5	2.5	1.25	12750
185	80	15.5	7.5	116.25	14800
165	75	-4.5	2.5	-11.25	12375
140	50	-29.5	-22.5	663.75	7000
180	70	10.5	-2.5	-26.25	12600
150	60	-19.5	-12.5	243.75	9000
200	100	30.5	27.5	838.75	20000
160	65	-9.5	-7.5	71.25	10400
175	80	5.5	7.5	41.25	14000
170	70	0.5	-2.5	-1.25	11900
$\bar{X}=169.5$	$\bar{Y}=72.5$	$\sum(X-\bar{X})=0$	$\sum(Y-\bar{Y})=0$	$\sum(X-\bar{X})(Y-\bar{Y})$=1937.5	$\sum$XY=124,825

$$S_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}=\frac{1937.5}{9}=215.28$$

$$S_{XY}=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{n-1}=\frac{124825-10(169.5)(72.5)}{9}=215.28$$

Calculation of Covariance

- Divide the X value by 100 → X'=X/100

X'	Y	$X'-\overline{X'}$	$Y-\overline{Y}$	$(X'-\overline{X'})(Y-\overline{Y})$	X'Y
1.7	75	0.005	2.5	0.0125	127.5
1.85	80	0.155	7.5	1.1625	148
1.65	75	-0.045	2.5	-0.1125	123.75
1.4	50	-0.295	-22.5	6.6375	70
1.8	70	0.105	-2.5	-0.2625	126
1.5	60	-0.195	-12.5	2.4375	90
2	100	0.305	27.5	8.3875	200
1.6	65	-0.095	-7.5	0.7125	104
1.75	80	0.055	7.5	0.4125	140
1.7	70	0.005	-2.5	-0.0125	119
$\overline{X'}=1.695$	$\overline{Y}=72.5$	$\sum(X'-\overline{X'})=0$	$\sum(Y-\overline{Y})=0$	$\sum(X'-\overline{X'})(Y-\overline{Y})=19.375$	$\sum X'Y=1248.25$

- The sample covariance is reduced by a factor of 100

$$S_{X'Y}=\frac{\sum_{i=1}^{n}(X_i'-\overline{X}')(Y_i-\overline{Y})}{n-1}=\frac{19.375}{9}=2.1528$$

$$S_{X'Y}=\frac{\sum_{i=1}^{n}X_i'Y_i-n\overline{X}'\overline{Y}}{n-1}=\frac{1248.25-10(1.695)(72.5)}{9}=2.1528$$

Disadvantage of Covariance

- One disadvantage of covariance is that it is dependent on the units used to measure X and Y
  - Its value does not indicate the strength of the linear relationship of the two variables
  - Its value cannot be directly compared for different variables

X'=X/100

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{1937.5}{9} = 215.28$$

$$S_{XY} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{n-1} = \frac{124825 - 10(169.5)(72.5)}{9} = 215.28$$

$$S_{X'Y} = \frac{\sum_{i=1}^{n}(X_i' - \bar{X}')(Y_i - \bar{Y})}{n-1} = \frac{19.375}{9} = 2.1528$$

$$S_{X'Y} = \frac{\sum_{i=1}^{n} X_i' Y_i - n\bar{X}'\bar{Y}}{n-1} = \frac{1248.25 - 10(1.695)(72.5)}{9} = 2.1528$$

Coefficient of Correlation

- The coefficient of correlation measures the strength and direction of the linear relationship between two numerical variables, which is not affected by the variables' measurement scale
  - It adjusts the covariance by the standard deviations of X and Y so that the resulting measure is unit-free
  - It is a "standardized score" of the covariance

Population covariance

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Sample covariance

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Population coefficient of correlation

$$\rho_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)/N}{\sqrt{(\sum_{i=1}^{N}(X_i - \mu_X)^2/\mathrm{N})\ (\sum_{i=1}^{N}(Y_i - \mu_Y)^2/N)}}$$

Sample coefficient of correlation

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{\sqrt{\{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)/(n-1)\}\ \{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)/(n-1)\}}}$$

Coefficient of Correlation

Population covariance

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)}{N}$$

Sample covariance

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{n-1}$$

Population coefficient of correlation

pronounced rho

$$\rho_{XY} = \frac{\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{N}(X_i-\mu_X)^2\sum_{i=1}^{N}(Y_i-\mu_Y)^2}} == \frac{\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)\Big/N}{\sqrt{\left(\sum_{i=1}^{N}(X_i-\mu_X)^2\Big/N\right)\left(\sum_{i=1}^{N}(Y_i-\mu_Y)^2\Big/N\right)}}$$

Sample coefficient of correlation

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)}} = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2\right)}}$$

$$= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\Big/n-1}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\Big/n-1}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\Big/n-1}} = \frac{S_{XY}}{S_X S_Y}$$

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1};\ S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}};\ S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$$

- The sign of r_{XY} is the same as that of S_{XY}
  - the denominator of r_{XY} is the product of standard deviation of X and Y, which are always non-negative

Calculation of Coefficient of Correlation

- Consider the sample data regarding X and Y again

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	X^2	Y^2	XY
170	75	0.5	2.5	1.25	0.25	6.25	28900	5625	12750
185	80	15.5	7.5	116.25	240.25	56.25	34225	6400	14800
165	75	-4.5	2.5	-11.25	20.25	6.25	27225	5625	12375
140	50	-29.5	-22.5	663.75	870.25	506.25	19600	2500	7000
180	70	10.5	-2.5	-26.25	110.25	6.25	32400	4900	12600
150	60	-19.5	-12.5	243.75	380.25	156.25	22500	3600	9000
200	100	30.5	27.5	838.75	930.25	756.25	40000	10000	20000
160	65	-9.5	-7.5	71.25	90.25	56.25	25600	4225	10400
175	80	5.5	7.5	41.25	30.25	56.25	30625	6400	14000
170	70	0.5	-2.5	-1.25	0.25	6.25	28900	4900	11900
$\bar{X}=$ 169.5	$\bar{Y}=$ 72.5	$\sum(X-\bar{X})=0$	$\sum(Y-\bar{Y})=0$	$\sum(X-\bar{X})(Y-\bar{Y})$ =1937.5	$\sum(X-\bar{X})^2$ =2672.5	$\sum(Y-\bar{Y})^2$ = 1612.5	$\sum X^2$ =289975	$\sum Y^2$ =54175	$\sum XY$ =124825

$$r_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)}}=\frac{1937.5}{\sqrt{(2672.5*1612.5)}}=0.933$$

$$=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)\left(\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2\right)}}=\frac{124825-10(169.5)(72.5)}{\sqrt{(289975-10*169.5^2)(54175-10*72.5^2)}}=0.933$$

Calculation of Coefficient of Correlation

- What if X value is divided by 100? → X'=X/100

X'	Y	$X'-\overline{X'}$	$Y-\overline{Y}$	$(X'-\overline{X'})(Y-\bar{Y})$	$(X'-\overline{X}')^2$	$(Y-\overline{Y})^2$	X'^2	Y^2	X'Y
1.7	75	0.005	2.5	0.0125	0.000025	6.25	2.89	5625	127.5
1.85	80	0.155	7.5	1.1625	0.024025	56.25	3.4225	6400	148
1.65	75	-0.045	2.5	-0.1125	0.002025	6.25	2.7225	5625	123.75
1.4	50	-0.295	-22.5	6.6375	0.087025	506.25	1.96	2500	70
1.8	70	0.105	-2.5	-0.2625	0.011025	6.25	3.24	4900	126
1.5	60	-0.195	-12.5	2.4375	0.038025	156.25	2.25	3600	90
2	100	0.305	27.5	8.3875	0.093025	756.25	4	10000	200
1.6	65	-0.095	-7.5	0.7125	0.009025	56.25	2.56	4225	104
1.75	80	0.055	7.5	0.4125	0.003025	56.25	3.0625	6400	140
1.7	70	0.005	-2.5	-0.0125	0.000025	6.25	2.89	4900	119
$\overline{X}=$ 1.695	$\overline{Y}=$ 72.5	$\sum(X'-\overline{X'})=0$	$\sum(Y-\overline{Y})=0$	$\sum(X'-\overline{X'})(Y-\overline{Y})$ =19.375	$\sum(X'-\overline{X}')^2$ =0.26725	$\sum(Y-\overline{Y})^2$ = 1612.5	$\sum X'^2$ =28.9975	$\sum Y^2$ =54175	$\sum X'$Y =1248.25

$$r_{XY}=\frac{\sum_{i=1}^{n}(X_i'-\bar{X}')(Y_i-\bar{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i'-\bar{X}')^2\right)\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)}}=\frac{19.375}{\sqrt{(0.26725*1612.5)}}=0.933$$

$$=\frac{\sum_{i=1}^{n}X_i'Y_i-n\bar{X}'\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n}X_i'^2-n\bar{X}'^2\right)\left(\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2\right)}}=\frac{1248.25-10(1.695)(72.5)}{\sqrt{(28.9975-10*1.695^2)(54175-10*72.5^2)}}=0.933$$

- The sample correlation remains unchanged although the sample covariance has been reduced by a factor of 100

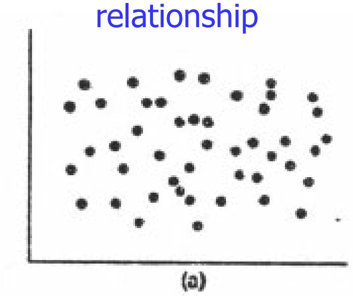
Interpretation of Coefficient of Correlation

- $-1 \leq r_{XY} \leq 1$
  - When $r_{XY} = -1$,
    - all sample values of X and Y lie exactly on a straight line having a negative slope
    - We say that X and Y are perfectly negatively linearly related
  - When r_{XY} is closer to -1, strong negative linear relationship
  - When r_{XY} is closer to 0 but negative, weak negative linear relationship
  - When $r_{XY} = 0$
    - We say that X and Y are not linearly related (uncorrelated)
  - When r_{XY} is closer to 0 but positive, weak positive linear relationship
  - When r_{XY} is closer to 1, strong positive linear relationship
  - When $r_{XY} = 1$,
    - all sample values of X and Y lie exactly on a straight line having a positive slope
    - We say that X and Y are perfectly positively linearly related

Coefficient of Correlation

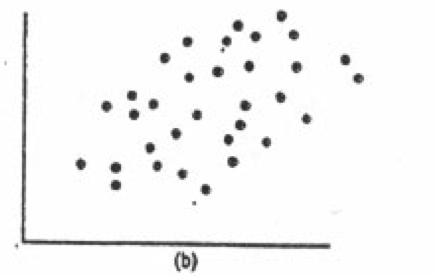
- Here are some diagrams illustrating different values of r_{XY}

No linear relationship



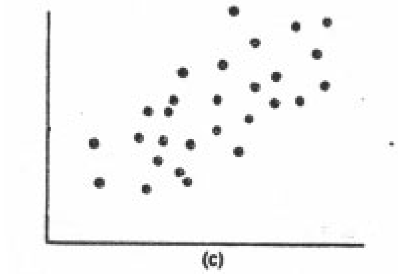
$r = 0$

Moderate positive linear relationship



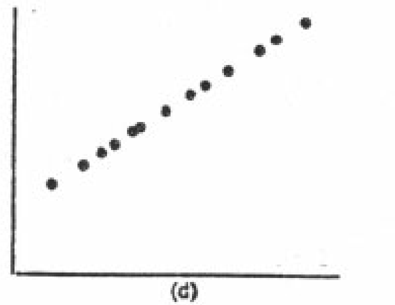
$r = 0.5$

Strong positive linear relationship



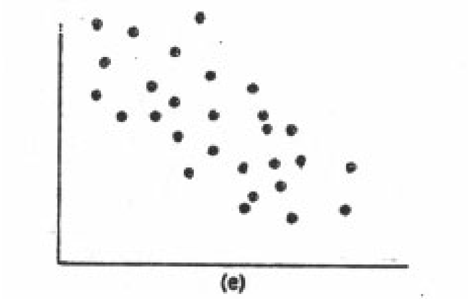
$r = 0.8$

Perfect positive linear relationship



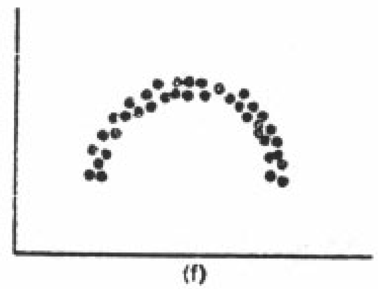
$r = 1$

Strong negative linear relationship



$r = -0.8$

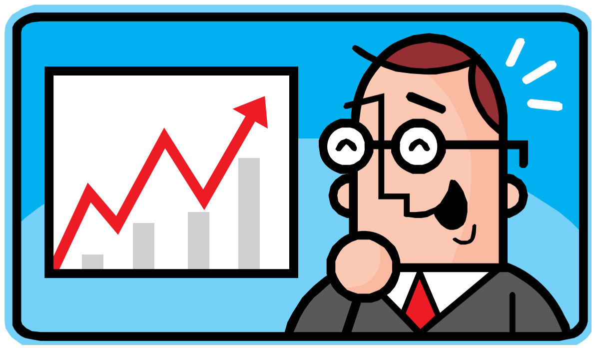
No linear relationship (non-linear relationship)



$r = 0$

Diversifying Your Investments

- One basic theory of investing is diversification
  - The idea is that you want to have a basket of stocks that do not all “move in the same direction’
  - If one investment goes down, you don’t want a second investment in your portfolio that is also likely to go down
- One hallmark of a good portfolio is a low correlation between investments



Diversifying Your Investments

- The following data represent the annual rates of return for various stocks

Year	Cisco Systems	Walt Disney	General Electric	Exxon Mobil	TECO Energy	Dell
1999	1.310	-0.015	0.574	0.151	-0.303	-0.319
2000	-0.286	-0.004	-0.055	0.127	0.849	-0.661
2001	-0.527	-0.277	-0.151	-0.066	-0.150	0.553
2002	-0.277	-0.203	-0.377	-0.089	-0.369	-0.031
2003	0.850	0.444	0.308	0.206	0.004	0.254
2004	-0.203	0.202	0.207	0.281	0.128	0.234
2005	0.029	-0.129	-0.014	0.118	0.170	-0.288
2006	0.434	0.443	0.093	0.391	0.051	-0.164
2007	0.044	-0.043	0.126	0.243	0.058	-0.033
2008	-0.396	-0.306	-0.593	-0.193	-0.355	-0.580
2009	0.459	0.417	-0.102	-0.171	0.249	0.393
2010	-0.185	0.155	0.053	0.023	0.044	-0.323

Source: Yohoo!Finance

Diversifying Your Investments

	Cisco Systems	Walt Disney	General Electric	Exxon Mobil	TECO Energy	Dell
Cisco Systems	1					
Walt Disney	0.5512	1				
General Electric	0.7461	0.5110	1			
Exxon Mobil	0.3625	0.4701	0.7024	1		
TECO Energy	-0.1211	0.3432	0.1477	0.2828	1	
Dell	0.0630	0.2906	0.1448	-0.0445	-0.1768	1

- If you only wish to invest in two stocks
  - Which two would you select if your goal is to have low correlation between the two investments?

    Dell and Exxon Mobil as their correlation is the nearest to 0
  - Which two would you select if your goal is to have one stock go up when the other goes down?

    Dell and TECO Energy as they have the strongest negative correlation

Part Three

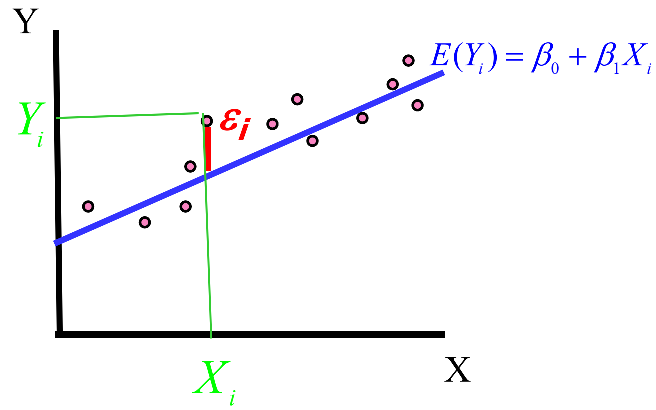
- Scatter Plot
- Covariance and the Coefficient of Correlation
- Simple Linear Regression

Simple Linear Regression Model

- Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, N$$

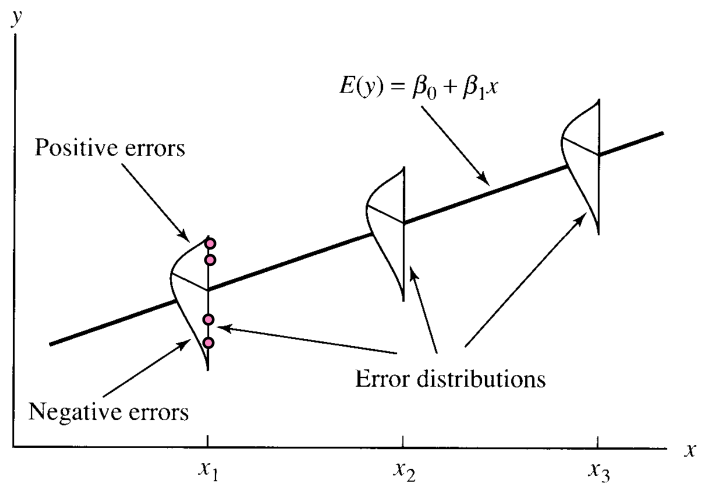
where
Y_i = dependent variable for observation *i*,
X_i = independent variable for observation *i*,
β_0 = Y intercept for the population,
β_1 = slope for the population,
ε_i = random error in Y for observation *i*,



Assumptions : $\varepsilon_i \sim N(0, \sigma^2)$

Variance is constant for all x values.

Error terms are independent.

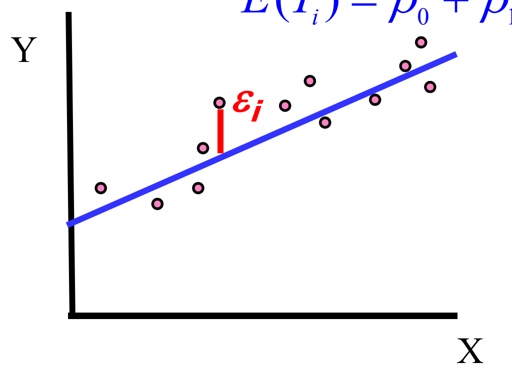


X

Simple Linear Regression Model

- **Simple Linear Regression Model**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, N$$



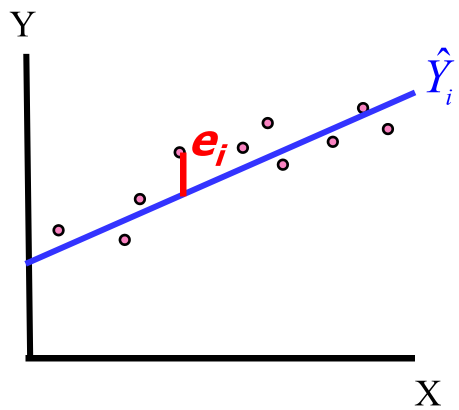
- **Estimated Simple Linear Regression Equation**

$$\hat{Y}_i = b_0 + b_1 X_i, \quad i = 1, \ldots, n$$

Sample	Y	X
1	y_1	x_1
2	y_2	x_2
…		
n	y_n	x_n

where
Y_i = actual value of Y for observation *i*,
$\hat{Y}_i$ = predicted value of Y for observation *i*,
X_i = value of X for observation *i*,

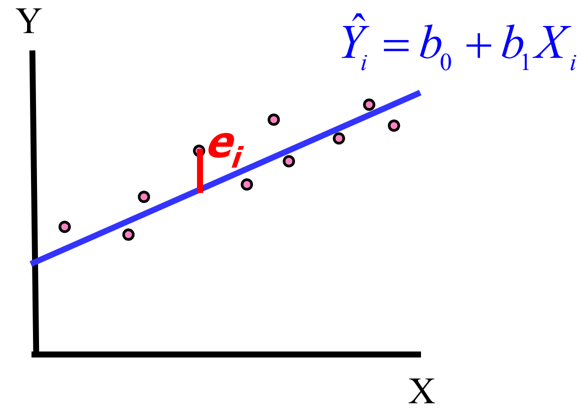
b_0 = sample Y intercept,
b_1 = sample slope
e_i = residual for observation *i*
= *actual* Y_i *- predicted* $\hat{Y}_i$



Least Squares Method

$$\hat{Y}_i = b_0 + b_1 X_i$$

- Residual $e_i = Y_i \ (Actual) - \hat{Y}_i \ (\text{Predicted})$
- Sum of the squared residuals $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- The parameter estimates b_0, b_1 are found by Least Square Method, which minimizes the sum of the squared residuals
- It can be shown that



$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Least Squares Method

- The parameter estimate b_1 is related to the sample coefficient of correlation r_{XY} as follows:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \Big/ n-1}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \Big/ n-1} = \frac{S_{XY}}{(S_x)^2}$$

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{S_X S_Y} \Rightarrow b_1 = \frac{r_{XY} S_X S_Y}{(S_x)^2} = \frac{r_{XY} S_Y}{S_X}$$

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}; S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}; S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

- Since S_X and S_Y (standard deviation of X and Y) are non-negative, b_1 will have the same sign as r_{XY}

Regression Analysis Example

- The following table gives data collected last year for seven employees of a company

  X = Number of years of service

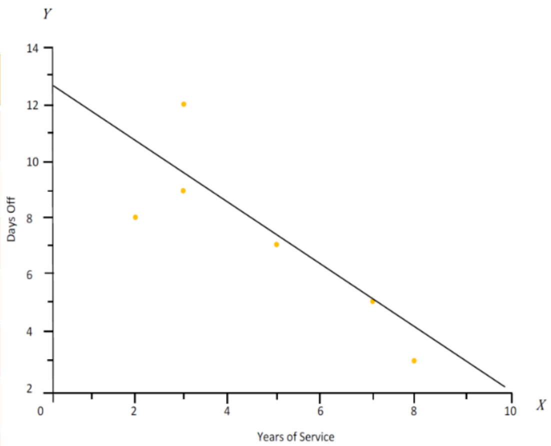
  Y = Number of days taken off work

- Find the relationship between X and Y

X	Y	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$	X^2	Y^2	XY
2	8	-3	9	1	1	-3	4	64	16
5	7	0	0	0	0	0	25	49	35
7	5	2	4	-2	4	-4	49	25	35
3	12	-2	4	5	25	-10	9	144	36
8	3	3	9	-4	16	-12	64	9	24
3	9	-2	4	2	4	-4	9	81	27
7	5	2	4	-2	4	-4	49	25	35
$\bar{X}=5$	$\bar{Y}=7$	$\sum=0$	$\sum=34$	$\sum=0$	$\sum=54$	$\sum=-37$	$\sum=209$	$\sum=397$	$\sum=208$

Regression Analysis Example

X	Y	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$	X^2	Y^2	XY
2	8	-3	9	1	1	-3	4	64	16
5	7	0	0	0	0	0	25	49	35
7	5	2	4	-2	4	-4	49	25	35
3	12	-2	4	5	25	-10	9	144	36
8	3	3	9	-4	16	-12	64	9	24
3	9	-2	4	2	4	-4	9	81	27
7	5	2	4	-2	4	-4	49	25	35
$\bar{X}=5$	$\bar{Y}=7$	$\sum=0$	$\sum=34$	$\sum=0$	$\sum=54$	$\sum=-37$	$\sum=209$	$\sum=397$	$\sum=208$



$$r_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{-37}{\sqrt{34*54}}=-0.864$$

$$b_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{-37}{34}=-1.09,$$

The least-squares regression line is

$$\hat{Y}=12.45-1.09X$$

$$\text{or } b_1=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}=\frac{(208-7*5*7)}{(209-7*5^2)}=-1.09,\ \text{or } b_1=r_{XY}\frac{S_Y}{S_X}=-0.864\frac{\sqrt{54}}{\sqrt{34}}=-1.09$$

$$b_0=\bar{Y}-b_1\bar{X}=7-(-1.09)*5=12.45$$

Regression Analysis Example

$\hat{Y} = 12.45 - 1.09X$

- Interpreting b_0:
  - We should not interpret b_0 = 12.45 as the predicted number of days off for an employee with 0 years of service
  - The level X = 0 is beyond the range of data studied
  - Linearity assumption seems reasonable in the range of 2 and 8 years of service as shown by the data, it would be dangerous to extrapolate far outside that range
- Interpreting b_1:
  - b_1 is the change in the estimated number of days off for an additional year's service
    - Subtracting the prediction for X = 5 (i.e. $\hat{Y}$ = 7) from the prediction for X = 6 (i.e. $\hat{Y}$ = 5.91) gives b_1 = -1.09
  - We are estimating that each 1 year increase in service leads, on average, to a decrease of 1.09 days off work

X	Y
2	8
5	7
7	5
3	12
8	3
3	9
7	5
$\bar{X} = 5$	$\bar{Y} = 7$

Regression Analysis Example

$\hat{Y} = 12.45 - 1.09X$

- Suppose we want to predict the number of days off work this year for employees with 0, 5, 6, 8 and 14 years of service
- All we have to do is to substitute these given X values into the estimated regression equation $\hat{Y} = 12.45 - 1.09X$
  - For X = 0, $\hat{Y} = 12.45 - 1.09(0) = 12.45$ days off work (extrapolation)
  - For X = 5, $\hat{Y} = 12.45 - 1.09(5) = 7$ days off work
  - For X = 6, $\hat{Y} = 12.45 - 1.09(6) = 5.91$ days off work
  - For X = 8, $\hat{Y} = 12.45 - 1.09(8) = 3.73$ days off work
  - For X = 14, $\hat{Y} = 12.45 - 1.09(14) = -2.81$ days off work (extrapolation)
  - The relationship between X and Y is approximately linear over the range covered by the sample
  - Once we go beyond the sample range, the relationship may cease to be approximately linear
  - We should only predict within the range of observed X values

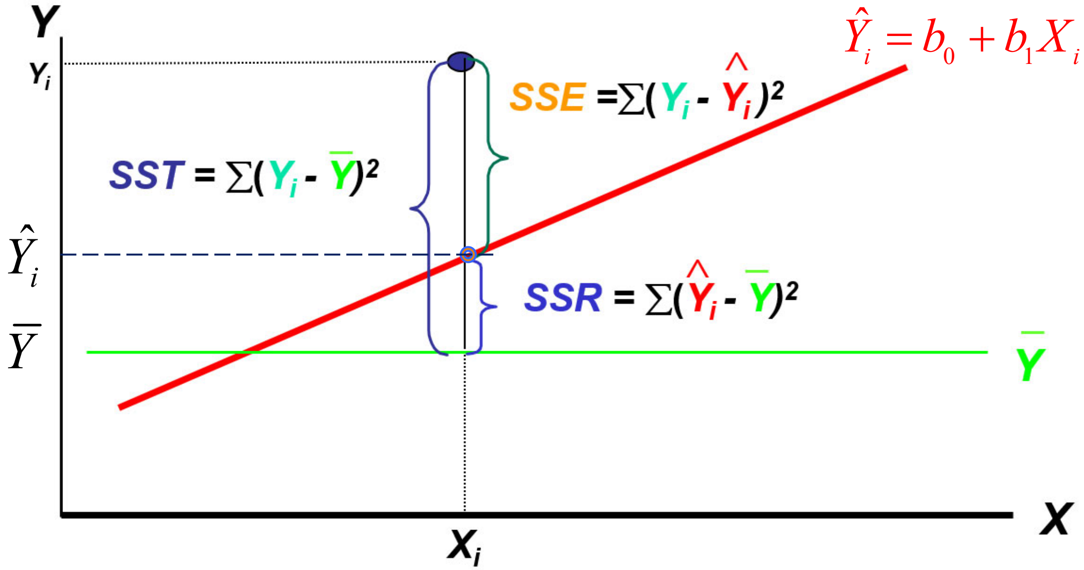
X	Y
2	8
5	7
7	5
3	12
8	3
3	9
7	5
$\bar{X} = 5$	$\bar{Y} = 7$

Three Sum of Squares

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

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + e_i$$

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



SST = total sum of squares

- Measures the total variation of the Y_i values around the mean

SSR = regression sum of squares

- Measures the variation of the $\hat{Y}_i$ values around the mean
- Explained the variation in Y by the linear relationship between *X* and *Y* (*explained variation*)

SSE = error sum of squares

- Measures the variation in Y that cannot be explained by the linear relationship between *X* and *Y* (*unexplained variation*)

SST=SSR+SSE

Coefficient of Determination

SST = total sum of squares
SSR = regression sum of squares
SSE = error sum of squares

SST=SSR+SSE

$$SST = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

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

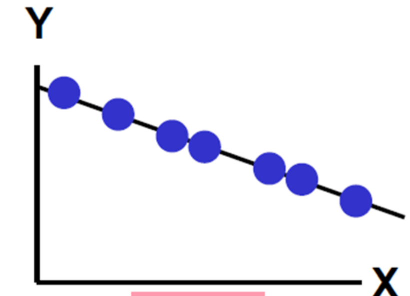
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

- Coefficient of determination (R^2) measures the proportion of the total variation in Y that can be explained by the regression

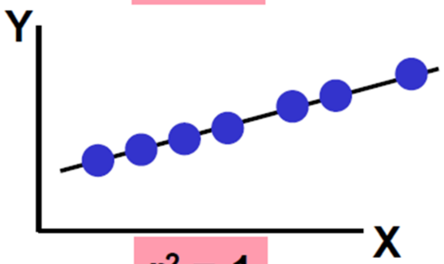
  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2 is unit-free with value in between 0 and 1 inclusive
- The higher the R^2, the better the fitting (the stronger linear association between X and Y)
- However, it does not mean that X causes Y

- In a regression model containing only one X variable, $R^2 = (r_{XY})^2$



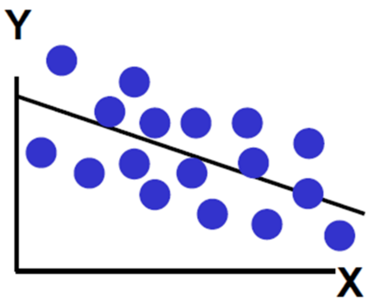
$r^2 = 1$

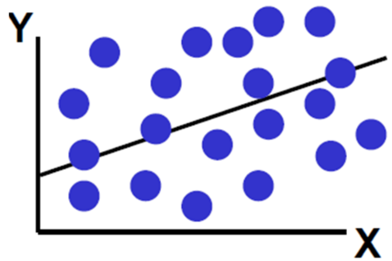


$r^2 = 1$

Perfect linear relationship between X and Y.

100% of the variation in Y is explained by variation in X.

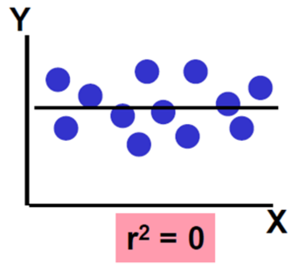




$0 < r^2 < 1$

Weaker linear relationships between X and Y.

Some but not all of the variation in Y is explained by variation in X.



$r^2 = 0$

No linear relationship between X and Y.

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X.)

Coefficient of Determination – Example

X	Y	$\hat{Y}$	e	e^2	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$
2	8	10.27	-2.27	5.1529	1	10.6929
5	7	7	0	0	0	0
7	5	4.82	0.18	0.0324	4	4.7524
3	12	9.18	2.82	7.9524	25	4.7524
8	3	3.73	-0.73	0.5329	16	10.6929
3	9	9.18	-0.18	0.0324	4	4.7524
7	5	4.82	0.18	0.0324	4	4.7524
$\bar{X}=5$	$\bar{Y}=7$	$\sum = 49$	$\sum = 0$	SSE=$\sum = 13.7354$	$SST = \sum = 54$	SSR=$\sum = 40.2646$

SST = 54
SSR=40.2646
SSE = 13.7354

$R^2 = \frac{SSR}{SST} = \frac{40.2646}{54}$=*0.7456*

R^2 =1- $\frac{SSE}{SST}$
=1 $- \frac{13.7354}{54} = 0.7456$

$(r_{XY})^2 = (-0.864)^2$
=0.7456

- The coefficient of determination is interpreted as
  - 74.56% of the sample variability in Y is explained by its linear dependency on X
  - Or, alternatively, by taking the linear dependence on X into account, the total variability in Y is reduced by 74.56%

Inference about the Slope

- **Simple Linear Regression Model**

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

$$\varepsilon_i \sim N(0, \sigma^2)$$

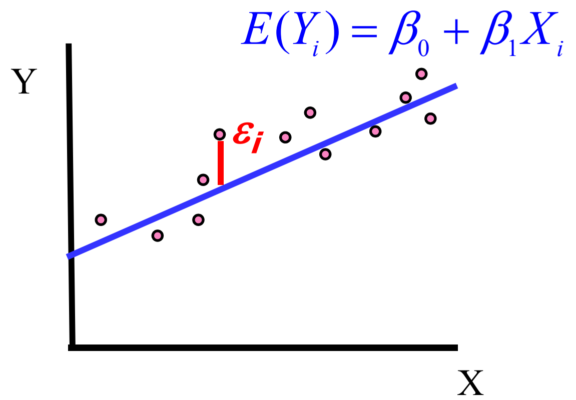
- **Estimated Simple Linear Regression Equation**

$$\hat{Y}_i = b_0 + b_1 X_i$$

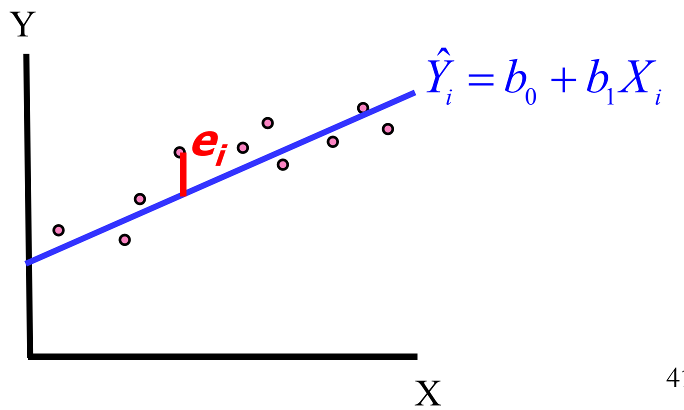
$$E(b_1) = \beta_1$$

$$Var(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

$$b_1 \sim N(\beta_1, Var(b_1)) \Rightarrow Z = \frac{b_1 - \beta_1}{sd(b_1)} \sim N(0,1)$$



Sample	Y	X
1	y_1	x_1
2	y_2	x_2
…		
n	y_n	x_n

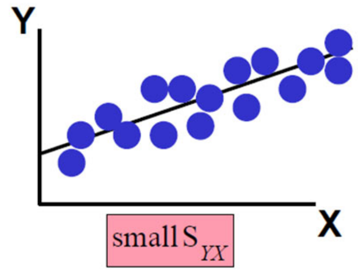


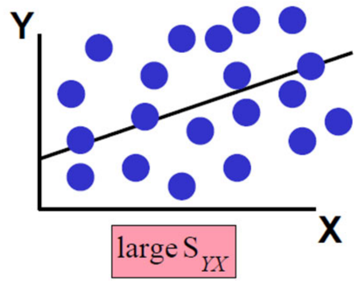
Inference about the Slope

$$E(b_1) = \beta_1$$

$$Var(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

$$b_1 \sim N(\beta_1, Var(b_1)) \Rightarrow Z = \frac{b_1 - \beta_1}{sd(b_1)} \sim N(0,1)$$





Standard error of the estimate
The standard deviation of the variation of observations around the regression line

$$\hat{\sigma} = S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

Estimated variance of b_1
Estimate the variability in the slope of regression lines arising from different possible samples

$$\hat{V}ar(b_1) = \frac{S_e^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

Standard error of the estimate for the slope

$$S_{b_1} = se(b_1) = \frac{S_e}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}}$$

The statistic $t = \frac{b_1 - \beta_1}{S_{b_1}}$ follows a t distribution with n-2 degrees of freedom

$$t = \frac{b_1 - \beta_1}{S_{b_1}} \sim t_{n-2}$$

Confidence interval for Regression Slope

- 100(1-α)% confidence interval for the population regression slope β_1 is given by

  $$\left[b_1 - t_{\alpha/2,n-2}\, S_{b_1}, b_1 + t_{\alpha/2,n-2}\, S_{b_1}\right]$$

  where $t_{\alpha/2,n-2}$ is the value corresponding to an upper-tail probability of α / 2 from the t distribution at degrees of freedom n −2

- The confidence interval for the population regression slope is interpreted as
  - The 100(1-α)% confidence interval for the expected change in Y resulting from one-unit increase in X is between $\left[b_1 - t_{\alpha/2,n-2}\, S_{b_1}, b_1 + t_{\alpha/2,n-2}\, S_{b_1}\right]$

Confidence interval for Regression Slope

In the example on number of days taken off work ,

X = Number of years of service

Y = Number of days taken off work

X	Y	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$	X^2	Y^2	XY
2	8	-3	9	1	1	-3	4	64	16
5	7	0	0	0	0	0	25	49	35
7	5	2	4	-2	4	-4	49	25	35
3	12	-2	4	5	25	-10	9	144	36
8	3	3	9	-4	16	-12	64	9	24
3	9	-2	4	2	4	-4	9	81	27
7	5	2	4	-2	4	-4	49	25	35
$\bar{X}=5$	$\bar{Y}=7$	$\sum=0$	$\sum=34$	$\sum=0$	$\sum=54$	$\sum=-37$	$\sum=209$	$\sum=397$	$\sum=208$

Table A.2

t Distribution: Critical Values of t

		Significance level					
Degrees of freedom	*Two-tailed test:*	10%	5%	2%	1%	0.2%	0.1%
	One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408

- $b_1 = -1.09$
- $S_{b_1}^2 = \dfrac{S_e^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \dfrac{SSE/(n-2)}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \dfrac{13.7354/(7-2)}{34} = 0.08-796 \Rightarrow S_{b_1} = 0.2842$
- 95% CI for β_1

  $= b_1 \pm t_{\alpha/2,n-2}S_{b_1} = -1.09 \pm 2.571 \times 0.2842 = [-1.821, -0.359]$

  The 95% CI for the expected decrease in the number of days taken off work resulting from one additional year of service is between 0.359 and 1.821

Hypothesis Testing for Regression Slope β_1

- H_0: $\beta_1 = 0$ (no linear relationship)
- H_1: $\beta_1 \neq 0$ (linear relationship exists)
- test statistic $t = \frac{b_1 - \beta_1}{se(b_1)}$ follows a *t* distribution with *df = n-2*
- Critical value approach
  - Reject H_0 if $t < -t_{\frac{\alpha}{2},n-2}$ or $t > t_{\frac{\alpha}{2},n-2}$ at a significance level of α
- p-value approach
  - p-value = $P(t \leq -|t|) + P(t \geq |t|)$
  - Reject H_0 if p-value $< \alpha$
- The same t can also be used for testing the hypotheses

  $H_0 : \beta_1 \leq 0$ vs $H_1 : \beta_1 > 0$, or $H_0 : \beta_1 \geq 0$ and $H_1 : \beta_1 < 0$

Example for Hypothesis Testing about the Slope

In the example on number of days taken off work ,

X = Number of years of service

Y = Number of days taken off work

Is years of service linearly influencing the number of days taken off work? Test at 5% level of significance

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

At $\alpha = 0.05$

$n = 7 \quad df = 5$

Critical Value = ± 2.571

Reject H_0 if $t < -2.571$ or $t > +2.571$

Given $b_1 = -1.09$ and $S_{b_1} = 0.2842$,

$$t = \frac{b_1}{S_{b_1}} = \frac{-1.09}{0.2842} = -3.835$$

p-value = $P(t \leq -3.835) + P(t \geq 3.835)$

0.01 < p-value < 0.02

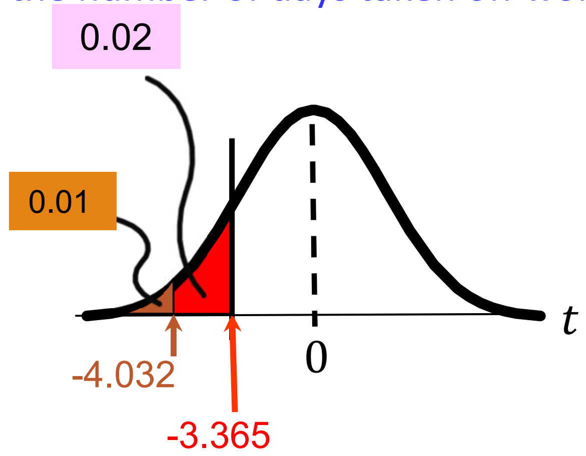
At $\alpha = 0.05$, reject H_0

There is evidence that years of service is linearly relating to the number of days taken off work

TABLE A.2

t Distribution: Critical Values of *t*

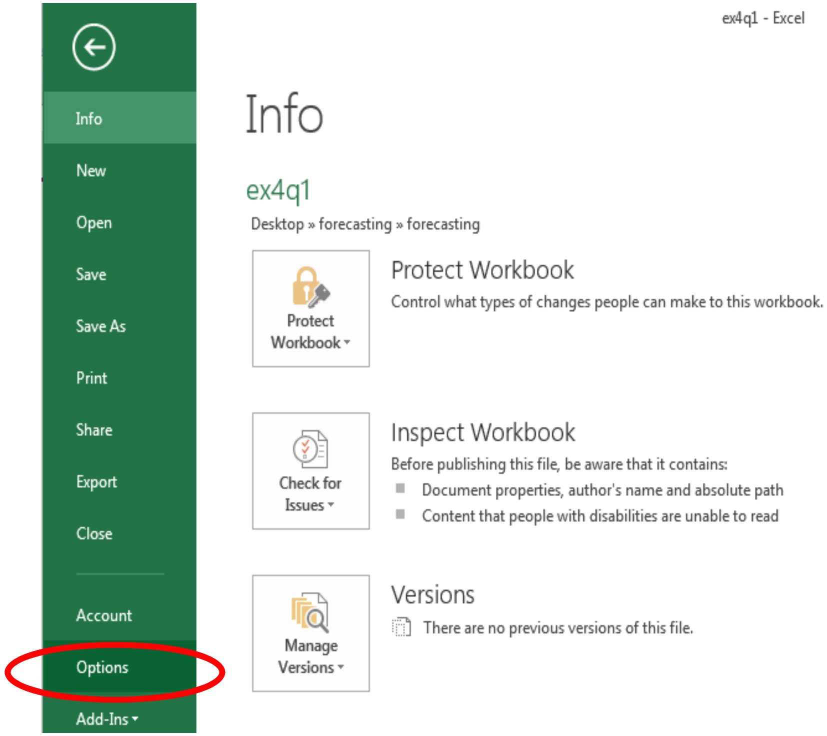
Degrees of freedom	Two-tailed test: / One-tailed test: (Significance level)	10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408

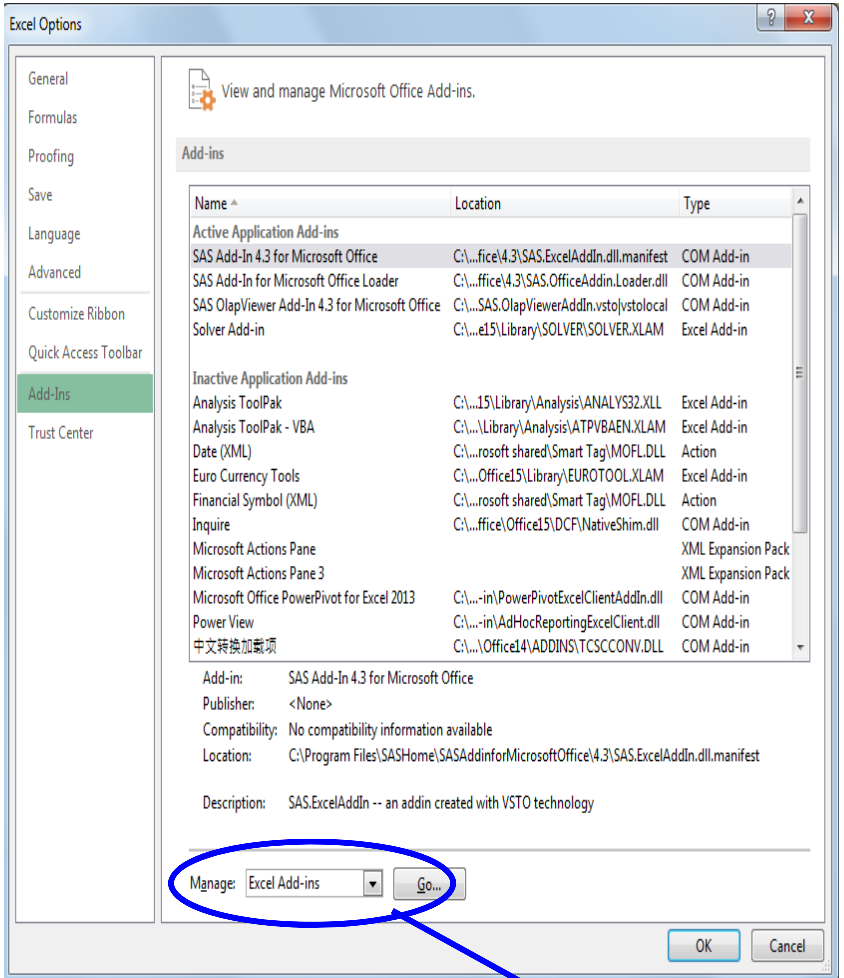


Developing Regression Model in Excel

To install Data Analysis package:

- Go to **File** tab and choose **Options**

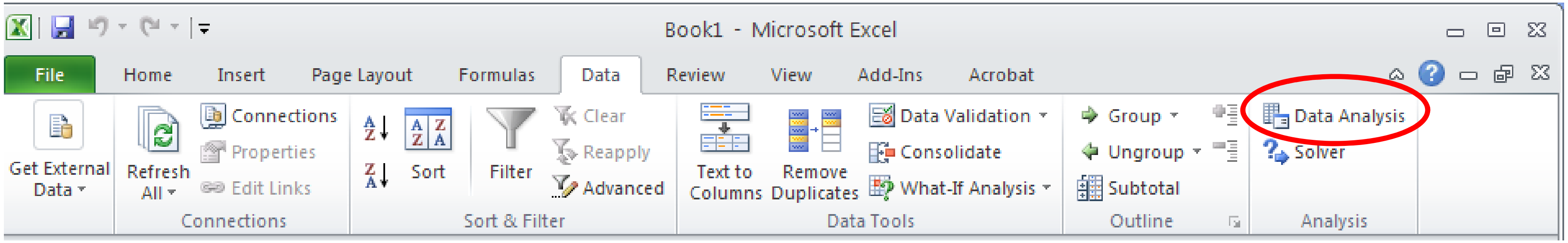




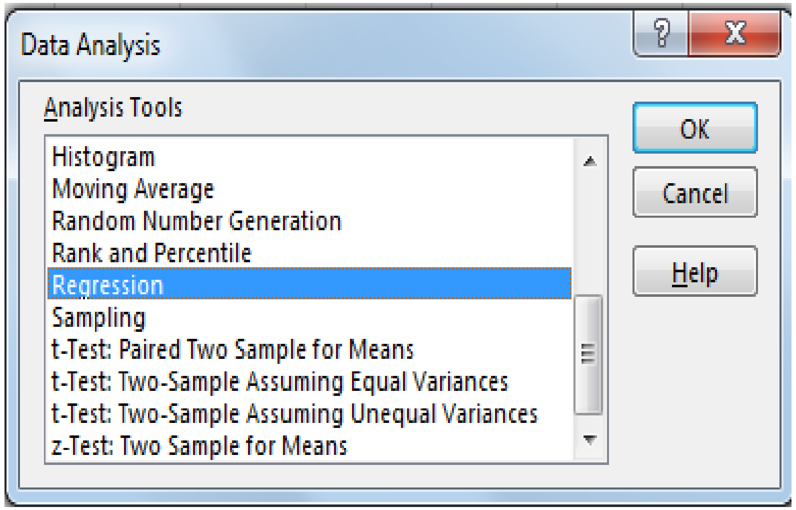
- In the page of Add-Ins within the Excel Options, Select Excel add-ins in Manage button and Click **Go**

Developing Regression Model in Excel

- In the Add-Ins dialog box, select the Analysis ToolPak and then click OK.
- Find "Data Analysis" in the "Data" menu bar



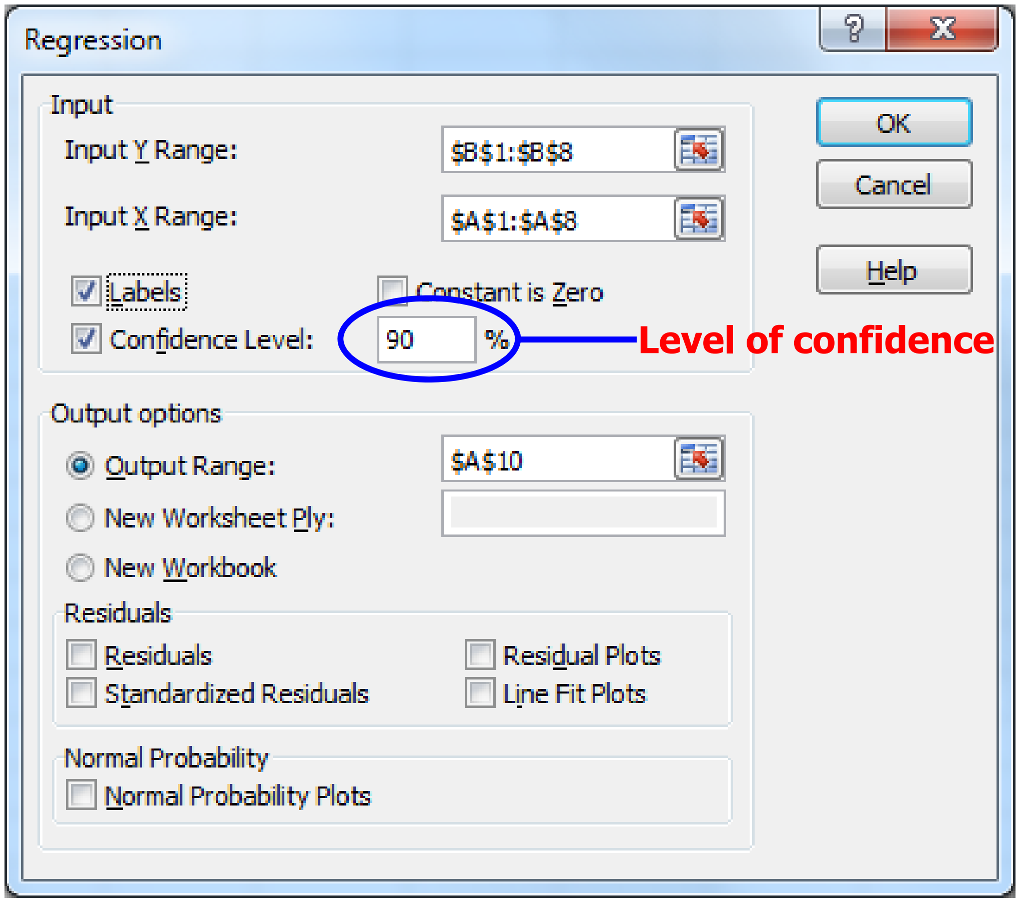
- Select Data Analysis. In the Data Analysis dialog box, choose Regression and click OK.



Developing Regression Model in Excel

- Data

	A	B
1	X	Y
2	2	8
3	5	7
4	7	5
5	3	12
6	8	3
7	3	9
8	7	5



Developing Regression Model in Excel

- Output

SUMMARY OUTPUT

Regression Statistics		
Multiple R	$\lvert r_{XY} \rvert$	0.8635
R Square	R^2	0.7456
Adjusted R Square		0.6948
Standard Error	S_e	1.6574
Observations	n	7

ANOVA

	df	SS	MS	F	Significance F
Regression	1	40.2647	40.2647	14.6574	0.0123
Residual	5	SSE 13.7353	2.7471		
Total	6	SST 54			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 90.0%	Upper 90.0%
Intercept	b_0 12.4412	1.5532	8.0102	0.0005	8.4486	16.4337	9.3115	15.5709
X	b_1 -1.0882	S_{b_1} 0.2842	-3.8285	0.0123	-1.8189	-0.3576	-1.6610	-0.5155

t for β_1 p-value for β_1 95% CI for β_1 90% CI for β_1