GE2262 Business Statistics

Topic 8
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

Outline

- Scatter Plot
- Covariance and the Coefficient of Correlation
- Simple Linear Regression
  - Least Squares Estimation
  - Predictions in Regression Analysis
  - Coefficient of Determination
  - Inferences about the Slope
- Applications of Linear Regression

Association Between Two Numerical Variables

- To visualize the relationship between two numerical variables
  - Using scatter plot
- To measure the degree of linear association
  - Using coefficient of correlation
- To forecast one variable for given values of the other
  - Using regression model
- Examples
  - Apartment price vs. Gross floor area
  - Weekly sales for chain stores vs. Number of customers

Association Between Two Numerical Variables

Cont'd

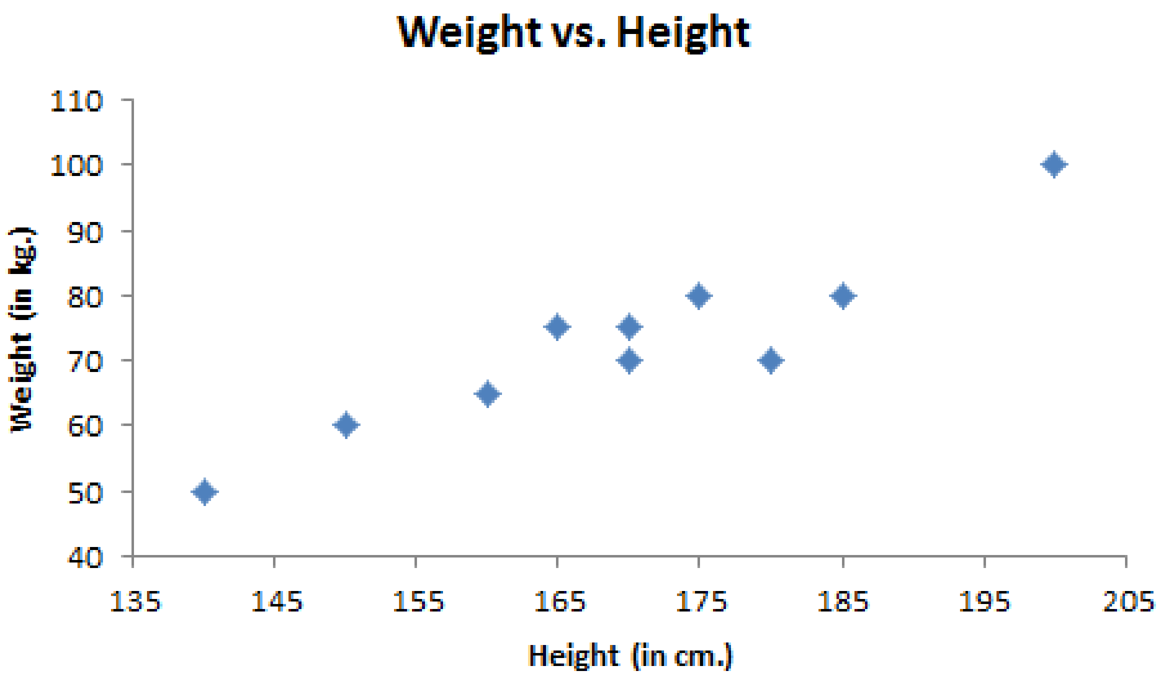
- We will look at two variables measuring different characteristics of some population of individuals
- Usually consist of paired sample data corresponding to pairs of observations on the two variables for n members of a sample taken from the population
- If two variables are related, then the nature of the relationship may be indicated by plotting paired samples of observations for both variables on a scatter plot

Association Between Two Numerical Variables – Example

Cont'd

- Consider the following data for variables from a sample of 10 students
  - X = Height (in cm.)
  - Y = Weight (in kg.)

X	Y
170	75
185	80
165	75
140	50
180	70
150	60
200	100
160	65
175	80
170	70



Association Between Two Numerical Variables – Example

Cont'd

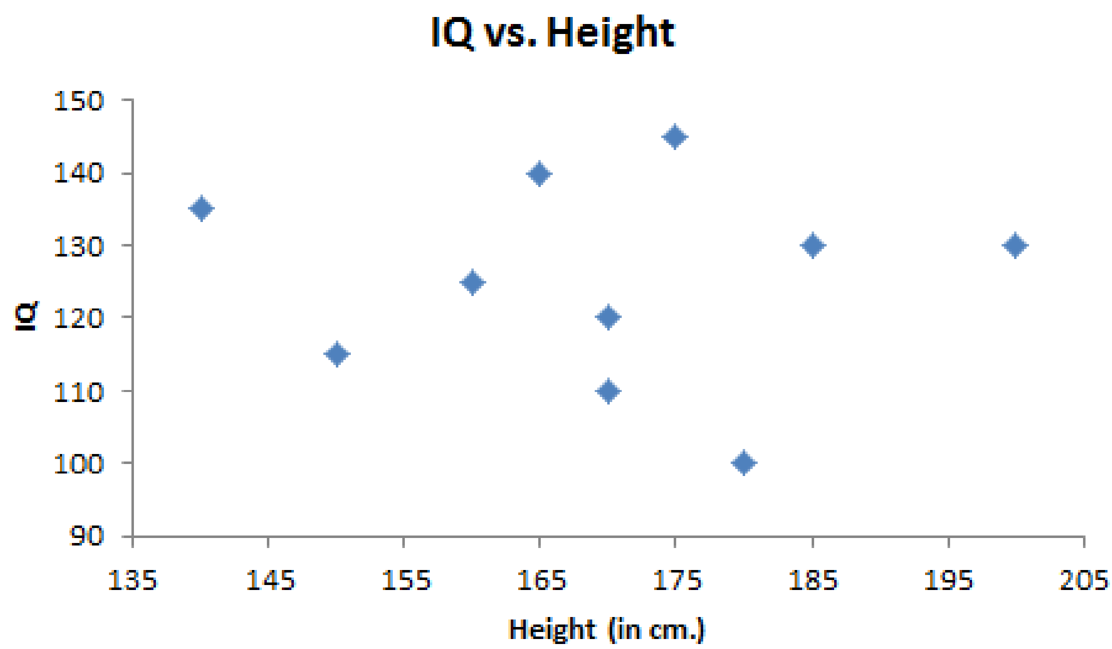
- There is a clear tendency for small values of X to be associated with small values of Y, and large X with large Y
- The dots on the scatter plot lie "close to" a straight line with a positive slope
- We say that these two variables, height and weight, have a positive linear association

Association Between Two Numerical Variables – Example

Cont'd

- Consider the scatter plot between Height (X) and IQ (Z) for the same 10 students

X	Z
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, as you might well expected, since there is no known relationship between height and IQ

Association Between Two Numerical Variables – Example

Cont'd

- If the dots on the scatter plot lie "close to" a straight line with negative slope, we say that the variables exhibit a negative linear association

Covariance

- How do we measure the degree of linear association between two variables X and Y?
- The answer to this question is the covariance
  - A quantity that measures the linear association
- Population covariance

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

- Sample covariance

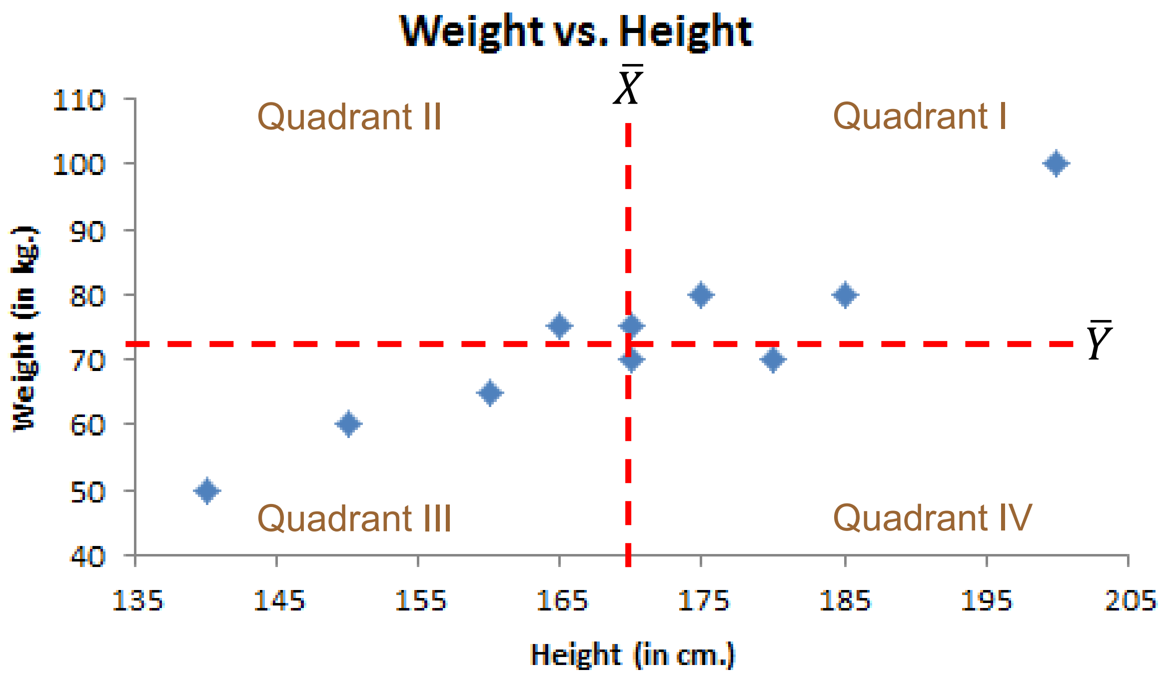
$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

  - An estimator of σ_{XY} based on n pairs of sample values

Covariance

Cont'd

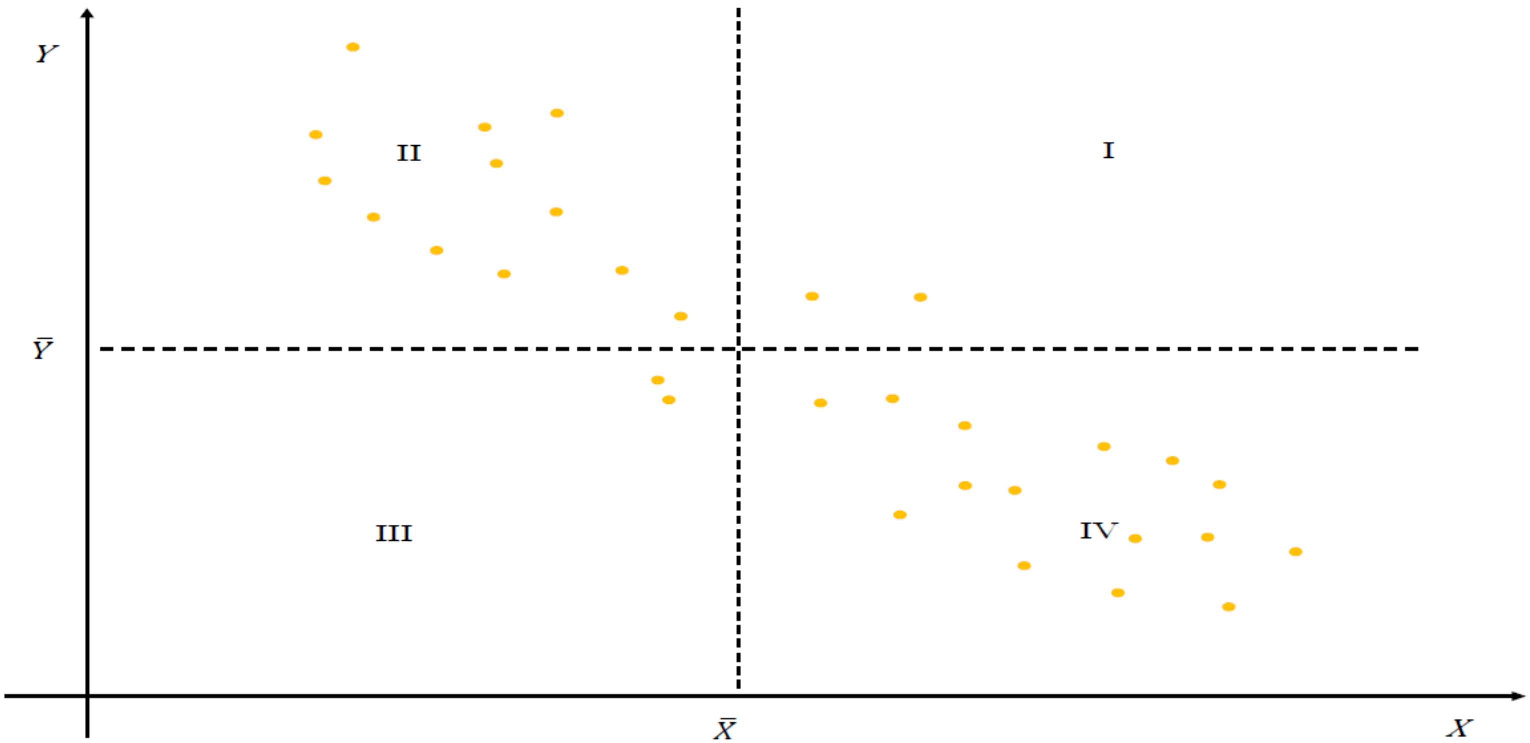
- The cross product term $(X_i - \mu_X)(Y_i - \mu_Y)$ will be positive in quadrants I and III, and negative in quadrants II and IV
- With positive linear association, there is a tendency for the dots to lie predominantly in quadrants I and III



Covariance

Cont'd

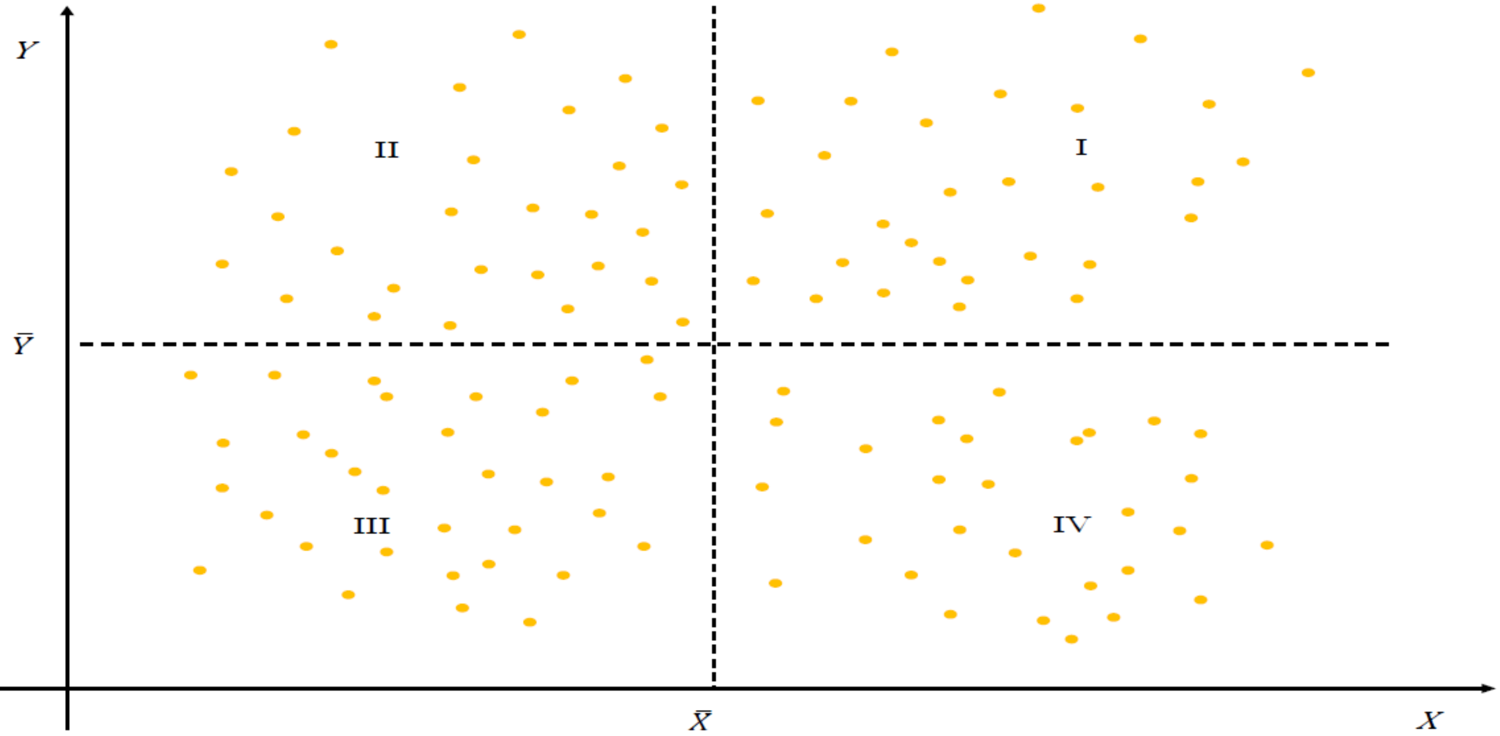
- On the other hand, with negative linear association, there is a tendency for the dots to lie predominantly in quadrants II and IV



Covariance

Cont'd

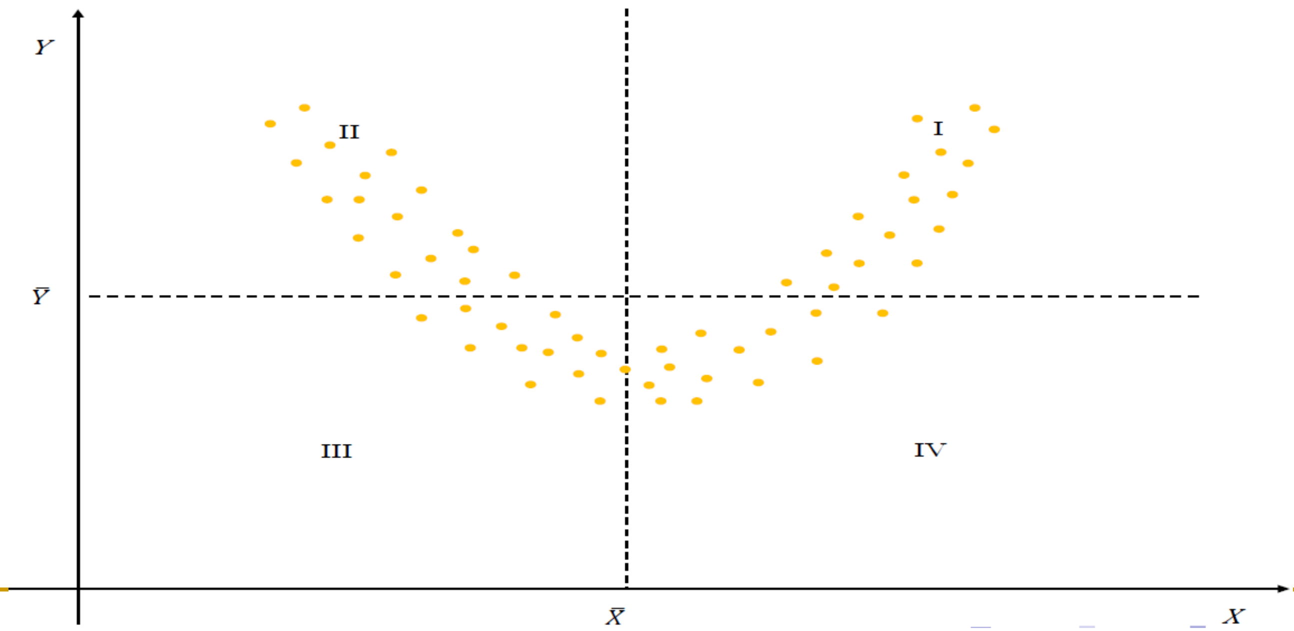
- If there is no or very weak linear association, then there is a tendency for the dots to scatter across all four quadrants



Covariance

Cont'd

- The covariance only measures linear association
- A covariance of zero does not necessarily imply that X and Y have no association because they may be related in a non-linear way



Covariance – Example

Cont'd

- Consider the sample data regarding Height (X) and Weight (Y)

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
170	75	0.5	2.5	1.25
185	80	15.5	7.5	116.25
165	75	-4.5	2.5	-11.25
140	50	-29.5	-22.5	663.75
180	70	10.5	-2.5	-26.25
150	60	-19.5	-12.5	243.75
200	100	30.5	27.5	838.75
160	65	-9.5	-7.5	71.25
175	80	5.5	7.5	41.25
170	70	0.5	-2.5	-1.25
$\bar{X}=$ **169.5**	$\bar{Y}=$ **72.5**			$S_{XY}=$ **215.28**

Covariance – Example

Cont'd

- Let's convert the height of the students from cm. to m.

X'	Y	$X' - \overline{X'}$	$Y - \overline{Y}$	$(X' - \overline{X'})(Y - \overline{Y})$
1.7	75	0.005	2.5	0.0125
1.85	80	0.155	7.5	1.1625
1.65	75	-0.045	2.5	-0.1125
1.4	50	-0.295	-22.5	6.6375
1.8	70	0.105	-2.5	-0.2625
1.5	60	-0.195	-12.5	2.4375
2	100	0.305	27.5	8.3875
1.6	65	-0.095	-7.5	0.7125
1.75	80	0.055	7.5	0.4125
1.7	70	0.005	-2.5	-0.0125
$\overline{X'} = \mathbf{1.695}$	$\overline{Y} = \mathbf{72.5}$			$S_{X'Y} = \mathbf{2.1528}$

- The sample covariance is reduced by a factor of 100

Covariance

Cont'd

- One problem with the 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

Coefficient of Correlation

- The coefficient of correlation measures the relative strength of a linear association between two variables that is not affected by the variables' units of measure
  - 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

Coefficient of Correlation

Cont'd

- Population coefficient of correlation

pronounced rho

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

- Sample coefficient of correlation

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

  - An estimator of ρ_{XY}
- The sign of ρ_{XY} (r_{XY}) is the same as that of σ_{XY} (S_{XY})
  - As the denominator of ρ_{XY} is always non-negative

Coefficient of Correlation – Example

- Consider the sample data regarding Height (X) and Weight (Y) again

X	Y	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})(Y-\overline{Y})$	$(X-\overline{X})^2$	$(Y-\overline{Y})^2$
170	75	0.5	2.5	1.25	0.25	6.25
185	80	15.5	7.5	116.25	240.25	56.25
165	75	-4.5	2.5	-11.25	20.25	6.25
140	50	-29.5	-22.5	663.75	870.25	506.25
180	70	10.5	-2.5	-26.25	110.25	6.25
150	60	-19.5	-12.5	243.75	380.25	156.25
200	100	30.5	27.5	838.75	930.25	756.25
160	65	-9.5	-7.5	71.25	90.25	56.25
175	80	5.5	7.5	41.25	30.25	56.25
170	70	0.5	-2.5	-1.25	0.25	6.25
$\overline{X}=169.5$	$\overline{Y}=72.5$			$S_{XY}=215.28$	$S_X=17.232$	$S_Y=13.385$

- $r_{XY}=S_{XY}/S_XS_Y=215.28/(17.232\times 13.385)=0.933$

Coefficient of Correlation – Example

Cont'd

- What if the height is measured in m.?

X'	Y	$X' - \overline{X'}$	$Y - \overline{Y}$	$(X' - \overline{X'})(Y - \overline{Y})$	$(X' - \overline{X'})^2$	$(Y - \overline{Y})^2$
1.7	75	0.005	2.5	0.0125	0.000025	6.25
1.85	80	0.155	7.5	1.1625	0.024025	56.25
1.65	75	-0.045	2.5	-0.1125	0.002025	6.25
1.4	50	-0.295	-22.5	6.6375	0.087025	506.25
1.8	70	0.105	-2.5	-0.2625	0.011025	6.25
1.5	60	-0.195	-12.5	2.4375	0.038025	156.25
2	100	0.305	27.5	8.3875	0.093025	756.25
1.6	65	-0.095	-7.5	0.7125	0.009025	56.25
1.75	80	0.055	7.5	0.4125	0.003025	56.25
1.7	70	0.005	-2.5	-0.0125	0.000025	6.25
$\overline{X'} = 1.695$	$\overline{Y} = 72.5$			$S_{X'Y} = 2.1528$	$S_{X'} = 0.1723$	$S_Y = 13.385$

- $r_{X'Y} = S_{X'Y}/S_{X'}S_Y = 2.1528/(0.1723 \times 13.385) = 0.933$
  - The sample correlation remains unchanged although the sample covariance has been reduced by a factor of 100

Coefficient of Correlation

Cont'd

- It can be shown it is always the case that

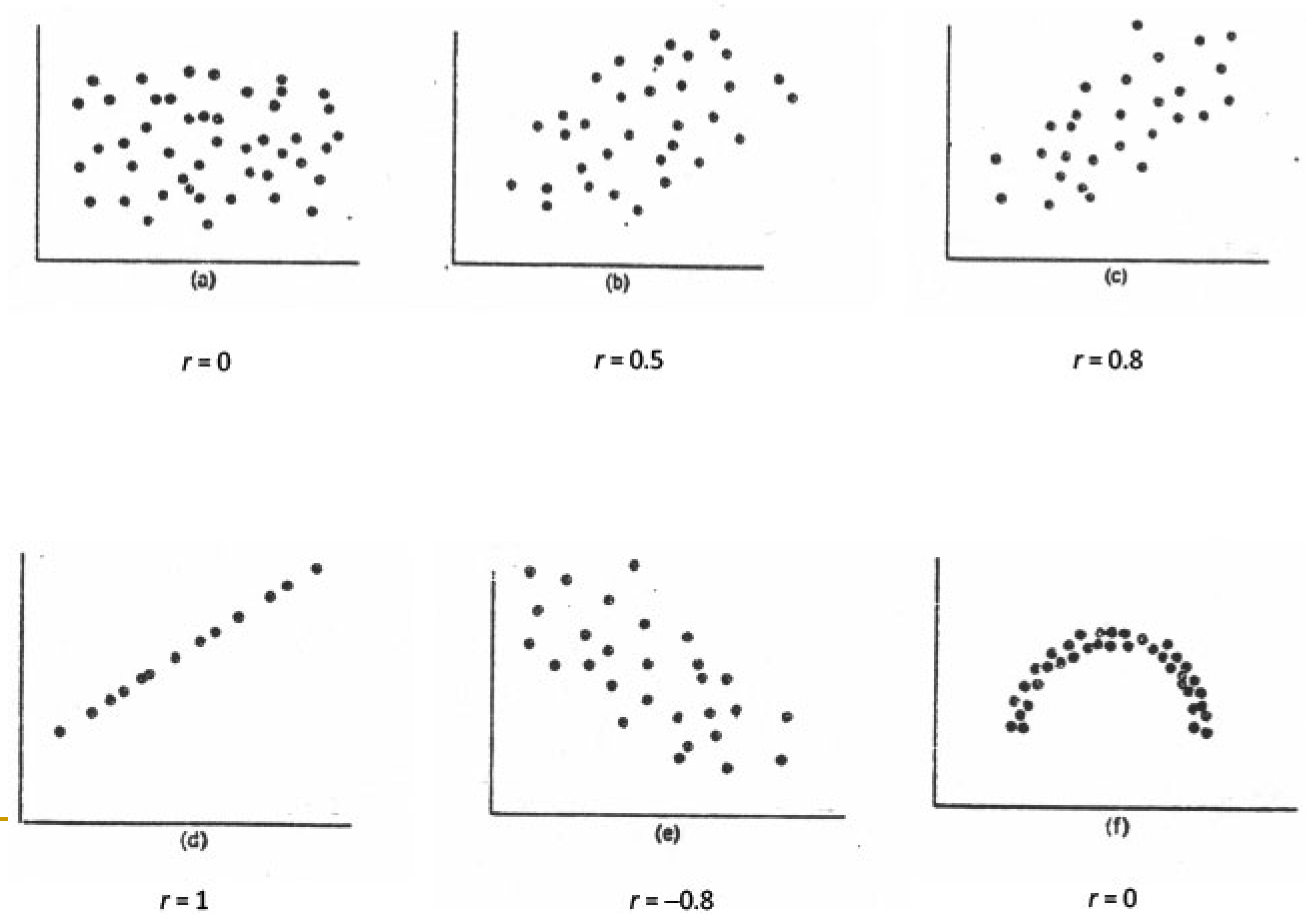
  $$-1 \le \rho_{XY} \le 1 \qquad \text{and} \qquad -1 \le r_{XY} \le 1$$

- Three special values of ρ_{XY} and r_{XY} are of interest
  - When $\rho_{XY} = 0$ $(r_{XY} = 0)$, X and Y are not linearly related, and we say that X and Y are uncorrelated in the population (sample)
  - When all population (sample) values of X and Y lie exactly on a straight line having a positive slope, then $\rho_{XY} = 1$ $(r_{XY} = 1)$
  - When all population (sample) values of X and Y lie exactly on a straight line having a negative slope, then $\rho_{XY} = -1$ $(r_{XY} = -1)$
- If the population (sample) values of X and Y lie close to a straight line, then ρ_{XY} (r_{XY}) will be close to 1 or -1

Coefficient of Correlation

Cont'd

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



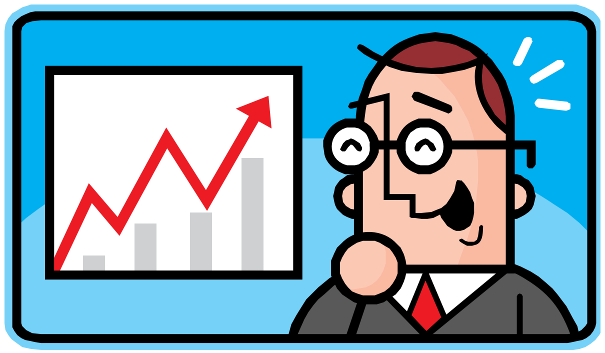
Coefficient of Correlation

Cont'd

- Correlation alone cannot prove that there is a causation effect
  - Causation effect means that the change in the value of one variable caused the change in the other variable

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

Cont'd

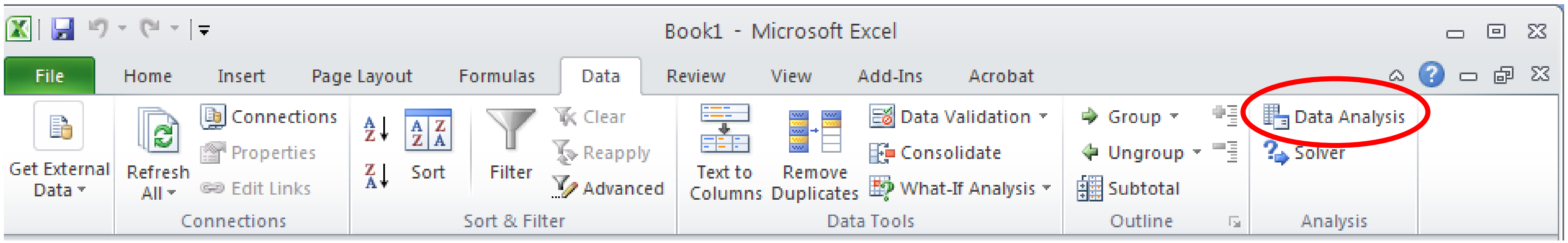
- The following data represent the annual rates of return for various stocks

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

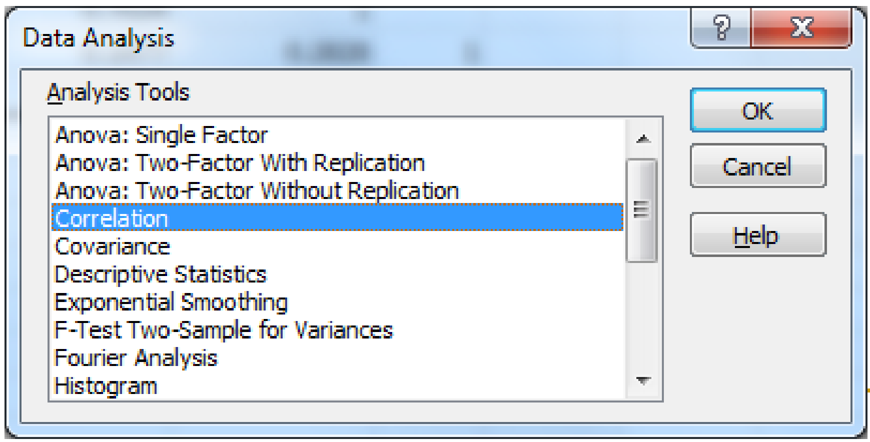
Source: Yohoo!Finance

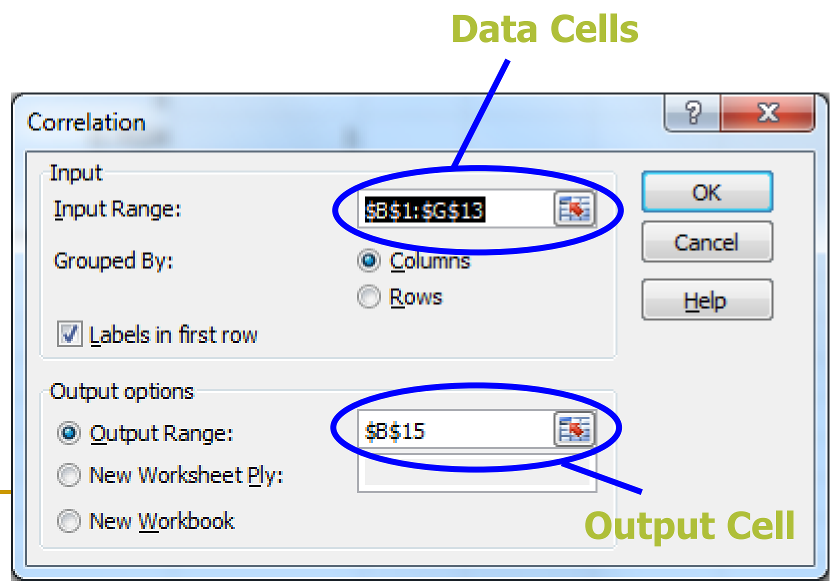
Calculating Coefficient of Correlation in Excel

- Find “Data Analysis” in the “Data” menu bar



- Choose “Correlation” at “Data Analysis” browser





Diversifying Your Investments

Cont'd

	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?
  - Which two would you select if your goal is to have one stock go up when the other goes down?

Diversifying Your Investments

Cont'd

	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

Inferences about the Slope – Exercise

Cont'd

- Refer to the example our example on number of days taken off work, given $b_1 = -1.09$ and $S_{b_1} = 0.2842$
- A 95% CI for β_1 is

95% CI for β_1
$= b_1 \pm t_{\alpha/2,n-2} S_{b_1}$
$= -1.09 \pm 2.5706 \times 0.2842$
$= [-1.821, -0.359]$

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

Inferences about the Slope – Exercise

Cont'd

- In the example on number of days taken off work , test at 5% level of significance, is years of service linearly influencing the number of days taken off work?

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
At $\alpha = 0.05$
$n = 7 \quad df = 5$
Critical Value = ± 2.5706
Reject H_0 if $t < -2.5706$ or $t > +2.5706$

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

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

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

Hong Kong Population

Cont'd

1. $r_{XY} = 0.9914$ is very close to +1, indicating X and Y have a very strong positive linear relationship
2. $\hat{Y} = 3332.2934 + 79.5741X$
   - So, $b_0 = 3332.2934$ is the predicted Hong Kong population size for the year 1960 (X = 0)
   - $b_1 = 79.5741$ is the predicted average annual increment in population size
3. $R^2 = 0.9829$ indicating that the estimated regression line has the ability to capture 98.29% of the variation in Y in the sample

Linear Regression Model

- Suppose that a scatter plot or the coefficient of correlation indicates linear association between two variables, then it is quite easy
  - To fit a straight line to the scatter plot, and
  - Use the fitted straight line to forecast values of one variable (indicated as Y variable or dependent variable) given values of the other (indicated as X variable or independent variable)
    - In other words, given that the variable X takes a specific value, we expect a response in the variable Y
    - This can be thought of as a dependency of Y on X

Linear Regression Model

Cont'd

- Our concern is with the value taken by the variable Y, when the variable X takes a specific value
- The variable Y could take many different values for a specific X value
  - For example, we may be interested in the value of retail sales per household in a year in which disposable income per household is $12,000. At that income, the retail sales value per household in the population could be $5,800, or $5,900, or $6,000, etc. It is not reasonable to think of just a single possible retail sales level resulting from a particular value for disposable income

Linear Regression Model

Cont'd

- It is more realistic to consider a distribution of possible Y values resulting from each possible X value
- A crucial characteristic of this distribution is the population mean, or the expected value, of Y when X takes a specific value
  - For example, we can ask what would be the average (population mean) retail sales per household in which disposable income per household was $12,000

Linear Regression Model

Cont'd

- In general, we will denote the expected value of the variable Y, when the variable X takes the specific value of x by

$$E(Y|X = x)$$

- Our assumption of linearity is the assumption that this conditional expectation depends linearly on x
- This implies that

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

  where the fixed numbers β_0 and β_1 determine a specific straight line

  - The true values of β_0 and β_1 are unknown to us

Linear Regression Model

Cont'd

- We hypothesize that the conditional expected value of Y_i depends linearly on X_i
  - Such hypothesis will not hold exactly in the real world
  - In addition, we do not actually observe the expected value of Y_i for the X_i
- Denote the discrepancy between the observed Y_i and its conditional expected value $E(Y_i|X = X_i)$ by ε_i such that

$$\varepsilon_i = Y_i - E(Y_i|X = X_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

Linear Regression Model

Cont'd

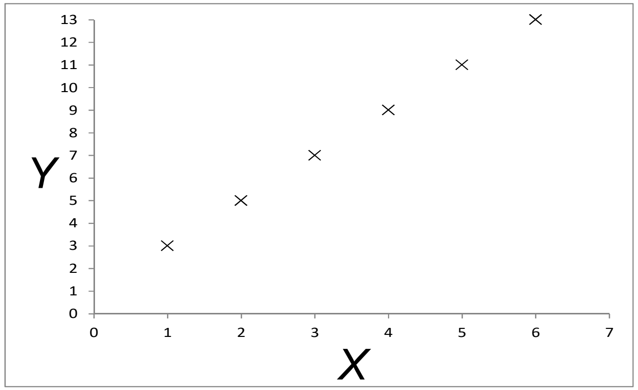
- The population (or true) regression line is defined as

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

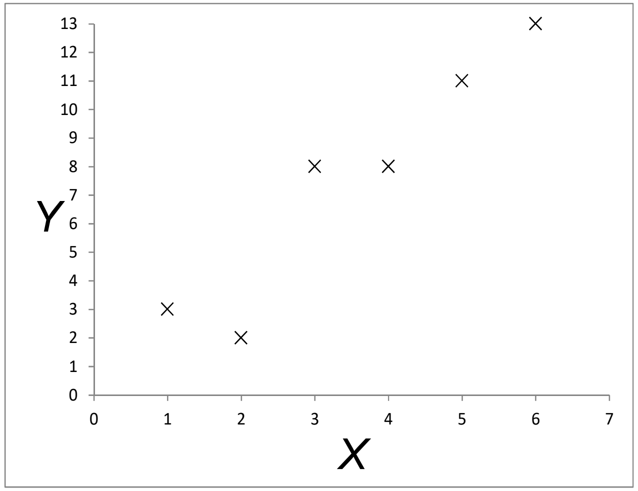
- The response of Y_i to a particular value X_i will be the sum of two parts
    - An expectation ($\beta_0+\beta_1 X_i$) reflecting their systematic relationship
    - A discrepancy ε_i from the expectation, often called the error term
- Since the population regression line involves on one independent variable (X_i), the line is sometimes called simple linear regression model

Least Squares Estimation

- If the random error term ε_i = 0 for all i, it implies $Y_i = \beta_0 + \beta_1 X_i$ exactly



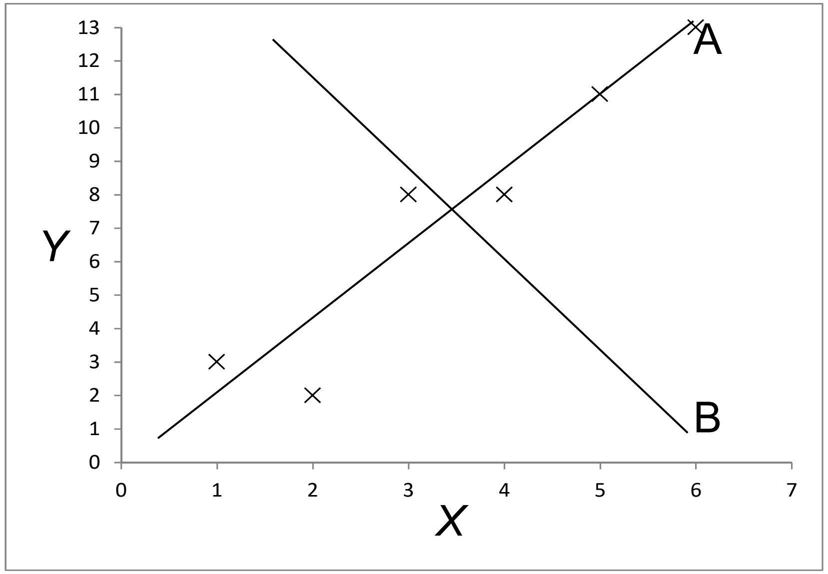
- If the random error term ε_i , i = 1, ..., n, are not all equal to 0, then the n observed pairs (X_i, Y_i), i = 1, ..., n, cannot be drawn on a straight line
  - It is possible to find a straight line that will fit the set as accurately as possible, i.e. the fitting errors should be minimized



Least Squares Estimation

Cont'd

- Consider two lines A and B, both are fitted to the same set of (X_i, Y_i) pairs

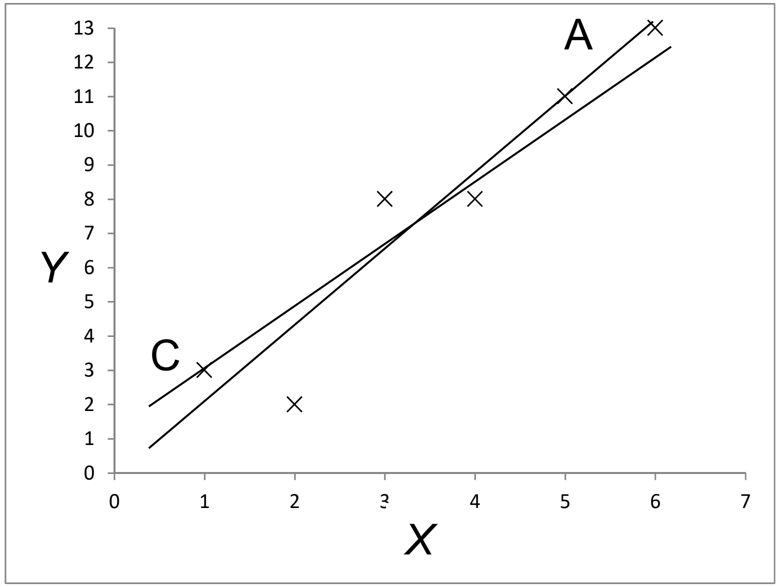


  - Which line seems to fit the set of points better? Why?

Least Squares Estimation

Cont'd

- Consider anther two possible lines A and C, both are fitted to the same set of (X_i, Y_i) pairs

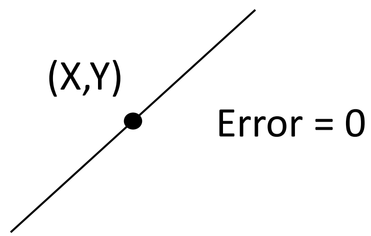


- Which line seems to fit the set of points better? Why?

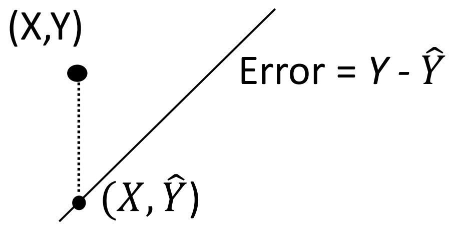
Least Squares Estimation

Cont'd

- For a given X, when a line pass through the point $(X\ ,Y)$ exactly, we say there is no error



- When a line does not pass through the point, we say there is an error
- The amount of error is represented by the distance between the actual value (Y) and the fitted (or predicted) value $(\hat{Y})$ given by the straight line for the same X



- That is, $error = Y - \hat{Y}$
  - This error is also called residual in regression analysis, and denoted as e

Least Squares Estimation

Cont'd

- We must consider the entire set of (X_i, Y_i), i = 1, …, n, for determining the goodness of fit
- Consider an observed set of (X_i, Y_i), i = 1, …, n, suppose there exists a straight line

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

  such that it minimizes the sum of squared errors (SSE)

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

  - Least-squares criterion is about finding such for b_0 and b_1
  - The resulting line is often called the least-squares regression line

Least Squares Estimation

Cont'd

- It is possible to show using calculus that the least-squares form of b_0 and b_1 can be determined as

$$b_0 = \bar{Y} - b_1\bar{X} \qquad \text{and} \qquad b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

  - b_0 and b_1 are the least squares estimates for β_0 and β_1 respectively

Least Squares Estimation

Cont'd

- The estimated b_1 is also related to the sample coefficient of correlation r_{XY} as follows

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

- Since S_X and S_Y are non-negative, b_1 will have the same sign as r_{XY}

Least Squares Estimation – 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

X	Y	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
2	8	-3	9	1	1	-3
5	7	0	0	0	0	0
7	5	2	4	-2	4	-4
3	12	-2	4	5	25	-10
8	3	3	9	-4	16	-12
3	9	-2	4	2	4	-4
7	5	2	4	-2	4	-4
$\bar{X}=5$	$\bar{Y}=7$	$\sum=0$	$\sum=34$	$\sum=0$	$\sum=54$	$\sum=-37$

Least Squares Estimation – Example

Cont'd

- Therefore

$$r_{XY} = -37/\sqrt{34 \times 54} = -0.864$$

$$b_1 = -37/34 = -1.09 \quad \text{or} \quad b_1 = -0.864\frac{\sqrt{54}}{\sqrt{34}} = -1.09$$

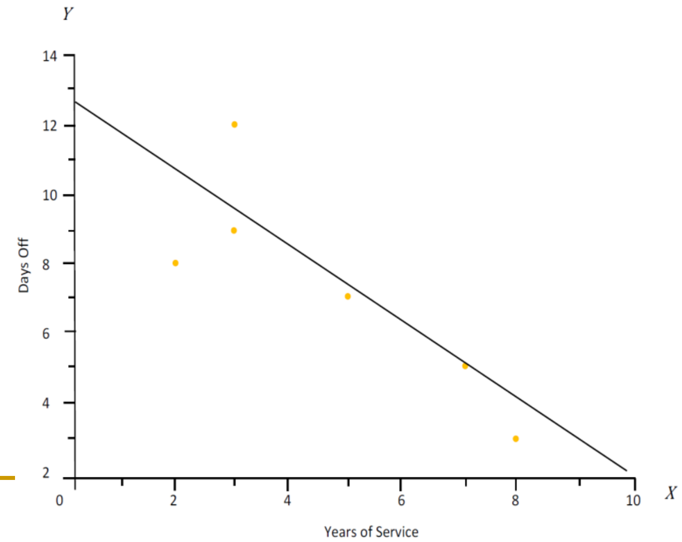
$$b_0 = 7 - (-1.09)5 = 12.45$$

- The least-squares regression line is

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

where $\hat{Y}$ = predicted or fitted value of Y for a given value of X

  - $\hat{Y} = Y$ for all sample values if and only if $|r_{XY}| = 1$



Predictions in Regression Analysis – Example

Cont'd

- 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
  - 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

What???

Interpreting the Estimated Coefficients – Example

Cont'd

- Interpreting b_0: From the prediction of Y for X = 0, we see that b_0 = 12.45 is the predicted number of days off for an employee with 0 years of service
  - We should not take this interpretation seriously as this probably would never happen
  - 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 our conclusions far outside that range

Interpreting the Estimated Coefficients – Example

Cont'd

- Interpreting b_1: 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, thus b_1 is the change in the estimated number of days off for an additional year's service
  - We are estimating that each 1 year increase in service leads, on average, to a decrease of 1.09 days off work

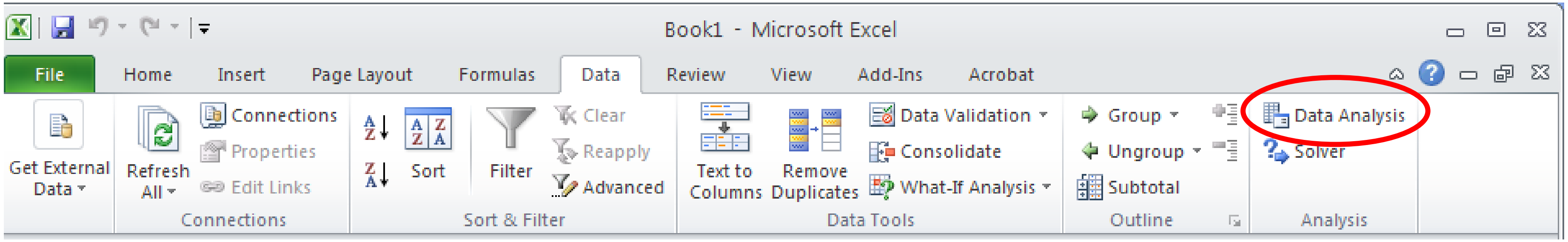
Interpreting the Estimated Coefficients – Example

Cont'd

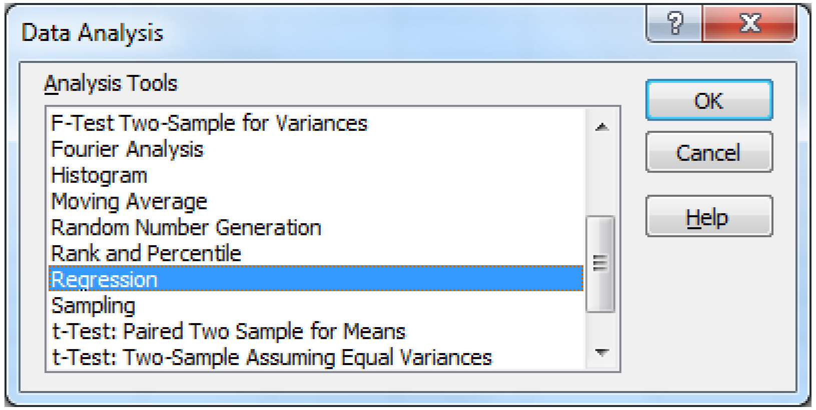
- The regression line gives a non-sense prediction of -2.81 days off work for X = 14 years of service, because
  - The relationship between X and Y is approximately linear over the range covered by the sample, but the regression line cannot be extended indefinitely without cutting the X-axis
  - Once 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

Developing Regression Model in Excel

- Find “Data Analysis” in the “Data” menu bar



- Choose “Regression” at “Data Analysis” browser



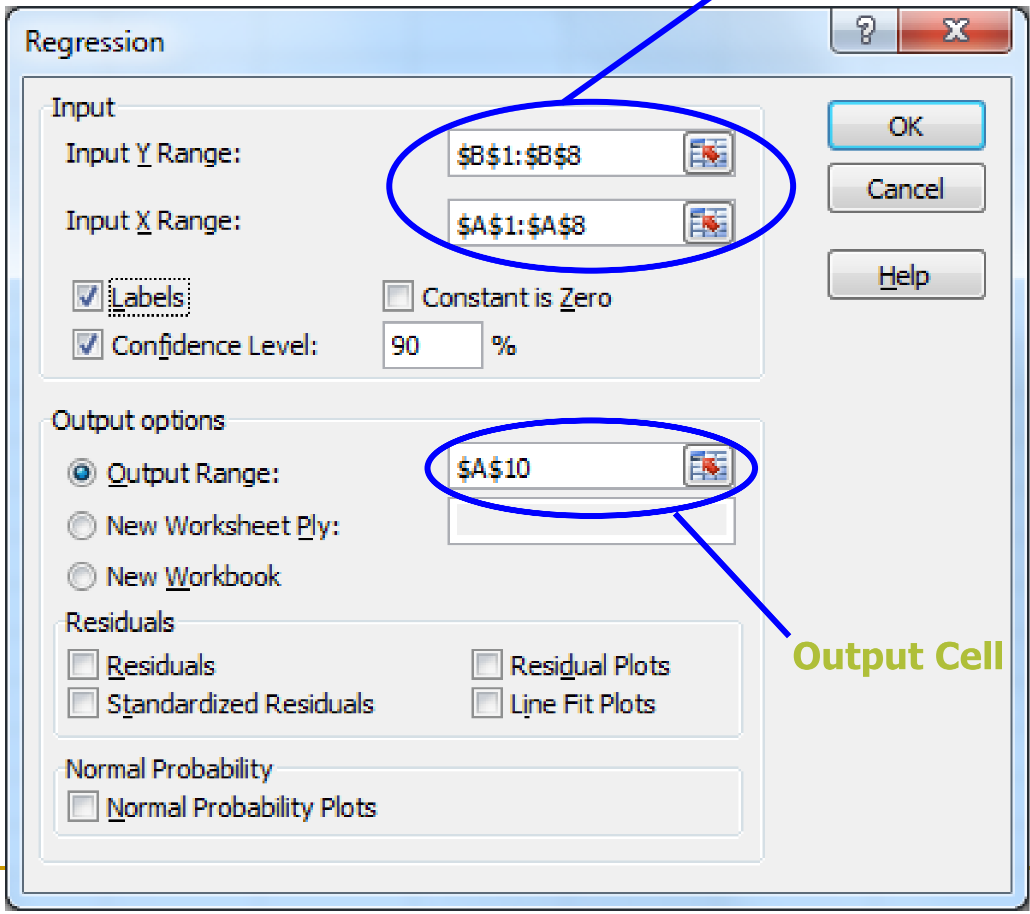
Developing Regression Model in Excel

Cont'd

- 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

Data Cells for *Y* and *X* variables



Output Cell

Developing Regression Model in Excel

Cont'd

- Output

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	$\lvert r_{XY} \rvert$ 0.8635							
R Square	0.7456							
Adjusted R Square	0.6948							
Standard Error	1.6574							
Observations	n 7							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	40.2647	40.2647	14.6574	0.0123			
Residual	5	13.7353	2.7471					
Total	6	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	0.2842	-3.8285	0.0123	-1.8189	-0.3576	-1.6610	-0.5155

Coefficient of Determination

- By comparing the actual against predicted Y values, we obtain the errors ($e = Y - \hat{Y}$)
  - When $X = 5, Y = 7, \hat{Y} = 7, e = 0$
  - When $X = 8, Y = 3, \hat{Y} = 3.73, e = -0.73$
    - It over-estimates the number of days off work
- This does not mean our model is bad as the regression line can never make a precise prediction without errors unless the linear association is perfect

Coefficient of Determination

Cont'd

- The least-squares regression line minimizes the sum of squared errors, $SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$. In theory, no other straight line will give a smaller value of SSE for the same set of data
- In general, the smaller the amount of SSE, the better the data fit to a straight line
- However, SSE is scale dependent, it can be made as large or as small by adjusting the scale of Y

Coefficient of Determination

Cont'd

- A better way to measure the goodness of fit for a least-squares regression line is to compare its SSE value to that of another regression line based on the same set of Y
- A natural second line to be compared with is $\hat{Y}_i^{\,*} = \bar{Y}$, that is, estimating the mean value of Y without using X
- The corresponding SSE is

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

  - SST is called the total variation in Y or the total sum of squares

Coefficient of Determination

Cont'd

- The goal is to determine by how much the SSE is smaller than SST
  - Or, the amount of improvement in using the regression line and the independent variable X rather than just the sample mean to predict Y
- This measure is provided through a statistic called the coefficient of determination (R^2)

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

  - 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

Coefficient of Determination – Example

Cont'd

- Thus, in our example on number of days taken off work

X	Y	$\hat{Y}$	e	e^2
2	8	10.27	-2.27	5.1529
5	7	7	0	0
7	5	4.82	0.18	0.0324
3	12	9.18	2.82	7.9524
8	3	3.73	-0.73	0.5329
3	9	9.18	-0.18	0.0324
7	5	4.82	0.18	0.0324
$\bar{X} = 5$	$\bar{Y} = 7$		$\sum = 0$	$\sum = 13.7354$

SSE = 13.7354

SST = 54

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

Coefficient of Determination – Example

Cont'd

- Commonly, 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 SSE is reduced by 74.56%

Coefficient of Determination

Cont'd

- In a regression model containing only one X variable, $R^2 = (r_{XY})^2$
- Hence, in our example, the sample correlation coefficient between X and Y is $r_{XY} = -\sqrt{0.7456} = -0.8635$
  - We know r_{XY} has a negative sign because b_1 is negative
  - r_{XY} would have a positive sign if b_1 was positive

Inferences about the Slope

- At times, tests concerning β_1 are of interest, particularly one of the forms: H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$
- If $\beta_1 = 0$, there is no linear relationship between X and Y
  - The means of the probability distribution of Y are all equal, namely $E(Y|X = x) = \beta_0 + 0x = \beta_0$ for all levels of X
  - A change in X does not induce any change in Y
- Similar to those discussed in Topics 6 & 7, we need to consider the sampling distribution of b_1, the least squares point estimate of β_1, in order to perform the inferences on β_1

Inferences about the Slope

Cont'd

- The population regression line is defined as

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

- It is very common to assume that the error terms ε_i are independent and normally distributed with mean 0 and variance σ^2, i = 1, …, n
  - This assumption can be relaxed, but it will make the inference on the slope parameter (and others) more complicated
- Under this assumption, the dependent variables Y_i are also independent and normally distributed with mean $E(Y_i) = \beta_0 + \beta_1 X_i$ and variance σ^2, i = 1, …, n
  - We are treating X_i as known constants

Inferences about the Slope

Cont'd

- Sampling distribution of b_1
  - Since the Y_i are normal, the estimator b_1 is also normal. It can be shown that b_1 has mean and variance

$$E(b_1) = \beta_1 \qquad \sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

  - The variance $\sigma_{b_1}^2$ can be estimated by $S_{b_1}^2$ as

$$S_{b_1}^2 = \frac{S_e^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{SSE/(n-2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n-2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

    - S_{b_1} measures the variability in the slope of regression lines arise from different possible samples
    - S_e^2 is called the mean squared error (MSE) of the regression model. It measures the variance of the errors around the regression line. It is an unbiased estimator of σ^2

Inferences about the Slope

Cont'd

- Confidence intervals for the population regression slope
  - Since b_1 is normally distributed, when σ_{b_1} is estimated by S_{b_1}, the statistic

    $\frac{b_1-\beta_1}{S_{b_1}} \sim t$ with n-2 degrees of freedom
  - If the error term ε_i are normally distribution as assumed, a 100(1-α)% confidence interval for the population regression slope β_1 is given by

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

    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 $n-2$

Inferences about the Slope

Cont'd

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

Inferences about the Slope – Exercise

Cont'd

- Refer to the example on number of days taken off work, given $b_1 = -1.09$ and $S_{b_1} = 0.2842$
- A 95% CI for β_1 is

95% CI for β_1
$= b_1 \pm t_{\alpha/2,n-2} S_{b_1}$

Diversifying Your Investments

Cont'd

	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

Inferences about the Slope – Exercise

Cont'd

- Refer to the example our example on number of days taken off work, given $b_1 = -1.09$ and $S_{b_1} = 0.2842$
- A 95% CI for β_1 is

95% CI for β_1
$= b_1 \pm t_{\alpha/2,n-2}S_{b_1}$
$= -1.09 \pm 2.5706 \times 0.2842$
$= [-1.821, -0.359]$

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

Inferences about the Slope – Exercise

Cont'd

- In the example on number of days taken off work , test at 5% level of significance, is years of service linearly influencing the number of days taken off work?

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
At $\alpha = 0.05$
$n = 7 \quad df = 5$
Critical Value = ± 2.5706
Reject H_0 if $t < -2.5706$ or $t > +2.5706$

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

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

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

Hong Kong Population

Cont'd

1. $r_{XY} = 0.9914$ is very close to +1, indicating X and Y have a very strong positive linear relationship
2. $\hat{Y} = 3332.2934 + 79.5741X$
   - So, $b_0 = 3332.2934$ is the predicted Hong Kong population size for the year 1960 (X = 0)
   - $b_1 = 79.5741$ is the predicted average annual increment in population size
3. $R^2 = 0.9829$ indicating that the estimated regression line has the ability to capture 98.29% of the variation in Y in the sample

Inferences about the Slope

Cont'd

- Hypothesis testing for β_1
  - For hypotheses H_0: $\beta_1 = 0$ and H_1: $\beta_1 \neq 0$, the t test statistic is $t = \frac{b_1}{s_{b_1}}$
  - Critical value approach
    - At α significance level, reject H_0 if $t < critical\ value_L$ or $t > critical\ value_U$ where the critical values are obtained from the t distribution table at n – 2 degrees of freedom
  - 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 $H_1 : \beta_1 < 0$

Inferences about the Slope – Exercise

Cont'd

- In the example on number of days taken off work , test at 5% level of significance, is years of service linearly influencing the number of days taken off work?

H_0:

H_1:

At $\alpha = 0.05$

$n = 7$ $df = 5$

Critical Value =

Reject H_0 if

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

$$t = \frac{b_1}{S_{b_1}} =$$

At $\alpha = 0.05$,

Diversifying Your Investments

Cont'd

	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

Inferences about the Slope – Exercise

Cont'd

- Refer to the example our example on number of days taken off work, given $b_1 = -1.09$ and $S_{b_1} = 0.2842$
- A 95% CI for β_1 is

95% CI for β_1
$= b_1 \pm t_{\alpha/2,n-2}S_{b_1}$
$= -1.09 \pm 2.5706 \times 0.2842$
$= [-1.821, -0.359]$

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

Inferences about the Slope – Exercise

Cont'd

- In the example on number of days taken off work , test at 5% level of significance, is years of service linearly influencing the number of days taken off work?

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
At $\alpha = 0.05$
$n = 7 \quad df = 5$
Critical Value = ± 2.5706
Reject H_0 if $t < -2.5706$ or $t > +2.5706$

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

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

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

Hong Kong Population

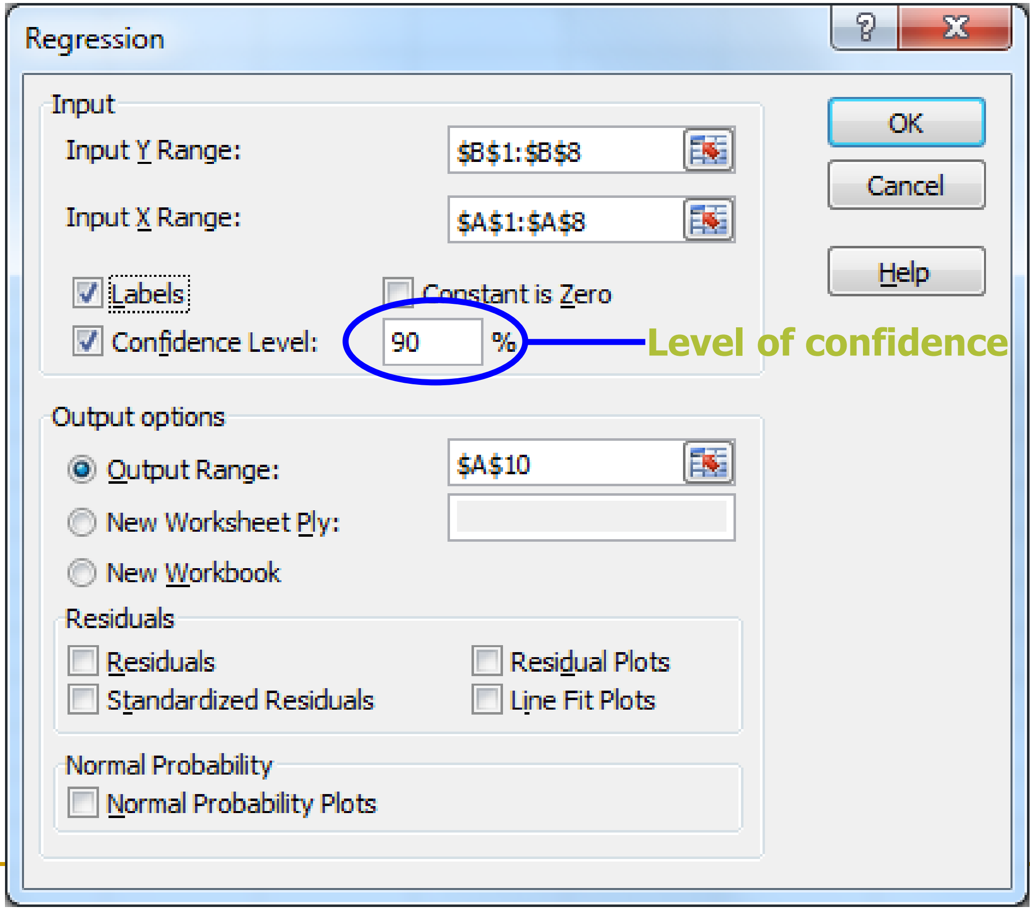
Cont'd

1. $r_{XY} = 0.9914$ is very close to +1, indicating X and Y have a very strong positive linear relationship
2. $\hat{Y} = 3332.2934 + 79.5741X$
   - So, $b_0 = 3332.2934$ is the predicted Hong Kong population size for the year 1960 (X = 0)
   - $b_1 = 79.5741$ is the predicted average annual increment in population size
3. $R^2 = 0.9829$ indicating that the estimated regression line has the ability to capture 98.29% of the variation in Y in the sample

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

Cont'd

- 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**

Calculating Correlation and Regression Coefficients in Calculator (For Casio fx-50F)

1. Calculator Mode: Lin

   MODE MODE 5 1

2. Clear Previous Data

   SHIFT CLR 1 EXE

Data Set:

Shelf space, X	5	5	5	10	10	10	15	15	15	20	20	20
Weekly sales, Y	1.6	2.2	1.4	1.9	2.4	2.6	2.3	2.7	2.8	2.6	2.9	3.1

3. Input Data

5	,	1.6	M+
5	,	2.2	M+
…	…	…	…
20	,	2.9	M+
20	,	3.1	M+

4. Calculating Regression Data

   Regression line, y-intercept, A = SHIFT 2 ▶ ▶ 1 EXE = 1.45

   Regression line, slope, B = SHIFT 2 ▶ ▶ 2 EXE = 0.074

   Coefficient of correlation, r = SHIFT 2 ▶ ▶ 3 EXE = 0.827

Applications of Linear Regression

- Hong Kong Population
  - Time Series Model
- Centa-City Index
  - Multiple Linear Regression

Time Series Model

- Attempt to predict future by using a stream of historical data
- Assume what happened in the recent past will continue in the near future
- Time is used as the only independent variable
- $\hat{Y}_t = b_0 + b_1 t$
  - Where $\hat{Y}_t$ = Predicted value at time period t
- For time series data exhibit some trend in a long-range time horizon

Hong Kong Population

Census and Statistics Department
The Government of the Hong Kong Special Administrative Region

- Census and Statistics Department (C&S) published the *Hong Kong Annual Digest of Statistics* so as to provide detailed annual statistical series on various aspects of the social and economic developments of Hong Kong
- Yearly data on Hong Kong's population from 1961 to 2013, totalling 53 observations are downloaded from C&S's website
- Let Y denote the population size (in thousands)

  X = 1, 2, 3… denote the sequence of time

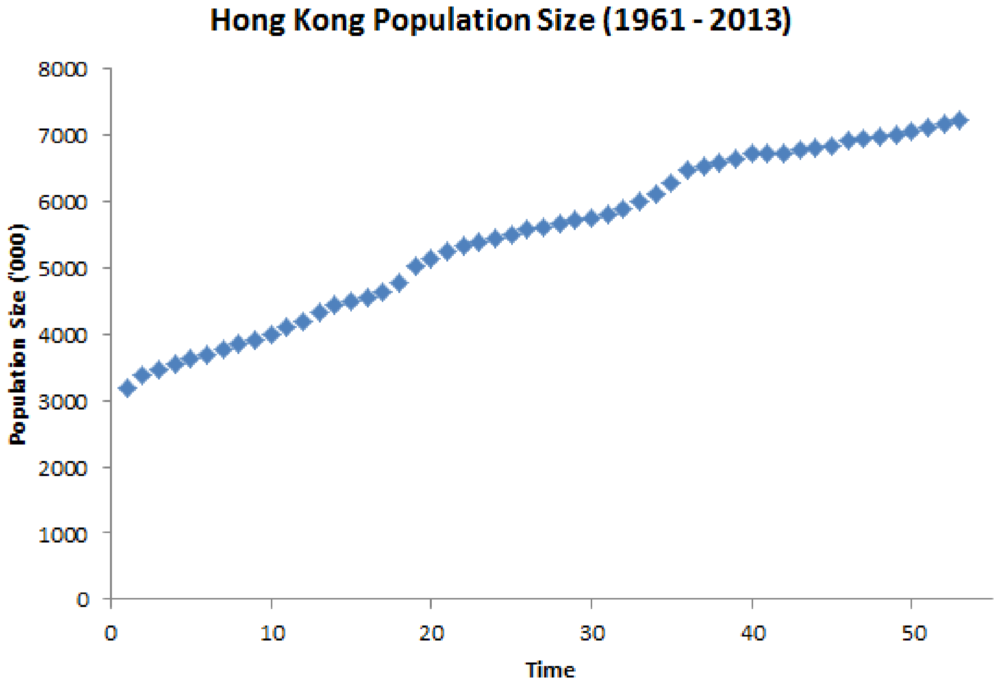
  with X = 1 representing the year 1961

  X = 2 representing 1962, etc.

Hong Kong Population

Cont'd

- A scatter plot of Y vs. X reveals the following



- The association between X and Y appears to be approximately linear
- It therefore makes sense to write $Y = \beta_0 + \beta_1 X + \varepsilon$

Hong Kong Population

Cont'd

- Using the aforementioned least squares method and Excel, the following regression output has been obtained

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9914
R Square	0.9829
Adjusted R Square	0.9826
Standard Error	163.4630
Observations	53

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	78529945.5136	78529945.5136	2938.9781	9.18598E-47
Residual	51	1362727.8347	26720.1536		
Total	52	79892673.3483			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3332.2934	45.5498	73.1571	2.5641E-53	3240.8483	3423.7385
X	79.5741	1.4678	54.2123	9.18598E-47	76.6273	82.5209

- What can you tell from this output?

Hong Kong Population

Cont'd

1. $r_{XY} =$

2. $\hat{Y} =$
   - So, $b_0 = 3332.2934$ is the predicted Hong Kong population size for the year 1960 (X = 0)
   - $b_1 = 79.5741$ is the predicted average annual increment in population size

3. $R^2 =$

Diversifying Your Investments

Cont'd

	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

Inferences about the Slope – Exercise

Cont'd

- Refer to the example our example on number of days taken off work, given $b_1 = -1.09$ and $S_{b_1} = 0.2842$
- A 95% CI for β_1 is

95% CI for β_1
$= b_1 \pm t_{\alpha/2,n-2} S_{b_1}$
$= -1.09 \pm 2.5706 \times 0.2842$
$= [-1.821, -0.359]$

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

Inferences about the Slope – Exercise

Cont'd

- In the example on number of days taken off work , test at 5% level of significance, is years of service linearly influencing the number of days taken off work?

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
At $\alpha = 0.05$
$n = 7 \quad df = 5$
Critical Value = ± 2.5706
Reject H_0 if $t < -2.5706$ or $t > +2.5706$

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

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

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

Hong Kong Population

Cont'd

1. $r_{XY} = 0.9914$ is very close to +1, indicating X and Y have a very strong positive linear relationship
2. $\hat{Y} = 3332.2934 + 79.5741X$
   - So, $b_0 = 3332.2934$ is the predicted Hong Kong population size for the year 1960 (X = 0)
   - $b_1 = 79.5741$ is the predicted average annual increment in population size
3. $R^2 = 0.9829$ indicating that the estimated regression line has the ability to capture 98.29% of the variation in Y in the sample

Hong Kong Population

Cont'd

4. X has a high significant linearly relationship to Y, as $t = 54.2132$ and p-value is close to zero for testing H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$
5. The predicted Hong Kong population sizes for 2014 – 2019 are
   - 2014 ($X = 54$): $\hat{Y} = 3332.2934 + 79.5741(54) =$
   - 2015 ($X = 55$): $\hat{Y} =$
   - 2016 ($X = 56$): $\hat{Y} =$
   - 2017 ($X = 57$): $\hat{Y} =$
   - 2018 ($X = 58$): $\hat{Y} =$
   - 2019 ($X = 59$): $\hat{Y} =$

- By the end of 2019, the Hong Kong population size is expected to excess 8 millions

Hong Kong Population

Cont'd

4. X has a high significant linearly relationship to Y, as $t = 54.2132$ and p-value is close to zero for testing H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$
5. The predicted Hong Kong population sizes for 2014 – 2019 are
   - 2014 ($X = 54$): $\hat{Y} = 3332.2934 + 79.5741(54) = 7629.2948$ thousands
   - 2015 ($X = 55$): $\hat{Y} = 7708.8689$ thousands
   - 2016 ($X = 56$): $\hat{Y} = 7788.4430$ thousands
   - 2017 ($X = 57$): $\hat{Y} = 7868.0171$ thousands
   - 2018 ($X = 58$): $\hat{Y} = 7947.5912$ thousands
   - 2019 ($X = 59$): $\hat{Y} = 8027.1653$ thousands

- By the end of 2019, the Hong Kong population size is expected to excess 8 millions

Hong Kong Population

Cont'd

- Of course, the accuracy of these forecasts depends, among other things, on the legitimacy to extend the linear relationship established based on the sample values beyond the estimation period
  - It is a commonly used method for predicting time series data
- These forecasts are called "ex-ante" forecasts since the actual values of the variable being predicted are unknown at the time of prediction

Multiple Linear Regression

- In many situations, two or more independent variables may be included in a regression model to provide an adequate description of the process under study or to yield sufficiently precise inferences
- For example a regression model for predicting the demands for a firm's product in different countries uses socioeconomic variables (mean household income, average years of schooling of head of household), demographic variables (average family size, percentage of retired population), and environmental variables (mean daily temperature, pollution index), etc.

Multiple Linear Regression

Cont'd

- Linear regression models containing two or more independent variables are called multiple linear regression models
- The simple linear regression model can be extended to include k independent variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Centa-City Index

即時新聞 2014年06月20日 請選擇

樓價繼續升！ 中原指數7個月新高

13,827

讚 3

建立時間: 0620 16:22

中原地產研究部高級聯席董事黃良昇指出，中原城市領先指數CCL最新報119.89點，創29週新高，按週上升0.55％。中原城市大型屋苑領先指數CCL Mass報119.81點，創32週新高，按週上升0.59％。二大指數連續2週上升，CCL累升1.29％，CCL Mass累升1.50％。港島樓價率先上揚，料升勢逐步蔓延至九龍及新界。顯示微調DSD後，二手市況向好，預期樓價繼續反覆向上。

Centa-City Index 中原城市指數

20/06/2014 16:50

【樓市不落】樓價指數按周升０．５５％，創２９周新高

《經濟通通訊社２０日專訊》中原地產表示，中原城市領先指數ＣＣＬ最新報１１９．８９點，創２９周新高，按周上升０．５５％。其中，中原城市大型屋苑領先指數ＣＣＬ　Ｍａｓｓ報１１９．８１點，創３２周新高，按周上升０．５９％。二大指數連續兩周上升，ＣＣＬ累升１．２９％，ＣＣＬ　Ｍａｓｓ累升１．５％。

港島樓價率先上揚，料升勢逐步蔓延至九龍及新界。顯示微調ＤＳＤ後，二手市況向好，預期樓價繼續反覆向上。

至於６月１９日美國聯儲局宣布維持超低利率及資產規模，利好香港樓市，對香港樓價的影響，有待７月上旬公布的ＣＣＬ開始反映。

四區大型屋苑樓價指數方面，港島、九龍及新界東升，新界西跌。港島區指數報１３０．０９點，創３２周新高，按周升０．７４％，連升３周共３．２７％。九龍區指數報１１９．１５點，創三周新高，按周升０．９３％，連升２周共３．０２％。新界東區指數報１２２．５９點，按周升１．８２％。新界西區指數報１００．４１點，創７周新低，按周跌１．０１％。（ｗｉ）

Centa-City Index

Cont'd

- Why Property Price Indices?
  - Investors and potential homebuyers are in need of indicators to study the current movement of property prices in Hong Kong
  - The creation of the "Centa-City Index" aims to provide such information to the public as a source of reference on trends in Hong Kong's property market
- How are the Index constructed?
  - Regression analysis is used to determine the effect of various attributes on property price
  - Attributes such as floor area, years of occupancy, location, direction, view, floor level, etc. are considered

Centa-City Index

Cont'd

(July 1997 = 100)

Centa-City Leading Index CCL

Announced every Friday, **latest on 2014/06/20**; reflecting secondary private residential property price from 2014/06/09 to 2014/06/15 (based on scheduled formal sale & purchase date; on average, formal S&P are signed within 14 days after preliminary S&P)

	This Week	Previous Week	Previous Month
[Centa-City Leading Index]	119.89	↑0.55 %	↑1.12 %
[Mass Centa-City Leading Index]	119.81	↑0.59 %	↑1.75 %

[Centa-City Leading Sub-index]

	This Week	Previous Week	Previous Month
HK	130.09	↑0.74 %	↑2.83 %
KLN	119.15	↑0.93 %	↑2.21 %
NT (East)	122.59	↑1.82 %	↑1.7 %
NT (West)	100.41	↓1.01 %	↓0.49 %

Centa-City Index

Cont'd

Centa-City Index
中 原 城 市 指 數

Constituent Estates	Adjusted Unit Price (gross area basis) (This week)	*Adjusted Unit Price (net area basis) (This week)	Comparison (Previous month)
[Hong Kong Island]			
The Belcher's	13,305.19	17,027.26	↑ 0.20 %
The Merton	11,486.89	15,373.31	↑ 1.66 %
Queen's Terrace	10,533.38	15,144.35	↓ 4.18 %
Robinson Place	14,072.46	17,111.95	↑ 2.50 %
Tregunter	18,964.23	24,018.67	↓ 1.31 %
Dynasty Court	25,091.52	32,055.96	↑ 0.20 %
Clovelly Court	23,841.69	28,553.1	↑ 0.20 %
Convention Plaza Apartments	14,214.42	19,271.38	↑ 0.06 %
The Zenith	12,733.74	17,175.53	↑ 2.37 %
The Leighton Hill	25,197.17	32,954.59	↑ 0.06 %
Beverly Hill	15,841.56	19,427.66	↑ 0.06 %
Cavendish Heights	20,211.7	25,405.45	↑ 0.06 %
Illumination Terrace	11,661.35	14,461.64	↑ 1.14 %
City Garden	10,648.81	12,010.75	↓ 3.01 %

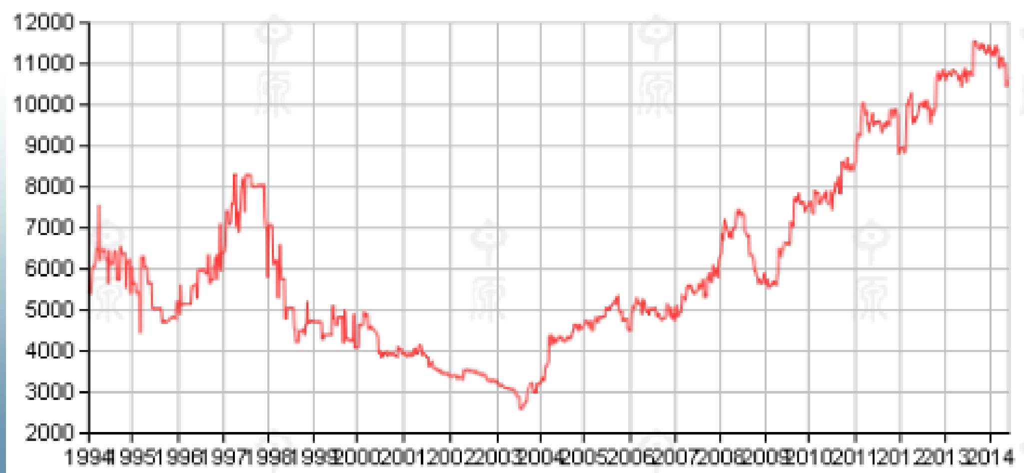
Centa-City Index

Cont'd

CITY GARDEN

Adjusted Unit Price: HK$ 10648.81 Announced on 2014/06/20

Adjusted Unit Price Chart



- More information
  - http://www.cb.cityu.edu.hk/ms/work/hkcci/
  - http://hk.centadata.com/cci/cci_e.htm