CNP 553 - IT for PM

PART II ESSENTIAL SOFTWARE PACKAGES FOR CONST. PROJECT MANAGEMENT

QUT Queensland University of Technology
Brisbane Australia

Week 3 – Outline

1. Regression analysis (forecasting)
2. Use of solver function
3. Use of other functions (Goal Seek)

Week 2 Review – Topics Covered

- **EXCEL's Statistical Functions**
  - **Descriptive/summary Statistics**
  - **Frequency Distribution**
- **EXCEL's Modelling Tools**
  - **What-if Analysis (Data Table)**

Week 2 Review: Statistics

- A **statistic** is a summary measure computed from a sample to describe a characteristic of the population

Statistics **is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data to assist in making more effective decisions.**

Week 2 Review: Statistical Methods

- **Descriptive statistics**
  - Collecting, summarizing, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

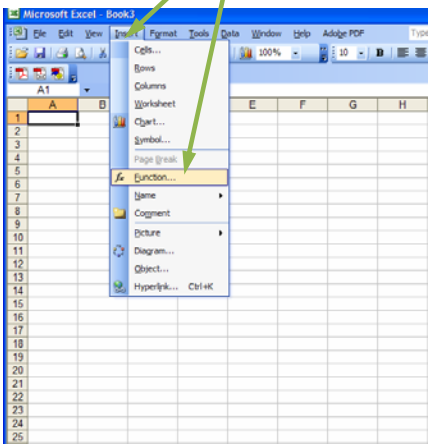
Week 2 Review: … in Excel

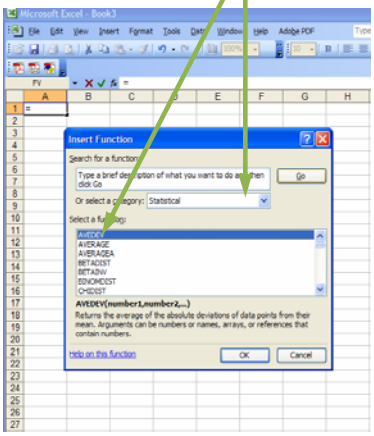
- Excel's statistical features are built into its
  - **Worksheet functions**
  - **Array functions**
  - **Data Analysis tools**
- Excel Statistical functions calculate all the standard measures.

Week 2 Review: 80 Functions

Insert Function Button

80 Functions





Week 2 Review: 16 Data Analysis

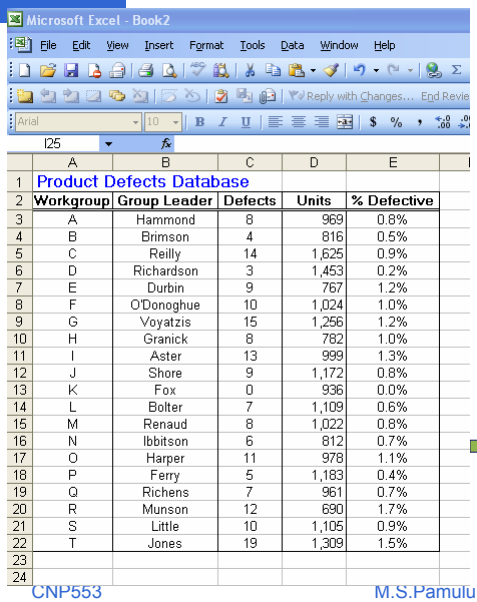
1. ANOVA-Single Factors Analysis of variance for two or more samples
2. ANOVA-Two factor with replication
3. ANOVA-Two factor without replication
4. Correlation
5. Descriptive Statistics
6. Exponential Smoothing
7. Exponential Smoothing
8. F-Test: Two sample for variance
9. Histogram
10. Moving Average
11. Random Number Generation
12. Rank & Percentile
13. Regression
14. Sampling
15. T-Test: Two sample
16. Z-Test: Two sample for Means

Week 2 Review: Using Excel

- Descriptive Statistics (**Summary Statistics**) can be obtained from Microsoft® Excel
  - Use menu choice:

    tools / data analysis / descriptive statistics
  - Enter details in dialog box

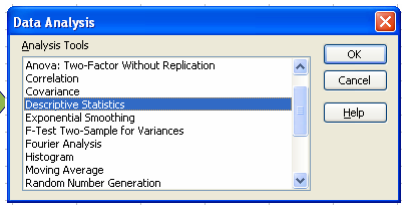
Descriptive Statistics Example



	A	B	C	D	E
1	Product Defects Database				
2	Workgroup	Group Leader	Defects	Units	% Defective
3	A	Hammond	8	969	0.8%
4	B	Brimson	4	816	0.5%
5	C	Reilly	14	1,625	0.9%
6	D	Richardson	3	1,453	0.2%
7	E	Durbin	9	767	1.2%
8	F	O'Donoghue	10	1,024	1.0%
9	G	Voyatzis	15	1,256	1.2%
10	H	Granick	8	782	1.0%
11	I	Aster	13	999	1.3%
12	J	Shore	9	1,172	0.8%
13	K	Fox	0	936	0.0%
14	L	Bolter	7	1,109	0.6%
15	M	Renaud	8	1,022	0.8%
16	N	Ibbitson	6	812	0.7%
17	O	Harper	11	978	1.1%
18	P	Ferry	5	1,183	0.4%
19	Q	Richens	7	961	0.7%
20	R	Munson	12	690	1.7%
21	S	Little	10	1,105	0.9%
22	T	Jones	19	1,309	1.5%

- Use menu choice:

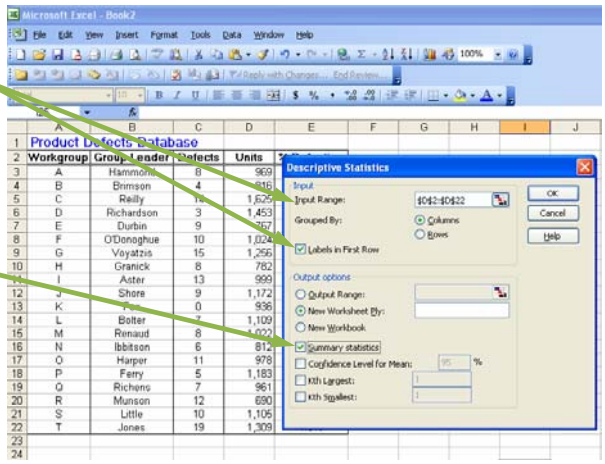
  tools / data analysis / descriptive statistics



Week 2 Review: Using Excel

(continued)

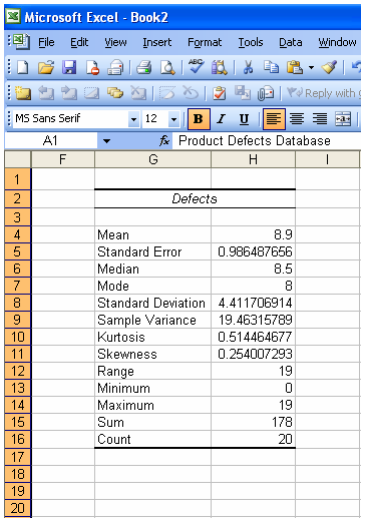
1. Enter dialog box details
2. Check box for summary statistics
3. Click OK



Week 2 Review: Excel output

Microsoft Excel descriptive statistics output, using the product defect data:

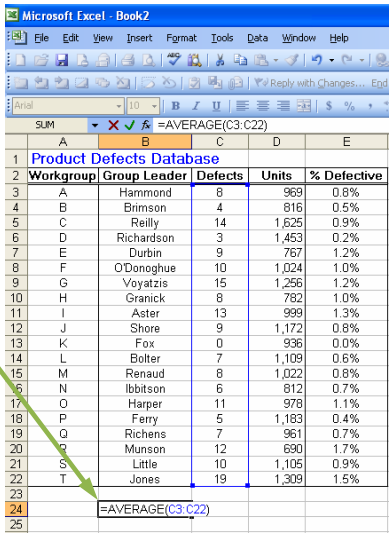
Workgroup	Group Leader	Defects	Units
A	Hammond	8	969
B	Brimson	4	816
C	Reilly	14	1,625
D	Richardson	3	1,453
E	Durbin	9	767
F	O'Donoghue	10	1,024
G	Voyatzis	15	1,256
H	Granick	8	782
I	Aster	13	999
J	Shore	9	1,172
K	Fox	0	936
L	Bolter	7	1,109
M	Renaud	8	1,022
N	Ibbitson	6	812
O	Harper	11	978
P	Ferry	5	1,183
Q	Richens	7	961
R	Munson	12	690
S	Little	10	1,105
T	Jones	19	1,309



	Defects
Mean	8.9
Standard Error	0.986487656
Median	8.5
Mode	8
Standard Deviation	4.411706914
Sample Variance	19.46315789
Kurtosis	0.514464677
Skewness	0.254007293
Range	19
Minimum	0
Maximum	19
Sum	178
Count	20

Descriptive Statistics Example

- =AVERAGE(MEAN)
- =MODE()
- =MEDIAN()
- =VAR()
- =MAX() =MIN()
- =KURT()
- =SKEW()
- =MAX – MIN (RANGE)
- =SUM()
- =COUNT()



Week 2 Review: Functions vs. Formulas

- In general, use functions instead of formulas
  - Functions are adjusted as rows or columns are deleted or added within the range referenced by the function
  - With formulas
    - Adding a row adjusts the cell references in the formula, but does not include the new row in the formula
    - Deleting a row causes a #REF error message

Week 2 Review: Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a list or a table …containing class groupings (categories or ranges within which the data falls) ...and the corresponding frequencies with which data falls within each grouping or category

Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a list or a table …
- containing class groupings (categories or ranges within which the data falls) ...
- and the corresponding frequencies with which data falls within each grouping or category

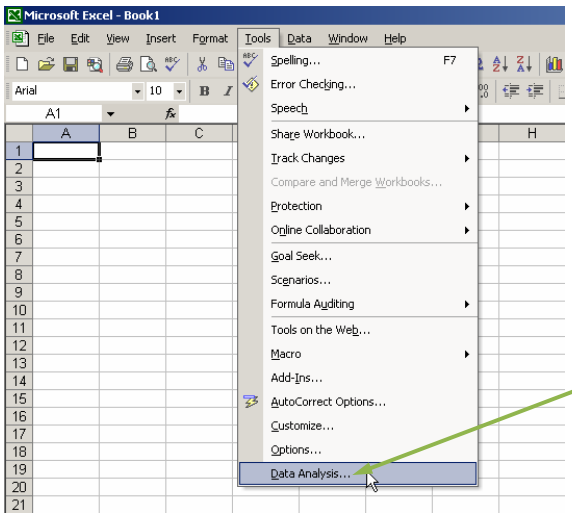
Week 2 Review: The Histogram

- A graph of the data in a frequency distribution is called a **histogram**
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis
- the vertical axis is either **frequency, relative frequency,** or **percentage**
- Bars of the appropriate heights are used to represent the number of observations within each class

Week 2 Review: The Polygon

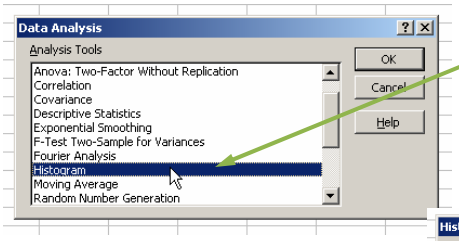
- A **percentage polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages
- The **cumulative percentage polygon**, or **ogive,** displays the variable of interest along the *X* axis, and the cumulative percentages along the *Y* axis.

Week 2 Review: Histograms in Excel

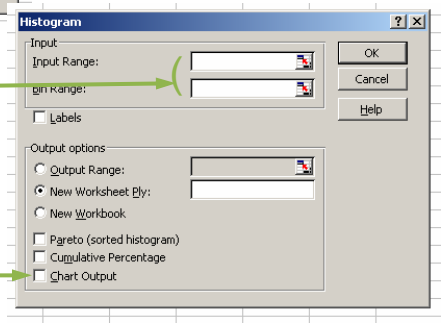


1 Select Tools/Data Analysis

Week 2 Review: Histograms in Excel



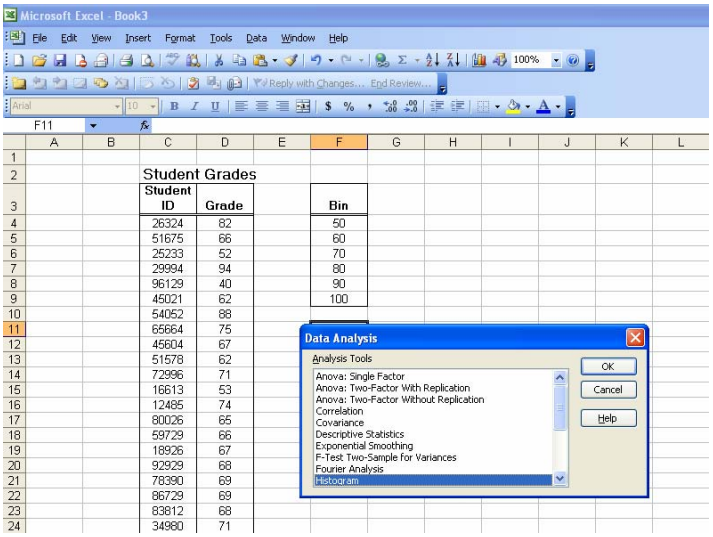
2 Choose Histogram



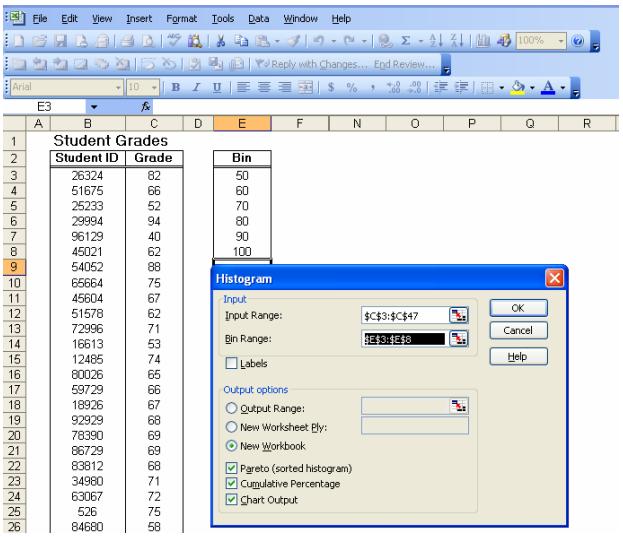
3 Input data range and bin range (bin range is a cell range containing the upper class boundaries for each class grouping)

4 Select Chart Output and click "OK"

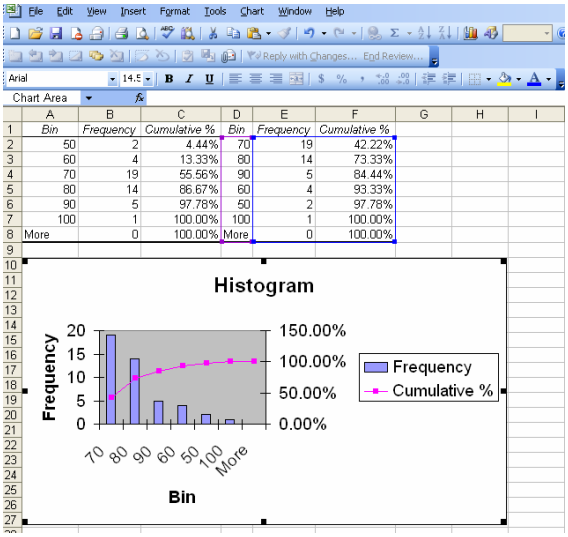
Histogram in Excel Example



Histogram in Excel Example

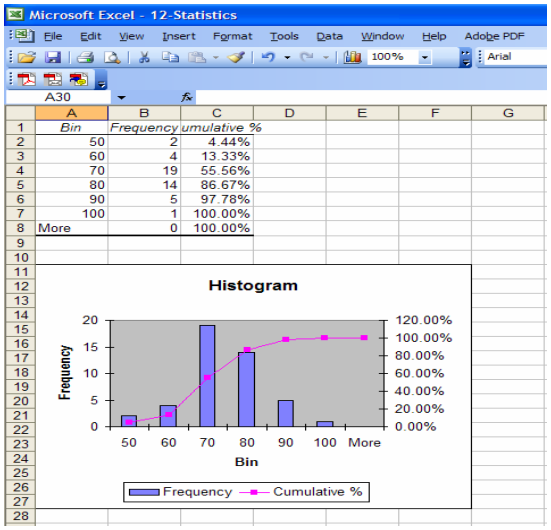


Histogram in Excel Example



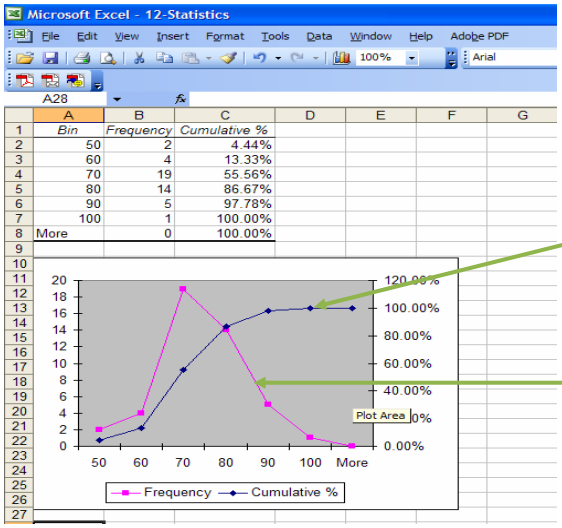
Bin	Frequency	Cumulative %	Bin	Frequency	Cumulative %
50	2	4.44%	70	19	42.22%
60	4	13.33%	80	14	73.33%
70	19	55.56%	90	5	84.44%
80	14	86.67%	60	4	93.33%
90	5	97.78%	50	2	97.78%
100	1	100.00%	100	1	100.00%
More	0	100.00%	More	0	100.00%

Histogram in Excel Example



Bin	Frequency	umulative %
50	2	4.44%
60	4	13.33%
70	19	55.56%
80	14	86.67%
90	5	97.78%
100	1	100.00%
More	0	100.00%

Polygon in Excel Example



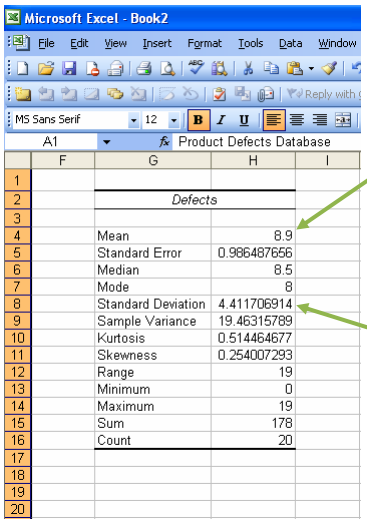
Cumulative percentage polygon

Percentage polygon

Normal Frequency Distributions

- The simple way to know the frequency distribution is at or close to a **NORMAL DISTRIBUTION** (bell shape) is to consider value of **MEAN** (=0) and **STANDAR DEVIATION** (=1)
- Another way to find out how close the frequency distribution is to a **NORMAL DISTRIBUTION** is to consider the **SHAPE** (skewness) and **FLATTNESS** (kurtosis) of the curve/polygon or histogram

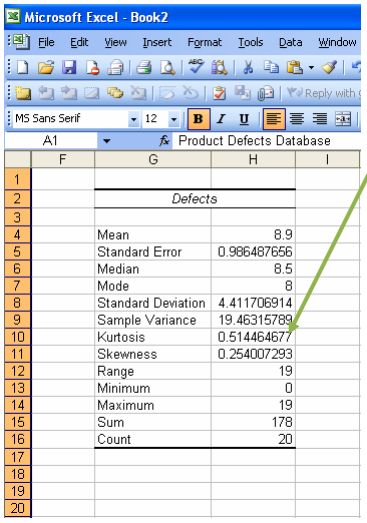
Normal Frequency Distributions



	F	G	H	I
1				
2		*Defects*		
3				
4		Mean	8.9	
5		Standard Error	0.986487656	
6		Median	8.5	
7		Mode	8	
8		Standard Deviation	4.411706914	
9		Sample Variance	19.46315789	
10		Kurtosis	0.514464677	
11		Skewness	0.254007293	
12		Range	19	
13		Minimum	0	
14		Maximum	19	
15		Sum	178	
16		Count	20	
17				
18				
19				
20				

- **MEAN ≠ 0**
- **STANDARD DEVIATION ≠ 1**

Normal Frequency Distributions



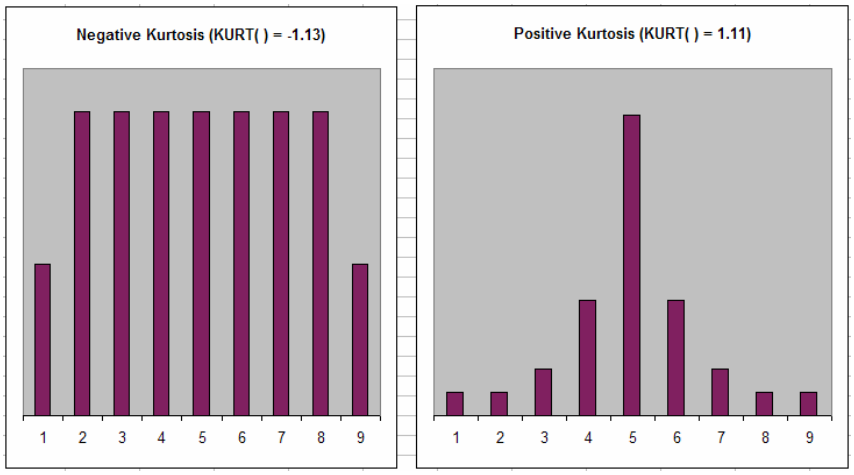
	F	G	H	I
1				
2		*Defects*		
3				
4		Mean	8.9	
5		Standard Error	0.986487656	
6		Median	8.5	
7		Mode	8	
8		Standard Deviation	4.411706914	
9		Sample Variance	19.46315789	
10		Kurtosis	0.514464677	
11		Skewness	0.254007293	
12		Range	19	
13		Minimum	0	
14		Maximum	19	
15		Sum	178	
16		Count	20	
17				
18				
19				
20				

- Histogram is Leptokurtic **KURTOSIS = +**
- Histogram is Platykurtic **KURTOSIS = -**

PEAKED = the values are distributed evenly across all or most of the bins)

FLAT = the values are clustered around a narrow range of values)

Normal Frequency Distributions



Normal Frequency Distributions

Microsoft Excel - Book2

File Edit View Insert Format Tools Data Window

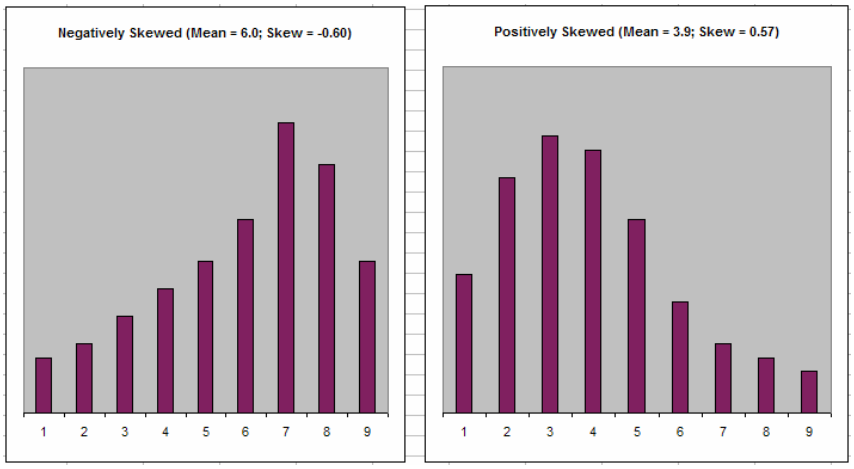
MS Sans Serif 12 B I U

A1 Product Defects Database

	F	G	H	I
1				
2		*Defects*		
3				
4		Mean	8.9	
5		Standard Error	0.986487656	
6		Median	8.5	
7		Mode	8	
8		Standard Deviation	4.411706914	
9		Sample Variance	19.46315789	
10		Kurtosis	0.514464677	
11		Skewness	0.254007293	
12		Range	19	
13		Minimum	0	
14		Maximum	19	
15		Sum	178	
16		Count	20	
17				
18				
19				
20				

- For Symmetric Histogram, **SKEWNESS = 0 (NORMAL DIST.)**
- Histogram skewed to the right **SKEWNESS = + (positive)**
- Histogram skewed to the right **SKEWNESS = - (negative)**

Normal Frequency Distributions



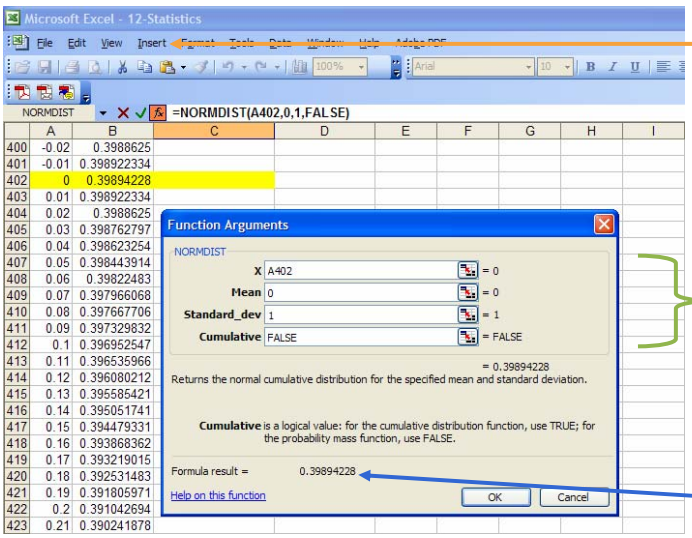
Normal Frequency Distributions

- To generate this normal distribution, use Excel's **NORMDIST()** function, which returns the probability that a given value exists within a population

NORMDIST (*x, mean, standard_dev, cumulative*)

- ***x*** = the value want to work with
- ***mean*** = the arithmetic mean of the distribution
- ***standard_dev*** = the standard deviation of the distribution
- ***cumulative*** = a logical value that determines how the function results are calculated. TRUE if the function returns the cumulative prob. of the observation that occur or below x; FALSE if the function returns the probability associated with x.

Normal Distributions in EXCEL



Menu Choice

Insert Function

Statistical

NORMDIST

X = A2

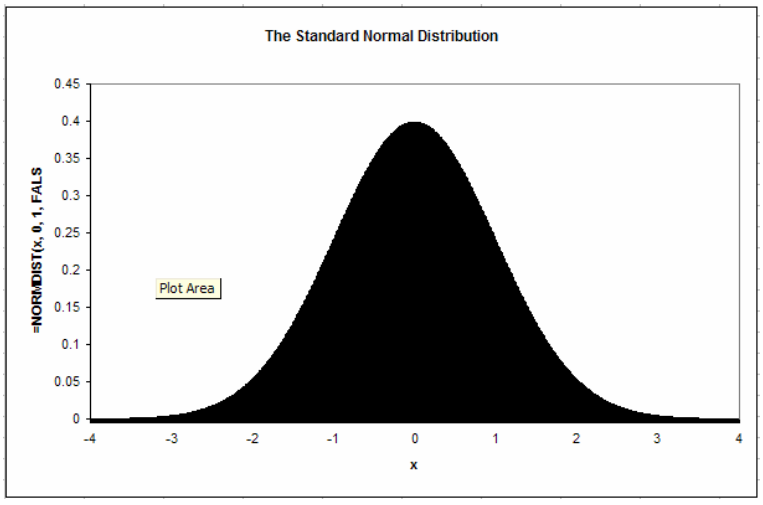
Mean = 0

Standard_dev = 1

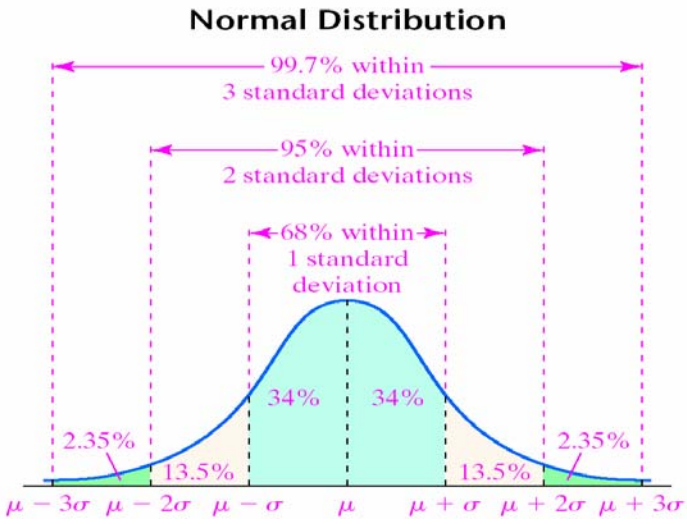
Cumulative = FALSE

RESULT

Normal Frequency Distributions



Normal Frequency Distributions



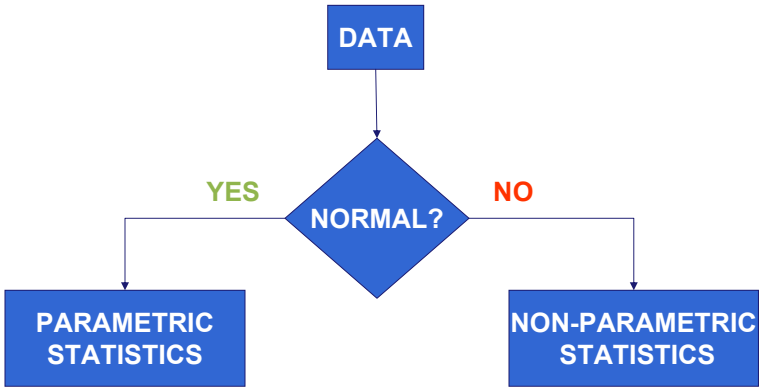
Normal Frequency Distributions

- Approximately 68% of all the value within one standard deviation of the mean (that is, either one standard deviation above or one standard deviation below)
- Approximately 95% of all the value within two standard deviation of the mean
- Approximately 99.7% of all the value within three standard deviation of the mean

Normal Frequency Distributions

- This time, the result is **0.39894228**. In other words, in this distribution, about **3.99%** of all the values in the population are 0
- If the cumulative argument is set to TRUE, this formula returns **0.5**, which makes intuitive sense because, in this distribution, half of values fall below 0. In other words, the probabilities of all the values below 0 add up to **50%.**

Normal Frequency Distributions



Statistical Procedures

- **Nonparametric statistical procedures** are inferential procedures that are not based upon parameters and require fewer assumptions be satisfied in order to perform the tests. They do not require that the population follow a specific type of distribution (such as the normal distribution), and therefore, are often referred to as **distribution-free procedures**.

Statistical Procedures

- For example: Correlation/relationship analysis between variables
- **Parametric Statistics**: Linear Correlation, and Linear/multi-linear Regression test.
- **Non-parametric Techniques**: Spearman Rank Correlation, and Kendall Correlation test.

Week 2 Review - Online

- Learn about using statistical functions and formulas in Microsoft Office Excel 2003 (online training)
  - Excel statistical functions http://office.microsoft.com/training/training.aspx?AssetID=RC010919231033

Week 2 Review: What-If Analysis

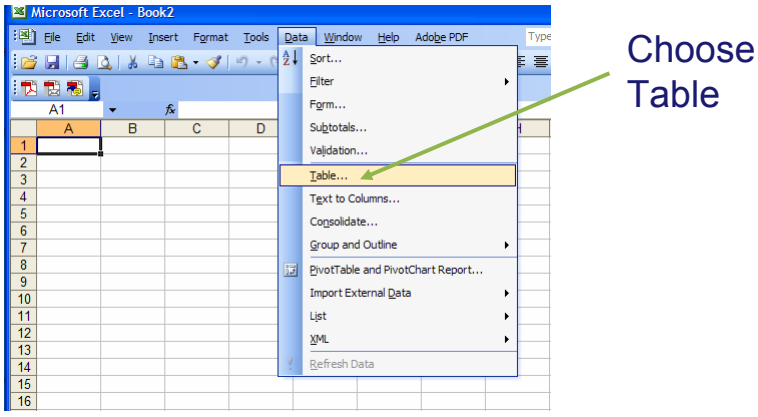
- **Enables decision making in a worksheet**
- What-If analysis is perhaps the most basic method for interrogating the worksheet data.
- The simplest what-if analysis involves changing worksheet variables and watching the result.

Week 2 Review: What-If Analysis

- With what-if analysis, we first calculate a formula D, based on the input variables A, B, and C. We then say, **"what if we change variable A? Or B or C?** What happens to the result?"

What-IF Analysis in Excel

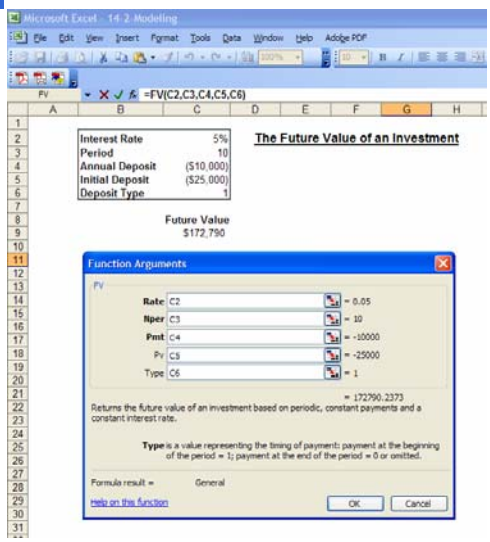
- It is **Data Table** Function in Excel



What-IF in Excel Example

- **For Example**, Following figure shows a worksheet that calculates the future value of an investment based on the five variables: the interest rate, period, annual deposit, initial deposit, and deposit type. Cell C9 shows the result of the financial FV() function.

What-IF in Excel Example



What-IF in Excel Example

- Now the question begin:
  - What if you deposited $8,000 per year? Or $12,000?
  - What if you reduced the initial deposit?
- Answering these questions is straightforward matter of changing the appropriate variables and watching the effect on the result

What-IF in Excel Example

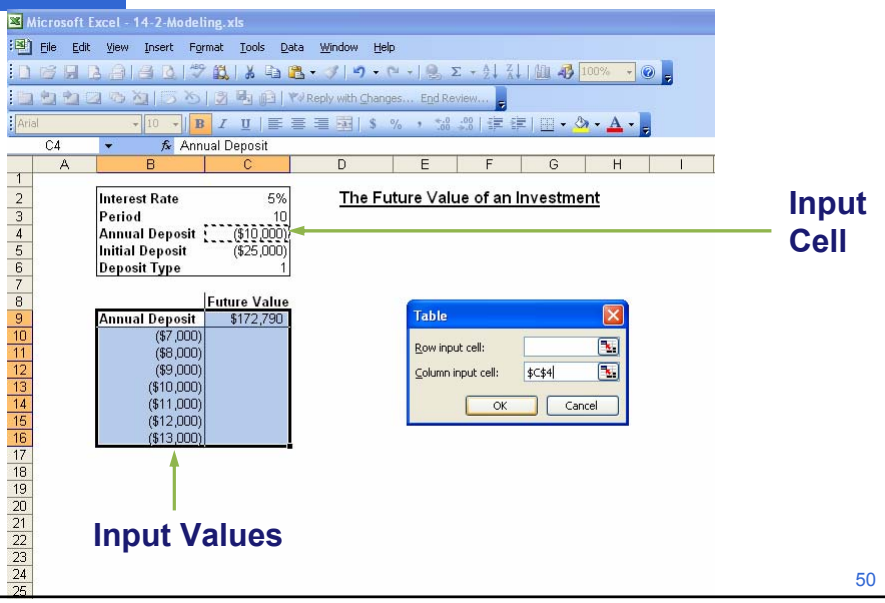
- **Setting Up a One-Input Data Table**

1. Add to the worksheet the values you want to input into the formula. You have 2 choices for the placement of these values:
   - If you want to enter the values in a row, start the row one cell up and one cell to the right of the formula
   - If you want to enter the values in a column, start the column one down and one cell to the left of the cell containing the formula, as shown in Slide 49.
2. Select the range that include the input value and the formula (In the next, this is B9:B16)

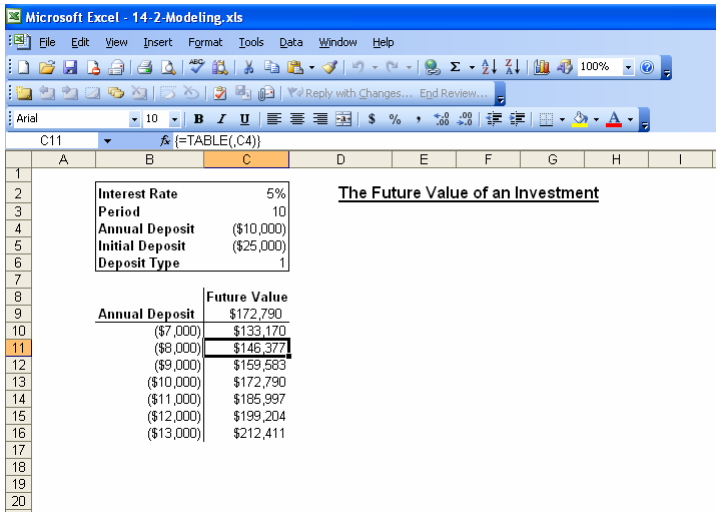
What-IF in Excel Example

3. Choose Data, Table. Excel displays the Table dialog box
4. How you fill in this dialog box depends on how you set up your data table:
   - If you enter the input values in a row, use the Row Input Cell text box to enter the cell address of the input cell
   - If you enter the input values in a column, enter the Column Input Cell text box. In the investment analysis example, you enter C4 in the Column Input Cell, as shown in Slide 49.
5. Click OK. Excel places each of the input value in the input cell; Excel then displays the results in the data table, as shown in Slide 50.

What-IF in Excel Example



What-IF in Excel Example

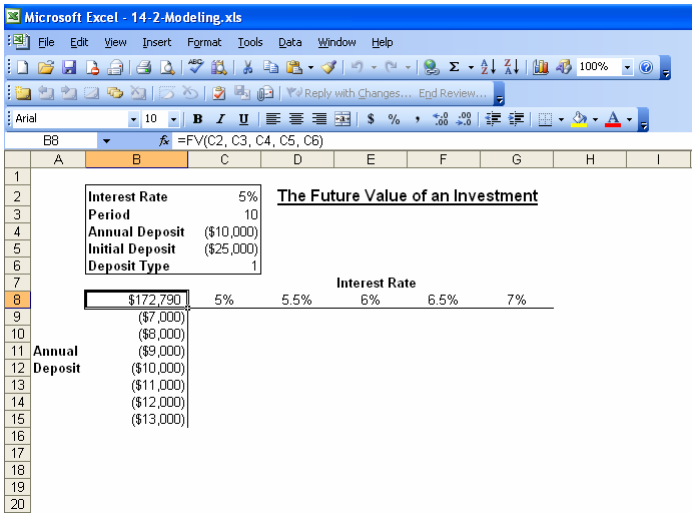


What-IF in Excel Example

- **Setting Up a Two-Input Data Table**

1. Enter one set of values in a column below the formula and the second set of values to the right of the formula in the same row, as shown in Slide 52
2. Select the range that include the input value and the formula (In the next, this is B8:G15)
3. Choose Data, Table. Excel displays the Table dialog box

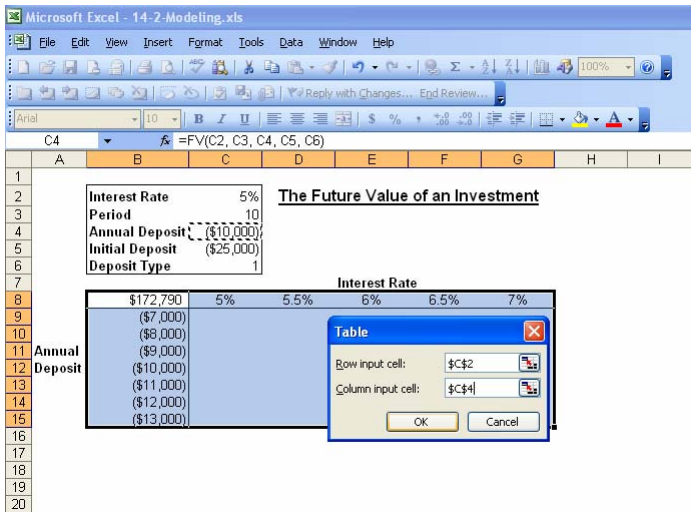
What-IF in Excel Example



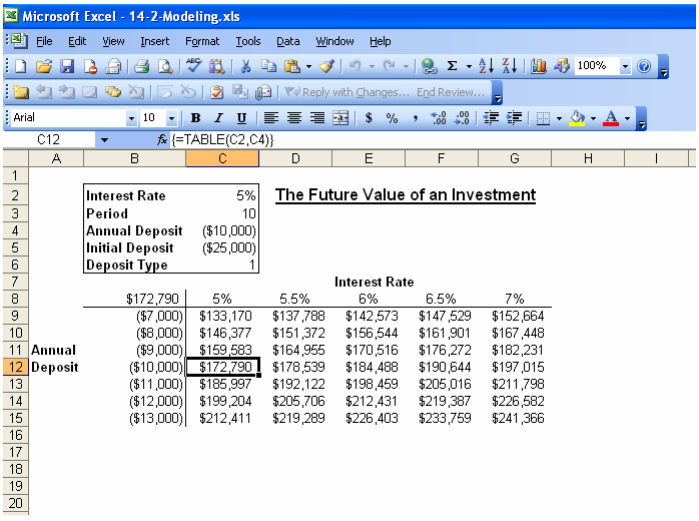
What-IF in Excel Example

4. In the Row Input Cell text box, enter the cell address of the input cell that corresponds to the row values you entered (C2 in Slide 54 – the Interest Rate variable).
5. In the Column Input Cell text box, enter the cell address of the input cell that corresponds to the column values you entered (C4 in Slide 54 – the Annual Deposit variable).
6. Click OK. Excel runs through the various input combinations and then displays the results in the data table, as shown in Slide 55.

What-IF in Excel Example



What-IF in Excel Example



Models

- Representation of Some Phenomenon
- Mathematical Model Is a Mathematical Expression of Some Phenomenon
- Often Describe Relationships between Variables
- Types
  - Deterministic Models
  - Probabilistic Models

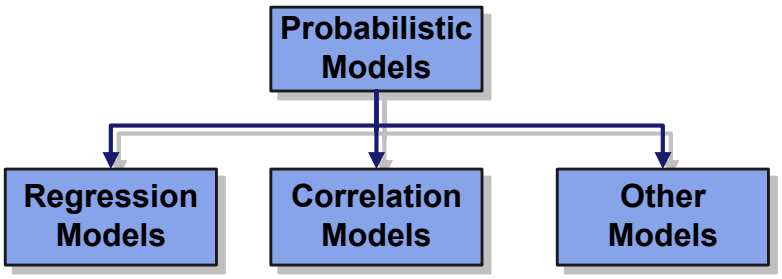
Deterministic Models

- Hypothesize Exact Relationships
- Suitable When Prediction Error is Negligible
- Example: Force Is Exactly Mass Times Acceleration
  - $F = m \cdot a$

Probabilistic Models

- Hypothesize 2 Components
  - Deterministic
  - Random Error
- Example: Sales Volume Is 10 Times Advertising Spending + Random Error
  - $Y = 10X + \varepsilon$
  - Random Error May Be Due to Factors Other Than Advertising

Probabilistic Models



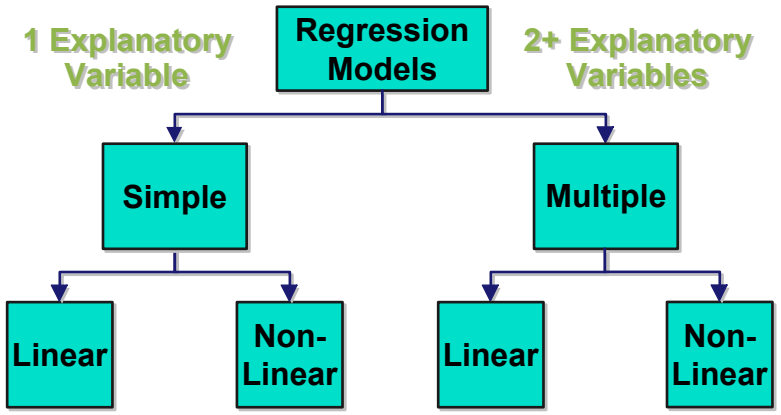
Week 3 - Overview

- A scatter plot (or scatter diagram) can be used to show the relationship between two numerical variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation

Week 3 - Overview

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable: the variable you wish to explain
- Independent variable: the variable used to explain the dependent variable

Regression Models

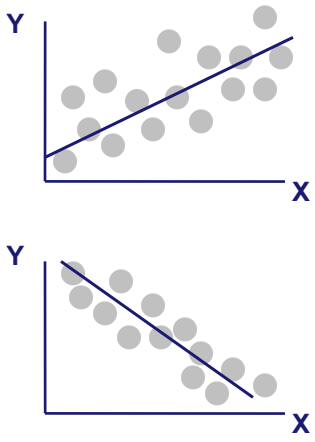


Simple Linear Regression Model

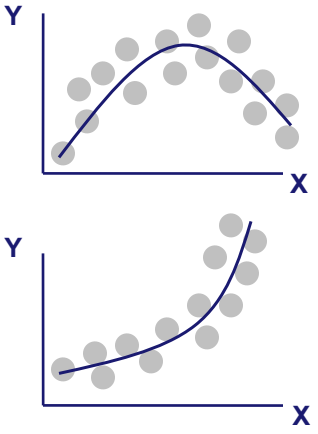
- Only **one** independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are related to changes in X

Types of Relationships

Linear relationships



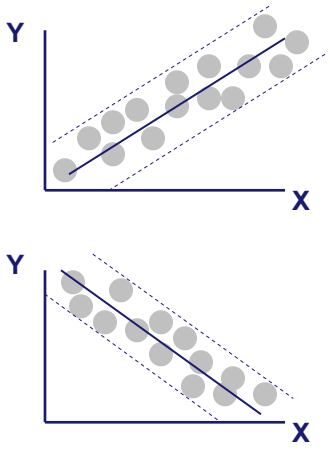
Curvilinear relationships



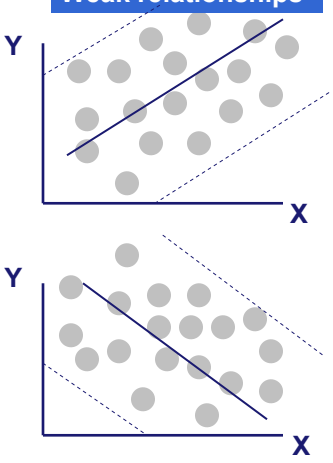
CNP553 M.S.Pamulu, 2007. 65

Types of Relationships

Strong relationships

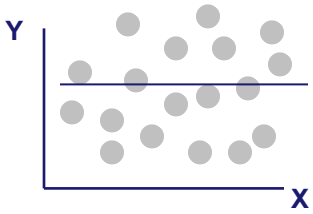


Weak relationships

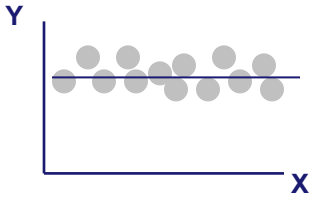


CNP553 M.S.Pamulu, 2007. 66

Types of Relationships



No relationship



The Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

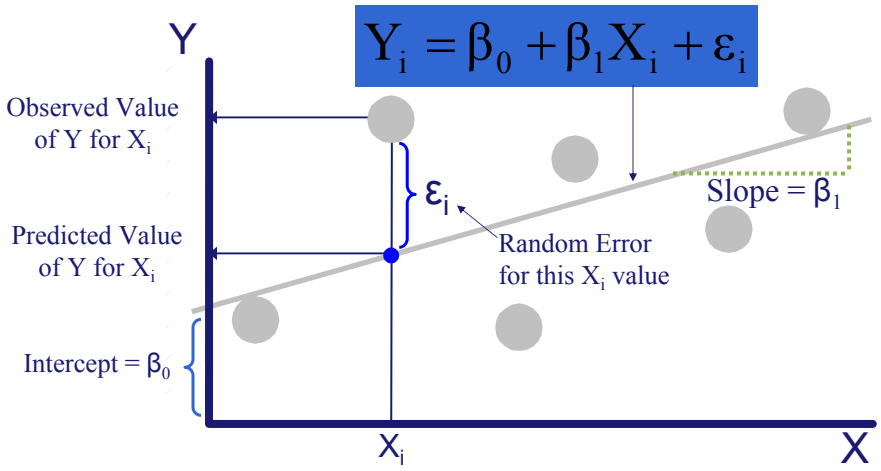
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component

The population regression model:

The Linear Regression Model



Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

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

The Least Squares Method

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Finding the Least Squares Equation

- The coefficients b_0 and b_1, and other regression results in this chapter, will be found using Excel

Formulas are shown in the text for those who are interested

Interpretation of the Intercept and the Slope

- b_0 is the estimated mean value of Y when the value of X is zero
- b_1 is the estimated change in the mean value of Y for every one-unit change in X

Linear Regression Example

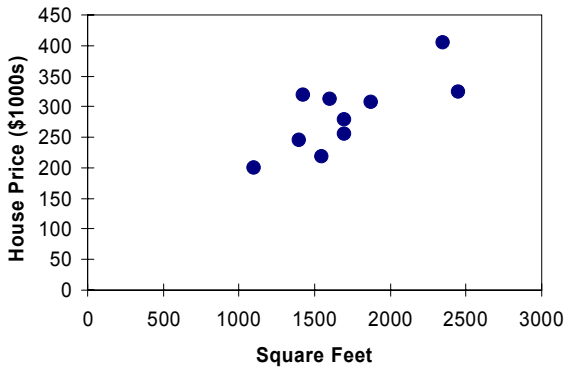
- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Linear Regression Example Data

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Linear Regression Example Scatterplot

- House price model: scatter plot

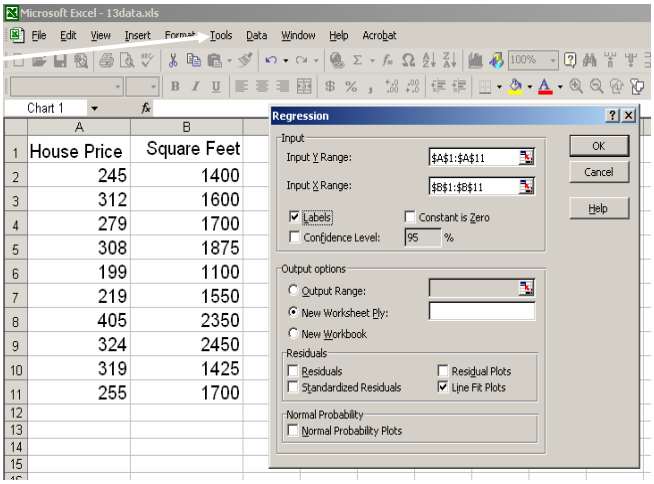


Linear Regression Example Using Excel

Tools

Data Analysis

Regression



	A	B
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700

Linear Regression Example Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

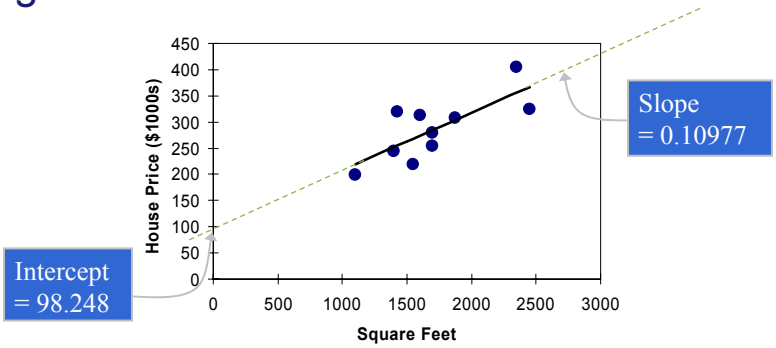
house price = 98.24833 + 0.10977 (square feet)

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Linear Regression Example Graphical Representation

- House price model: scatter plot and regression line



$$\text{house price} = 98.24833 + 0.10977\ (\text{square feet})$$

Linear Regression Example Interpretation of b_0

$$\text{house price} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_0 is the estimated mean value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- Because the square footage of the house cannot be 0, the Y intercept has no practical application.

Linear Regression Example Interpretation of b_1

$$\text{house price} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_1 measures the mean change in the average value of Y as a result of a one-unit change in X
- Here, b_1 = .10977 tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Linear Regression Example Making Predictions

Predict the price for a house with 2000 square feet:

$$\text{house price} = 98.25 + 0.1098\ (\text{sq.ft.})$$

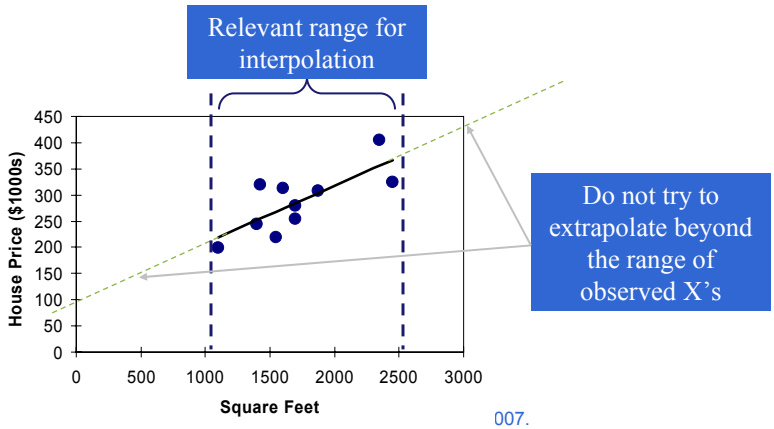
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Linear Regression Example Making Predictions

- When using a regression model for prediction, only predict within the relevant range of data



007. 83

Measures of Variation

Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of	Regression Sum of	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

where:

$\bar{Y}$ = Mean value of the dependent variable

Y_i = Observed values of the dependent variable

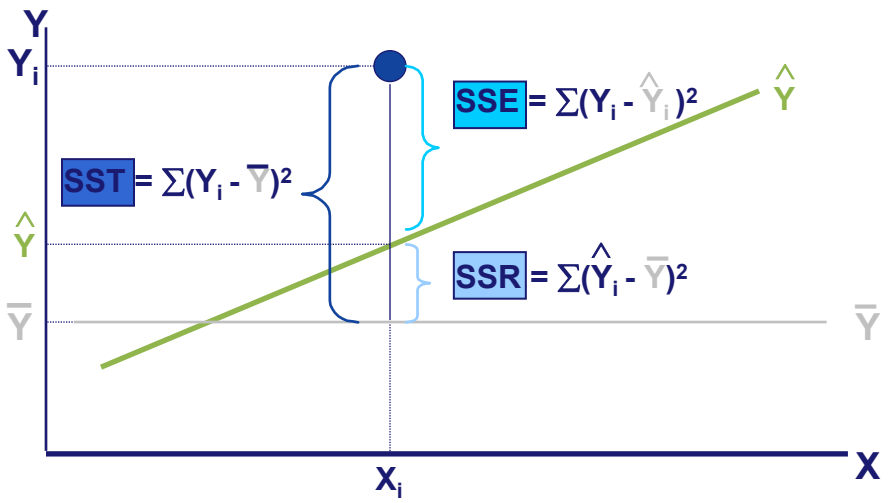
$\hat{Y}_i$ = Predicted value of Y for the given X_i value

CNP553 M.S.Pamulu, 2007. 84

Measures of Variation

- SST = total sum of squares
  - Measures the variation of the Y_i values around their mean Y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between X and Y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between X and Y

Measures of Variation



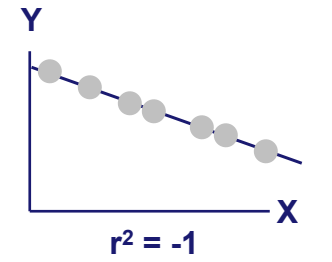
Coefficient of Determination, r^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

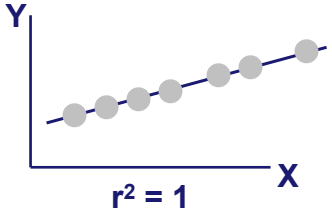
$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

$$0 \le r^2 \le 1$$

Coefficient of Determination, r^2



$r^2 = -1$



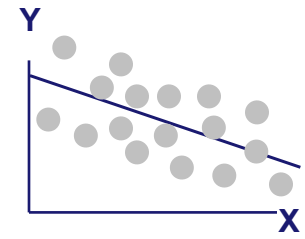
$r^2 = 1$

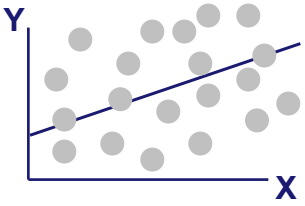
$r^2 = 1$

Perfect linear relationship between X and Y:

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

Coefficient of Determination, r^2



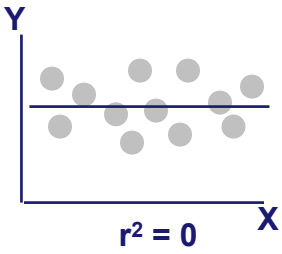


$0 < r^2 < 1$

Weaker linear relationships between X and Y:

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

Coefficient of Determination, r^2



$r^2 = 0$

No linear relationship between X and Y:

The value of Y is not related to X. (None of the variation in Y is explained by variation in X)

Linear Regression Example
Coefficient of Determination, r^2

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.934	11.084	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Where

SSE = error sum of squares

n = sample size

Linear Regression Example Standard Error of Estimate

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

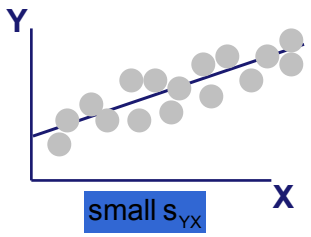
$S_{YX} = 41.33032$

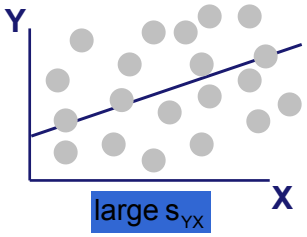
ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.934 8	11.084 8	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.1289 2	-35.57720	232.0738 6
Square Feet	0.10977	0.03297	3.32938	0.0103 9	0.03374	0.18580

Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line





The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

Assumptions of Regression L.I.N.E

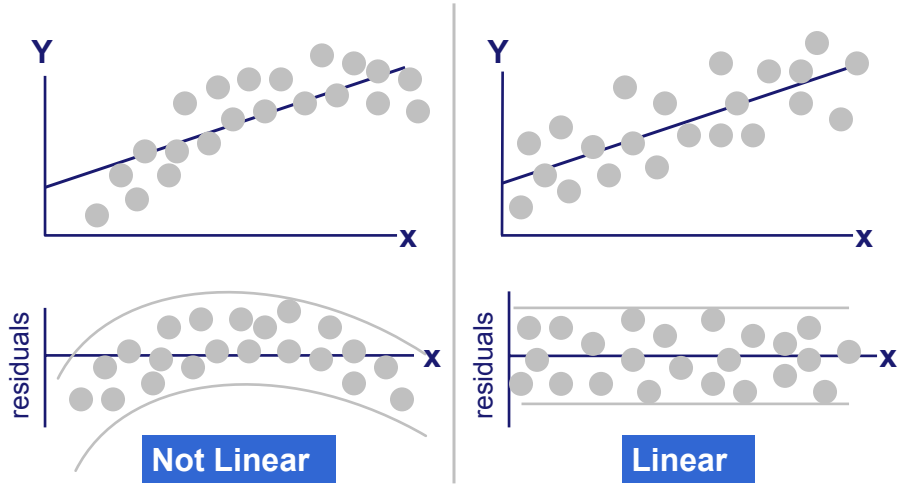
- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

Residual Analysis

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

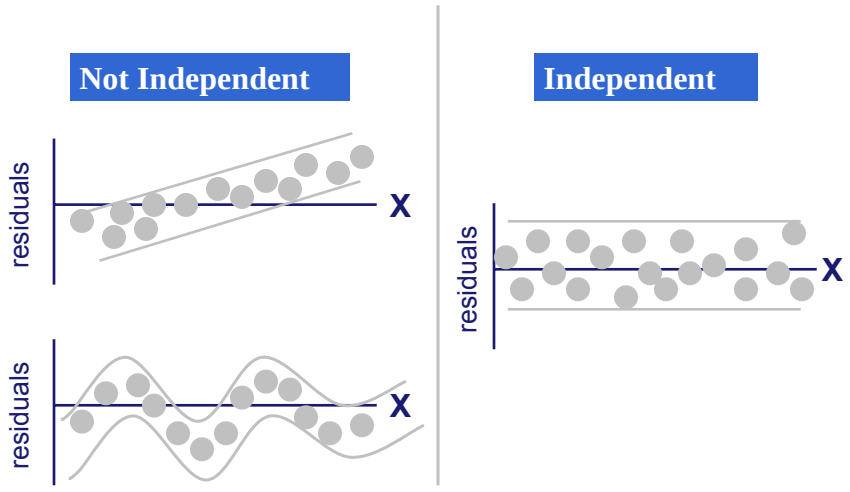
- The **residual** for observation i, $\mathbf{e_i}$, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for Linearity assumption
  - Evaluate Independence assumption
  - Evaluate Normal distribution assumption
  - Examine Equal variance for all levels of X
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

Residual Analysis for Linearity



CNP553 M.S.Pamulu, 2007. 97

Residual Analysis for Independence

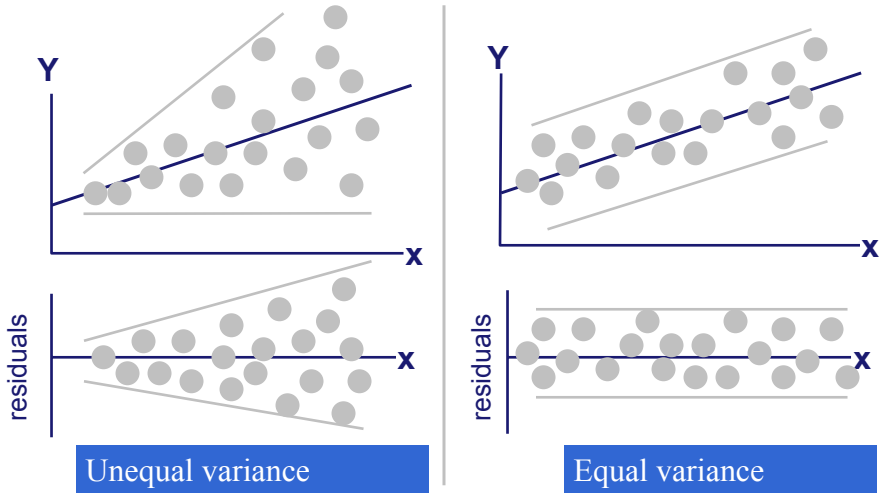


CNP553 M.S.Pamulu, 2007. 98

Checking for Normality

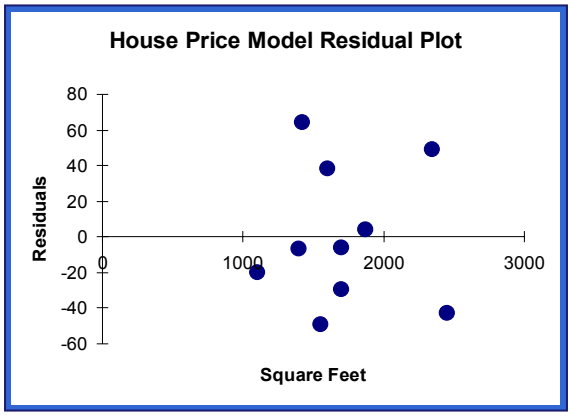
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Box-and-Whisker Plot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Residual Analysis for Equal Variance



Linear Regression Example Excel Residual Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



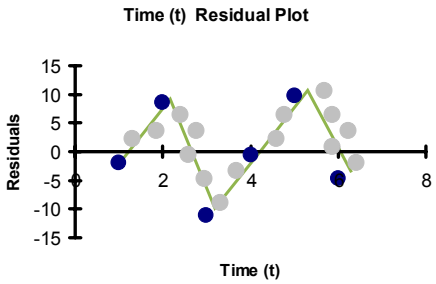
Does not appear to violate any regression assumptions

Measuring Autocorrelation: The Durbin-Watson Statistic

- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

Autocorrelation

- Autocorrelation is correlation of the errors (residuals) over time



- Here, residuals suggest a cyclic pattern, not random
- Violates the regression assumption that residuals are statistically independent

The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation

H_0: residuals are not correlated
H_1: autocorrelation is present

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- The possible range is $0 \leq D \leq 4$
- D should be close to 2 if H_0 is true
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

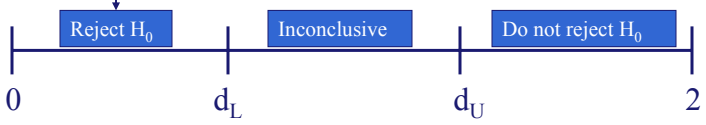
The Durbin-Watson Statistic

H_0: positive autocorrelation does not exist

H_1: positive autocorrelation is present

- Calculate the Durbin-Watson test statistic = D
  (The Durbin-Watson Statistic can be found using Excel)
- Find the values d_L and d_U from the Durbin-Watson table
  (for sample size **n** and number of independent variables **k**)

Decision rule: reject H_0 if $D < d_L$

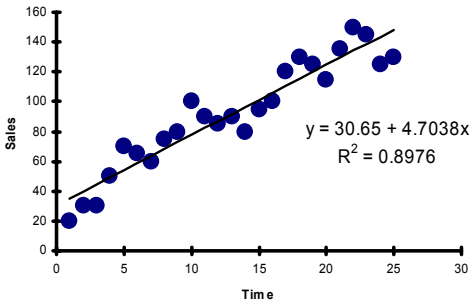


The Durbin-Watson Statistic

❖ Example with n = 25:

Excel output:

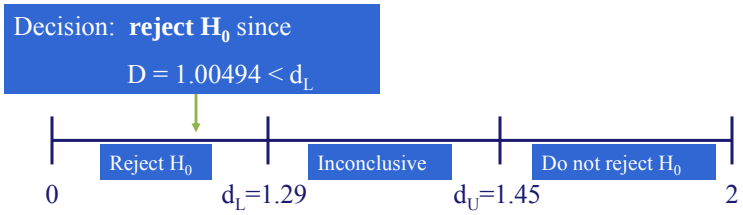
Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

The Durbin-Watson Statistic

- Here, n = 25 and there is k = 1 independent variable
- Using the Durbin-Watson table, d_L = 1.29 and d_U = 1.45
- D = 1.00494 < d_L = 1.29, so reject H_0 and conclude that significant positive autocorrelation exists
- Therefore the linear model is not the appropriate model to predict sales



Inferences About the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

Inferences About the Slope: t Test Example

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098\ (\text{sq.ft.})$$

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Inferences About the Slope: t Test Example

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

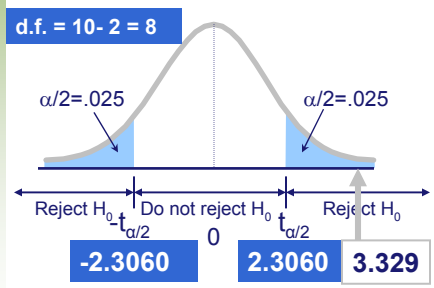
b_1 S_{b_1} t

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: t Test Example

Test Statistic: **t = 3.329**

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Inferences About the Slope: t Test Example

From Excel output:

P-Value

	Coefficients	Standard Error	t Stat	P-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

Decision: Reject H_0, since p-value < α

There is sufficient evidence that square footage affects house price.

F-Test for Significance

❖F Test statistic: $F = \frac{MSR}{MSE}$

where $MSR = \frac{SSR}{k}$

$MSE = \frac{SSE}{n-k-1}$

where F follows an F distribution with k numerator degrees of freedom and (n - k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

F-Test for Significance Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$

With 1 and 8 degrees of freedom

P-value for the F-Test

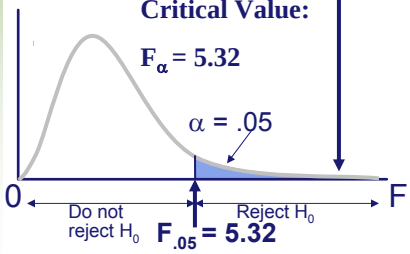
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

F-Test for Significance

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$
- $\alpha = .05$
- $df_1 = 1$ $df_2 = 8$

Critical Value:

$F_\alpha = 5.32$



Test Statistic:

$$F = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{n-2} S_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At the 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is $1000s, you are 95% confident that the mean change in sales price is between $33.74 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

t Test for a Correlation Coefficient

- Hypotheses
  - H_0: ρ = 0 (no correlation between X and Y)
  - H_1: ρ ≠ 0 (correlation exists)
- Test statistic

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

(with n – 2 degrees of freedom)

where

$r = +\sqrt{r^2}$ if $b_1 > 0$

$r = -\sqrt{r^2}$ if $b_1 < 0$

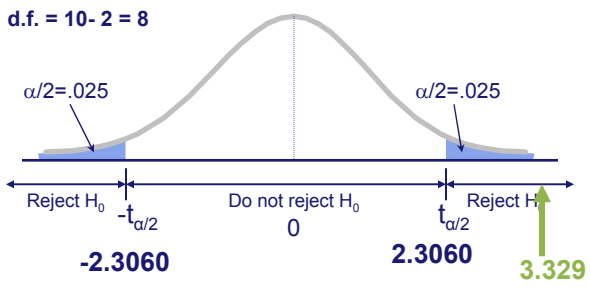
t Test for a Correlation Coefficient

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

H_0: ρ = 0 (No correlation)
H_1: ρ ≠ 0 (correlation exists)
α =.05 , df = 10 - 2 = 8

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.762 - 0}{\sqrt{\frac{1-.762^2}{10-2}}} = 3.329$$

t Test for a Correlation Coefficient



Decision:
Reject H_0

Conclusion:
There is evidence of a linear association at the 5% level of significance

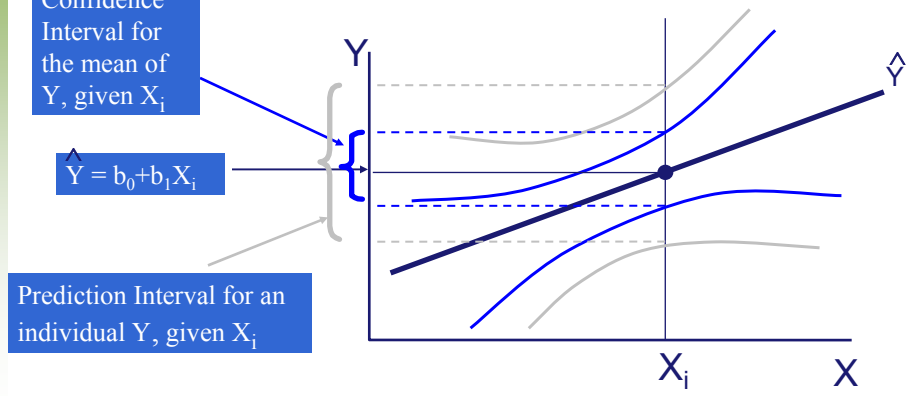
Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around Y to express uncertainty about the value of Y for a given X_i

Confidence Interval for the mean of Y, given X_i

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

Prediction Interval for an individual Y, given X_i



Confidence Interval for the Average Y, Given X

Confidence interval estimate for the mean value of Y given a particular X_i

Confidence interval for $\mu_{Y|X=X_i}$:

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{h_i}$$

Size of interval varies according to distance away from mean, $\bar{X}$

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

Prediction Interval for an Individual Y, Given X

Prediction interval estimate for an individual value of Y given a particular X_i

Prediction interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{1 + h_i}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90, or from $280,660 to $354,900

Estimation of Individual Values: Example

Prediction Interval Estimate for $Y_{X=X_i}$

Find the 95% prediction interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

$$\hat{Y} \pm t_{n-1} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints are 215.50 and 420.07, or from $215,500 to $420,070

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X on Y to observe possible relationship
- Perform residual analysis to check the assumptions
  - Plot the residuals vs. X to check for violations of assumptions such as equal variance
  - Use a histogram, stem-and-leaf display, box-and-whisker plot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding the Pitfalls of Regression

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range

Linear Regression Example Using Excel

- Analyzing Trends using Best-Fit Line
  - **Plotting a Best-Fit Trendline**

Linear Regression Example Using Excel

- Making Forecasts
  - **Forecasting with the Regression Equation**

Plotting a Best-Fit Trendline

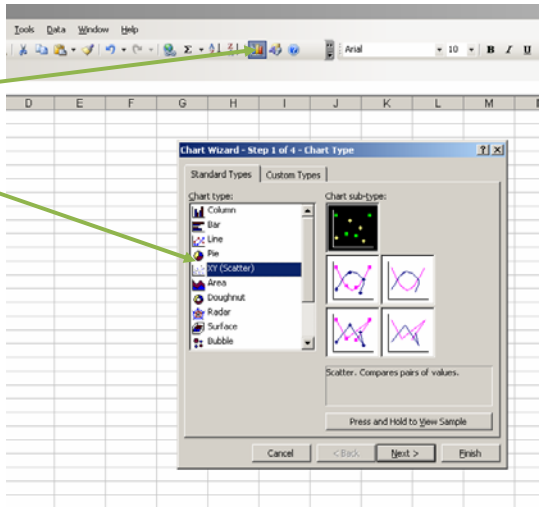
- The easiest way to see the best-fit line is to use a chart. Note, however, that this works only if your data is plotted using an XY (**scatter plot**) chart. For example, Slide 132 shows a worksheet with quarterly sales figures plotted on an XY chart. Here the quarterly sales is dependent variable (Y) and the period is the independent variable (X).

Scatter Plots

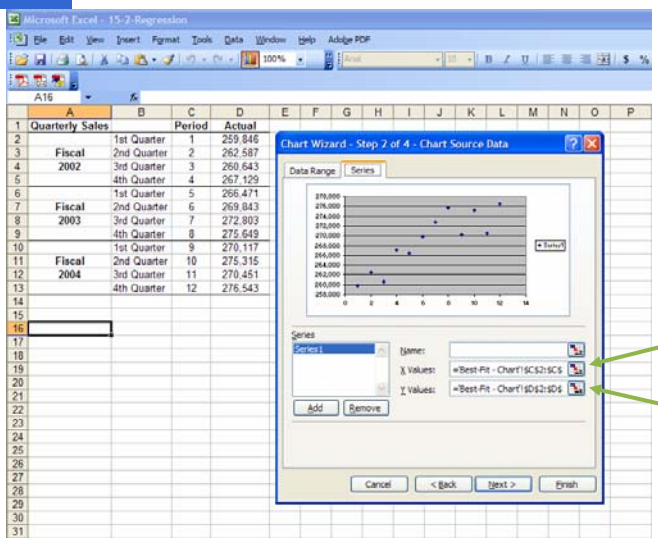
- **Scatter plots** are used for numerical data consisting of paired observations taken from two numerical variables
- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis

Scatter Plot in Excel

1. **Select the chart wizard**
2. **Select XY(Scatter) option, then click "Next"**
3. **When prompted, enter the data range, then click "Next".**
4. **Enter Title, Axis Labels, and Legend and click "Finish"**



Plotting a Best-Fit Trendline

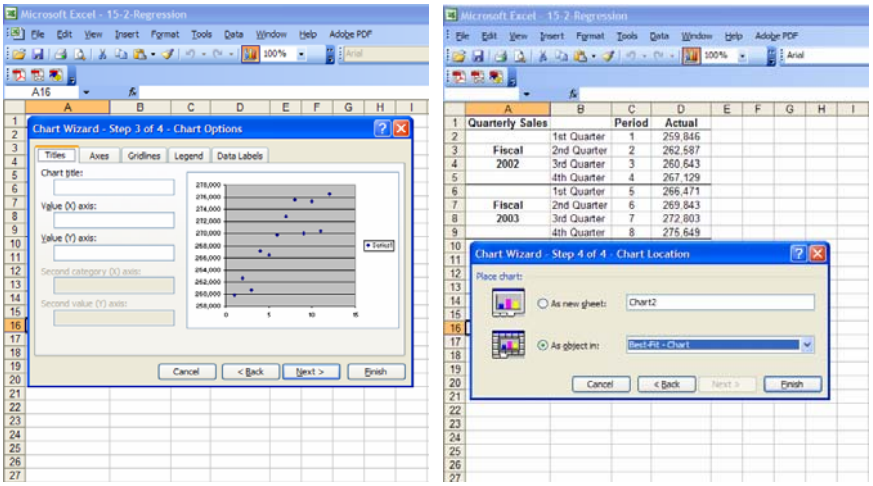


Data Range

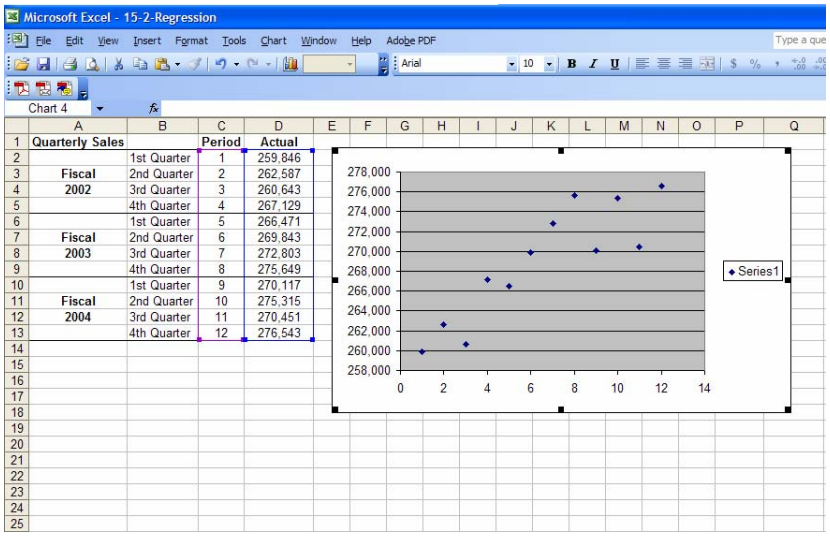
X = C1:C13

Y = D1:D13

Plotting a Best-Fit Trendline



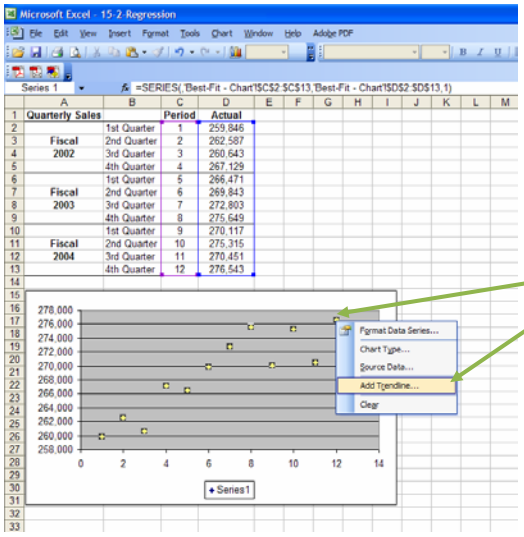
Plotting a Best-Fit Trendline



Plotting a Best-Fit Trendline

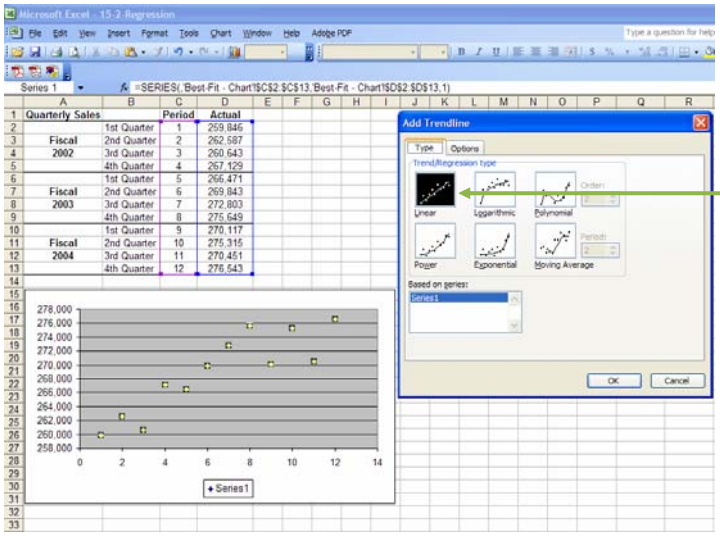
- The following steps show how to add a **trendline** to a chart:
  1. Activate chart and if more than one data series is plotted, click series you want to work with
  2. Choose Chart, Add Trendline. Excel displays the Add Trendline dialog box, shown in Slide 138
  3. On the Type tab, click Linear
  4. Select the Options tab
  5. Activate the **Display Equation** on Chart check box. (See "Regression Equation" later in this slide)
  6. Activate the **Display R-Squared Value** on Chart check box
  7. Click OK. Excel inserts the trendline

Plotting a Best-Fit Trendline

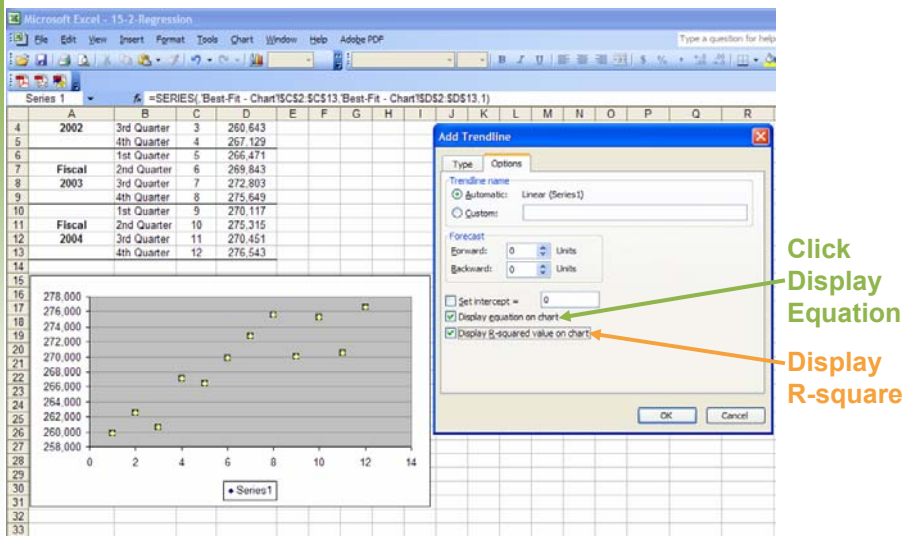


Click the series, and then right click to get the dialog box for Add Trendline

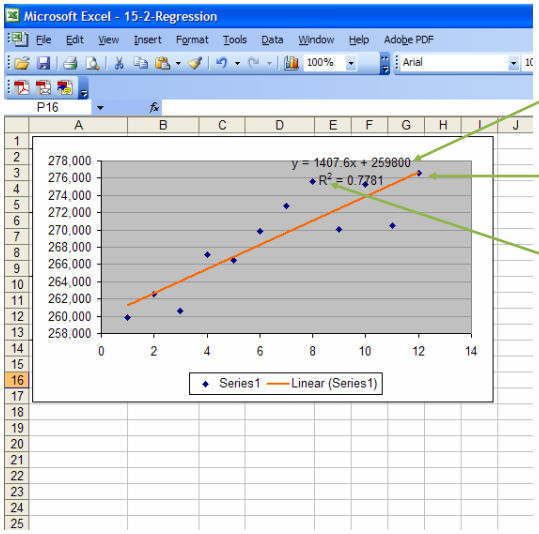
Plotting a Best-Fit Trendline



Plotting a Best-Fit Trendline



Plotting a Best-Fit Trendline



Regression Equation
y = mx + b

Trendline

R-squared Value,
Generally, the closer the result is to 1, the better the fit is. Value < 0.7 mean that the trendline is not very good fit for the data

Trendline Summary

- Knowing the overall trend exhibited by a data set is useful because it tells you the broad direction that sales or cost or employee acquisition is going, and it gives you a good idea of how related the dependent variable is on the independent variable.

Making Forecasts

- A trend is also useful for making forecasts in which you extend the trendline into future (what will sales be in the first quarter of next year?) or calculate the trend value given some new independent value (if we spend $25,000 on advertising, what will the corresponding sales be?)

Making Forecasts

- How accurate is such a prediction?
- A projection based on historical data assumes that the factors influencing the data over the historical period will remain constant.
- The best-fit extensions should be used only for short-term projections.

Plotting Forecast Values

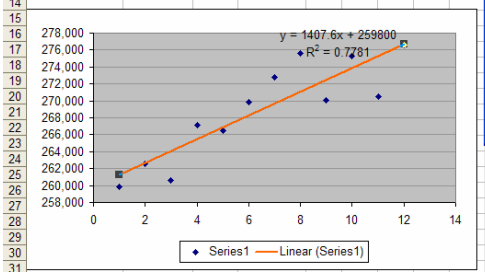
- The following steps show how to add a forecasting trendline to a chart:
  1. Activate chart and if more than one data series is plotted, click series you want to work with
  2. Choose Chart, Add Trendline to display the Add Trendline dialog box
  3. On the Type tab, click Linear
  4. Select the Options tab
  5. Activate the Display Equation on Chart check box.

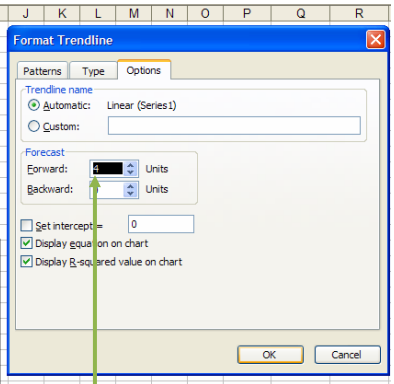
Plotting Forecast Values

6. Activate the Display R-Sq. on Chart check box
7. Use the **Forward** spinner to select the number of units you want to project the trendline into the future. (For example, to extend the quarterly sales number into the next year, you set Forward to 4 to extend the trendline by four quarters.)
8. Click OK. Excel inserts the trendline and extends it into the future.

Plotting Forecast Values

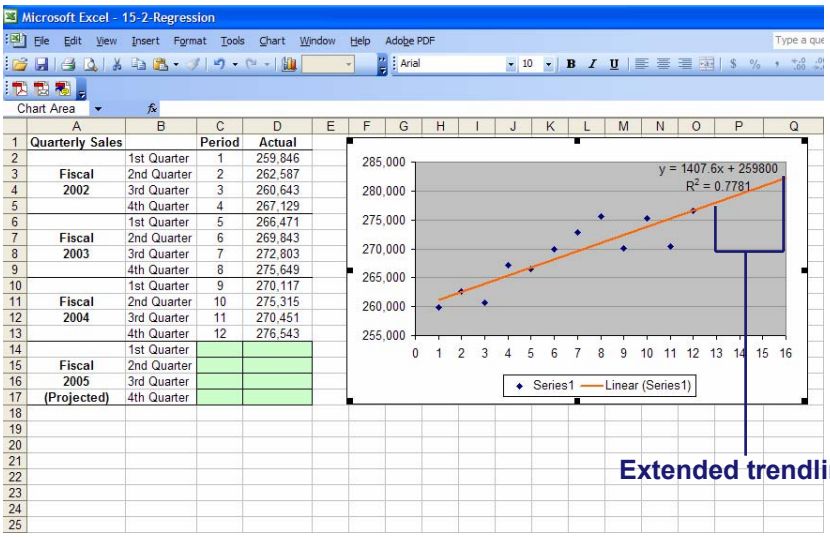
	A	B	C	D
1	Quarterly Sales		Period	Actual
2		1st Quarter	1	259,846
3	Fiscal	2nd Quarter	2	262,587
4	2002	3rd Quarter	3	260,643
5		4th Quarter	4	267,129
6		1st Quarter	5	266,471
7	Fiscal	2nd Quarter	6	269,843
8	2003	3rd Quarter	7	272,803
9		4th Quarter	8	275,649
10		1st Quarter	9	270,117
11	Fiscal	2nd Quarter	10	275,315
12	2004	3rd Quarter	11	270,451
13		4th Quarter	12	276,543





Set Forward Spinner to 4 to extend the trendline by four quarters

Plotting Forecast Values

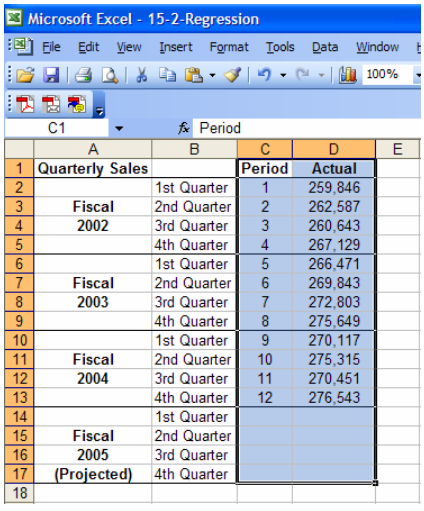


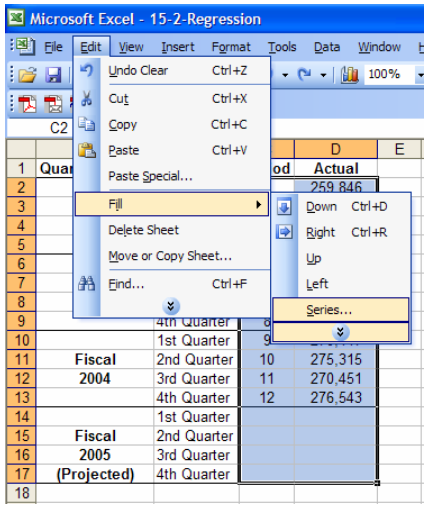
Extended trendline

Extending with the Series Command

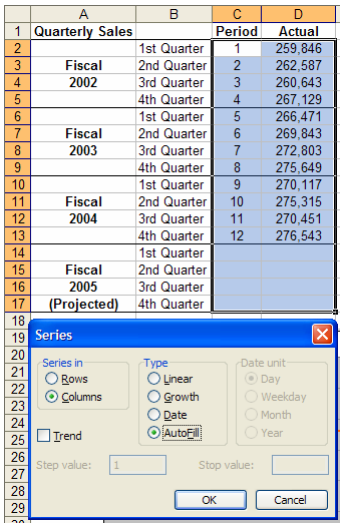
1. Select the range that includes both historical data and the cells that will contain the projections (make sure that the projection cells are blank)
2. Choose Edit, Fill, Series. Excel displays the Series dialog box
3. Activate AutoFill
4. Click OK. Excel fills in the blank cells with the best-fit projection.

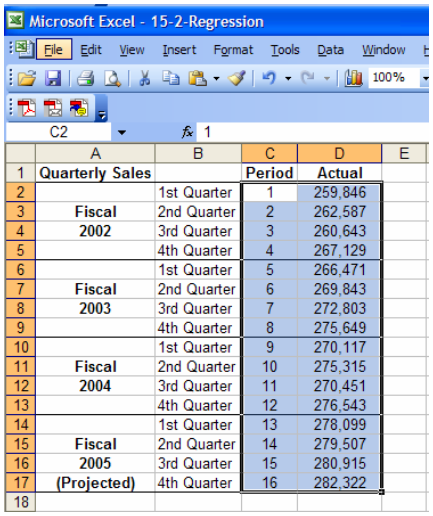
Extending with the Series Command





Extending with the Series Command





	A	B	C	D
1	Quarterly Sales		Period	Actual
2		1st Quarter	1	259,846
3	Fiscal	2nd Quarter	2	262,587
4	2002	3rd Quarter	3	260,643
5		4th Quarter	4	267,129
6		1st Quarter	5	266,471
7	Fiscal	2nd Quarter	6	269,843
8	2003	3rd Quarter	7	272,803
9		4th Quarter	8	275,649
10		1st Quarter	9	270,117
11	Fiscal	2nd Quarter	10	275,315
12	2004	3rd Quarter	11	270,451
13		4th Quarter	12	276,543
14		1st Quarter	13	278,099
15	Fiscal	2nd Quarter	14	279,507
16	2005	3rd Quarter	15	280,915
17	(Projected)	4th Quarter	16	282,322

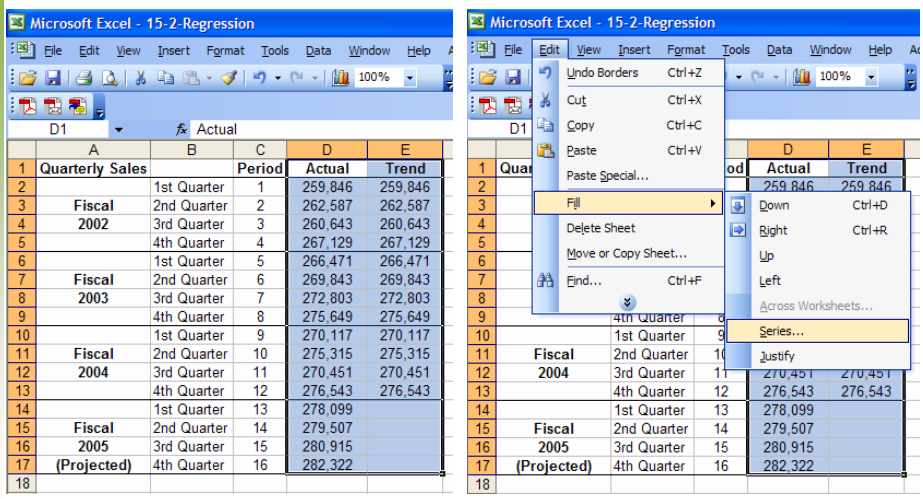
Extending with the Series Command

- The series command is also useful for producing the data that defines the full best-fit line so that you can see the actual trendline values:
  1. Copy the historical data into adjacent row or column
  2. Select the range that includes both the copied historical data and the cells that will contain the projection (again, make sure the projection cells are blank)

Extending with the Series Command

3. Choose Edit, Fill, Series. Excel displays the Series dialog box
4. Activate the Trend check box
5. Select the Linear option
6. Click OK. Excel replaces the copied historical data with the best-fit number sand projects the trend onto the blank cells.

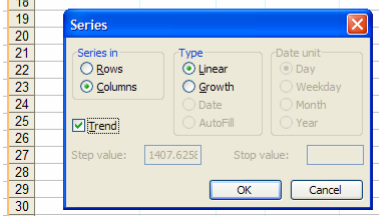
Extending with the Series Command

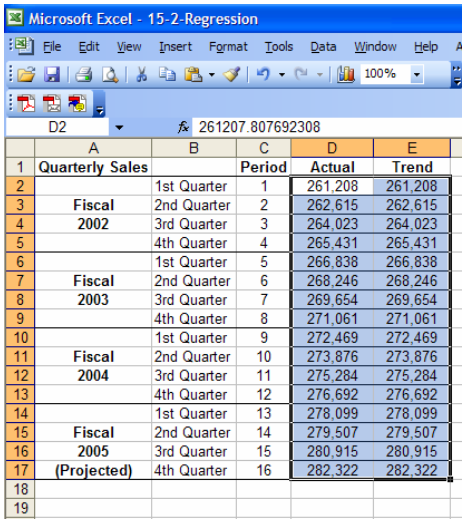


	A	B	C	D	E
1	Quarterly Sales		Period	Actual	Trend
2		1st Quarter	1	259,846	259,846
3	Fiscal	2nd Quarter	2	262,587	262,587
4	2002	3rd Quarter	3	260,643	260,643
5		4th Quarter	4	267,129	267,129
6		1st Quarter	5	266,471	266,471
7	Fiscal	2nd Quarter	6	269,843	269,843
8	2003	3rd Quarter	7	272,803	272,803
9		4th Quarter	8	275,649	275,649
10		1st Quarter	9	270,117	270,117
11	Fiscal	2nd Quarter	10	275,315	275,315
12	2004	3rd Quarter	11	270,451	270,451
13		4th Quarter	12	276,543	276,543
14		1st Quarter	13	278,099	
15	Fiscal	2nd Quarter	14	279,507	
16	2005	3rd Quarter	15	280,915	
17	(Projected)	4th Quarter	16	282,322	
18					

Extending with the Series Command

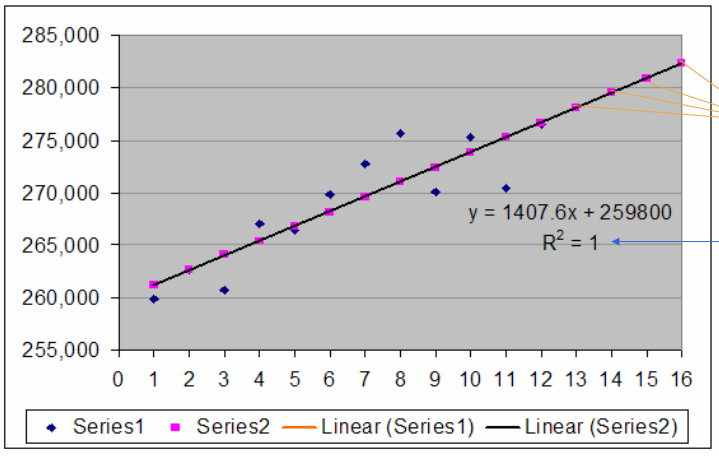
	A	B	C	D	E
1	Quarterly Sales		Period	Actual	Trend
2		1st Quarter	1	259,846	259,846
3	Fiscal	2nd Quarter	2	262,587	262,587
4		3rd Quarter	3	260,643	260,643
5		4th Quarter	4	267,129	267,129
6		1st Quarter	5	266,471	266,471
7	Fiscal	2nd Quarter	6	269,843	269,843
8	2003	3rd Quarter	7	272,803	272,803
9		4th Quarter	8	275,649	275,649
10		1st Quarter	9	270,117	270,117
11	Fiscal	2nd Quarter	10	275,315	275,315
12	2004	3rd Quarter	11	270,451	270,451
13		4th Quarter	12	276,543	276,543
14		1st Quarter	13	278,099	
15	Fiscal	2nd Quarter	14	279,507	
16	2005	3rd Quarter	15	280,915	
17	(Projected)	4th Quarter	16	282,322	





	A	B	C	D	E
1	Quarterly Sales		Period	Actual	Trend
2		1st Quarter	1	261,208	261,208
3	Fiscal	2nd Quarter	2	262,615	262,615
4	2002	3rd Quarter	3	264,023	264,023
5		4th Quarter	4	265,431	265,431
6		1st Quarter	5	266,838	266,838
7	Fiscal	2nd Quarter	6	268,246	268,246
8	2003	3rd Quarter	7	269,654	269,654
9		4th Quarter	8	271,061	271,061
10		1st Quarter	9	272,469	272,469
11	Fiscal	2nd Quarter	10	273,876	273,876
12	2004	3rd Quarter	11	275,284	275,284
13		4th Quarter	12	276,692	276,692
14		1st Quarter	13	278,099	278,099
15	Fiscal	2nd Quarter	14	279,507	279,507
16	2005	3rd Quarter	15	280,915	280,915
17	(Projected)	4th Quarter	16	282,322	282,322

Extending with the Series Command



Trend Values

Best Fit

Agenda – Week 4

1. Introduction to database
2. Use of databases for CM tasks
3. Use of forms for CM presentation
4. Sort and filter functions

Week 4 – Computer Lab.

- Where: L-315 (L Block, Third Floor)
- When: 5 – 8 pm (Tue, 14 August 2007)

References

- Levine, Stephan, Krehbiel & Berenson. 2005. ***Statistics for Managers Using Microsoft® Excel***. 4th Ed. New Jersey: Pearson Education
- McFedries, P. 2005. ***Formulas and Functions with Microsoft Excel 2003***. Indiana: Sam Publishing
- Schmuller, J. 2005. ***Statistical Analysis with Excel for Dummies***. New Jersey: Wiley

Q&A?

- **M. Sapri Pamulu (Sapri)**
  - GP O-Block **#0401**
  - Phone **3138-4186**
  - Fax **3138-1277**
  - Mobile **0402-155-808**
  - email **m.pamulu@qut.edu.au**
  - Web **http://www.unhas.ac.id/sapri**