Microsoft Excel Part 4: Data Analysis

Prakarn Unachak
Department of Computer Science
Faculty of Science
Chiang Mai University

Outline

- Installing Data Analysis ToolPak
- Correlation
- Histogram
- What-If Analysis (Goal Seek)
- Linear Regression

Installing Data Analysis ToolPak

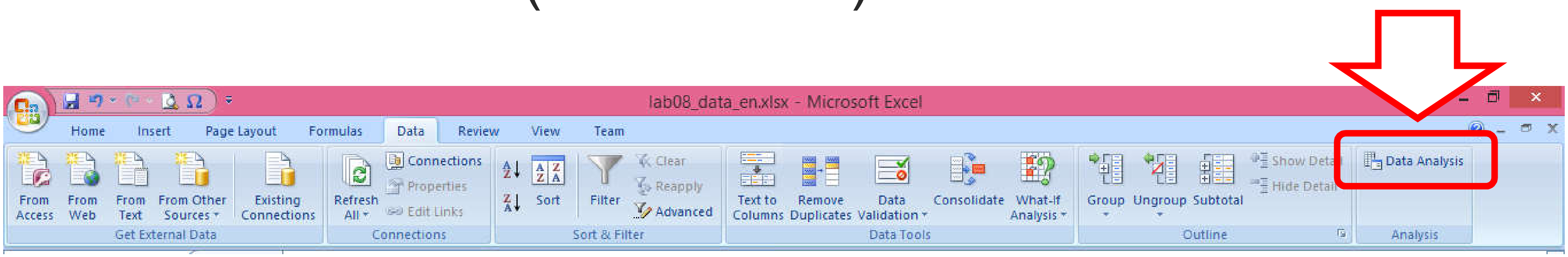
1. Check whether if *Data Analysis* command already appears under *Data* tab.

If the command does not appear yet:

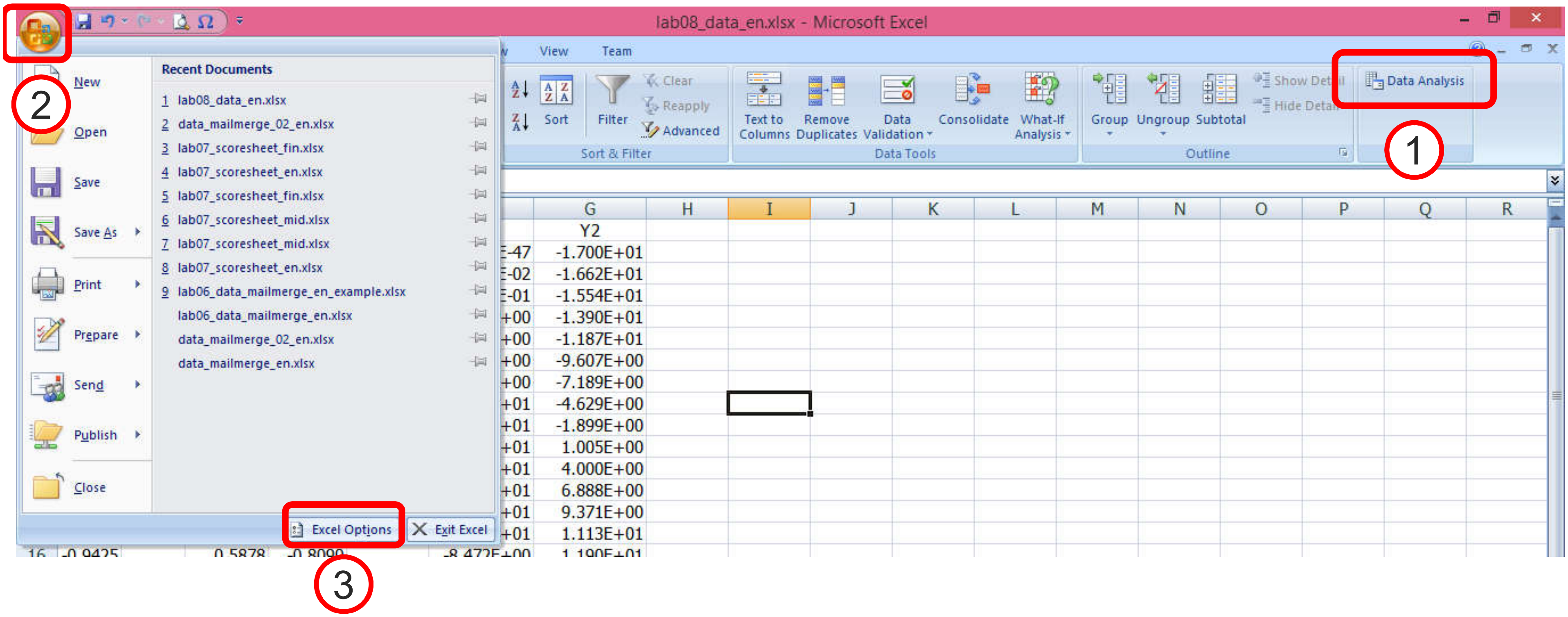
2. Click the office button.
3. Click *Excel Options*.
4. Select *Add-ins*.
5. Change Manage: option to *Excel Add-ins*.
6. Click *Go*.

Installing Data Analysis ToolPak

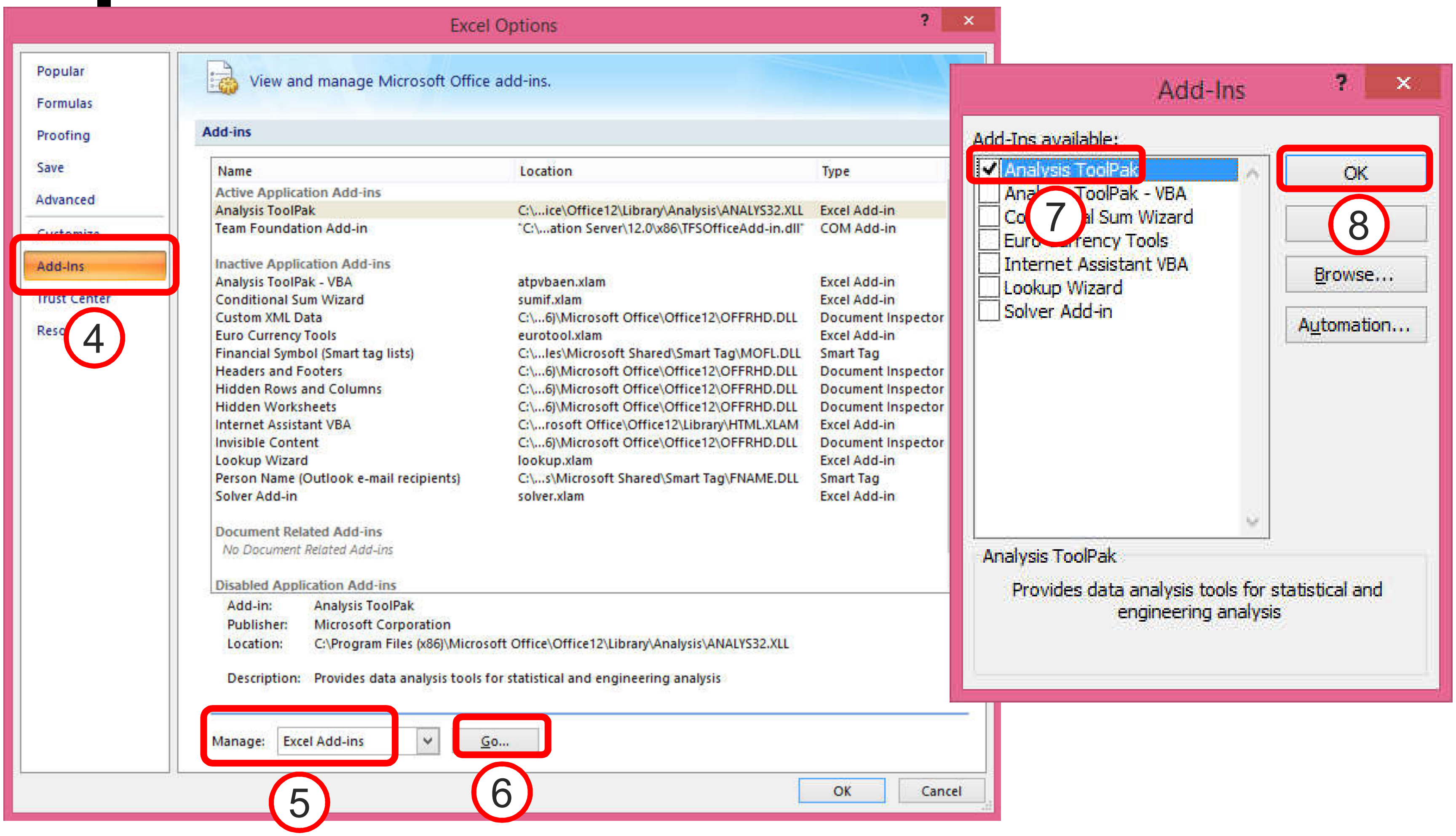
7. Under Add-Ins dialog box, make sure the Analysis ToolPak checkbox is checked.
8. Click OK (a few times).



Installing Data Analysis ToolPak (cont.)



Installing Data Analysis ToolPak (cont.)

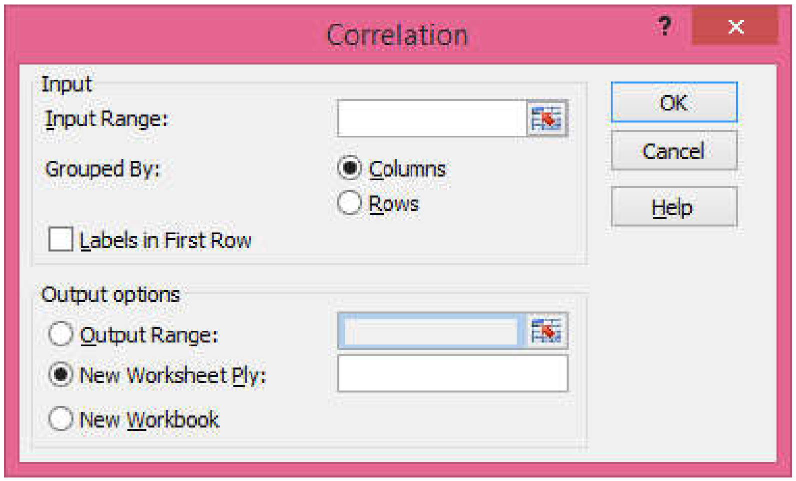


CORRELATION

Correlation

- Correlation tells you how close two set of data related to each other. The value is between -1 and 1.
- If the two data sets have positive correlation, when one increase/decrease, the other will go to the same direction.
- If the two data sets have negative correlation when one increase/decrease, the other will go to the other direction.
- If the correlation is zero, there is no relation between two data sets. Whether the one decrease/increase will not predict the other.

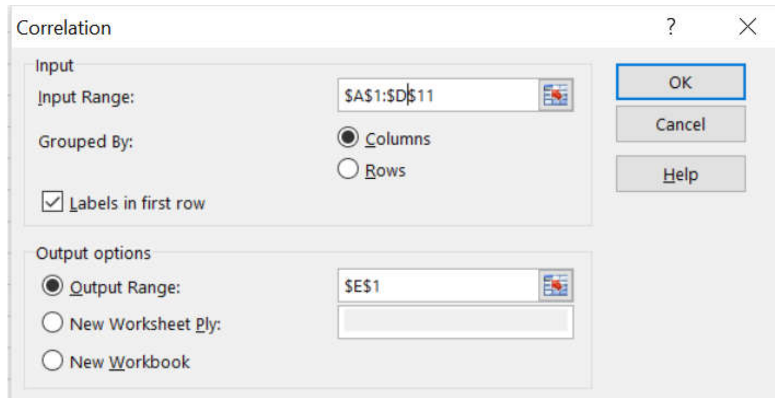
Calculating Correlation



1. Select *Data* → *Data Analysis* → *Correlation*
2. Select *Input Range*
3. Select Options
   - Grouped By: Columns/Rows.
   - Check if first row of data is a label.
4. Select *Output Range*.

Correlation Results

	A	B	C	D
1	W	X	Y	Z
2	1	1	10	1
3	2	1.5	9	1.5
4	3	2	8	2
5	4	2.5	7	2.5
6	5	3	6	3
7	6	3.5	5	3
8	7	4	4	3.5
9	8	4.5	3	4
10	9	5	2	4.5
11	10	5.5	1	5

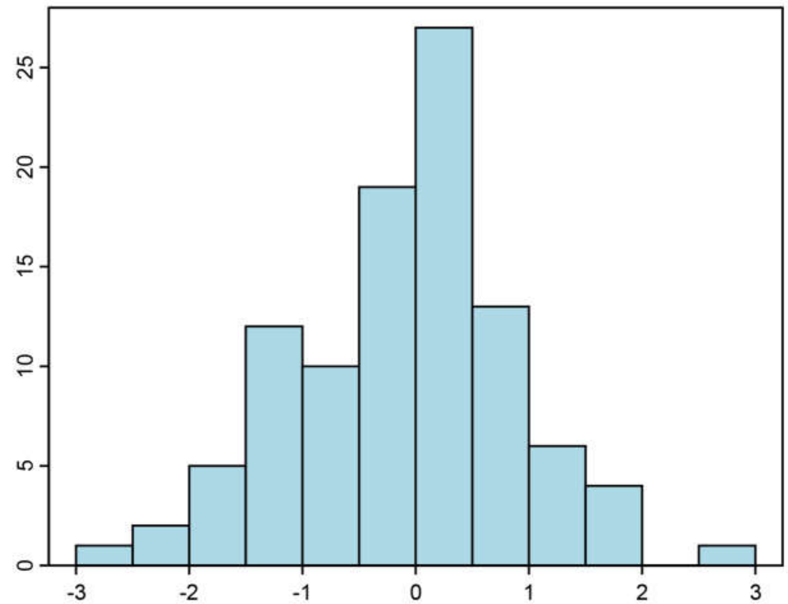


E	F	G	H	I
	W	*X*	*Y*	*Z*
W	1			
X	1	1		
Y	-1	-1	1	
Z	0.994937	0.994937	-0.99494	1

HISTOGRAM

Histogram

- A histogram is a way to display data by grouping each data point into ranges (bins) of their values
- Good for showing data distribution



source: Wikipedia

Histogram (cont.)

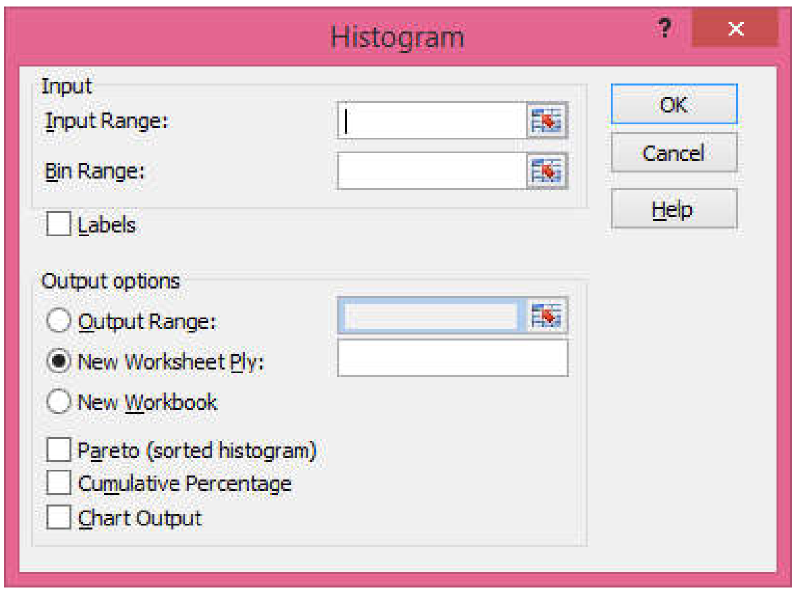
Bins

Data

96		20
35		40
62		60
21		80
87		
75		
20		
83		
10		
64		
20		
76		
41		
85		
57		
0		
67		
84		
2		
19		

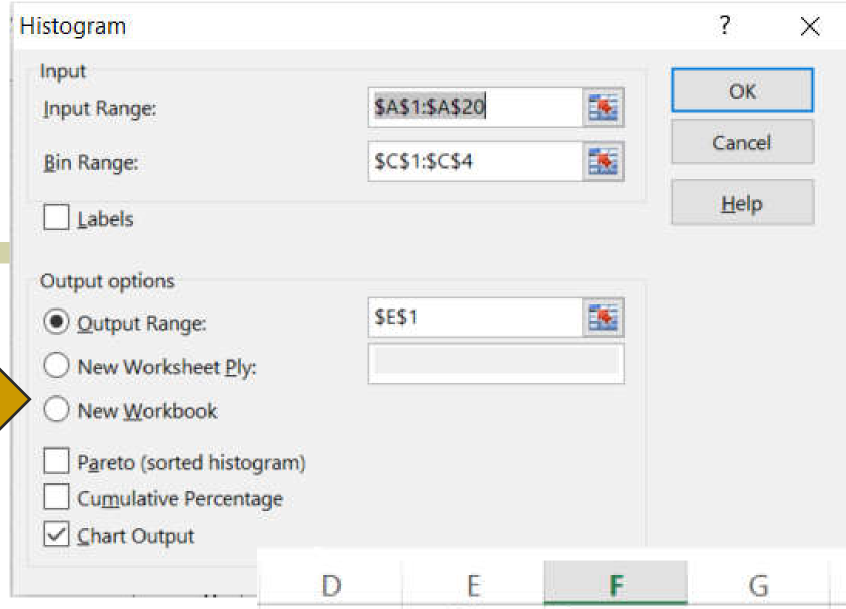
- To perform Histogram on Excel, you need data and bins.
- Bins are cells whose values used to determine which group a data point will go to.

Calculating Histogram

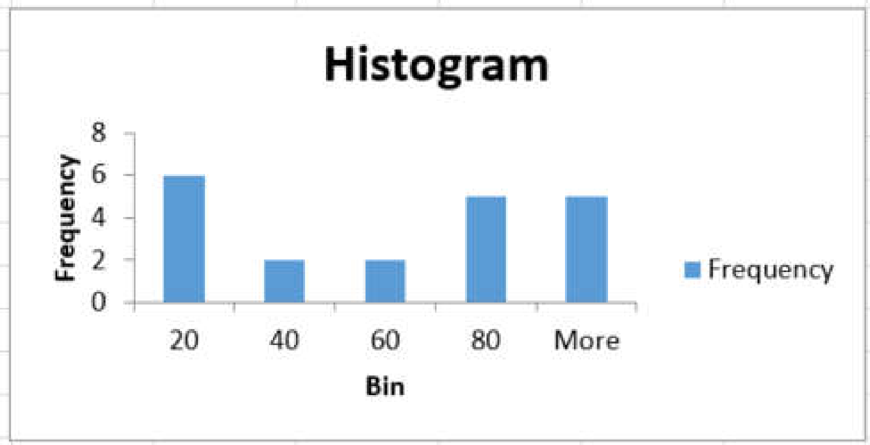


1. Select *Data* → *Data Analysis* → *Histogram*
2. Select *Input Range*
3. Select *Bin Range*
4. Select *Output Range*
5. Check other options
   - *Chart Output* if you want charted version of histogram.
6. Click *OK*

	A	B	C
1	96		20
2	35		40
3	62		60
4	21		80
5	87		
6	75		
7	20		
8	83		
9	10		
10	64		
11	20		
12	76		
13	41		
14	85		
15	57		
16	0		
17	67		
18	84		
19	2		
20	19		

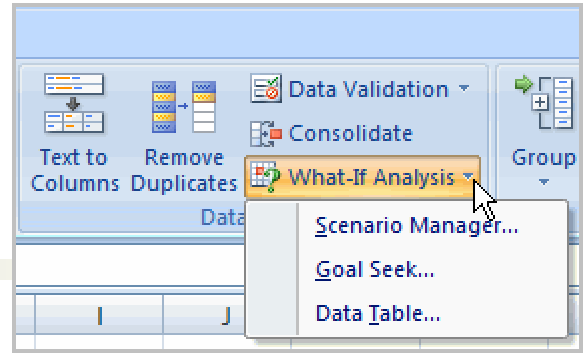


Bin	*Frequency*
20	6
40	2
60	2
80	5
More	5



WHAT-IF ANALYSIS (GOAL SEEK)

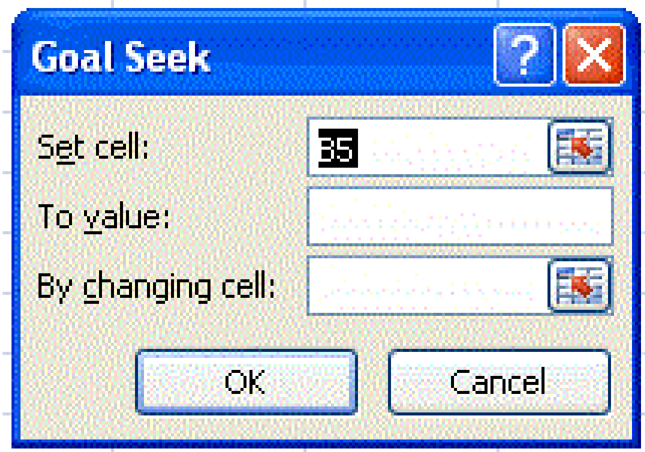
What-if Analysis



- Data → Data Tools → What-If Analysis
- Allow you to see the effect of different values on a (group of) formulas
- For this course, we will focus on Goal Seek
  - What should the input value be to get the result I need?

Goal Seek

- First, you need to set the formula on your spreadsheet
- Then you can use goal seek
  - Set cell: is the cell where the target output value is
  - To value: is the target output value
  - By changing cell: the input value



The answer will be here

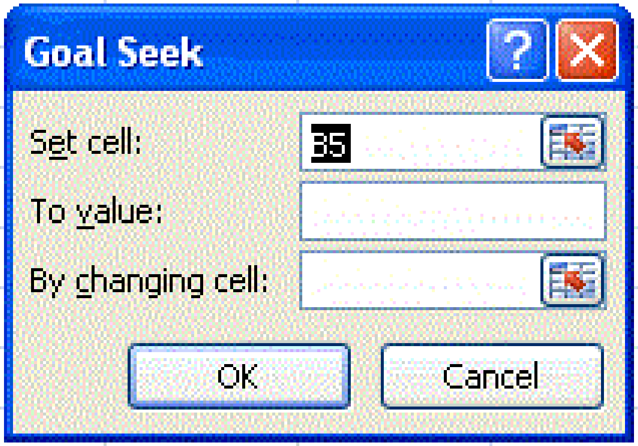
Goal Seek Example

You want to solve $x^2 - 16.5x + 35 = 0$

- Let Cell B1 be x. Set the value to 0.
- Let Cell A2:A4 contain the coefficients, 1, -16.5 and 35
- Let B2 = B1^2, B3 = B1,and B3 = 1
- Let C2 = A2*B2, C3 = A3*B3, C4 = A4*B4
- Let C5 = SUM(C2:C4)

Goal Seek (cont.)

- Start goal seek.
- Set...
  - *Set cell:* to C5
  - *To value*: to 0
  - And *By changing cell*: to B1
- Click OK. The answer should be at cell B1.
- You can try changing B2 to 20, and run goal seek again.



LINEAR REGRESSION

Linear Regression

- Linear regression is an analysis technique used to derive a relationship between dependent variable and one or more independent (explanatory) variables.
- Can be use both to explain data, and predict values of data in the future.

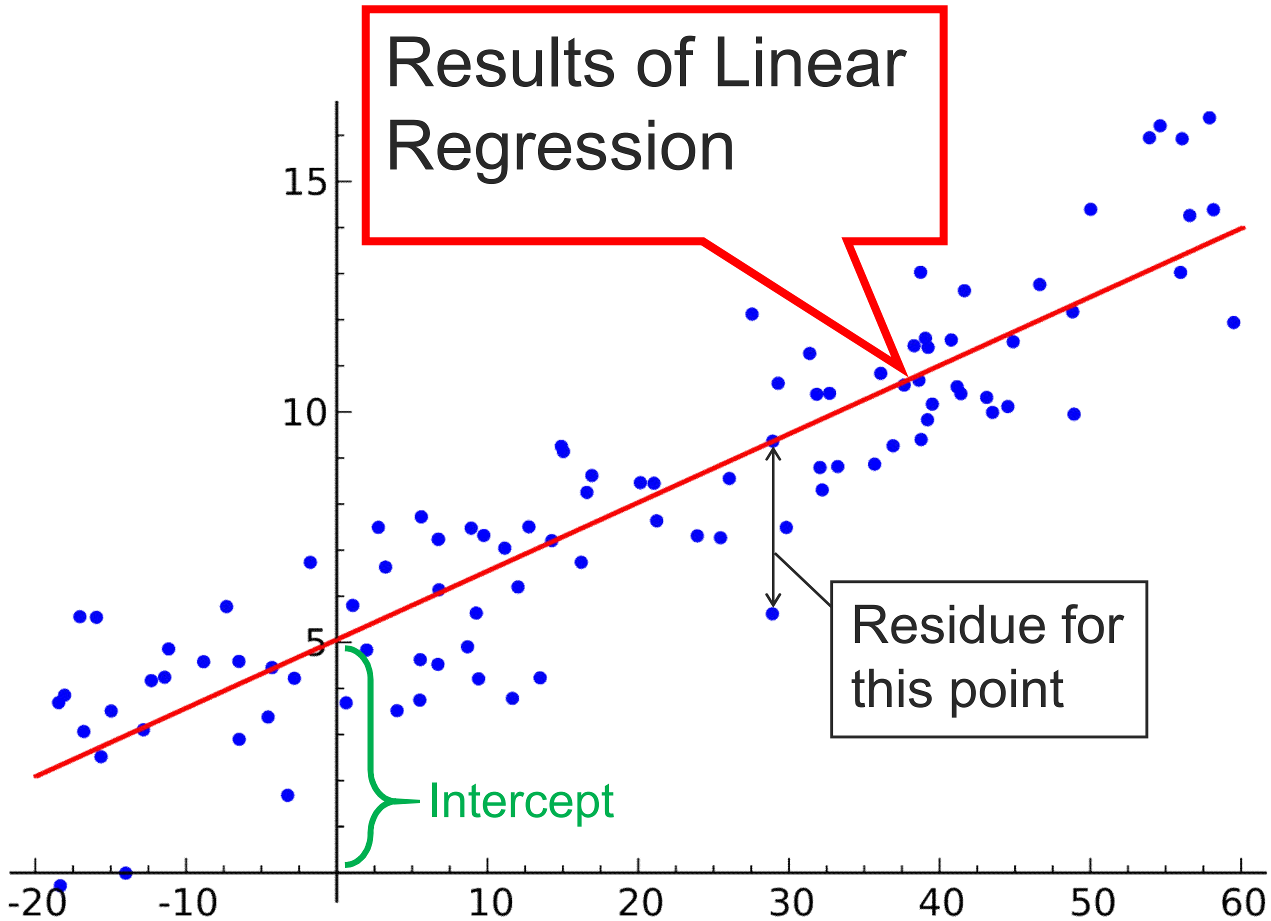
Linear Regression (cont.)

- Linear regression results will be in this form:

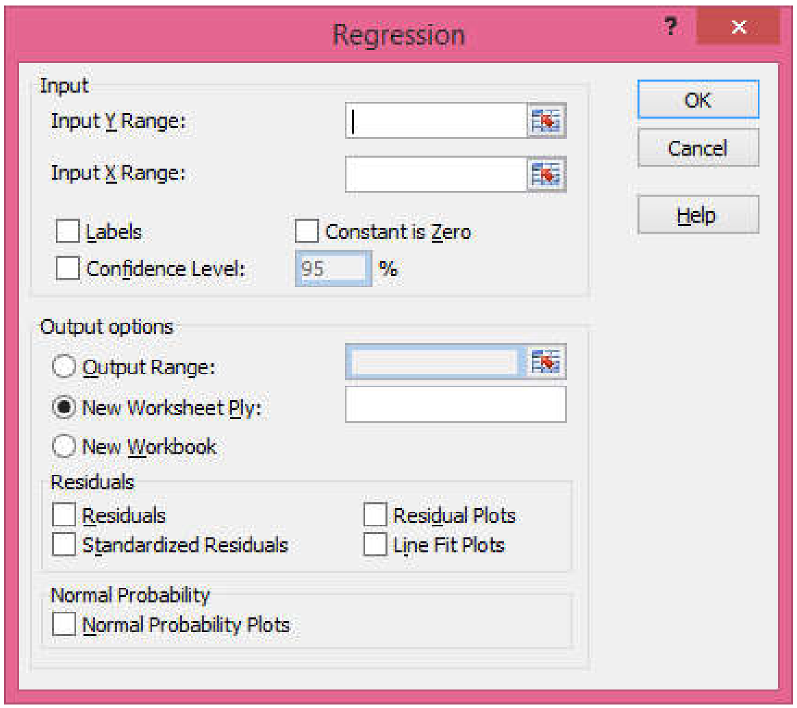
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- Y is dependent variable.
- $X_1..X_n$ are independent or explanatory variables.
- β_0 is an intercept or constant values of the equations.
- $\beta_1.. \beta_n$ are coefficients of each independent variables.
- ε is an error term or residue. This is <u>not</u> a constant.

Linear Regression, in Graph

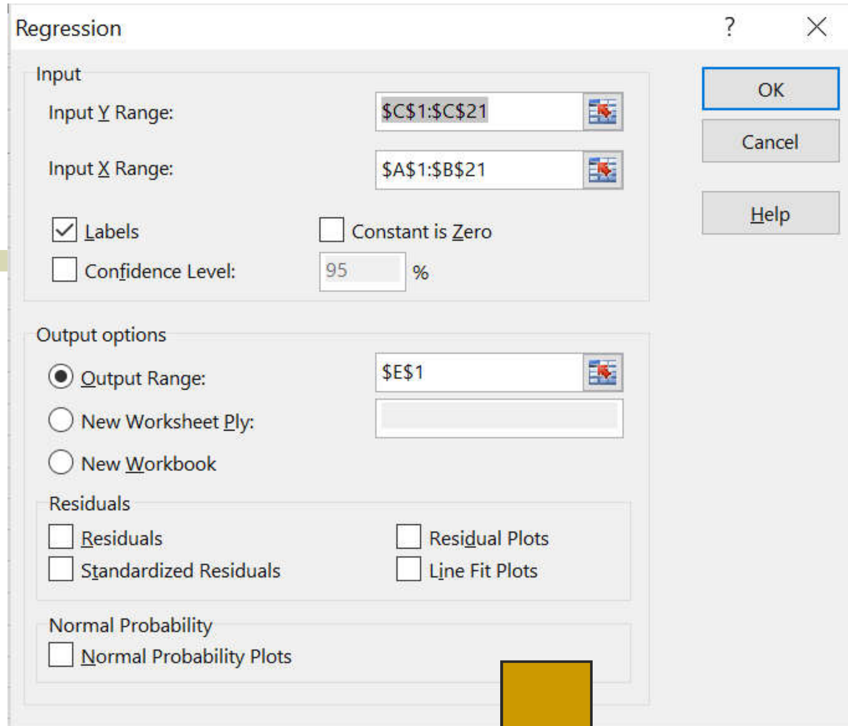


Linear Regression (cont.)



1. Select *Data* → *Data Analysis* → *Regression*
2. Select Y (dependent variable) Range
3. Select X (independent variable) Range
4. Select Output Range.
5. Select options.
6. Click *OK*

	A	B	C
1	X1	X2	Y
2	0	0	5.049
3	0.62	1.7	7.981
4	1.47	3.01	11.036
5	2.04	4.06	13.163
6	2.11	5.69	14.925
7	2.16	6.74	16.087
8	2.19	8.28	17.725
9	2.27	9.82	19.395
10	2.82	11	21.734
11	3.28	12.25	23.862
12	3.45	13.3	25.272
13	4.13	14.45	27.809
14	4.9	15.81	30.692
15	5.33	16.81	32.505
16	5.83	17.93	34.6
17	6.76	19.64	38.212
18	7.55	20.98	41.083
19	8.36	22.63	44.398
20	8.81	24.29	46.95
21	9.69	25.83	50.238



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99999831
R Square	0.99999661
Adjusted R Square	0.99999621
Standard Error	0.02599426
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3391.344686	1695.67234	2509499	3.1883E-47
Residual	17	0.011486927	0.0006757		
Total	19	3391.356173			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0*	*Upper 95.0%*
Intercept	5.04734708	0.011855951	425.722653	1.1E-35	5.02233321	5.07236095	5.02233	5.07236095
X1	1.97983667	0.009738943	203.290713	3.2E-30	1.9592893	2.00038405	1.95929	2.00038405
X2	1.00667981	0.003559745	282.795514	1.2E-32	0.9991694	1.01419021	0.99917	1.01419021

Reading the Results

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99999831
R Square	0.99999661
Adjusted R Square	0.99999621
Standard Error	0.02599426
Observations	20

- R-square, adjusted R-square measure accuracy of model as a whole. (The entire equation).
- Closer to 1 is better.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3391.344686	1695.67234	2509499	3.1883E-47
Residual	17	0.011486927	0.0006757		
Total	19	3391.356173			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	5.04734708	0.011855951	425.722653	1.1E-35	5.02233321
X1	1.97983667	0.009738943	203.290713	3.2E-30	1.9592893
X2	1.00667981	0.003559745	282.795514	1.2E-32	0.9991694

- t-Stat, P-value measure accuracy for that one variable.
- The higher the t-Stat, the better.
- The lower the P-value, the better

Coefficient for each variable + intercept
The regressed equation is Y = 1.98*X1 + 1.01*X2 + 5.05

Compare the Results

- You can then compute estimated

$$\hat{Y} = 1.98X_1 + 1.01X_2 + 5.05$$

and compare the difference

Y	Y^	diffY
5.049	5.05	0.02%
7.981	7.9946	0.17%
11.036	11.0007	0.32%
13.163	13.1898	0.20%
14.925	14.9747	0.33%
16.087	16.1342	0.29%
17.725	17.749	0.14%
19.395	19.4628	0.35%
21.734	21.7436	0.04%
23.862	23.9169	0.23%
25.272	25.314	0.17%
27.809	27.8219	0.05%
30.692	30.7201	0.09%
32.505	32.5815	0.24%
34.6	34.7027	0.30%
38.212	38.2712	0.15%
41.083	41.1888	0.26%
44.398	44.4591	0.14%
46.95	47.0267	0.16%
50.238	50.3245	0.17%