

Introduction to Excel for Statistical Analysis

Garrett Morrow, Laura Johnson, and Cara Marta Messina
Development Economics
Silvia Prina
Fall 2019



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Workshop Agenda

- Objectives
- About Excel
- Important Vocabulary and Functions
- Demonstration
- Activity: Practice Excel

Slides, handouts, and data available at

<http://bit.ly/dti-dev-econ-fall2019>



Workshop Objectives

- Understand the data structures of Excel
- Learn how to use basic Excel functions, such as =ADD and =SUM
- Learn how to analyze your data with pivot tables and charts
- Learn more advanced calculations like regression models



Excel

Excel is a program that is used to create and edit spreadsheets. In Excel, data are organized into rows and columns; data can be presented and analyzed using Excel's functions, such as pivot tables, charts, formulas, and more.



Why Excel?

Excel is an excellent way to store, organize, and analyze data. It is particularly useful for quantitative analysis because most of its functions are designed for numerical data.

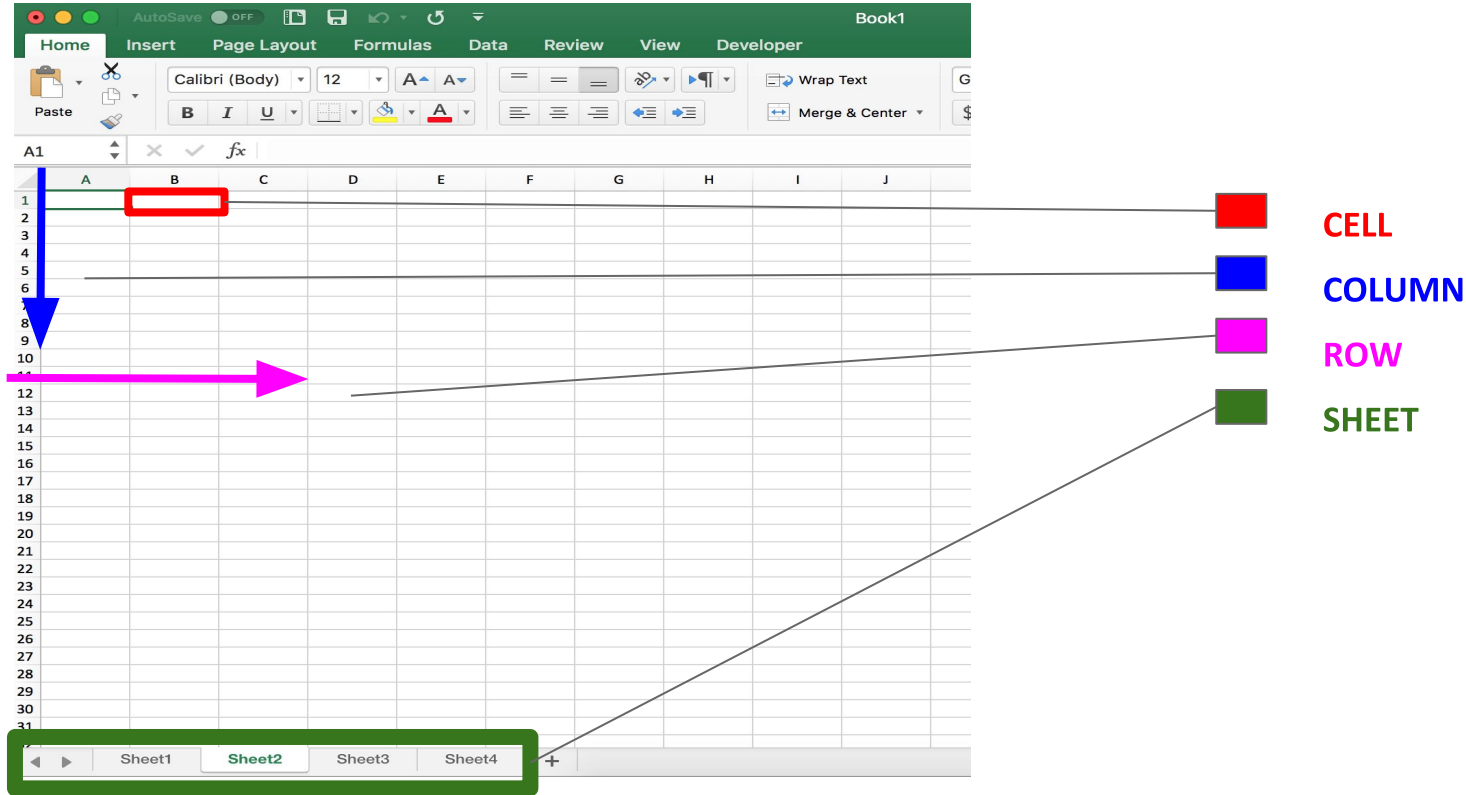


Important Vocabulary

- **Workbook:** the overall Excel file that you are creating
- **Sheet:** the different sheets inside the workbook that can be renamed
- **Row:** the horizontal and numerical (horizontal) rows
- **Column:** the vertical and alphabetical (vertical) columns
- **Cell:** the boxes that each have an ID based on its row and column placement (A1, A2, A3, etc).



Anatomy of Excel



Important Excel Features

- Function: Used to calculate and analyze numerical data using mean, median, standard deviation, addition, subtraction, and other forms of arithmetic
- Pivot Tables: Used to analyze and calculate numerical data and present different results based on functions and data chosen
- Charts: Used to visualize data with bar charts, scatter plots, and other formats.



How to Select Data

If you have a long dataset, it can be hard to drag your mouse down to the bottom of the dataset. Click

SHIFT + COMMAND/CONTROL + DOWN ARROW (or whatever direction)

The end of the data will be selected in the direction of the arrow you choose.



Basic Calculations

Using **pivot tables** or **functions**, you can find the:

- Average (Mean)
- Mode & Median
- Standard deviation
- Min/max values
- Correlation
- Results to other basic calculations such as addition, subtraction, division, multiplication



Functions for Excel

- In an empty cell, type = and then the proper calculation:
 - Correlation: CORREL(
 - Sum: SUM(
 - Average: AVERAGE(
 - Standard Deviation: STDEV(
- Select the range to calculate. If you are still in the function cell, the range will be automatically added for you as you select
 - Example: CORREL(B2:B20,C2:C20). B2:B20 is one range of values, while C2:C20 is another range.

D	E
hhe	
82.19051	
85.88746	=SUM(D2:D551
40.38055	SUM([number1],
40.68994	
49.5274	
67.08327	
42.86265	
65.08897	E7
36.52134	
47.16312	
37.96223	

The selected data (D column from rows 2-551)

The function (SUM) with the selected data



Your Turn!

Use the data emailed to you (also available the bit.ly link below) to calculate these for the “agehh”:

- Average
- Sum
- Median

Slides, handouts, and data available at

<http://bit.ly/dti-dev-econ-fall2019>



Pivot Tables for Calculations

- Select the data you want to be calculated (which can be more than one variable)
- Go to “Insert” > “Table” > “Pivot Table”
- Choose a new worksheet or add to your existing sheet. Creating a new worksheet is cleaner
- Go to “Pivot Table Analyze” to edit the table:
 - Go to “Field Settings” and choose the calculation (or right click the top of the table)



Example of Pivot Tables

Row Labels ▼	Average of hhe	Sum of hhe
34	67.40711229	38530.49088
99	72.46467868	
Grand Total	70.05543796	

Pivot table with **one** variable (looking at the average, but you can look at other calculations)

Pivot table with **two** variables (comparing one variable's values to another variable's values). This pivot table shows the average "hhe" for each of the variables in the "local" row.



Your Turn! Create your own pivot table

Find the average variables of the column “agehh” for each of the variables in the “eduhh” columns.

- Select the two columns (Shift+Command/Cntrl+Down Arrow)
- Click “Insert” then “Pivot Table”
- Use the PivotTable Fields to select both the “agehh” and “eduhh” columns
- Make “educhh” the pivot table’s rows and make the values the average of “agehh”



More Advanced Calculations - LINEST

LINEST is a statistical function that uses the least squares method to calculate a regression line. OLS Equation:

$$y = a + bx_1 \dots bx_n$$

- y = expected value
- a = intercept
- $bx_1 \dots bx_n$ = beta-coefficient (b) * value (x)



LINEST Excel Syntax

=LINEST(y_values, x_values, constant, additional_statistics)

- Note: x_values, constant, and additional_statistics are OPTIONAL, but we almost always use them.

What is the relationship between variable

“hhe” and variable “educhh?”

LINEST Steps

1. Select multiple rows + columns (2x2)
2. =Linest(D2:D551, G2:G551, TRUE, TRUE)
3. Control+Shift+Enter
4. =-2.0558007, 76.629212

Function Arguments

LINEST

Known_ys	D2:D551	= {82.19051;85.88746;40.38055;40.68994
Known_xs	G2:G551	= {3;3;2;2;2;2;4;4;5;5;5;5;6;6;4;4;3;0;0;3;
Const	True	= TRUE
Stats	True	= TRUE

Returns statistics that describe a linear trend matching known data points, by fitting a straight line using the least squares method.

Stats is a logical value: return additional regression statistics = TRUE; return m-coefficients and the constant b = FALSE or omitted.


Formula result = -2.055800695




[Help on this function](#)

OK Cancel



Example

File Home Insert Page Layout Formulas Data Review View Help  Search

L2    {=LINEST(D2:D551, G2:G551,TRUE,TRUE)}

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	hhid	round	local	hhe	treatcom	agehh	educhh	hssize	pscore	takeup			
2	15681	1	34	82.19051	1	53	3	3	865	0		-2.0558007	76.629212
3	15681	0	34	85.88746	1	52	3	3	865	0		0.50672529	2.07296412
4	15680	1	34	40.38055	1	51	2	7	602	1			
5	15680	0	34	40.68994	1	50	2	7	602	1			
6	15679	1	34	49.5274	1	43	2	5	653	1			
7	15679	0	34	67.08327	1	42	2	5	653	1			
8	15678	1	34	42.86265	1	29	4	3	619	1			
9	15678	0	34	65.08897	1	28	4	3	619	1			
10	15677	1	34	36.52134	1	46	5	6	525	1			
11	15677	0	34	47.16312	1	45	5	6	525	1			
12	15676	1	34	37.96223	1	27	5	4	686	1			
13	15676	0	34	53.53526	1	26	5	4	686	1			
14	15675	1	34	51.61393	1	22	6	3	622	1			
15	15675	0	34	58.82847	1	21	6	3	622	1			
16	15672	1	34	36.73437	1	41	4	7	635	1			
17	15672	0	34	39.0182	1	40	4	7	635	1			
18	15671	1	34	87.8801	1	53	3	2	735	1			
19	15671	0	34	85.22186	1	52	3	2	735	1			
20	15670	1	34	44.85114	1	31	0	5	549	1			
21	15670	0	34	44.4139	1	30	0	5	549	1			
22	15667	1	34	23.31059	1	45	3	2	667	1			
23	15667	0	34	74.36211	1	44	3	2	667	1			
24	15666	1	34	34.17051	1	42	0	4	594	1			
25	15666	0	34	59.11292	1	41	0	4	594	1			
26	15665	1	34	43.7287	1	36	3	6	513	1			
27	15665	0	34	43.28144	1	35	3	6	513	1			
28	15664	1	34	33.48979	1	32	0	5	542	1			



Alternative Excel Regression Method

- Use the “Analysis ToolPak” Add-in
 - Then Data → Data Analysis → Regression

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.170762315							
R Square	0.029159768							
Adjusted R Square	0.027388162							
Standard Error	30.32197098							
Observations	550							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	15133.23292	15133.23292	16.45950844	5.69217E-05			
Residual	548	503843.2142	919.4219238					
Total	549	518976.4472						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	76.62921204	2.072964121	36.96600981	5.6084E-151	72.55728372	80.70114036	72.55728372	80.70114036
educhh	-2.055800695	0.506725288	-4.057031975	5.69217E-05	-3.051162374	-1.060439016	-3.051162374	-1.060439016

Regression

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level: %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help



Multivariate LINEST

What is the relationship between “hhe” and “educhh” + “hhsiz

Similar syntax: =LINEST(D2:D551, G2:H551, TRUE, TRUE)

Select rows & columns - you need 1 more column than the number of variables because of the constant

Then press “Control+Shift+Enter”

The return of statistics is in **reverse order**



Example

FileHomeInsertPage LayoutFormulasDataReviewViewHelpSearch

L2

<



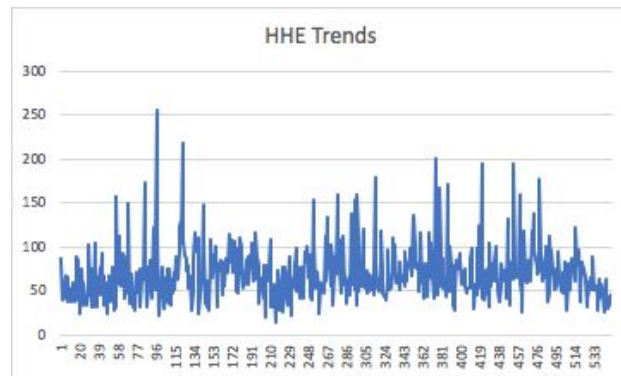
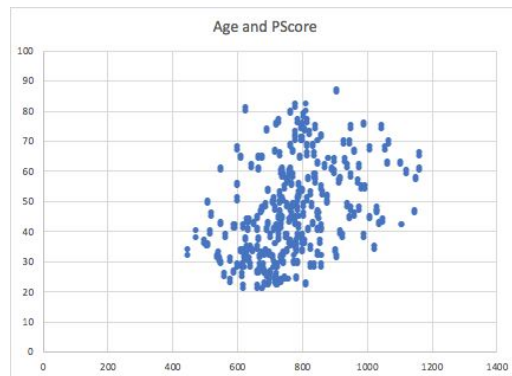
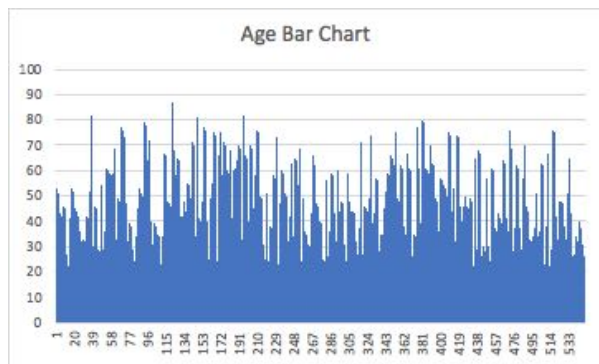
Add-In Example

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.498168082							
R Square	0.248171438							
Adjusted R Square	0.245422522							
Standard Error	26.70788953							
Observations	550							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	2	128795.1314	64397.5657	90.27974178	1.3137E-34			
Residual	547	390181.3158	713.3113634					
Total	549	518976.4472						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	111.548207	3.314526878	33.65433774	2.3048E-135	105.0374477	118.0589663	105.0374477	118.0589663
educhh	-2.202133341	0.446479123	-4.932220179	1.08015E-06	-3.079156886	-1.325109797	-3.079156886	-1.325109797
hhsz	-6.328687118	0.501355454	-12.62315403	3.069E-32	-7.313504805	-5.343869431	-7.313504805	-5.343869431



Charts

- Scatter plots: comparing **two** variables
- Bar charts/histograms: count of **one** variable
- Line charts: tracing **trends** of one or two variables



Inserting a Chart

- Similar to a pivot table, click the columns and variables you would like to include
 - For multiple columns, you may need to move the columns next to each other to be able to select multiple columns.
- Go to “Insert” and then “Charts” (often, “recommended charts” will suggest the option that you want)
- Use the “Chart Design” and “Format” toolbar at the top and/or the side toolbar to play with the formatting of the chart



Your Turn!

Create two charts.

- Histogram for “hhe”
- Scatterplot for “agehh” and “eduhh”

Slides, handouts, and data available at

<http://bit.ly/dti-dev-econ-fall2019>



Group Discussion

- First, does anyone have questions?
- How was using Excel? What are some easy features?
- What are some more difficult features, or aspects that you think will be challenging to work with?
- How might you use Excel in the future?



Thank you!

If you have any questions, contact us at:

Garrett Morrow

Digital Teaching Integration
Research Fellow

morrow.g@husky.neu.edu

Laura Johnson

Digital Teaching Integration
NULab Coordinator

johnson.lau@husky.neu.edu

Cara Marta Messina

Digital Teaching Integration
Assistant Director

messina.c@husky.neu.edu

Slides, handouts, and data available at <http://bit.ly/dti-dev-econ-fall2019>

Office Hours: **Tuesdays from 1–3PM in 401 Nightingale Hall**



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*