

Introduction to Excel for Statistical Analysis

Taught by Juniper Johnson & Ana Abraham
ENGL 7360 Topics in Rhetoric
Mya Poe
Fall 2022



Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Objectives
 - Understand relevant uses of Excel
 - Learn about basic Excel functions
 - Practice making table and chart
- About Excel
- Important vocabulary and functions
- Demonstration

Slides, handouts, and data available at: <https://bit.ly/fa22-poe-excel>





What is Excel?

Excel is a program used to create and edit **spreadsheets**. In Excel, data is organized into rows and columns; this 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 both data and metadata (data about data). Although it is particularly useful for budgeting and finance because many of its functions revolve around numerical data, Excel is used quite often across the disciplines.

In humanities and social science contexts, you might use Excel to pursue research interests, particularly for materials that are provided as spreadsheets (census data, bibliographies, and more).



Common Ways to Use Excel

- Tracking job applications
- Budgeting for events in your personal life
- Collaborative task-tracking (Google Sheets can also be helpful for this)
- Outlining content to be written for a website
- Analyzing data stored in .csv (comma separated value) files
- Collecting and analyzing survey information



Example dataset: Co-op application tracker

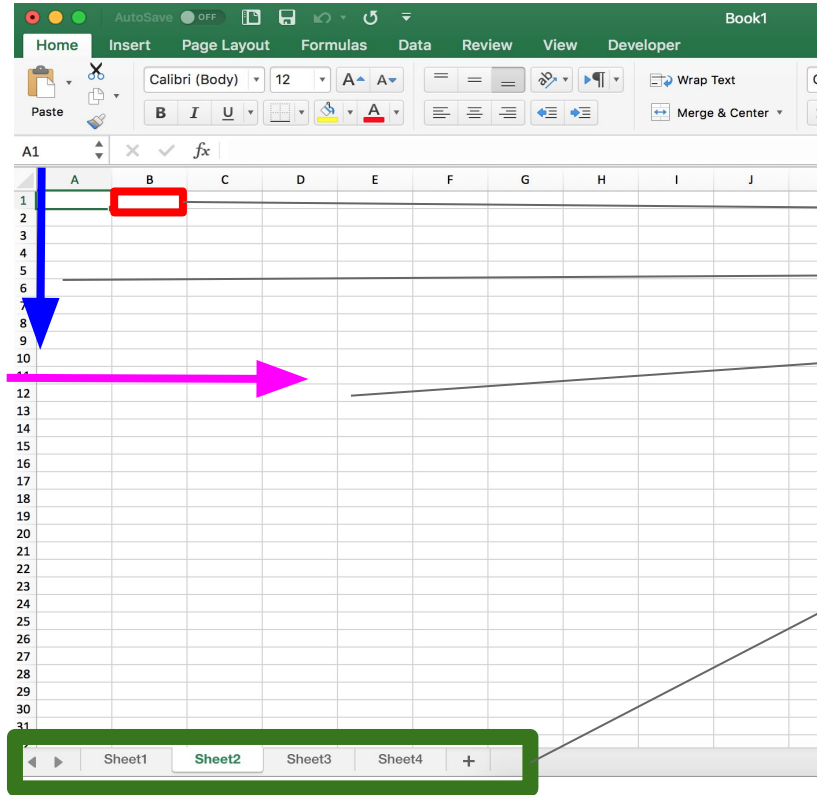
Home Insert Draw Page Layout Formulas Data Review View Acrobat Tell me						
Corbel (Body) 12 A^ A^ B I U A % Conditional Formatting Format as Table Cell Styles Cells Editing Analyz Data						
F17 x ✓ fx						
	A	B	C	D	E	F
1	Position	Organization	Pay (hourly wage)	Application Deadline	Applied (Y/N)	Interview?
2	Research Assistant	Ford's Theatre	18.00	19-Mar	Y	Pending
3	Museum Educator	The Spy Museum	16.00	31-Mar	Y	Declined
4	Digital Intern	White House Historical Association	20.00	22-Feb	Y	Declined
5	Museum Intern	Department of the Interior Museum	15.00	28-Feb	Y	Interview on 10-Mar
6	Library Assistant	DC History Center	19.25	19-Mar	Y	Interview on 15-Mar
7	Historic Preservation Intern	National Park Service	15.00	15-Feb	Y	Interview on 10-Apr
8	Curatorial Assistant	Maryland Historical Society	17.50	1-Apr	N	Pending
9						
10						
11						
12	note: these positions are fictional					

Important Vocabulary

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



Anatomy of Excel



CELL



COLUMN



ROW



SHEET



Important Excel Features

- **Functions:** Used to calculate and analyze numerical data, for example with: mean, median, standard deviation, addition, subtraction, and other forms of arithmetic.
- **Tables and Pivot Tables:** Used to filter, analyze, and summarize 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 **functions**, you can find the:

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



Writing Excel Functions

- In an empty cell, type = and then the calculation you want to do:
 - Sum: SUM()
 - Average: AVERAGE()
 - Median: MEDIAN()
 - Standard Deviation: STDEV()
- Select the range to calculate. You can enter the names of the cells or manually select the cells you want included. Then close the brackets.
- You can also write functions referencing other worksheets by using the sheet name and '!'. Example:
 - =AVERAGE(Sheet1!C2:C8)

C	D
Pay (hourly wage)	
\$17.50	
\$20.00	
\$15.00	
\$19.25	
\$16.00	
\$15.00	
\$18.00	=SUM(C2:C8)

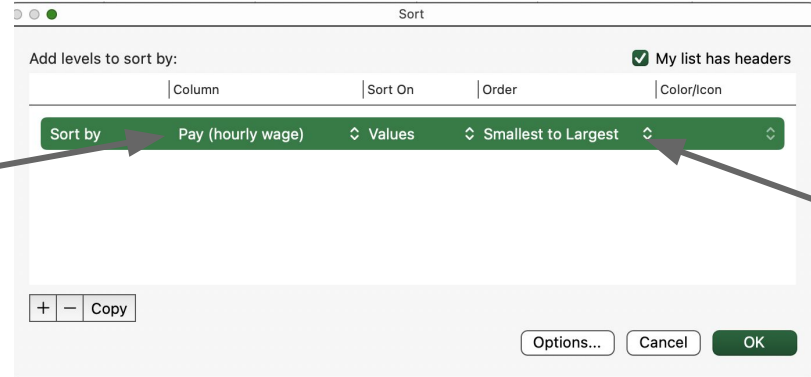
The selected data (C column from rows 2–8)

The function (SUM) with the selected data



Sorting Data

- **Sorting** allows you to organize your data by a certain value
- Select your dataset.
- Select “Sort” under the “Data” tab. Once you click, a pop-up window will appear.
- Choose which column you would like to sort values by, and how you would like to order the sort. The entire dataset will be sorted accordingly.
- If your list has headers (column titles) make sure to tick the right-upper box. Otherwise, Excel will automatically sort the column labels or titles as well.
- You can use the + button to add multiple sort requests.



Select column
from
drop-down

The order will vary
depending on the
variable's category
(number, text)



Your Turn!

Using the demo data provided, try out the following functions in Excel:

- **Sort data** in one column, either alphabetically or by numerical value
 - You can try sorting your data by the title of the Co-op position, by hourly pay, application deadline, or any other category that makes sense.
- **Calculate** the average and median hourly wage for the positions listed.



Adding Data Validation

- **Data validation** allows you to set a limited range of responses (either numbers or words/letters) for a selected group of cells.
- Highlight the cells to which you want to apply the data validation
- Select “Data Validation” under the “Data” tab
- Change “Allow” from “Any value” to “List” in the drop-down menu
- Type the responses you want to allow, separated by commas and spaces
- When applying data validation to filled-in cells, Excel will automatically overwrite the cell content. You can avoid this by creating new columns to apply data validation to.

Change from
“Any value” to
“List”

The screenshot shows the 'Data Validation' dialog box with the 'Settings' tab selected. Under 'Validation criteria', the 'Allow' dropdown is set to 'List'. The 'Data' dropdown is set to 'between'. The 'Source' field contains the text 'Definitely, Probably, Maybe'. There are two checked options: 'Ignore blank' and 'In-cell dropdown'. At the bottom, there is a checkbox for 'Apply these changes to all other cells' which is unchecked, and a 'Clear All' button.

Enter the values
you wish to allow,
separated by
commas and spaces



Your Turn!

Using your sorted data:

- **Add Data Validation** to a column (either qualitative or quantitative data). Ideas include:
 - Whether the position is virtual or in person
 - Your interest level or qualification level
 - Whether you have submitted your application



Adding Conditional Formatting

- Conditional formatting adds automatic color-coding based on your data values
- Highlight the cells to which you want to apply the conditional formatting
- Select “Conditional Formatting” under the “Home” tab and choose from a range of color-coding options
 - Options include Highlighting Rules, Data Bars, Color Scales, and Icon Sets
 - You can visualize numerical variables with data bars (left) or set rules for specific text (right)

Pay (hourly wage)	Interview?
17.50	Pending
20.00	Declined
15.00	Interview on 10-Apr
19.25	Interview on 15-Mar
16.00	Declined
15.00	Interview on 10-Mar
18.00	Pending

Data Bar
visualization

Highlighting
rule



Your Turn!

Now, **add Conditional Formatting** to a column.

- Use Data Bars or Color Scales on your numerical columns, namely: pay, application dates, your interest or qualification level on a scale of 1-5, etc.
- Apply a Highlighting Rule to a column (it can be a text or numerical column).



Creating a Table

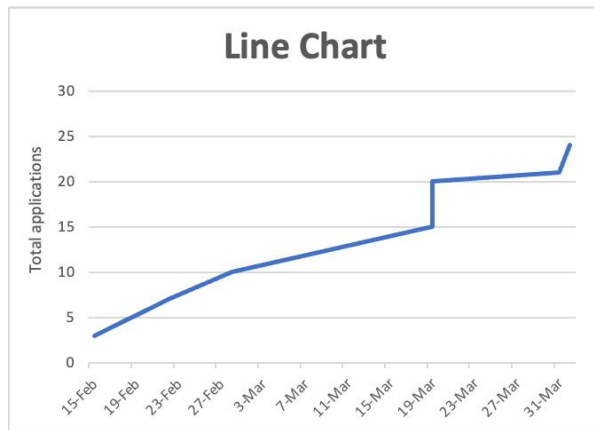
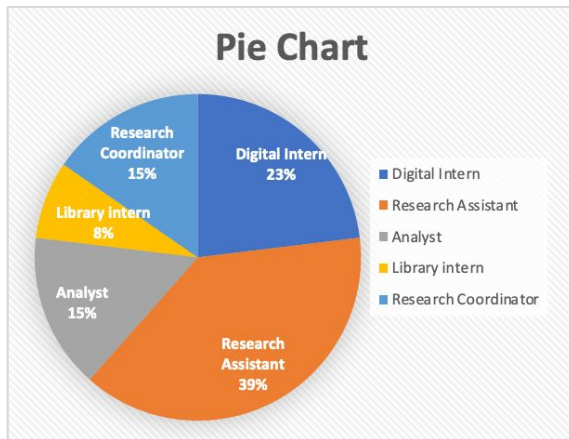
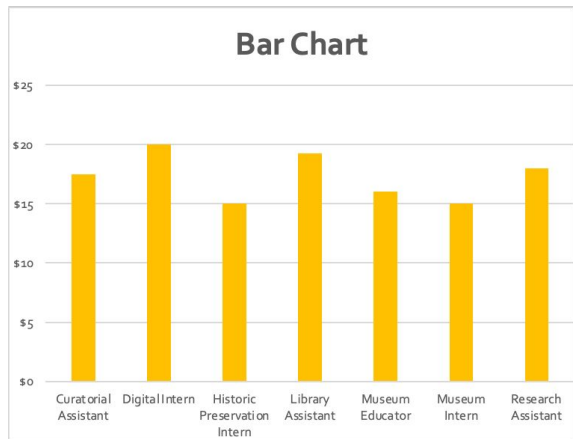
- **Tables** allow you to present your information in a more polished way. They also create a visual border between your data and the rest of the spreadsheet document.
- Select all the cells that you want included in your table
- Under the “Insert” tab, select “Table”
- You can customize the appearance of your table under the “Table” tab, much as you would in Microsoft Word
- You can still modify your data once it is in a table; although tables make your data look more presentable, they are not a “finished” form

Your Turn: using the same data, insert and customize a table.



Charts

- While tables represent data or information in rows and columns, a chart is the graphical representation of data in symbols like bars, lines, and slices.
- There are many types of charts you can create with Excel



Creating a Chart

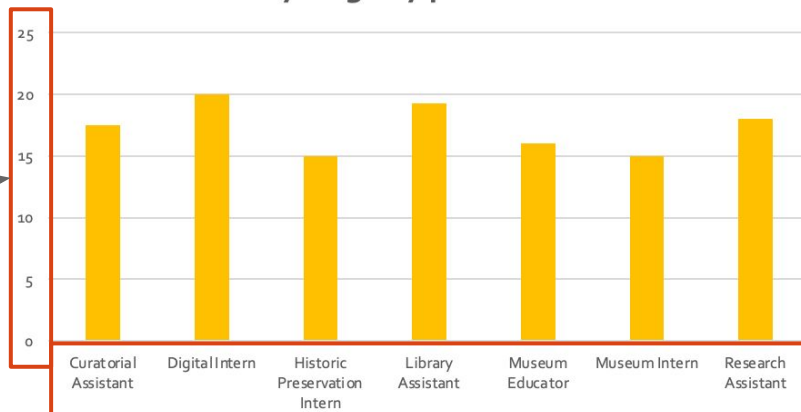
- First, you must **select a chart type** that suits your data:
 - **Bar Chart:** compares parts of a bigger set of data, highlights different categories, or shows change over time. Be careful not to overload your graph, avoid having more than 10 bars.
 - **Pie Chart:** shows relative proportions and percentages of a whole dataset. Best used with small datasets (up to 6 categories).
 - **Line Chart:** for continuous dataset that changes over time. Use it if your dataset is too big for a bar chart, or if you want to visualize trends instead of exact values.
 - **Scatter plots, bubble charts, area charts, etc...** You can learn more [here](#)



Creating a Chart (cont'd)

- Select the columns and variables you would like to include
 - For multiple columns, you may need to move the columns next to each other.
- Go to “Insert” and then “Charts”. Choose the chart type you want.
- If you are creating a bar or line chart, consider what your x-axis and y-axis should be.

Hourly wage by position



y-axis: should be numerical value

x-axis: should contain your categories



Formatting your chart

Once you have your chart, you may want to customize some aspects for more clarity and precision.

- **Format axis:** right-click either axis and select “Format Axis”. Under “Number” you can choose the appropriate category (number, currency, date), the number of decimal places, and the scale for the axis.
- **Add Elements:** under the “Chart Design” toolbar at the top, select “Add Elements” to add or delete chart and axis titles, data labels, gridlines, and legends.
- **Other formatting:** Change colors, font and size under “Chart Design” and “Format”.



Your Turn!

Using the demo data, try to make an Excel chart (make sure to use the Desktop version of Excel for all functionalities)

Select the columns you want to visualize. This may be the position and hourly wage, or application deadline for example.

- **Choose the type of chart.**
- **Format the axes.** Consider:
 - What type of data is your y-axis: Numerical? Dates? Currency?
 - What scale is appropriate?
 - Label your axes
- **Add a Title**
- **Change any formatting you want.** Ideas include:
 - Font type and size
 - Bar fill
 - Gridlines: keep or delete





Additional Resources

The Internet has a wealth of Excel tutorials. Some particularly useful ones are linked below, including a tutorial for pivot tables, which were not covered in this workshop.

- [Data Validation](#)
- [Conditional Formatting](#)
- [Creating Charts](#)
- [Pivot Tables](#)

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;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
educ hh	-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

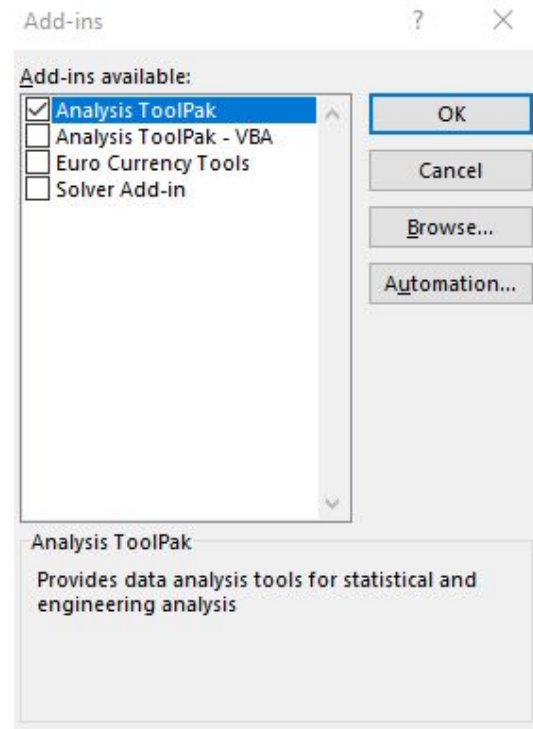
OK Cancel Help



Installing 'Analysis Toolpak'

- Analysis Toolpak “provides data analysis tools for statistical and engineering analysis.” It is an Excel add-in that allows for easy statistical analysis like bivariate and multivariate regression. We will also show you how to do regression analysis without the add-in.
- **For MacOS:** click on the “Tools” menu and select “Excel Add-ins”. In the “Add-ins available” box, check the “Analysis ToolPak” box. If you are unable to find this option, search for “Excel Add-ins” under the “Help” menu. If you have an older version of Excel, you may need to go to the Excel options in the “File” menu and find Add-ins there.
- **For Windows:** Click the “File” menu, then select “Options”, then the “Add-ins” category. In the “Manage” box, select “Excel Add-ins” and then click “Go.” The “Add-ins” box will appear, and there you can select “Analysis Toolpak” and click “Ok”.

Compatibility: Excel for Office 365, Excel for Office 365 for Mac, Excel 2019, Excel 2016, Excel 2019 for Mac, Excel 2013, Excel 2010, Excel 2007, Excel 2016 for Mac.



Multivariate LINEST

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

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

File Home Insert Page Layout Formulas Data Review View Help Search

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



Thank you!

If you have any questions, contact us at nulab.info@gmail.com

Schedule an appointment with DITI:

<https://calendly.com/diti-nu/>

We'd love your feedback! Please fill out a short survey here:

<https://bit.ly/diti-feedback>

Slides, handouts, and data available at

<https://bit.ly/fa22-poe-excel>

