Introduction to Excel for Statistical Analysis

Taught by Vaishali Kushwaha
Development Economics
Silvia Prina
Fall 2021

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

https://bit.ly/fall2021-prina

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

Workshop Objectives

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

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

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.

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

Installing Excel (reminder)

For more information on installing Excel, visit the guide published by Northeastern's ITS: http://bit.ly/2kDkYsL

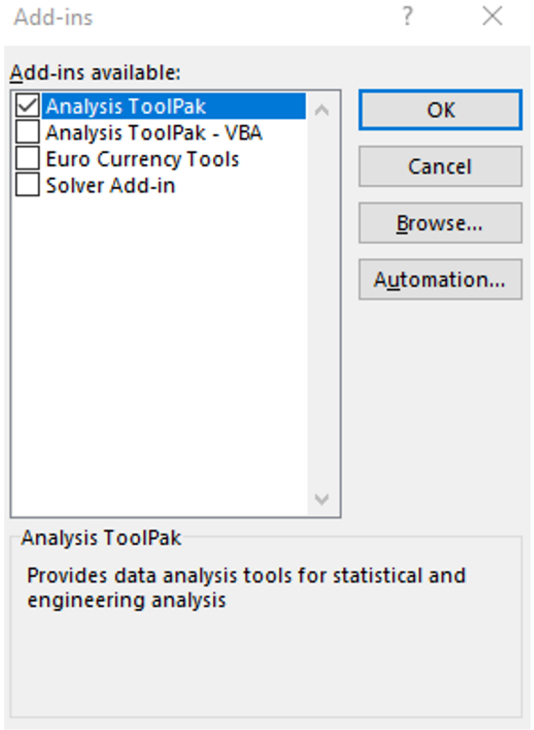
- Go to the Office 365 site (https://office.com) and enter your Northeastern email. You will be redirected to Northeastern's Office 365 portal, where you can enter your email and password (same as your university email)
- Once you're signed in, click "Install Office" in the top right corner and "Office 365". Excel is in this package
- Follow the directions to download and install the Microsoft Suite

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

Installing ‘Analysis Toolpak’ (reminder)

- 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.



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

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.

Please have Excel open now so you can follow along with this tutorial.

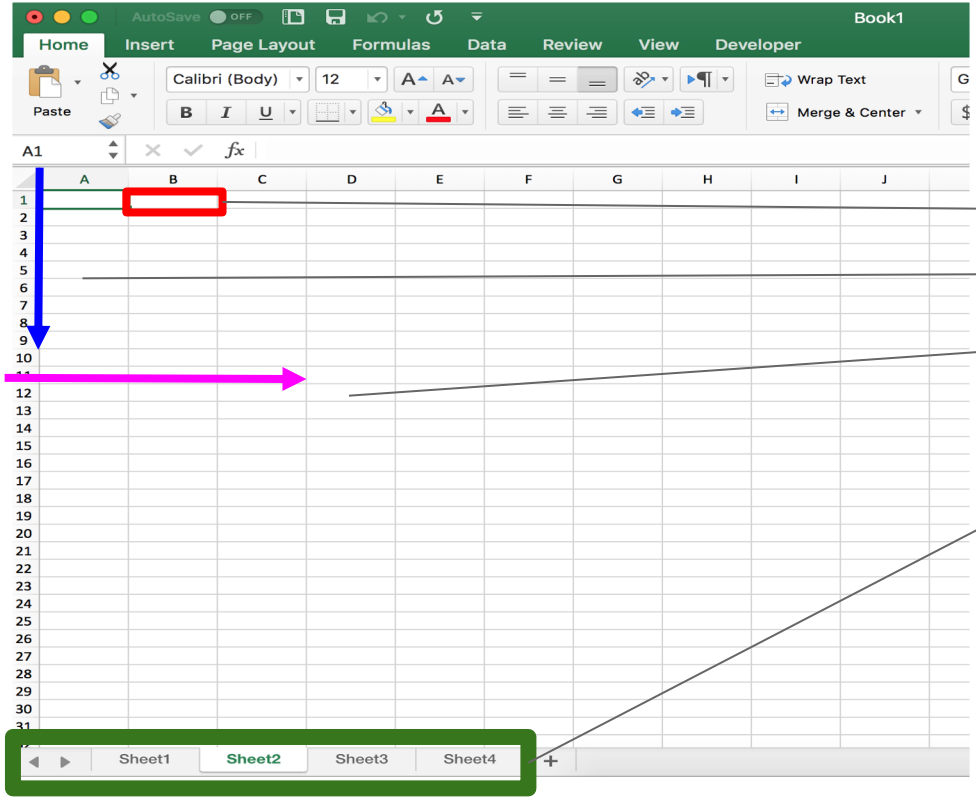
Feel free to ask questions at any point during the presentation!

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).

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

Anatomy of Excel



CELL

COLUMN

ROW

SHEET

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

Important Excel Features

- Function: Used to calculate and analyze numerical data using mean, median, standard deviation, addition, subtraction, and other forms of arithmetic.
- Tables and Pivot Tables: Used to filter, 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.

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

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.

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

Basic Calculations

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

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

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

Tutorial Scenario Goals

You are helping the Ministry of Health evaluate the effectiveness of a health program targeting poor households. The main objective of this program is to reduce the expenditure of low-income households in rural areas for health-related issues. A pilot has been carried out in several communities, and you have been asked to assess whether it actually generated a reduction in health expenditures for the beneficiary families.

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

Tutorial Dataset Variables

- **hhid:** a unique ID number for each household (constant across rounds)
- **round:** the year for which you have data (either 0 or 1)
- **local:** community identifier
- **hhe:** the household health expenditure
- **treatcom:** a dummy variable equal to one for the treatment community
- **agehh:** age of the household head
- **educhh:** years of education of the household head
- **hhsize:** household size at baseline
- **pscore:** the household poverty index
- **takeup:** a dummy variable equal to one if the household participated in the program

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

Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

- (a) How many observations are in the dataset?
- (b) How many households are in the dataset ?
- (c) How many localities are in the dataset?
- (d) How many localities are in the treatment group? And how many in the control?
- (e) How many households are in the treatment locality?
- (f) How many households are in the control locality?

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

Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

- **(a) How many observations are in the dataset?**
  - To answer this, think of observations as rows in our dataset.
    - How many rows do we have?
      - 551
  - However, remember that the first row is our variable headings
    - Therefore:
      - 551 - 1 = 550 observations

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

Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

- **(b) How many households are in the dataset ?**
  - Remember the context of this assignment: You are analyzing a pilot study to see if an intervention has had an effect on household health expenditure.
  - Therefore, we have two observations for each household: one observation before the health program, and one observation after the program. A baseline and an endline.
    - So, 550 / 2 = 275 households

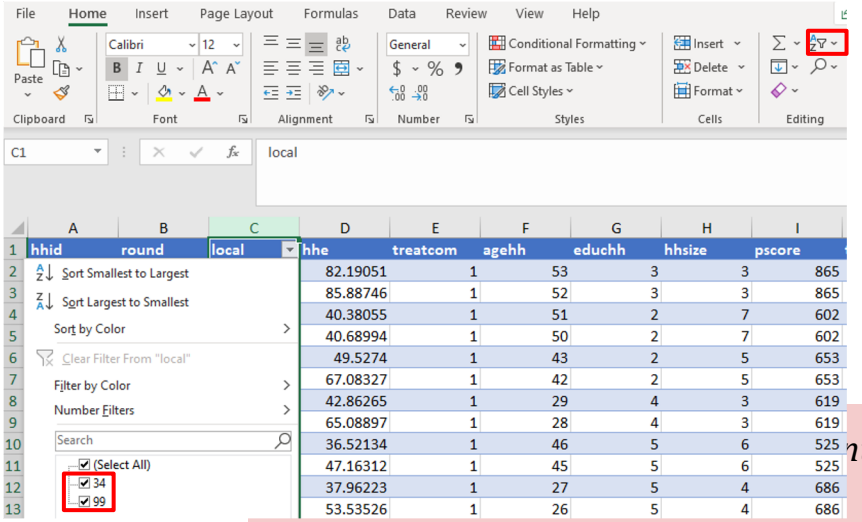
Feel free to ask questions at any point during the presentation!

Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

- **(c) How many localities are in the dataset?**
  - If we refer to our guide of what each variable is, we can see that 'local' is the community identifier, or the locality.
  - We can now look at this column in our dataset and see that we have two localities: 34 and 99.
  - Alternatively, we can also use the function:
    - =UNIQUE(C2:C551)

If we use the filter tool, we can also see the two localities.



Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

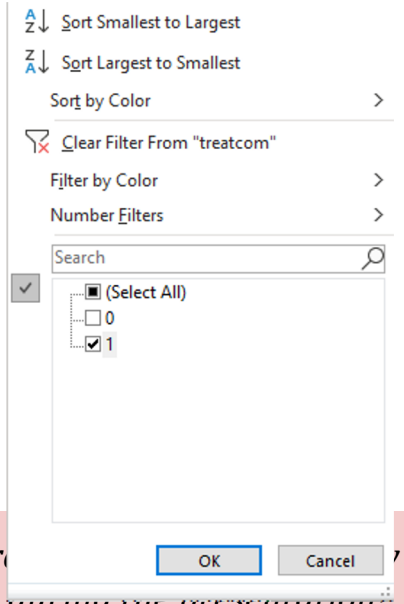
- **(d) How many localities are in the treatment group? And how many in the control?**
    - If we refer to our guide of what each variable is, we can see that 'treatcom' is the treatment identifier, meaning if there is a '1' that observation received the treatment, and if there is a '0' then that observation was in the control group.
    - Therefore, we can see that locality 99 is the control locality, and locality 34 is the treatment locality.
    - So, there is 1 locality in the treatment group, and 1 locality in the control group.

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

Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

- **(e) How many households are in the treatment locality?**
  - To find the answer to this question, we will use filtering and then counting rows.
  - Filter ‘treatcom’ by only 1 to mark the treatment group.
  - If we filter by ‘1’ we can see that there are 263 rows - header = 262.
    - 262 observations / 2 = 131 households

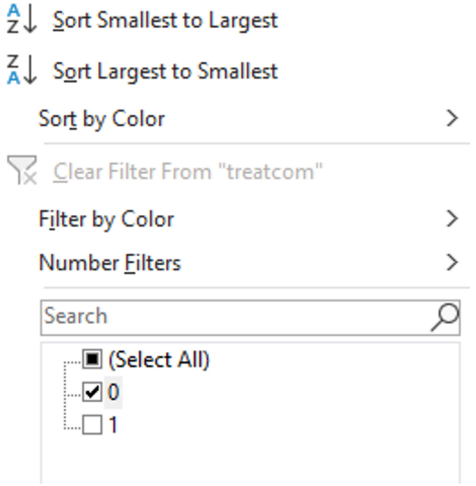


Feel fr… point during the presentation!

Mean, Standard Deviation, Min & Max

1. Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.

- **(f) How many households are in the control locality?**
  - To find the answer to this question, we will use filtering and then counting rows.
  - Filter ‘treatcom’ by only 0 to mark the control group.
  - If we filter by ‘0’ we can see that there are 289 rows - header = 288.
    - 288 observations / 2 = 144 households



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

Mean, Standard Deviation, Min & Max

1. **Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.**

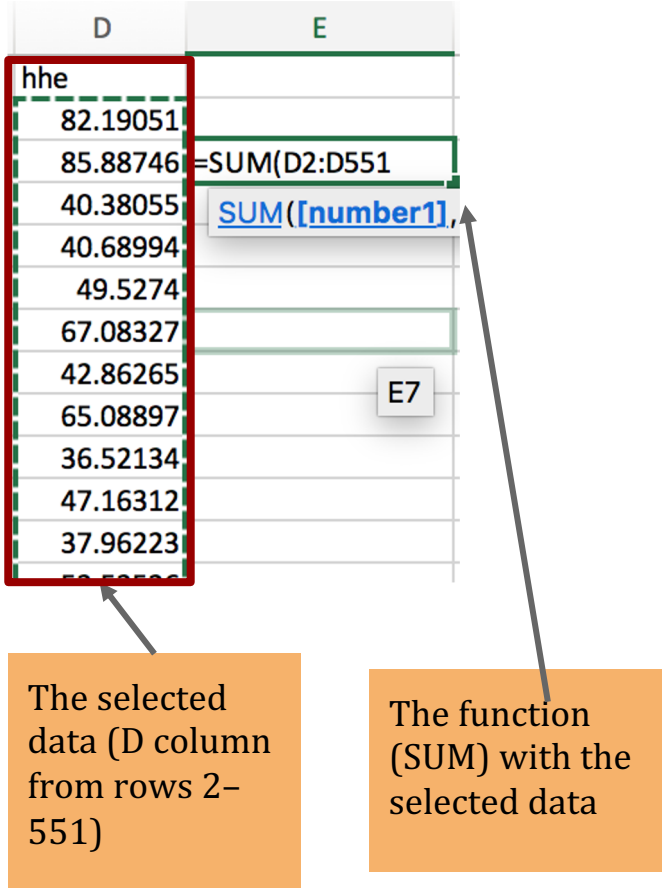
1. Create a new table by clicking on the '+' at the bottom of the page—this will be where we hold our results tables.
2. Title the columns: Variable, Mean, Std. Dev, Min, and Max
3. Under 'Variable' write the names of our variables: Household Health Expenditure, education, age, etc.

	A	B	C	D	E
1	Variable	Mean	Std. Dev	Min	Max
2	Household Health Expenditure				
3	Household Head Education				
4	Household Head Age				
5	Household Size				
6	Poverty Index				
7					

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

Writing Excel Functions

- 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.
- We can also write functions referencing other worksheets by using the sheet name and '!'. Example:
  - =AVERAGE(Sheet1!D2:D551)

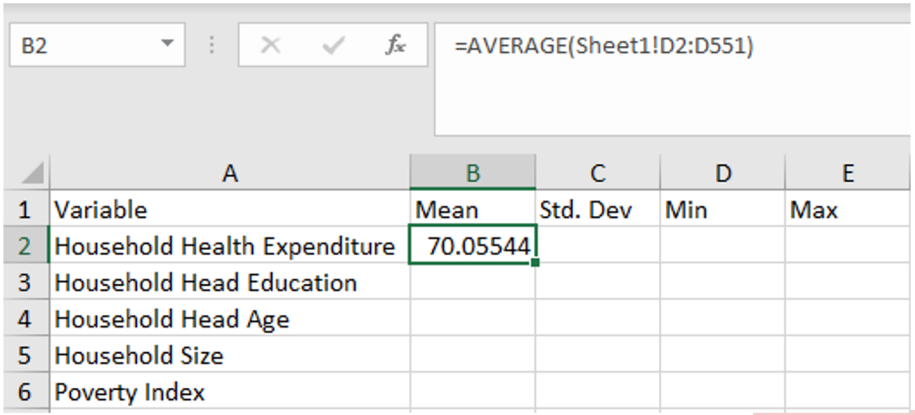


Feel free to ask questions at any point during the presentation!

Mean, Standard Deviation, Min & Max

1. **Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.**

1. In cell B2, write the following:
    a. =AVERAGE(Sheet1!D2:D551)
    b. This function is calculating the average of cells D2 through D551 on sheet 1.
2. We can now do the same for our other variables (Columns F,G,H, and I : Rows 2:551)



	A	B	C	D	E
1	Variable	Mean	Std. Dev	Min	Max
2	Household Health Expenditure	70.05544			
3	Household Head Education				
4	Household Head Age				
5	Household Size				
6	Poverty Index				

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

Mean, Standard Deviation, Min & Max

1. **Make a table that reports the means, standard deviations, and minimum and maximum values of all the household health expenditure, household head education and age, household size and poverty index in the dataset.**

1. Arithmetic Mean/Average function (for health expenditure): =AVERAGE(Sheet1!D2:D551)
2. Standard Deviation function (for health expenditure): =STDEV.S(Sheet1!D2:D551)
3. Minimum function (for health expenditure): =MIN(Sheet1!D2:D551)
4. Maximum function (for health expenditure): =MAX(Sheet1!D2:D551)

	A	B	C	D	E
1	Variable	Mean	Std. Dev	Min	Max
2	Household Health Expenditure	70.05544	30.74593	14.67067	255.0213
3	Household Head Education				
4	Household Head Age				
5	Household Size				
6	Poverty Index				

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

Your Turn!

Use the data to calculate the mean, standard deviation, minimum, and maximum for the rest of the variables.

Slides, handouts, and data available at **https://bit.ly/fall2021-prina**

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

Final Results

Did anyone end up with different results?

	A	B	C	D	E
1	Variable	Mean	Std. Dev	Min	Max
2	Household Health Expenditure	70.05544	30.74593	14.67067	255.0213
3	Household Head Education	3.197671	2.553869	0	14
4	Household Head Age	46.04182	15.66615	21	87
5	Household Size	5.443636	2.274333	1	12
6	Poverty Index	772.3322	134.5378	452	1167

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

New Variables and Calculations

2. Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.

- (a) How many households in the treatment locality are eligible to be in the program?
- (b) What fraction of households in the treatment locality are eligible to be in the program?
- (c) Consider now the variable ‘takeup’: how many eligible household in the treatment locality participated in the program?
- (d) What fraction of households in the treatment locality participated in the program?

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

New Variables and Calculations

2. **Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.**

1. In column k, cell 1, name the variable “eligible”.
2. In K2, write the formula:
   a. =IF(l2<=750,1,0)
   b. =IF(l2<=750,1,0) is logical function saying if the cell in column l is less than or equal to 750, make the cell a 1, if not then, make it a zero.
3. Now you can copy the formula to every cell in the column by dragging the square in the bottom-right of the selection box, or by copying & pasting the formula into the rest of the cells in the column.

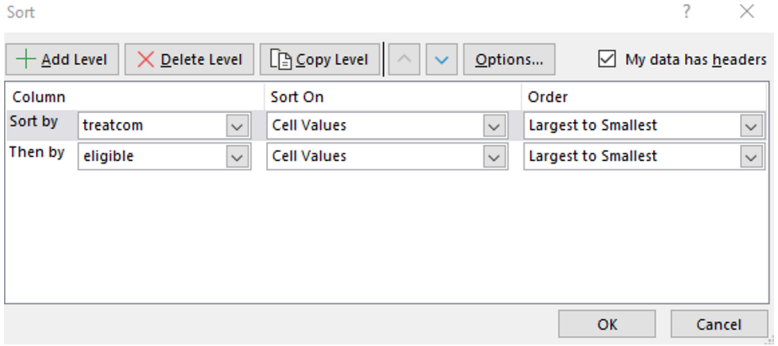
H	I	J	K
hhsize	pscore	takeup	eligible
3	865	0	0
3	865	0	0
7	602	1	1
7	602	1	1
5	653	1	1
5	653	1	1

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

New Variables and Calculations

2. Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.

- **(a) How many households in the treatment locality are eligible to be in the program?**
  - We can now calculate this by filtering our data by doing a custom sort (see screenshot below) and counting the 1's for the treatment community.
  - Alternatively, we can filter by the treatment group (treatcom = 1), selecting column K, and due to the simple math of 0's + 1's we can use the sum calculated and shown at the bottom of Excel.
  - Either way, we have: 120 observations / 2 = 60 households eligible.



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

New Variables and Calculations

2. Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.

- **(b) What fraction of households in the treatment locality are eligible to be in the program?**
  - Simple mathematics:
    - 60 households (120 observations) / 131 households (262 observations) * 100 = 45.8%

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

Your Turn!

Try to find the answers to c) and d) on your own.

2. Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.

- (a) How many households in the treatment locality are eligible to be in the program?
- (b) What fraction of households in the treatment locality are eligible to be in the program?
- (c) Consider now the variable 'takeup': how many eligible household in the treatment locality participated in the program?
- (d) What fraction of households in the treatment locality participated in the program?

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

New Variables and Calculations

2. Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.

- **(c) Consider now the variable 'takeup': how many eligible household in the treatment locality participated in the program?**
  - Looking at the 'takeup' variable, we can see that this is also just 0's and 1's, so while we are still filtering for the treatment community (treatcom = 1) we can look at the sum at the bottom of Excel.
  - So, 118 observations or 59 households in the treatment locality participated in the program.

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

New Variables and Calculations

2. Create a new variable called eligible, which is equal to one if the household has a poverty index lower or equal to 750.

- **(d) What fraction of households in the treatment locality participated in the program?**
  - Simple mathematics:
    - 59 participated (118 observations) / 60 total households (120 observations) * 100 = 98.3%

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

Treatment and Control Baseline Tables

3. Make a table in which you report the average:

- (a) household health expenditure (hhe);
- (b) level of education of the household head (educhh);
- (c) household size (hhsize);
- (d) poverty index (pscore)

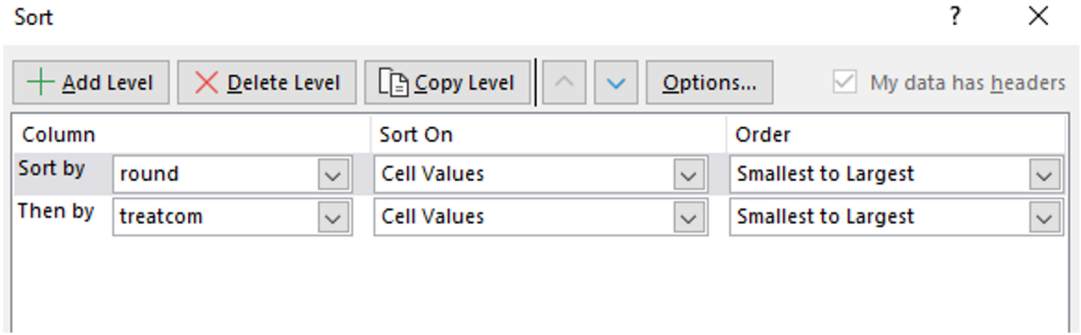
at baseline (round=0) for households in treatment and control localities separately.

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

Treatment and Control Baseline Tables

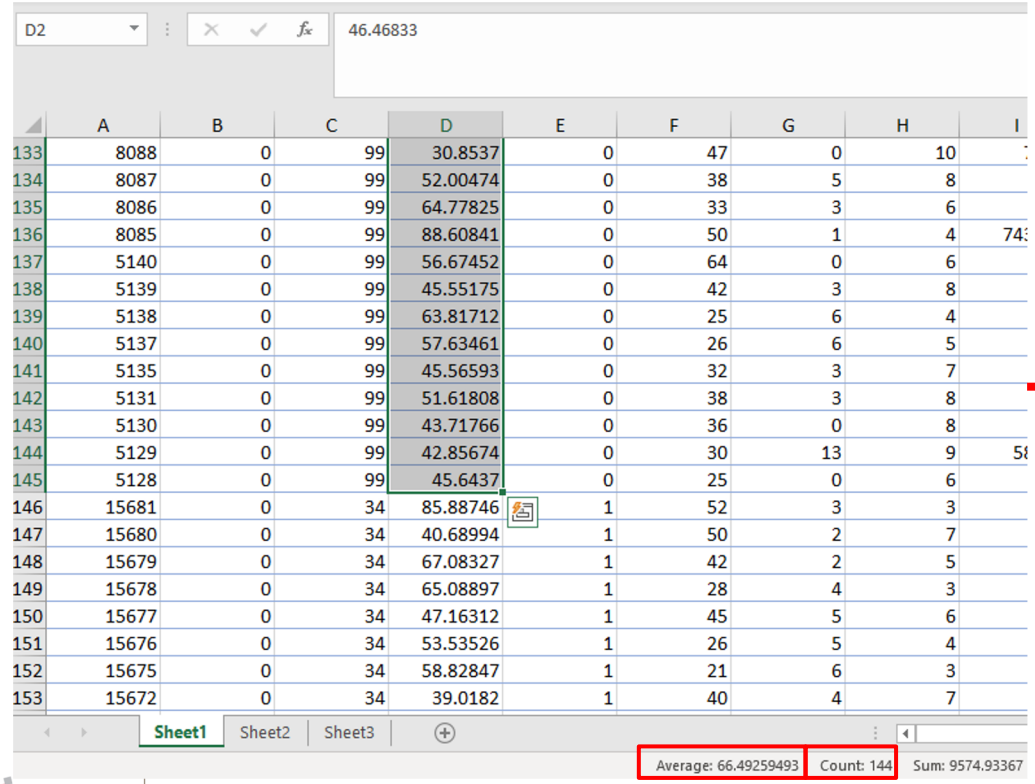
The easiest way to make this table is to create a new sheet with the headings: 'Variable,' 'Treatment,' and 'Control' as seen below. Then on sheet 1, use custom sort (see screenshot below) where we can then do quick calculations by selecting the observations in the columns, looking at the bottom toolbar of Excel, and typing the figures manually into our table.

	A	B	C
1	Variable	Treatment	Control
2	Household Health Expenditure		
3	Household Head Education		
4	Household Size		
5	Observations		



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

Results Example



D2 | 46.46833

	A	B	C	D	E	F	G	H	I
133	8088	0	99	30.8537	0	47	0	10	
134	8087	0	99	52.00474	0	38	5	8	
135	8086	0	99	64.77825	0	33	3	6	
136	8085	0	99	88.60841	0	50	1	4	74
137	5140	0	99	56.67452	0	64	0	6	
138	5139	0	99	45.55175	0	42	3	8	
139	5138	0	99	63.81712	0	25	6	4	
140	5137	0	99	57.63461	0	26	6	5	
141	5135	0	99	45.56593	0	32	3	7	
142	5131	0	99	51.61808	0	38	3	8	
143	5130	0	99	43.71766	0	36	0	8	
144	5129	0	99	42.85674	0	30	13	9	5
145	5128	0	99	45.6437	0	25	0	6	
146	15681	0	34	85.88746	1	52	3	3	
147	15680	0	34	40.68994	1	50	2	7	
148	15679	0	34	67.08327	1	42	2	5	
149	15678	0	34	65.08897	1	28	4	3	
150	15677	0	34	47.16312	1	45	5	6	
151	15676	0	34	53.53526	1	26	5	4	
152	15675	0	34	58.82847	1	21	6	3	
153	15672	0	34	39.0182	1	40	4	7	

Sheet1 | Sheet2 | Sheet3

Average: 66.49259493 | Count: 144 | Sum: 9574.93367

Variable	Treatment	Control
Household Health Expenditure	71,5	66,49
Household Head Eduction	3,44	2,98
Household Size	4,93	5,9
Poverty Index	782,27	763,3
Observations	131	144

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

Treatment and Control Baseline Tables

4. Make a table in which you report the average:

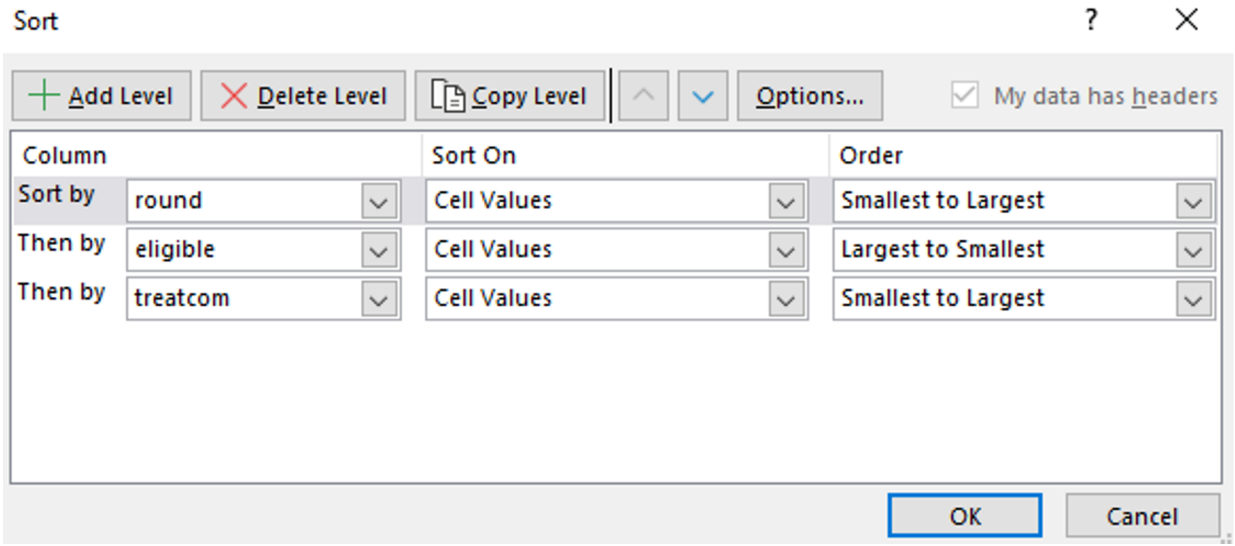
- (a) household health expenditure (hhe);
- (b) level of education of the household head (educhh);
- (c) household size (hhsize);
- (d) poverty index (pscore)

at baseline (round=0) for **eligible** households in treatment and control localities separately.

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

Treatment and Control Baseline Tables

This table can be created in similar methods to before, please refer to the custom sort image below (the table will look the same as in question 3).



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

Your Turn!

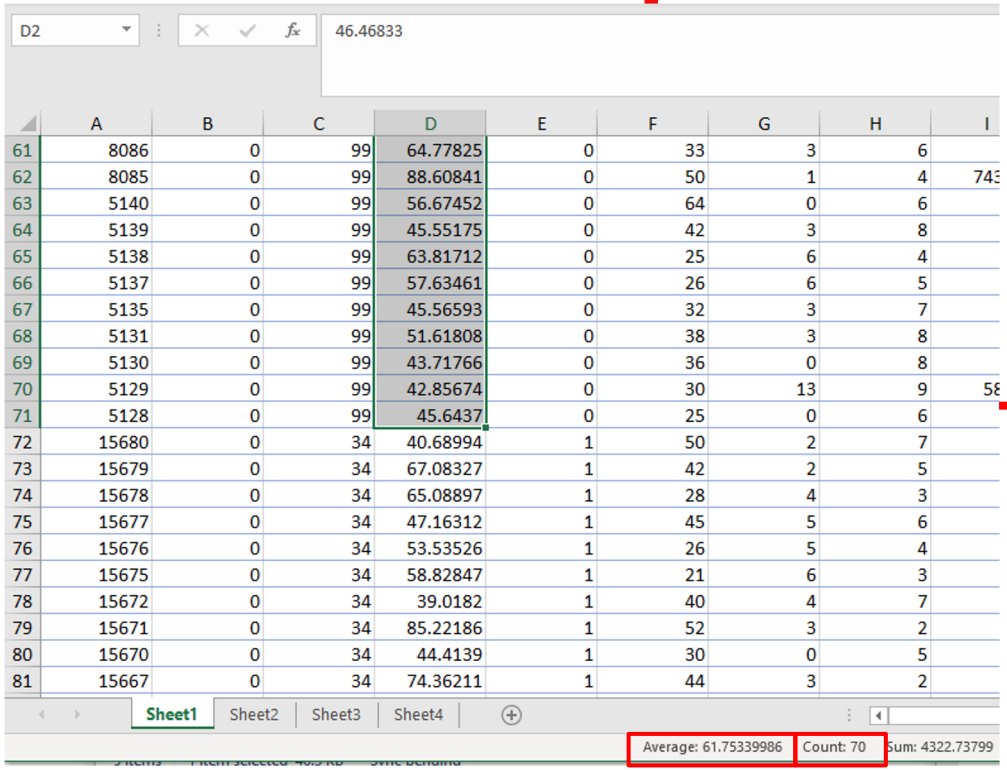
4. Make a table in which you report the average:

- (a) household health expenditure (hhe);
- (b) level of education of the household head (educhh);
- (c) household size (hhsize);
- (d) poverty index (pscore)

at baseline (round=0) for **eligible** households in treatment and control localities separately.

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

Results Example



D2 | 46.46833

	A	B	C	D	E	F	G	H	I
61	8086	0	99	64.77825	0	33	3	6	
62	8085	0	99	88.60841	0	50	1	4	74
63	5140	0	99	56.67452	0	64	0	6	
64	5139	0	99	45.55175	0	42	3	8	
65	5138	0	99	63.81712	0	25	6	4	
66	5137	0	99	57.63461	0	26	6	5	
67	5135	0	99	45.56593	0	32	3	7	
68	5131	0	99	51.61808	0	38	3	8	
69	5130	0	99	43.71766	0	36	0	8	
70	5129	0	99	42.85674	0	30	13	9	58
71	5128	0	99	45.6437	0	25	0	6	
72	15680	0	34	40.68994	1	50	2	7	
73	15679	0	34	67.08327	1	42	2	5	
74	15678	0	34	65.08897	1	28	4	3	
75	15677	0	34	47.16312	1	45	5	6	
76	15676	0	34	53.53526	1	26	5	4	
77	15675	0	34	58.82847	1	21	6	3	
78	15672	0	34	39.0182	1	40	4	7	
79	15671	0	34	85.22186	1	52	3	2	
80	15670	0	34	44.4139	1	30	0	5	
81	15667	0	34	74.36211	1	44	3	2	

Sheet1 | Sheet2 | Sheet3 | Sheet4

Average: 61.75339986 | Count: 70 | Sum: 4322.73799

Variable	Treatment	Control (99)
Household Health Expenditure	60,12	61,75
Household Head Eduction	4,27	3,47
Household Size	5,57	5,86
Poverty Index	672,12	659,61
Observations	60	70

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

Scatterplots

5. Make a scatter plot showing the relationship between household health expenditure at baseline and the household poverty index.

- Paste the image (of the scatter plot) into your problem set.

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

Scatterplots

5. Make a scatter plot showing the relationship between household health expenditure at baseline and the household poverty index.

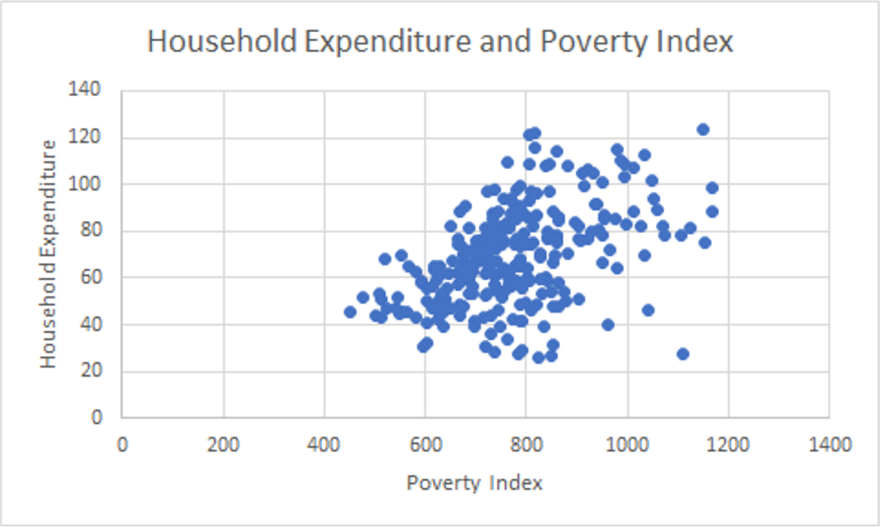
1. The first step is to clear any filters and sorting we have, and then filter by 'round' to make it only the baseline time (round = 0).
2. Once filtered, select columns D (hhe) and I (pscore).
3. To make a scatterplot, go to 'insert' and then under charts, click on this button:  and select the option in the top left:
4. Then, with the scatter plot  selected, click on the filter icon (the funnel), click on 'edit' on the lefthand side, and then insert the following values:
    a. Series name: =Sheet1!$D$1
    b. Series X Values: =Sheet1!$I$2:$I$276
    c. Series Y Values: =Sheet1!$D$2:$D$276
5. Note that this path is different across different versions (eg. right click on scatter plot, and then click 'Select Data')
6. Note that we use I276 and D276 for the max limits of our formula because we have 550 / 2 households, one observation at the baseline (round = 0) and one at the endline (round = 1).

Scatterplots

5. Make a scatter plot showing the relationship between household health expenditure at baseline and the household poverty index.

- **(a) Paste the image (of the scatter plot) into your problem set.**

1. We can also add axis titles and edit them by clicking on the '+' button while selecting our scatterplot.

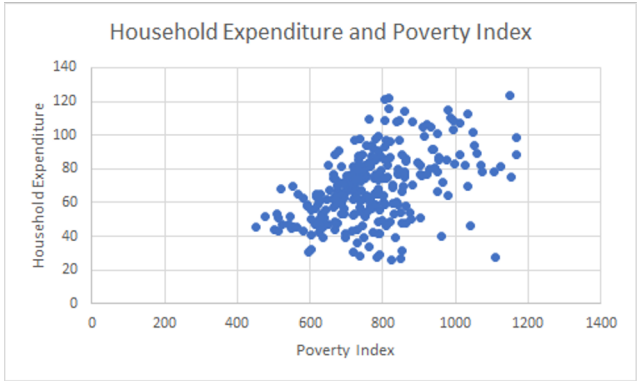


Scatterplots

5. Make a scatter plot showing the relationship between household health expenditure at baseline and the household poverty index.

- **Is the association positive or negative? In other words, does household health expenditure increase or decrease as the poverty index increases? Remember that the lower the poverty index, the poorer the household.**

1. Referring to our scatterplot, we can see that the association between household expenditure and poverty index is positive (the dots generally go up and to the right). So poorer households spend less in health expenditure.



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

Correlation Coefficients

6. Calculate the correlation between household health expenditure and education of the household head at baseline. Is the correlation positive or negative?

1. First we will filter by baseline (round = 0) if we are not already doing so. Again, remember that we now have 275 rows (550 / 2).
2. Then in a clear cell (M2 for instance) we will write the following function:
    a. =CORREL(D2:D276, G2:G276)
        i. Which gives us: -0.099, so the two variables are weakly negatively correlated.
3. We can also calculate the correlation between the correlation between household expenditure and poverty index:
    a. =CORREL(D2:D276, I2:I276)
        i. Which gives us: 0.4691, so the two variables are positively correlated.

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

Bivariate Regression

7. Estimate the program's impact by regressing the health expenditure (hhe) variable on the treatcom variable using data from the year in which the intervention took place (round=1). Paste your Excel results into your write-up of the answers.

- (a) How large is the estimated impact of the program?

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

More Advanced Calculations - LINEST

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

y = a + bx1…bxn

- y = expected value
- a = intercept
- bx1…bxn = beta-coefficient (b) * value (x)

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

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

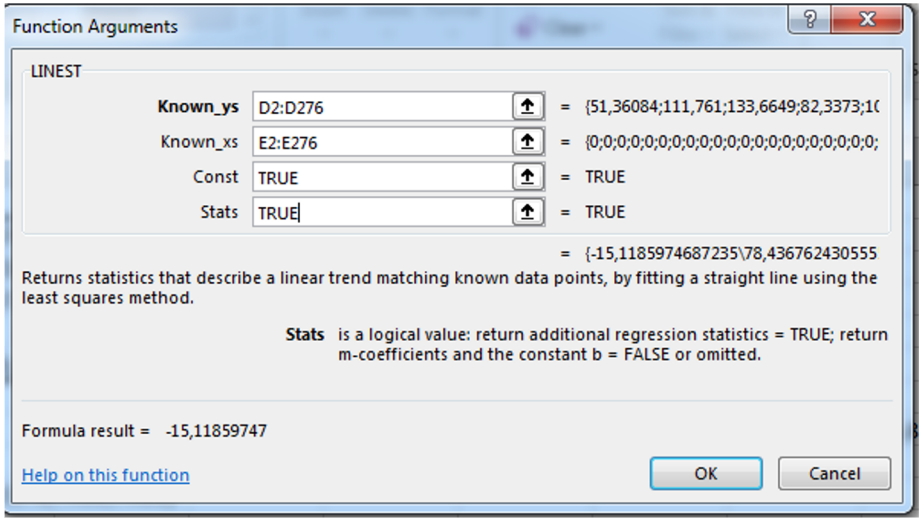
“hhe” and variable “treatcom” where

“round” = 1

LINEST Steps (after sorting for round)

1. Select multiple rows + columns (2x2)
2. =Linest(D2:D276, E2:E276, TRUE, TRUE)
3. =-15.12, 78.44 (Constant)

   Note that the output might be different in different versions of excel



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

Bivariate Regression Example

A	B	C	D	E	F	G	H	I	J	K	L	M	N
hhid	round	local	hhe	treatcom	agehh	educhh	hhsize	pscore	takeup	eligible			
8229	1	99	51,36084	0	31	0	6	639	0	1		-15,118597	78,4367624
8228	1	99	111,761	0	43	2	7	837,331	0	0		4,55631383	3,14472686
8227	1	99	133,6649	0	66	0	4	818	0	0		0,03876696	37,7367223
8225	1	99	82,3373	0	62	0	8	875	0	0		11,0102137	273
8224	1	99	102,7652	0	47	0	3	735	0	1		15679,2073	388768,438
8222	1	99	34,41784	0	30	6	5	553	0	1			
8221	1	99	88,74641	0	40	3	6	518,331	0	1			
8220	1	99	158,1178	0	25	6	4	750	0	1			

Note that the second row from above are the standard errors.

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

Bivariate Regression

7. Estimate of the program's impact by regressing the health expenditure (hhe) variable on the treatcom variable using data from the year in which the intervention took place (round=1). Paste your Excel results into your write-up of the answers.

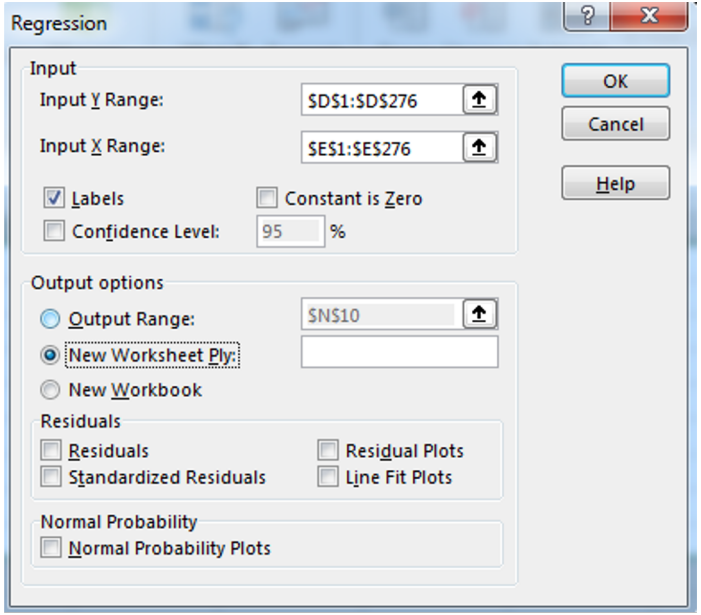
- **(a) How large is the estimated impact of the program?**

1. With our results of -15.12, that means that after the treatment, the households in the treatment locality (1) spend approximately 15 dollars less than households in the control locality.

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

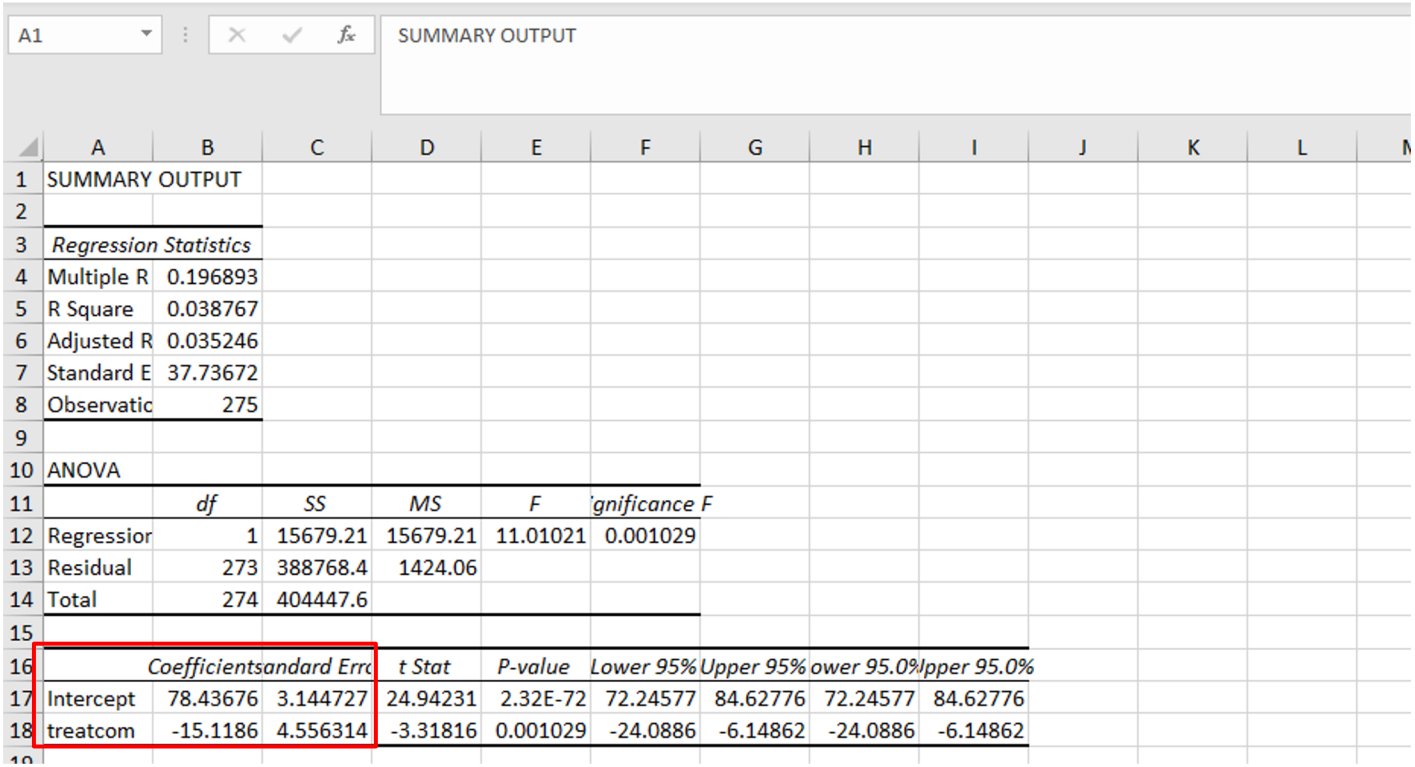
Regression with Analysis Toolpak

- Use the “Analysis ToolPak” Add-in
  - Data → Data Analysis → Regression
- The input data is similar to the LINEST method of regression:
  - Y range: D1:D276
  - X range: E1:E276
- Also enable ‘Labels’ (D1 & E1)
- Click ‘OK’



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

Analysis Toolpak Results



A1 | SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SUMMARY OUTPUT												
2													
3	*Regression Statistics*												
4	Multiple R	0.196893											
5	R Square	0.038767											
6	Adjusted R	0.035246											
7	Standard E	37.73672											
8	Observatic	275											
9													
10	ANOVA												
11		*df*	*SS*	*MS*	*F*	*gnificance F*							
12	Regressior	1	15679.21	15679.21	11.01021	0.001029							
13	Residual	273	388768.4	1424.06									
14	Total	274	404447.6										
15													
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*				
17	Intercept	78.43676	3.144727	24.94231	2.32E-72	72.24577	84.62776	72.24577	84.62776				
18	treatcom	-15.1186	4.556314	-3.31816	0.001029	-24.0886	-6.14862	-24.0886	-6.14862				

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

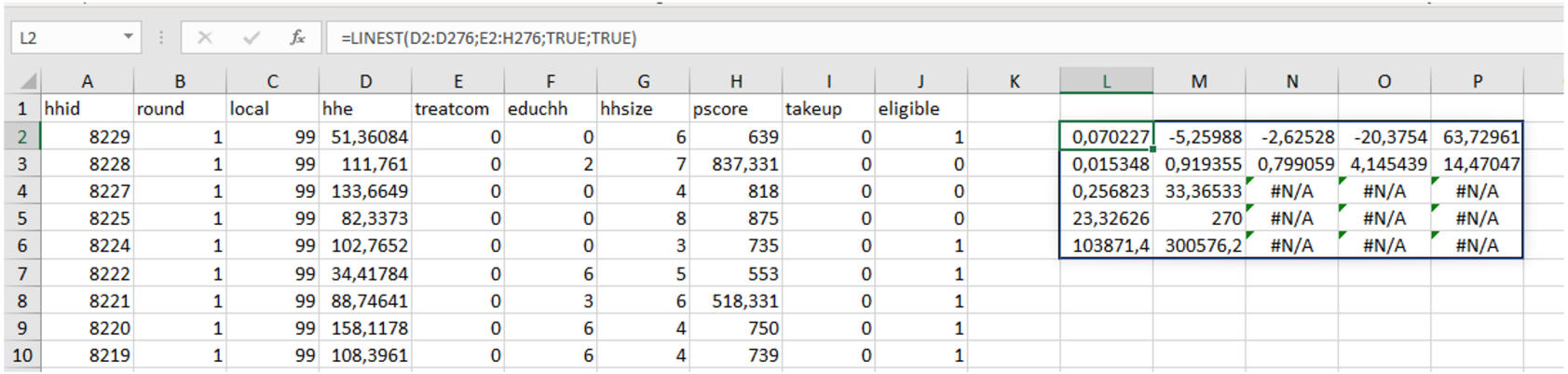
Multivariate Regression

8. Estimate of the program's impact by regressing the health expenditure (hhe) variable on the treatcom variable using data from the year in which the intervention took place (round=1) AND controlling for the level of education of the household head (educhh), the household size (hhsize), and the poverty index (pscore). Paste your Excel results into your write-up of the answers.

1. What is the relationship between “hhe” and “treatcom” while controlling for “hhsize”, “educhh”, and “pscore”
2. Be aware that the variables need to be next to each other, but “agehh” is in the way, so we will copy our data to a new sheet without that column.
3. Similar syntax: =LINEST(D2:D276, E2:H276, TRUE, TRUE)
4. Select rows & columns - you need 1 more column than the number of variables because of the constant - so we need 5x2
5. Note that the output of this is different from version to version

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

Multivariate Regression Example



L2 =LINEST(D2:D276;E2:H276;TRUE;TRUE)

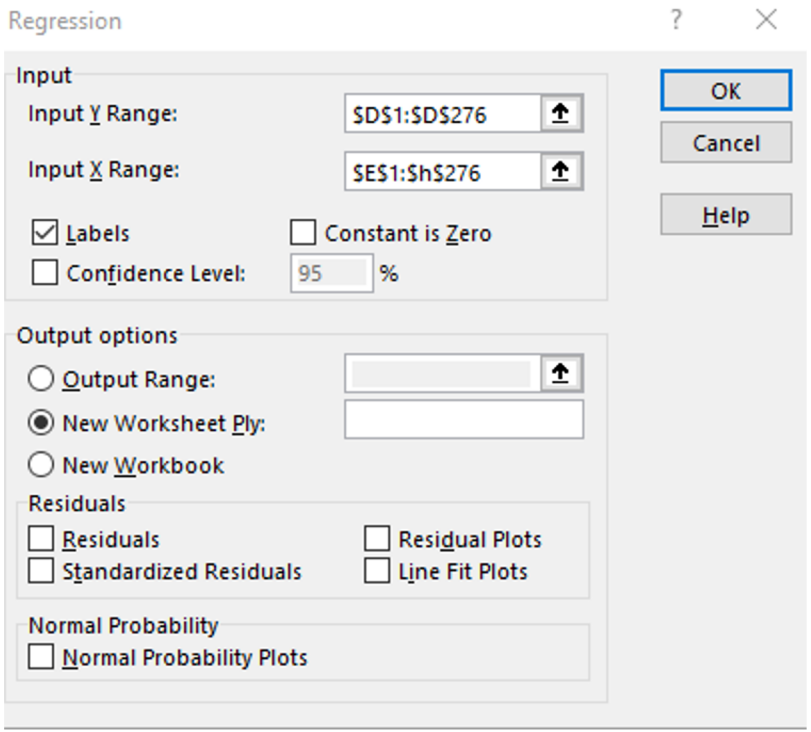
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	hhid	round	local	hhe	treatcom	educhh	hhsize	pscore	takeup	eligible						
2	8229	1	99	51,36084	0	0	6	639	0	1		0,070227	-5,25988	-2,62528	-20,3754	63,72961
3	8228	1	99	111,761	0	2	7	837,331	0	0		0,015348	0,919355	0,799059	4,145439	14,47047
4	8227	1	99	133,6649	0	0	4	818	0	0		0,256823	33,36533	#N/A	#N/A	#N/A
5	8225	1	99	82,3373	0	0	8	875	0	0		23,32626	270	#N/A	#N/A	#N/A
6	8224	1	99	102,7652	0	0	3	735	0	1		103871,4	300576,2	#N/A	#N/A	#N/A
7	8222	1	99	34,41784	0	6	5	553	0	1						
8	8221	1	99	88,74641	0	3	6	518,331	0	1						
9	8220	1	99	158,1178	0	6	4	750	0	1						
10	8219	1	99	108,3961	0	6	4	739	0	1						

So what do the results mean?

- While controlling for household head education (-2.63), household size (-5.26), and poverty index (.07), after the treatment, those in the treatment locality spend approximately 20 dollars less on health care expenditure

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

Add-In Example



K20

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.506777							
5	R Square	0.256823							
6	Adjusted R	0.245813							
7	Standard E	33.36533							
8	Observatic	275							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressior	4	103871.4	25967.85	23.32626	1.41E-16			
13	Residual	270	300576.2	1113.245					
14	Total	274	404447.6						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	63.72961	14.47047	4.404116	1.53E-05	35.24032	92.21891	35.24032	92.21891
18	treatcom	-20.3754	4.145439	-4.91514	1.54E-06	-28.5369	-12.2139	-28.5369	-12.2139
19	educhh	-2.62528	0.799059	-3.28546	0.001153	-4.19846	-1.0521	-4.19846	-1.0521
20	hhsize	-5.25988	0.919355	-5.72127	2.8E-08	-7.06989	-3.44986	-7.06989	-3.44986
21	pscore	0.070227	0.015348	4.575712	7.24E-06	0.040011	0.100444	0.040011	0.100444

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

Thank you!

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

Taught by Vaishali Kushwaha
Digital Integration Teaching Initiative
DITI Research Fellow

Slides, handouts, and data available at **https://bit.ly/fall2021-prina**

Schedule an appointment with us! **https://calendly.com/diti-nu**

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

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?

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