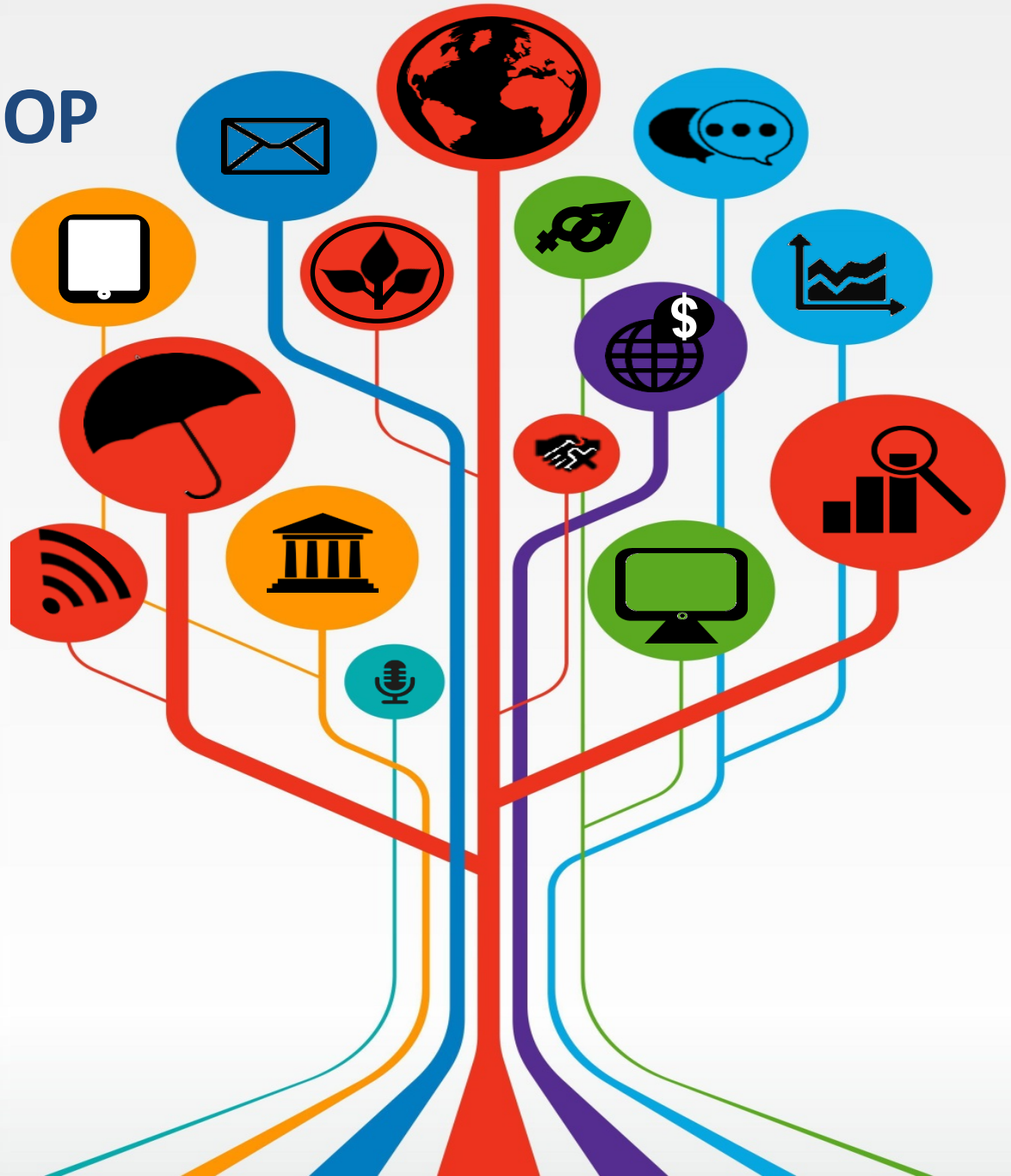
FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC



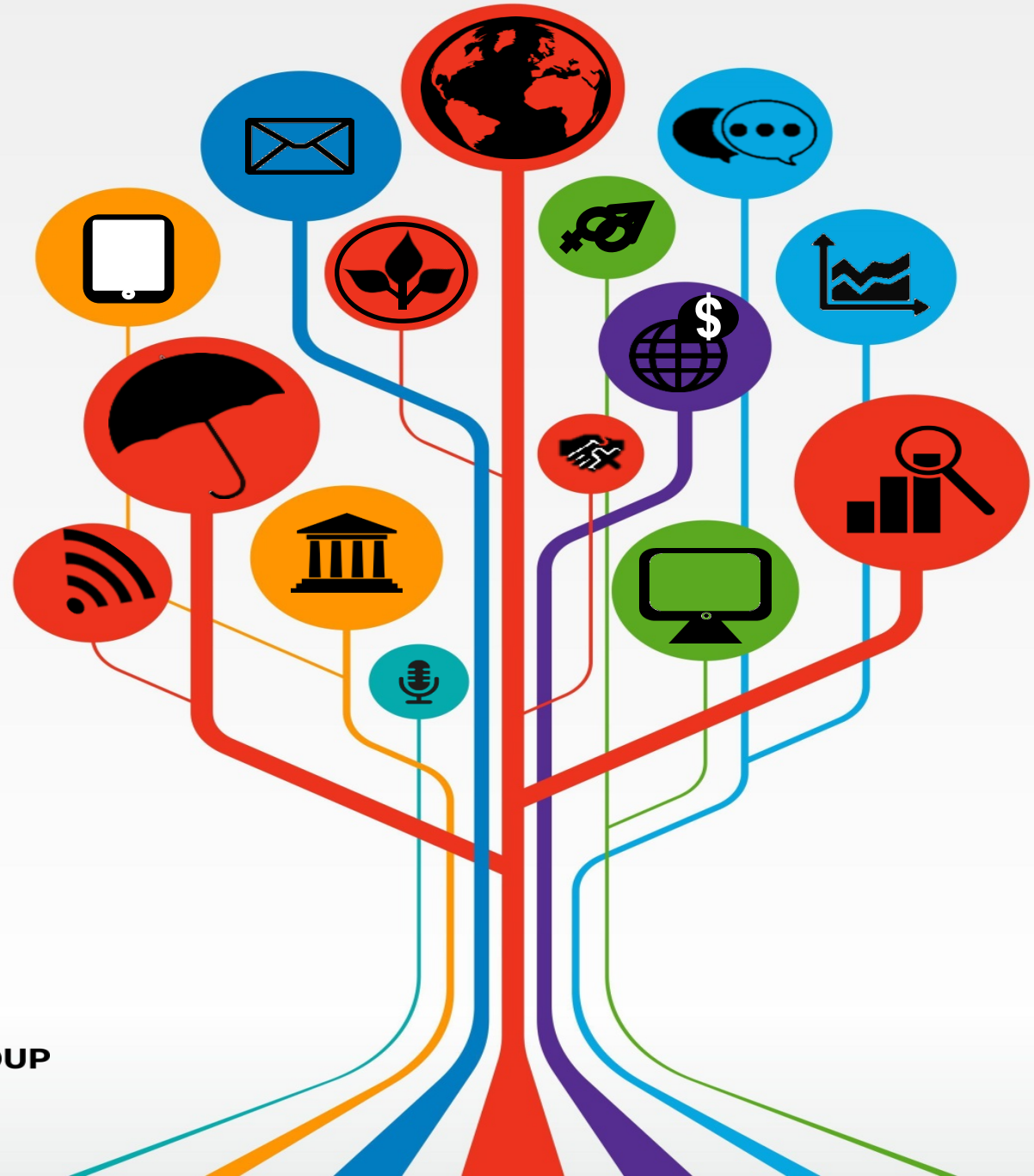
i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Descriptive Statistics: Creating Tables

Mrijan Rimal & Kristoffer Bjarkefur

June 21, 2018

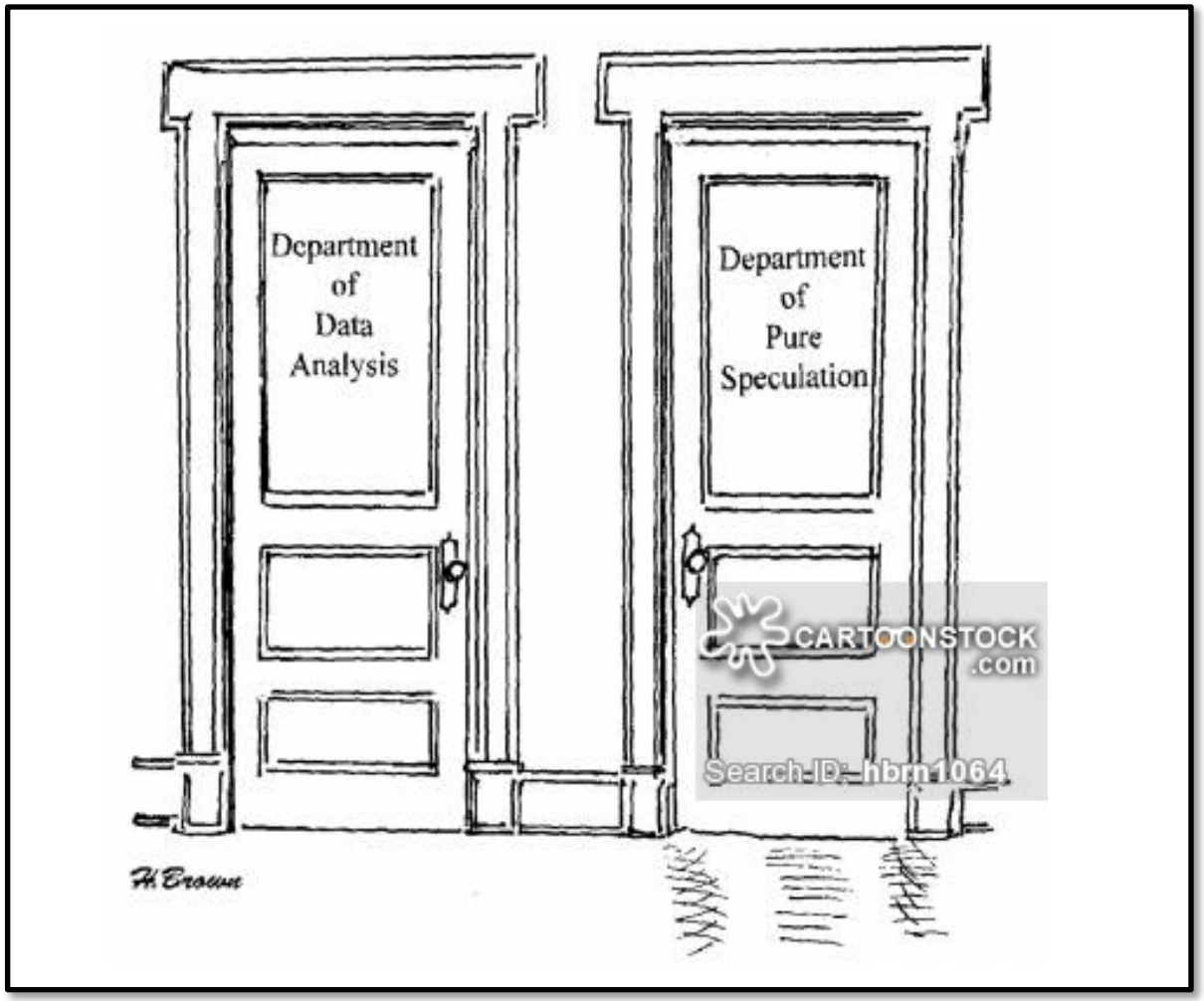


i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

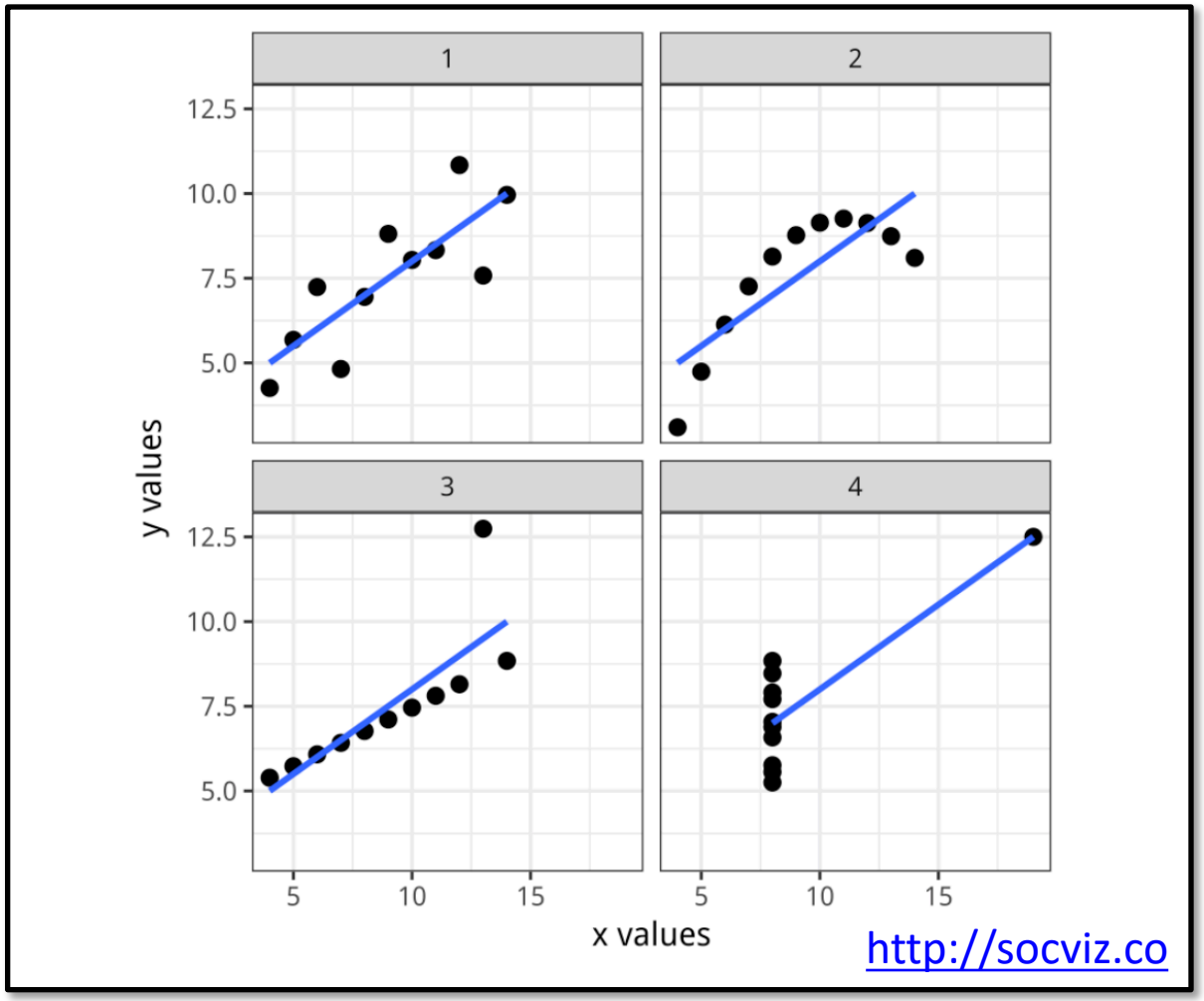
What are descriptive statistics?

- Numbers or figures that paint a picture of what a given dataset looks like
- They begin to help us understand the important features of our dataset, and can be useful in directing us towards areas of further analysis
- We will also show how to make basic regression outputs

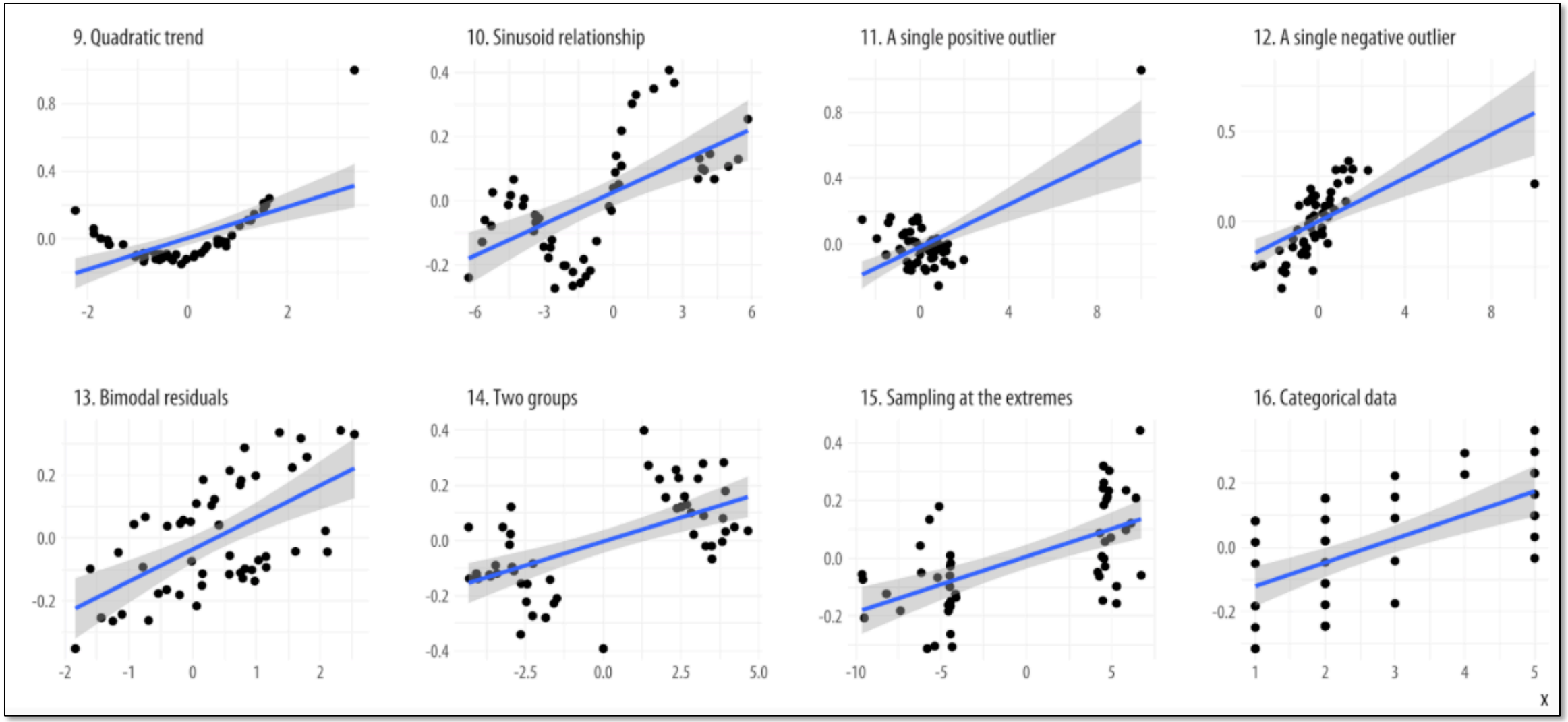


Descriptive statistics are NOT regressions

- This is “Anscombe’s Quartet”
- Every set here has:
  - The same means (x and y)
  - The same variances (x and y)
  - The same correlation coefficient
  - The same regression coefficient
  - The same regression R^2
- Regression analysis tells you *nothing* about the world if you don’t understand the shape of the world you are in!

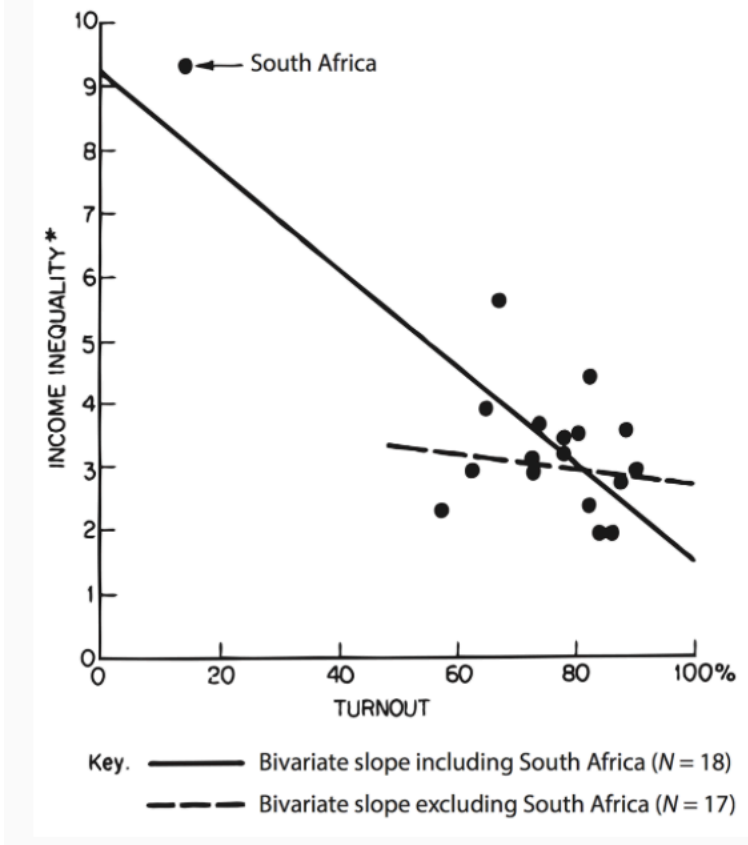


Data can take almost any shape



This really matters for impact analysis

- In this case, for example, simply running a regression on the data would give a very wrong impression about the strength of the relationship
- And real data has many more than two dimensions!



Task 1: Summarize and Tabulate

- These two commands are often used to get a quick idea of the data
- Summarize all the income variables, inc_01, inc_02, … inc_17
  - Do you see anything that stands out?
  - Use option detail for inc_02
- Use tabulate to see the distribution of categorical variables
  - Gender of the first person listed in the HH roster
  - tab pl_sex_1

Share what we learn about the data

- **summarize** and **tabulate** are great for your quick understanding, but you need a way to share that understanding with the rest of your team

- **summarize** and **tabulate** do not provide a great option for this, and that is why we want to introduce some more advanced tools for that

Three most common types of tables

- **Summary statistics**
  - Show an overview of variable distributions, possible for multiple groups
- **Balance tests**
  - Show a direct comparison of variable means across treatment arms

[*sumStats*]

https://github.com/worldbank/stata/tree/master/src/sumStats

[*iebaltab*]

https://worldbank.github.io/ietoolkit

Summary statistics with [*sumStats*]

- [*sumStats*] is a command that will output anything you can get from [*tabstat*]

statname	Definition	*statname*	Definition
mean	mean	p1	1st percentile
count	count of nonmissing observations	p5	5th percentile
n	same as count	p10	10th percentile
sum	sum	p25	25th percentile
max	maximum	median	median (same as p50)
min	minimum	p50	50th percentile (same as median)
range	range = max − min	p75	75th percentile
sd	standard deviation	p90	90th percentile
variance	variance	p95	95th percentile
cv	coefficient of variation (sd/mean)	p99	99th percentile
semean	standard error of mean (sd/$\sqrt{n}$)	iqr	interquartile range = p75 − p25
skewness	skewness	q	equivalent to specifying p25 p50 p75
kurtosis	kurtosis		

- It also allows multiple [*if*]-restrictions with different variable lists

sumStats

sumStats will produce requested statistics for any number and combination of variables and sample restrictions.

	A	B	C	D	E	F
1		mean	sd	p5	p95	N
2	Price	6,072.423	3,097.104	3,667.000	13,594.000	52.000
3	Mileage (mpg)	19.827	4.743	14.000	29.000	52.000
4	Repair Record 1978	3.021	0.838	2.000	4.000	48.000
5	Headroom (in.)	3.154	0.916	1.500	4.500	52.000
6	Trunk space (cu. ft.)	14.750	4.306	7.000	21.000	52.000
7	Price	6,384.682	2,621.915	3,798.000	11,995.000	22.000
8	Mileage (mpg)	24.773	6.611	17.000	35.000	22.000
9	Repair Record 1978	4.286	0.717	3.000	5.000	21.000
10	Headroom (in.)	2.614	0.486	2.000	3.500	22.000
11	Trunk space (cu. ft.)	11.409	3.217	6.000	16.000	22.000

```
wb_git_install sumStats
sysuse auto , clear
sumStats ///
  (price mpg rep78 headroom trunk if foreign == 0) ///
  (price mpg rep78 headroom trunk if foreign == 1) ///
  using "table_1.xls" ///
  , replace stats(mean sd p5 p95 N)
```

Task 2a

1. Run the code as it is in the 2a section and open the file generated. (You need to include the sections where the locals are defined)
   - Only the mean is in the table
2. Add the two outcome locals to the varlist
3. Add more statistics to the table. Type *help sumStats* and click the stats_list link to see all options you can use
   - Add, for example, mean, number of observation, standard deviation, median, max and min

Multiple levels or groups are easy

- Village statistics can be called with [*if tag_village == 1*]
- Treatment group can be called with [*if hh_treatment == 1*]
- And so on, with only one line of code in Stata

Table 1. Descriptive Statistics

	mean	sd	p25	p50	p75	N
Distance to Fault (km)	17.477	14.149	5.555	13.564	24.311	28,297.000
Distance to Epicenter (km)	36.373	17.490	25.115	35.161	48.013	28,297.000
Closest Faultline (km)	2.799	2.486	0.773	1.984	4.143	28,297.000
Death in HH During Quake	0.061	0.240	0.000	0.000	0.000	28,297.000
Home Destroyed	0.572	0.495	0.000	1.000	1.000	8,351.000
Home Damaged or Destroyed	0.911	0.285	1.000	1.000	1.000	8,350.000
Household Size	5.477	2.332	4.000	5.000	7.000	28,297.000
Total Annual Food Expenditu	83,207.844	88,160.971	37,500.000	62,280.000	98,805.000	2,456.000
Total Annual Nonfood Expen	84,207.025	109,511.091	26,786.500	46,182.500	93,035.000	2,456.000
Abbotabad	0.206	0.405	0.000	0.000	0.000	2,456.000
Bagh	0.175	0.380	0.000	0.000	0.000	2,456.000
Mansehra	0.276	0.447	0.000	0.000	1.000	2,456.000
Muzaffarabad	0.342	0.475	0.000	0.000	1.000	2,456.000
Family Size	6.123	2.689	4.000	6.000	8.000	2,455.000
Asset Index (PCA) (Pre-Quak	0.002	0.999	-0.551	-0.093	0.568	2,456.000
House Destroyed in Quake?	0.599	0.490	0.000	1.000	1.000	2,455.000
Eligible for death compensatic	0.149	0.433	0.000	0.000	0.000	2,455.000
Eligible for housing compensa	0.925	0.320	1.000	1.000	1.000	2,455.000
Eligible for injury compensati	0.156	0.438	0.000	0.000	0.000	2,455.000
Eligible for lcgs compensation	0.886	0.710	0.000	1.000	1.000	2,455.000
N children under 6 at EQ	0.971	1.132	0.000	1.000	2.000	2,456.000
Female Head of HH	0.100	0.300	0.000	0.000	0.000	2,456.000
Aid	0.668	0.471	0.000	1.000	1.000	2,456.000
Cash Aid	0.467	0.499	0.000	0.000	1.000	2,456.000
Aid Amount	121,130.068	105,991.244	0.000	143,000.000	175,000.000	1,899.000
Mean Slope of UC	21.143	6.690	16.893	22.150	26.140	98.000
Male	0.523	0.499	0.000	1.000	1.000	152,435.000
Age	23.984	18.351	10.000	20.000	35.000	152,435.000
In Utero - Age 11 During Eartl	0.328	0.470	0.000	0.000	1.000	152,435.000
In Utero	0.090	0.287	0.000	0.000	0.000	4,665.000
Age 0-2	0.257	0.437	0.000	0.000	1.000	4,665.000
Age 3+	0.653	0.476	0.000	1.000	1.000	4,665.000
Father Completed Primary Scl	0.573	0.495	0.000	1.000	1.000	4,379.000
Mother Completed Primary Sc	0.222	0.416	0.000	0.000	0.000	4,387.000
Mother's Age	37.425	8.433	31.000	37.000	42.000	4,387.000
Mother's Height (cm)	157.238	7.820	152.000	157.000	162.000	4,239.000
Name's height (in cm)?	117.460	22.330	101.000	119.000	132.000	4,096.000
Name's weight (in kg)?	25.631	9.319	18.000	24.000	31.000	4,097.000
Enrolled During Survey (Age	0.861	0.346	1.000	1.000	1.000	3,589.000
Private School Binary (Post-Q	0.217	0.412	0.000	0.000	0.000	3,089.000

Task 2b

1. In the Task 2b section, start by adding the outcome variables, and the additional statistics you added to the table in Task 2a
2. Restrict the first part of the table to control observations, and the second part of the table to treatment observations
3. Run the section for Task 2b and open the table. (Remember to include the section where the locals are defined)

Balance tables with [*iebaltab*]

- Balance tables feature in almost every impact evaluation
- We use balance tables to show that there was no difference between our control and treatment group in the baseline before the intervention
- To us [*iebaltab*], list all the variables you want to test balance in, and use the option [*grpvar*()] to indicate which group each observation is in.

Variable	(1) Control Mean/SE	(2) Treatment Mean/SE	T-test Difference (1)-(2)
Age in years	42.880	42.126	0.754
	(1.746)	(0.535)	
Respondent is male	0.538	0.479	0.059
	(0.050)	(0.008)	
Years of schooling	10.930	10.838	0.092
	(0.171)	(0.183)	
Respondent is employed	0.835	0.892	-0.057
	(0.060)	(0.041)	
Monthly earnings (number of minimum wages)	1.582	1.491	0.091
	(0.094)	(0.067)	
Average commuting distance	18.241	11.737	6.504***
	(1.078)	(0.233)	
N	158	167	
Clusters	6	6	
F-test of joint significance (F-stat)			9.892***

```
iebaltab    age d_male educ d_employed earnings distance, ///
            covariates(stratum) ///
            grpvar(tmt_status) ///
            vce(cluster neighborhood) ///
            savetex("$outputs/balance_table") ///
            replace onenrow ftest rowvarlabel
```

Task 3

1. Run the first iebaltab section
   – Click the link in the result window to see the table generated
2. Run the second iebaltab section, and see what is different
3. Add all the income variables, inc_01, inc_02 … inc_17, and re-run the first two graphs
4. Run the third iebaltab section
5. Write some manual labels for one or a few of the income variables you added, and re-run the last iebaltab section again

Helpful checklist before sending tables to PI

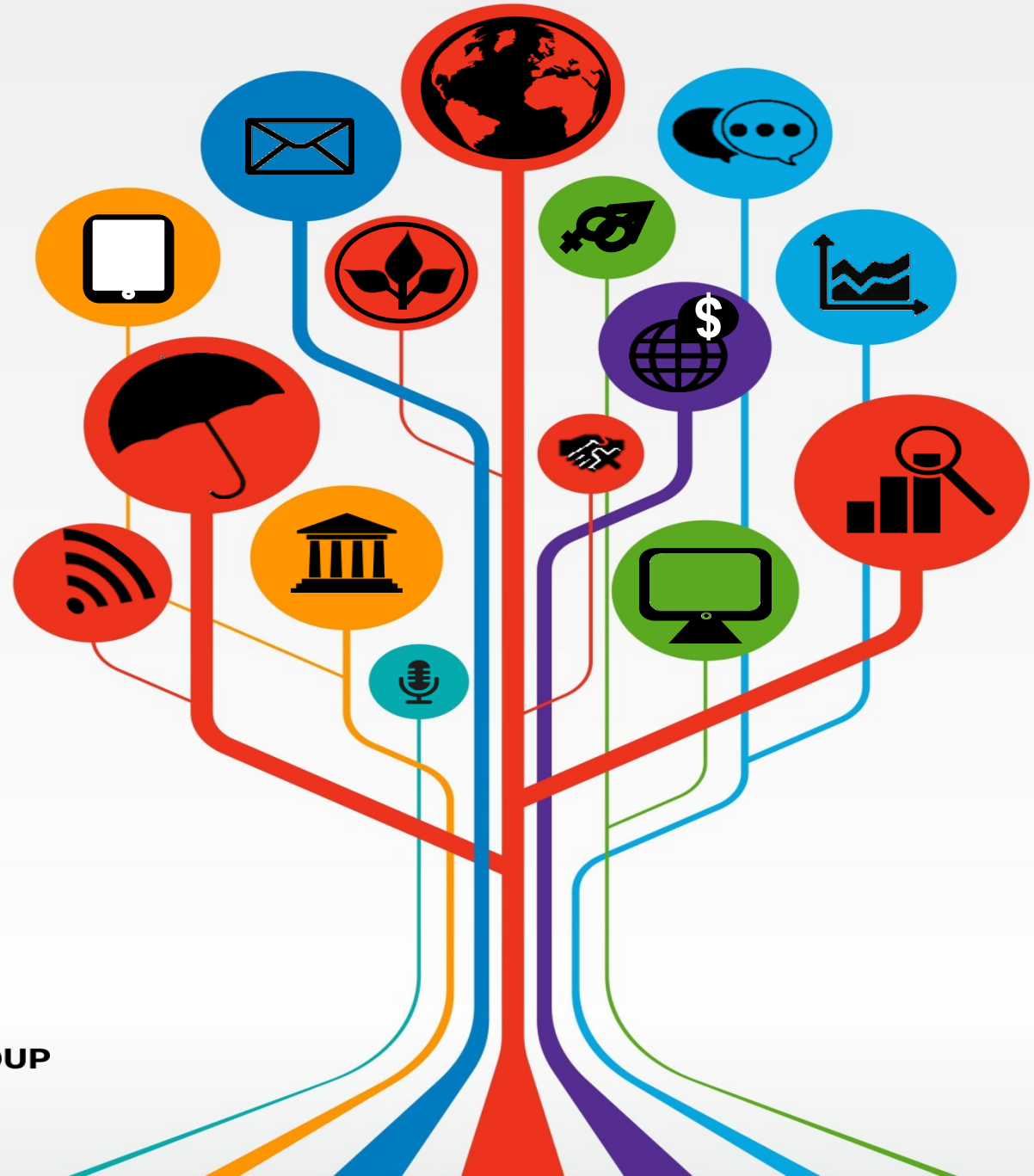
- Does the number of observations for each regression or summary statistic make sense?
- Do the magnitude and sign of each coefficient/summary statistic seem reasonable?
- Did you delete the constant term and add the control mean in the regression table?
- Did you check for joint significance of your covariates?
- Did you label the dependent variables/columns?
- Did you label the covariates/rows?
- Did you add a title?
- Is it clear what the estimation procedure is (e.g. regression vs. probit)?
- Are the column widths the right size so as not to cut off text?
- Is the bordering consistent with your other tables?
- Are the numbers rounded to an appropriate level, so you don't display too many decimal places?
- Do the notes to the table clearly indicate how standard errors have been estimated, and what control variables if any have been included but not shown?

https://dimewiki.worldbank.org/wiki/Checklist:_Submit_Table
https://blogs.worldbank.org/impactevaluations/generating-regression-and-summary-statistics-tables-stata-checklist-and-code

Thank you!

Mrijan Rimal & Kristoffer Bjarkefur

June 21, 2018



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP