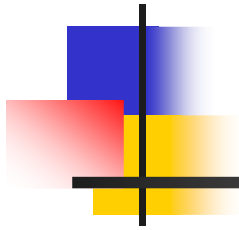


PUB – POS 316
Week 4-1



Two-way tables

Navid Ghaffarzadegan

navidg@gmail.com

Last updated - sep 24, 09



Agenda

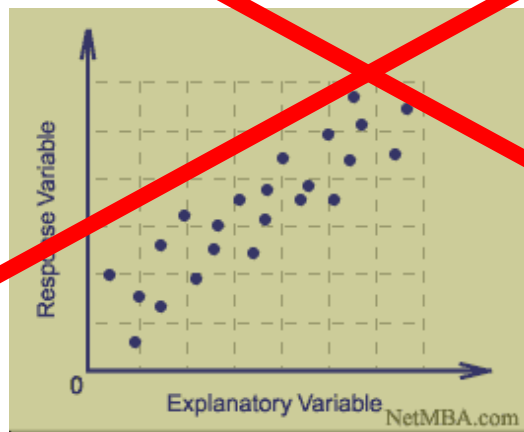
Two-way Table

- Data in categories
 - Graphs
 - Simpson Paradox
-
- It will be easiest to do these two-way table analyses using Excel (or some other spreadsheet) rather than JMP.]

Data in categories

What is "data in categories?"

- What is not?
 - Two quantitative variables → Scatter plot
 - SATm vs. SATv
 - SATv vs. Percent of students that take the test



Data in categories

- What is “data in categories?”

- Two Variables are categorical:

- X: Men or Women

- Y: Yes or No

	Gender	
Frequent of binge drinker	Men	Women
Yes	1630	1684
No	5550	8232

- How should we analyze this data?
- Joint Distribution, Conditional Distribution (look at Columns);
- e.g. conditional distribution of binge drinking for women

Data in categories

Frequent of binge drinker	Gender		total
	Men	Women	
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

Joint Distribution

Frequent of binge drinker	Gender	
	Men	Women
Yes	0.095344	0.0985026
No	0.324637	0.4815161

e.g. "Men-Yes" cell,
 $1630/17096=0.095$

Conditional Distribution

Frequent of binge drinker	Gender	
	Men	Women
Yes	0.227019	0.1698265
No	0.772981	0.8301735

e.g., "Men-Yes" cell
 $1630/7180=0.227$

- Which one is a better representation?
- You may also look at Marginal distribution in the book. We skip it.

Data in categories

- What is “data in categories?”
 - Two Variables are categorical
 - Each Variable can have more than two categories.
 - Categories of age vs. Categories of education

Data (1000's)	25-34	35-44	45-54	55-64	≥65	Total
Not complete HS	5836	4841	5230	7024	13183	36114
Completed HS	17889	13200	9860	8580	9412	58941
College, 1-3 yrs	9069	7309	3698	2793	2915	25784
College, ≥ 4 yrs	10174	9332	5008	3246	3018	30778
Total	42968	34682	23796	21643	28528	151617



Agenda

Two-way Table

- Data in categories
 - Graphs
 - Simpson Paradox
-
- It will be easiest to do these two-way table analyses using Excel (or some other spreadsheet) rather than JMP.]

Graphs

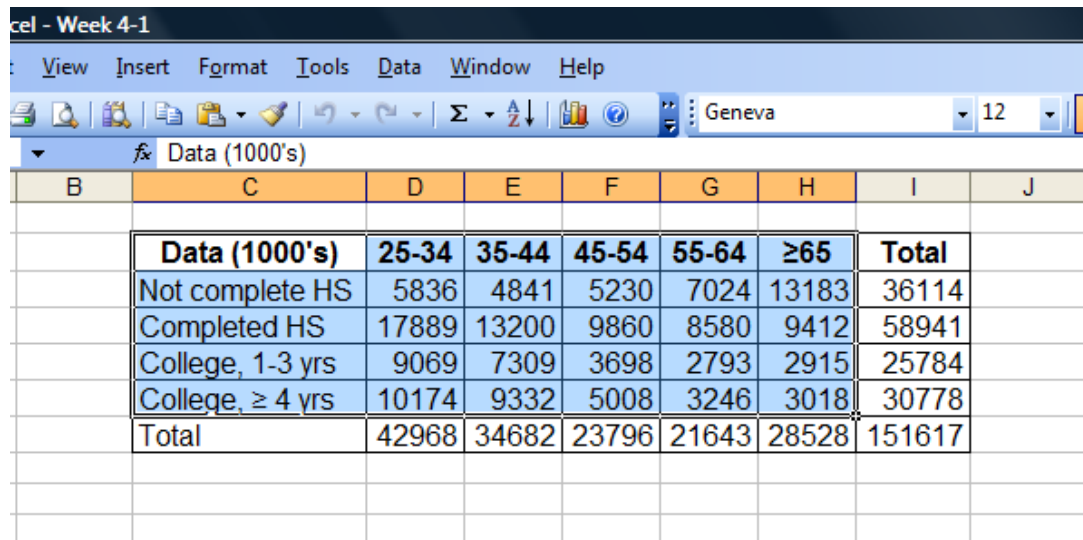
What kind of graph can we draw?

- Scatter plot?

Data (1000's)	25-34	35-44	45-54	55-64	≥65	Total
Not complete HS	5836	4841	5230	7024	13183	36114
Completed HS	17889	13200	9860	8580	9412	58941
College, 1-3 yrs	9069	7309	3698	2793	2915	25784
College, ≥ 4 yrs	10174	9332	5008	3246	3018	30778
Total	42968	34682	23796	21643	28528	151617

Graphs

1. select data



The screenshot shows a Microsoft Excel spreadsheet titled "Excel - Week 4-1". The spreadsheet contains a table with the following data:

	B	C	D	E	F	G	H	I	J
		Data (1000's)	25-34	35-44	45-54	55-64	≥65	Total	
		Not complete HS	5836	4841	5230	7024	13183	36114	
		Completed HS	17889	13200	9860	8580	9412	58941	
		College, 1-3 yrs	9069	7309	3698	2793	2915	25784	
		College, ≥ 4 yrs	10174	9332	5008	3246	3018	30778	
		Total	42968	34682	23796	21643	28528	151617	

Graphs

2- click on chart wizard

Chart Wizard - Step 1 of 4 - Chart Type

Standard Types Custom Types

Chart type:

- Column
- Bar
- Line
- Pie
- XY (Scatter)
- Area
- Doughnut
- Radar
- Surface
- Bubble

Chart sub-type:

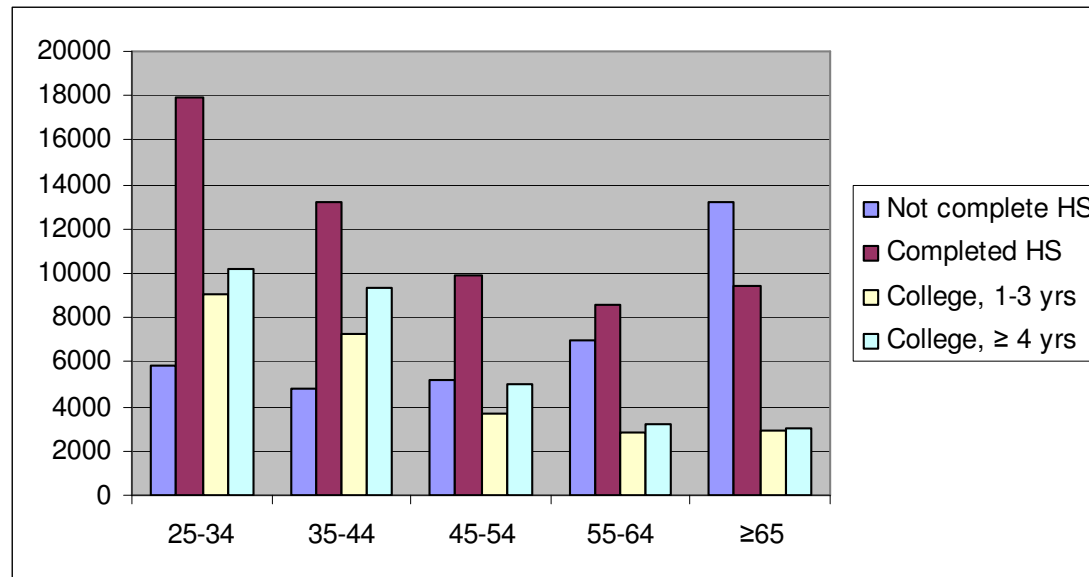
Clustered Column. Compares values across categories.

Press and Hold to View Sample

Cancel < Back Next > Finish

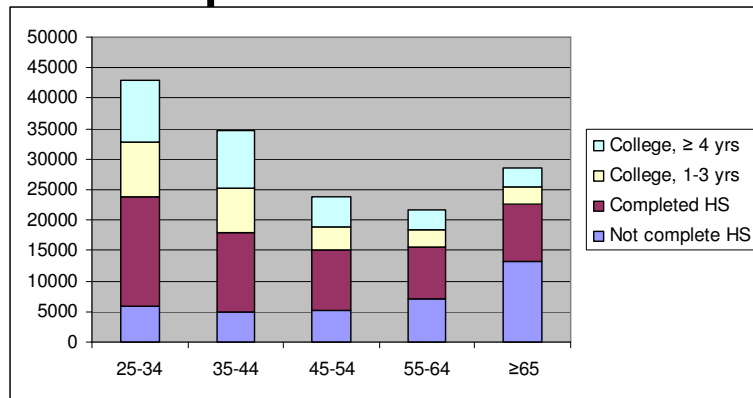
Graphs

3. click on "next" and follow the steps, or click on finish.

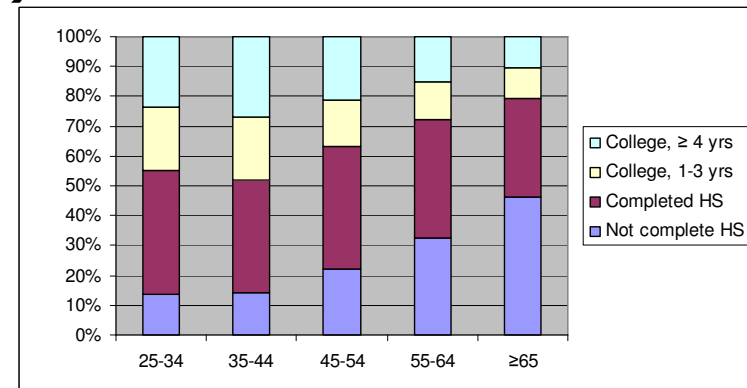


Graphs

In a similar way draw other graphs. (As explained in the class)



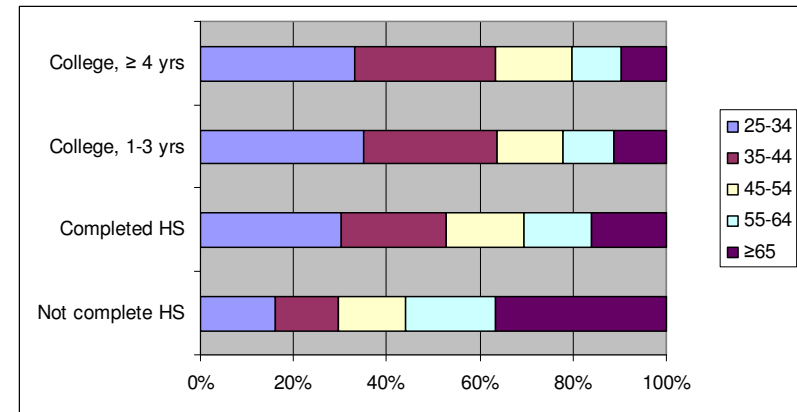
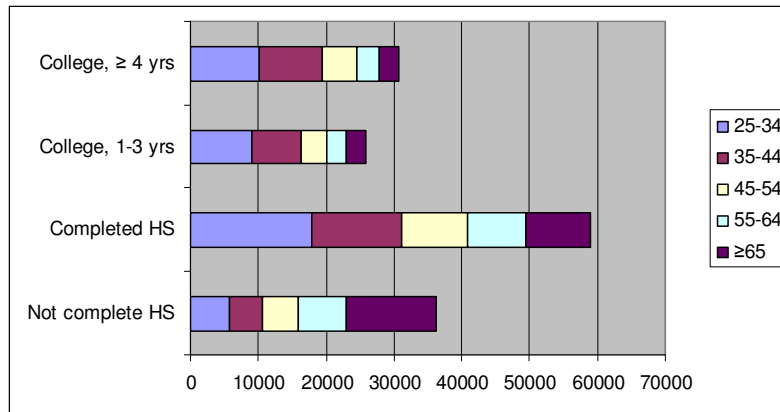
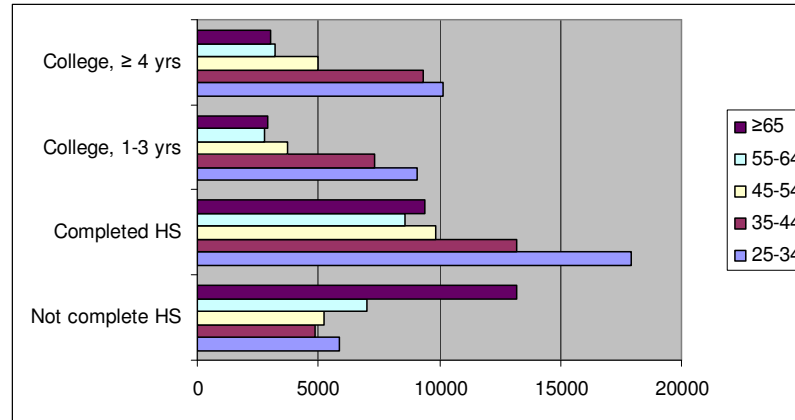
■ Stacked bar charts



■ percentage bar charts

- In a similar way draw other graphs where x is NOT age BUT education.

Graphs



■ Many times new graphs give new insights.



Graphs

What kind of graph can we draw?

- Bar charts
- Stacked bar charts
- Percents bar charts



Agenda

Two-way Table

- Data in categories
- Graphs
- Simpson Paradox

- It will be easiest to do these two-way table analyses using Excel (or some other spreadsheet) rather than JMP.]

Simpson Paradox

Flight Delay Example

	Los Angeles			Phoenix		
	On time	Delayed	Total	On time	Delayed	Total
Alaska Airlines	497	62	559	221	12	233
America West	694	117	811	4840	415	5255

- Alaska Airline percentage of delays = $(62+12)/(559+233) = 9.3\%$
- America West percentage of delays = $(117+415)/(811+5255) = 8.8\%$ → **America West is better**

- Let's look at each city separately: (percentage of delay in each city)
- Alaska Airline in LA = $62/559 = 11.1\%$, in Phoenix = $12/233 = 5.2\%$
- America West in LA = $117/811 = 14.4\%$, in Phoenix = $415/5255 = 7.9\%$ → **Alaska Airline is better in every city (WHY?)**

Simpson Paradox

Flight Delay Example

- Alaska Airline percentage of delays = $(62+12)/(559+233) = 9.3\%$
- America West percentage of delays = $(117+415)/(811+5255) = 8.8\%$ → **America West is better**

- Let's look at each city separately: (percentage of delay in each city)
- Alaska Airline in LA = $62/559 = 11.1\%$, in Phoenix = $12/233 = 5.2\%$
- America West in LA = $117/811 = 14.4\%$, in Phoenix = $415/5255 = 7.9\%$ → **Alaska Airline is better in every city (WHY?)**

- **The better choice is of course Alaska airline.** Looking at aggregated data sometimes can be misleading.
- Weighted average issue.
- We should break down the data and check for Simpson paradox.

PUB – POS 316
Week 4-2



Correlation and Causation

Navid Ghaffarzadegan

navidg@gmail.com



Agenda

- From Previous section – Two way tables
- Correlation vs Causation
 - Review of Correlation
 - Two examples



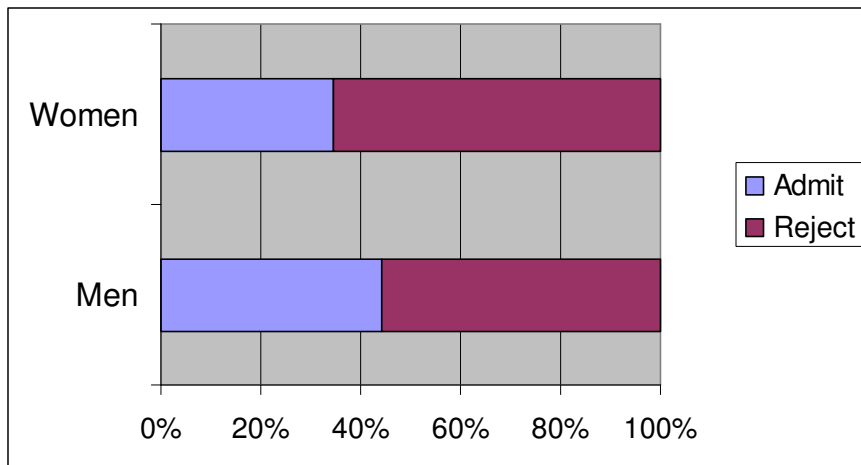
Two way tables

- Question: Is there any discrimination between female and male applicants in University A?
- Method?
- Berkeley Sex Bias Case

Two way tables

Question: Is there any discrimination between female and male applicants in Berkeley?

Observed:	Admit	Reject	Totals by gender
Men	3738	4704	8442
Women	1494	2827	4321
Totals by decision	5232	7531	12763



Percentage	Admit	Reject
Men	44.28%	55.72%
Women	34.58%	65.42%

Two way tables

Question: Is there any discrimination between female and male applicants in Berkeley?

- Data in Department Level

Departments	Men		Women	
	Applicants	% admitted	Applicants	% admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

- Only in major departments.. (Data doesn't add up)
- Simpson Paradox: women tended to apply to competitive departments with low rates of admission.



Agenda

- From Previous section – Two way tables
- Correlation vs Causation
 - Review of Correlation
 - Two examples



Correlation vs Causation

- A question:
 - How should we help students to perform better?
- How do you measure performance?
- Method?
 - survey?
 - Experiment?
 - Secondary data?



Correlation vs Causation – student performance

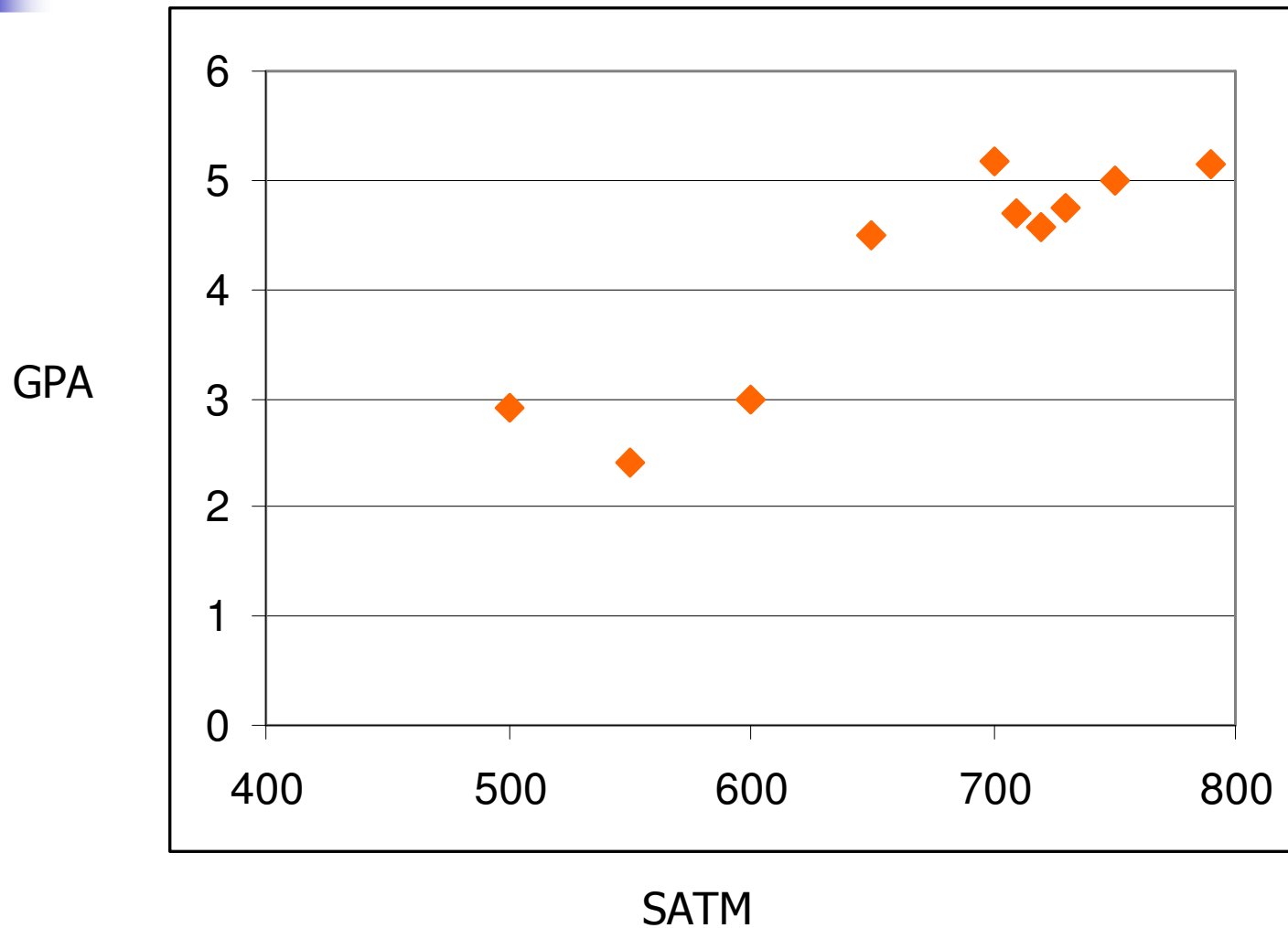
Observation	GPA	Male/Female	SATM	Hours of Study (per day)
1	2.4	M	550	1
2	2.91	F	500	1.5
3	4.5	F	650	4.6
4	5.16	M	700	5.8
5	5.14	F	790	7
6	4.75	F	730	8.2
7	4.58	M	720	9.4
8	4.7	M	710	5.5
9	5	M	750	6
10	3	F	600	4



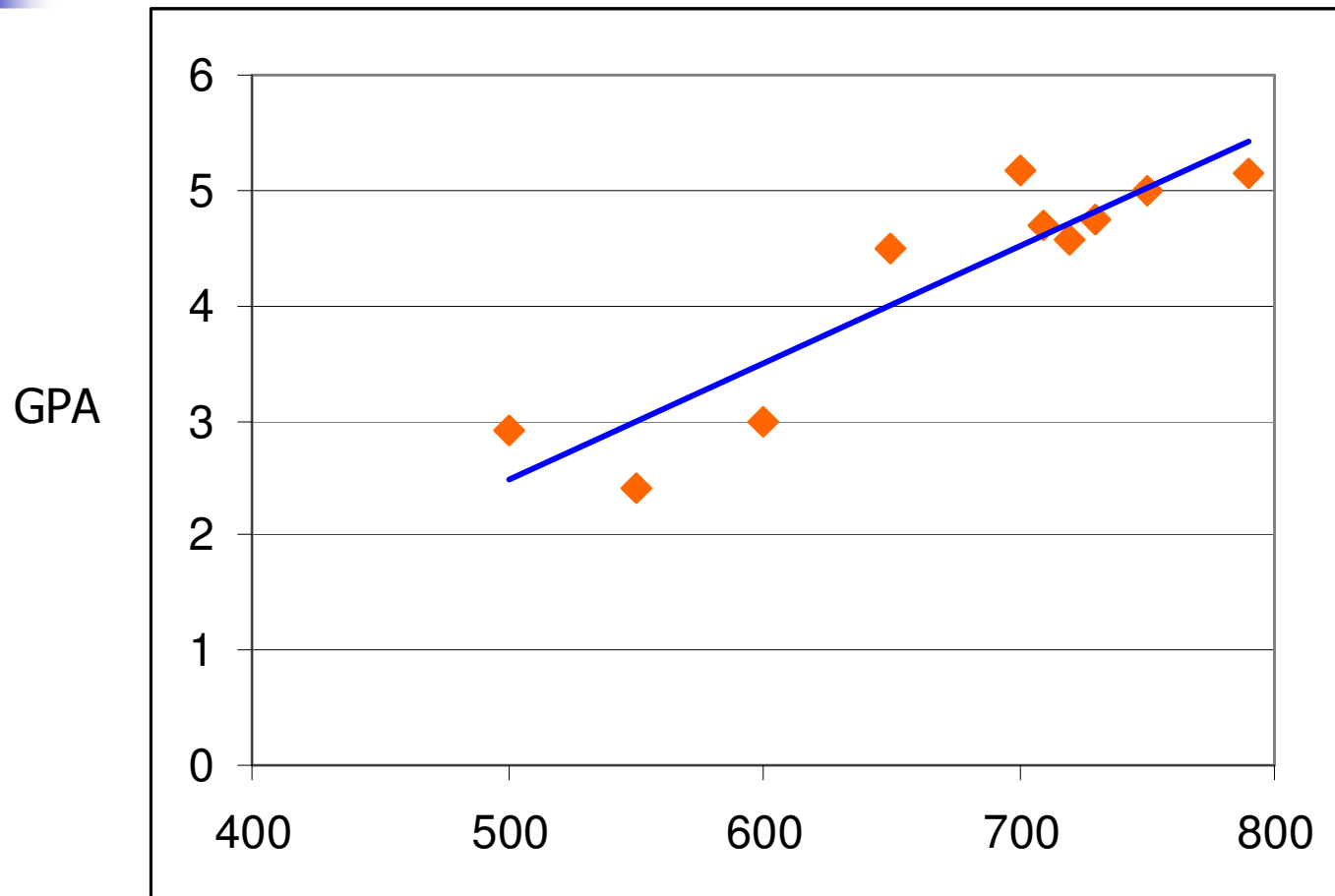
Correlation vs Causation – student performance

Observation	GPA	Male/Female	SATM	Hours of Study (per day)
1	2.4	M	550	1
2	2.91	F	500	1.5
3	4.5	F	650	4.6
4	5.16	M	700	5.8
5	5.14	F	790	7
6	4.75	F	730	8.2
7	4.58	M	720	9.4
8	4.7	M	710	5.5
9	5	M	750	6
10	3	F	600	4

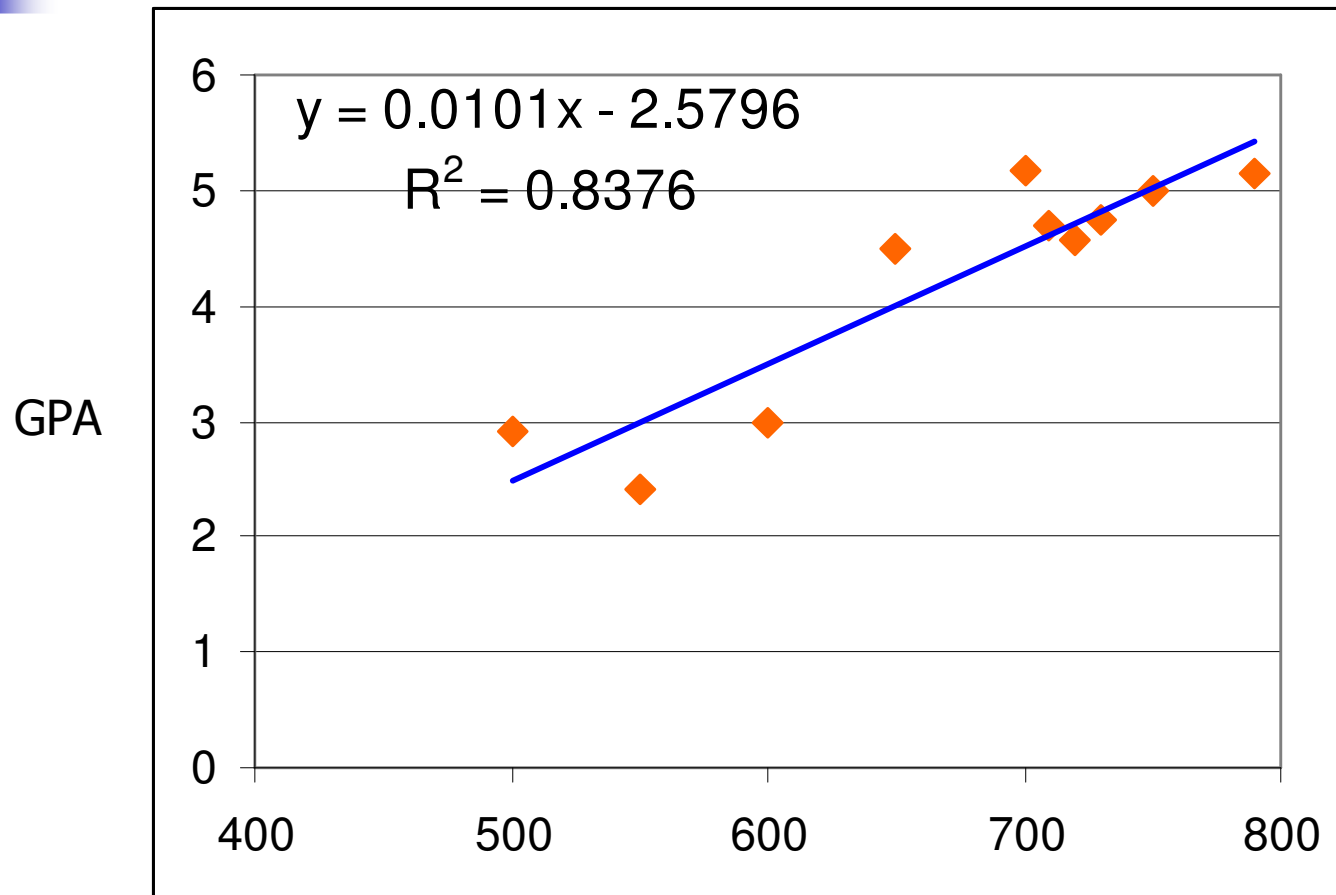
Correlation vs Causation – student performance



Correlation vs Causation – student performance



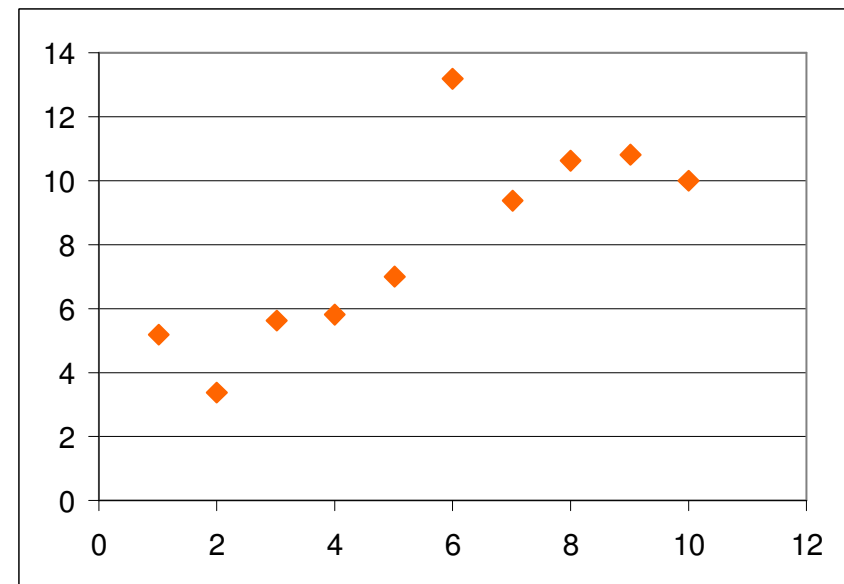
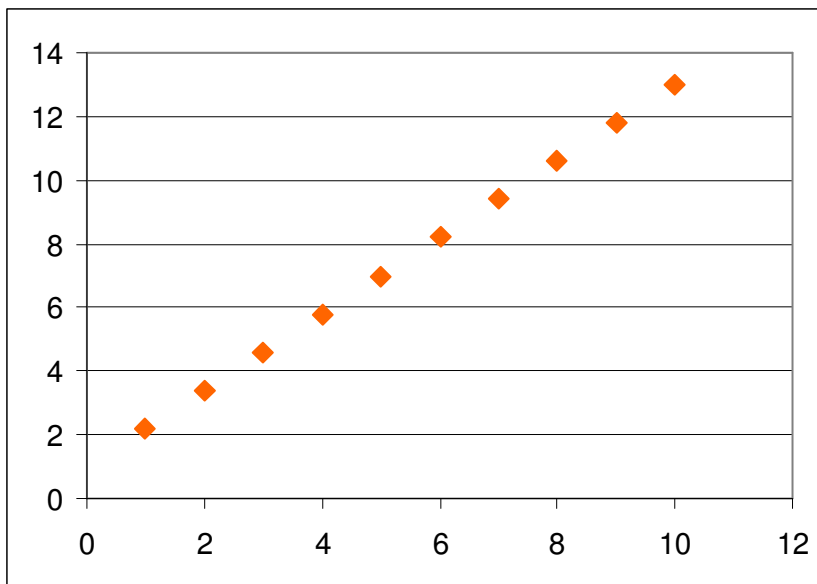
Correlation vs Causation – student performance



R² – A quick reminder

What was R² ?

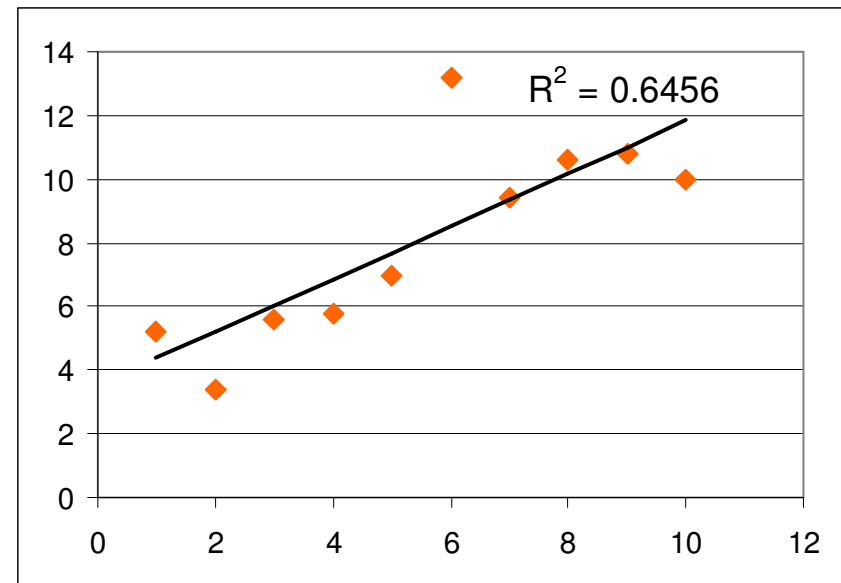
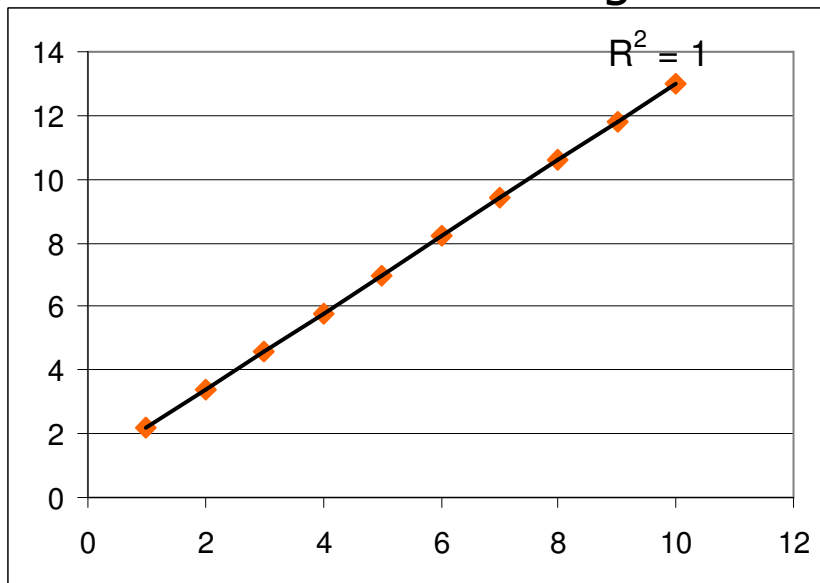
- Try to understand and keep in mind the technical definition of R². (= the fraction of variance in...)... Remind yourself. Practice.
- What is the maximum and minimum possible value for R² ?
 - $0 \leq R^2 \leq 1$
- Which one has a higher R² ?



R² – A quick reminder

What was R² ?

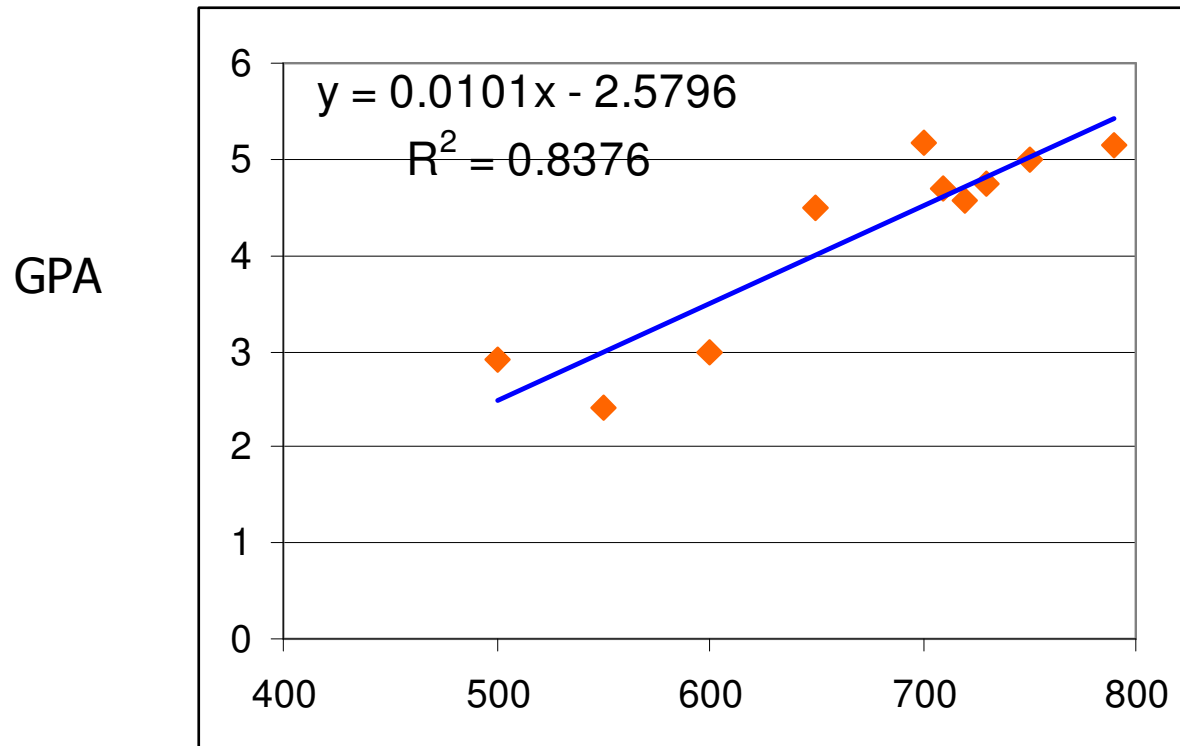
- Try to understand and keep in mind the technical definition of R². (the fraction of ...)... Remind yourself. Practice.
- What is the maximum and minimum possible value for R² ?
 - $0 \leq R^2 \leq 1$
- Which one has a higher R² ?



- What is correlation (r)? 1) take sq root of R², 2) take care of the sign, based on the slop of the line..

Back to our problem..

What can we conclude? Can we say higher SATM causes higher GPA?



AS MORE ICE CREAM IS EATEN . . . THE CRIME RATE GOES UP

The local police chief in a town observes:

- As ice cream consumption increases, crime rates increases.
- Even a scatter plot supports it!
- Even R^2 is high!!
- Even correlation is high!
 - So let's stop selling ice cream??
 - Let's arrest whoever eats more ice cream!
- The outside temperature is what they both have in common.
 - Warm temperature → more windows left open
 - Warm temperature → more ice cream

AS MORE ICE CREAM IS EATEN . . . THE CRIME RATE GOES UP

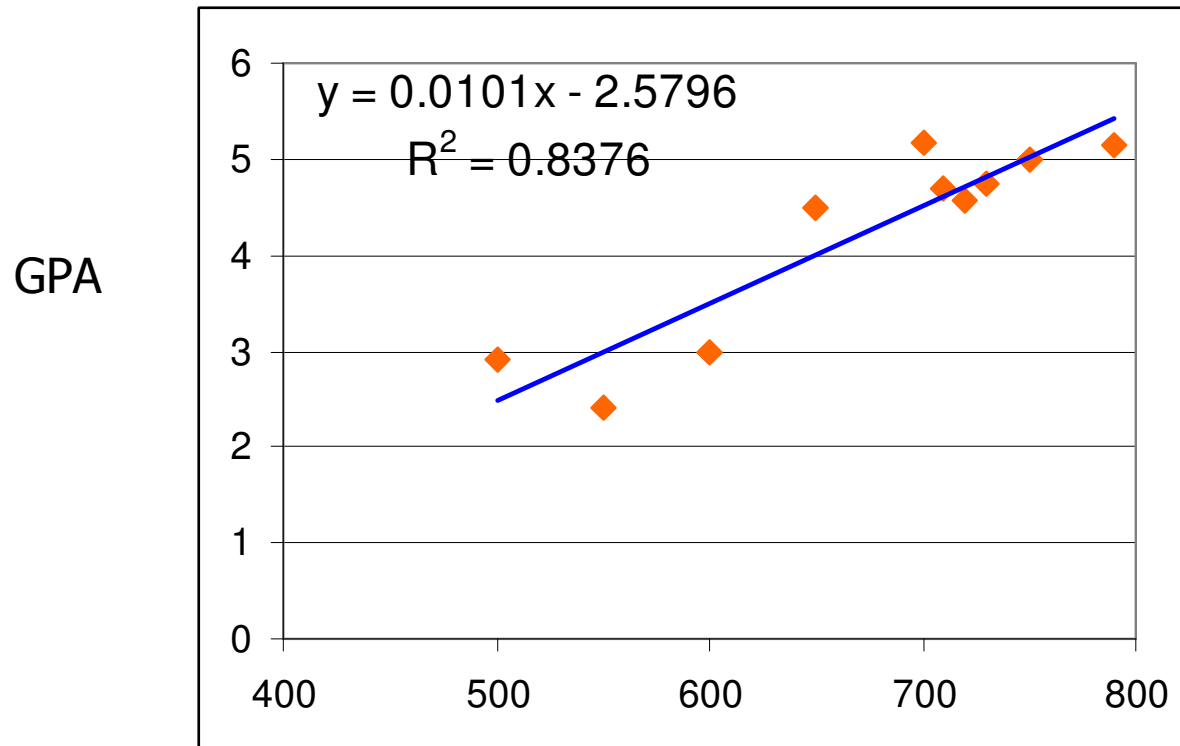
The local police chief in a town observes:

- As ice cream consumption increases, crime rates increase.
- Even a scatter plot supports it!
- Even R^2 is high!!
- Even correlation is high!
 - So let's stop selling ice cream??
 - Let's arrest whoever eats more ice cream!
- The outside temperature is what they both have in common.
 - Warm temperature \rightarrow more windows left open
 - Warm temperature \rightarrow more ice cream

Correlation is NOT causation

Back to our problem.. Student performance

What can we conclude? Can we say higher SATM causes higher GPA?



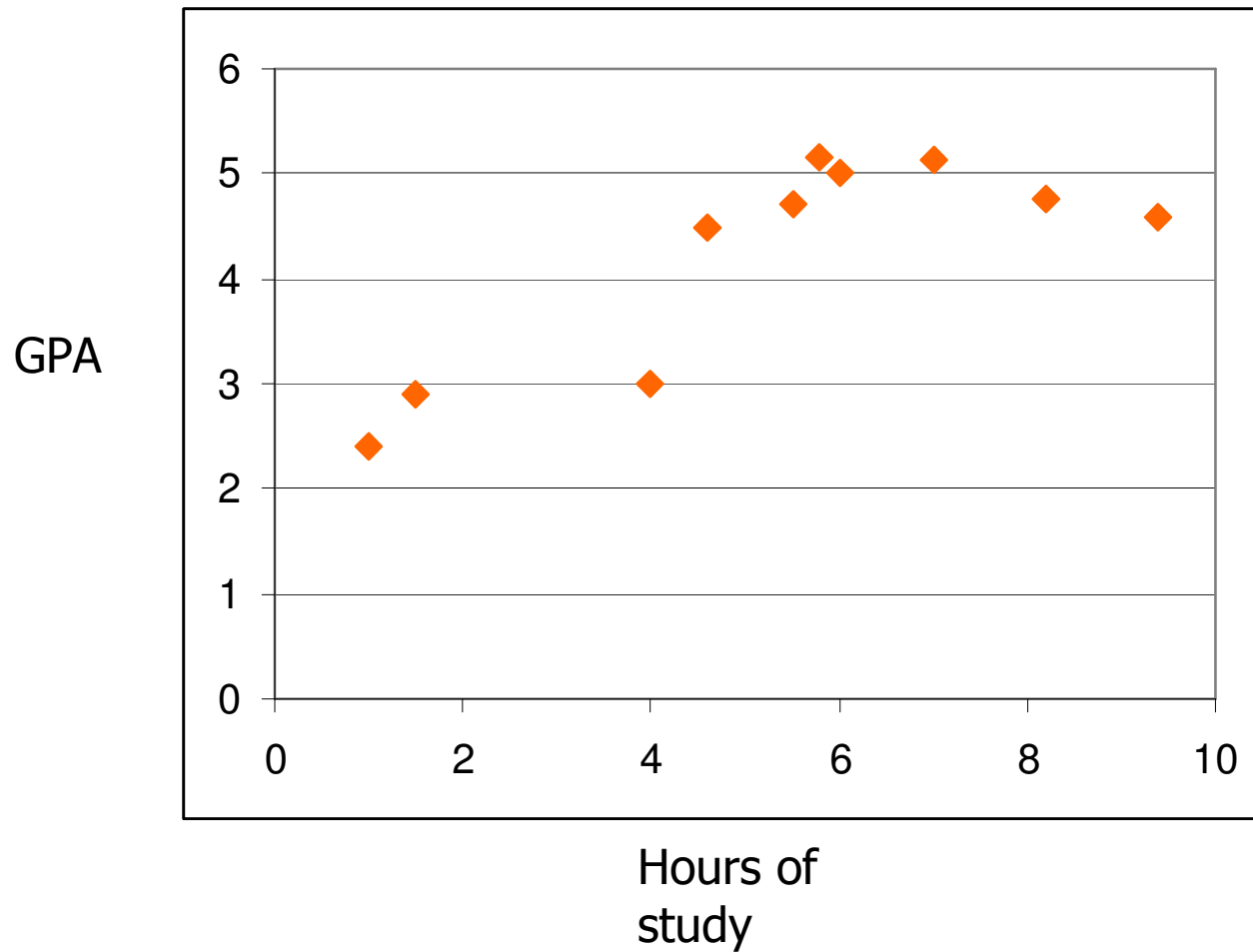
Correlation vs Causation – student performance

Observation	GPA	Male/Female	SATM	Hours of Study (per day)
1	2.4	M	550	1
2	2.91	F	500	1.5
3	4.5	F	650	4.6
4	5.16	M	700	5.8
5	5.14	F	790	7
6	4.75	F	730	8.2
7	4.58	M	720	9.4
8	4.7	M	710	5.5
9	5	M	750	6
10	3	F	600	4

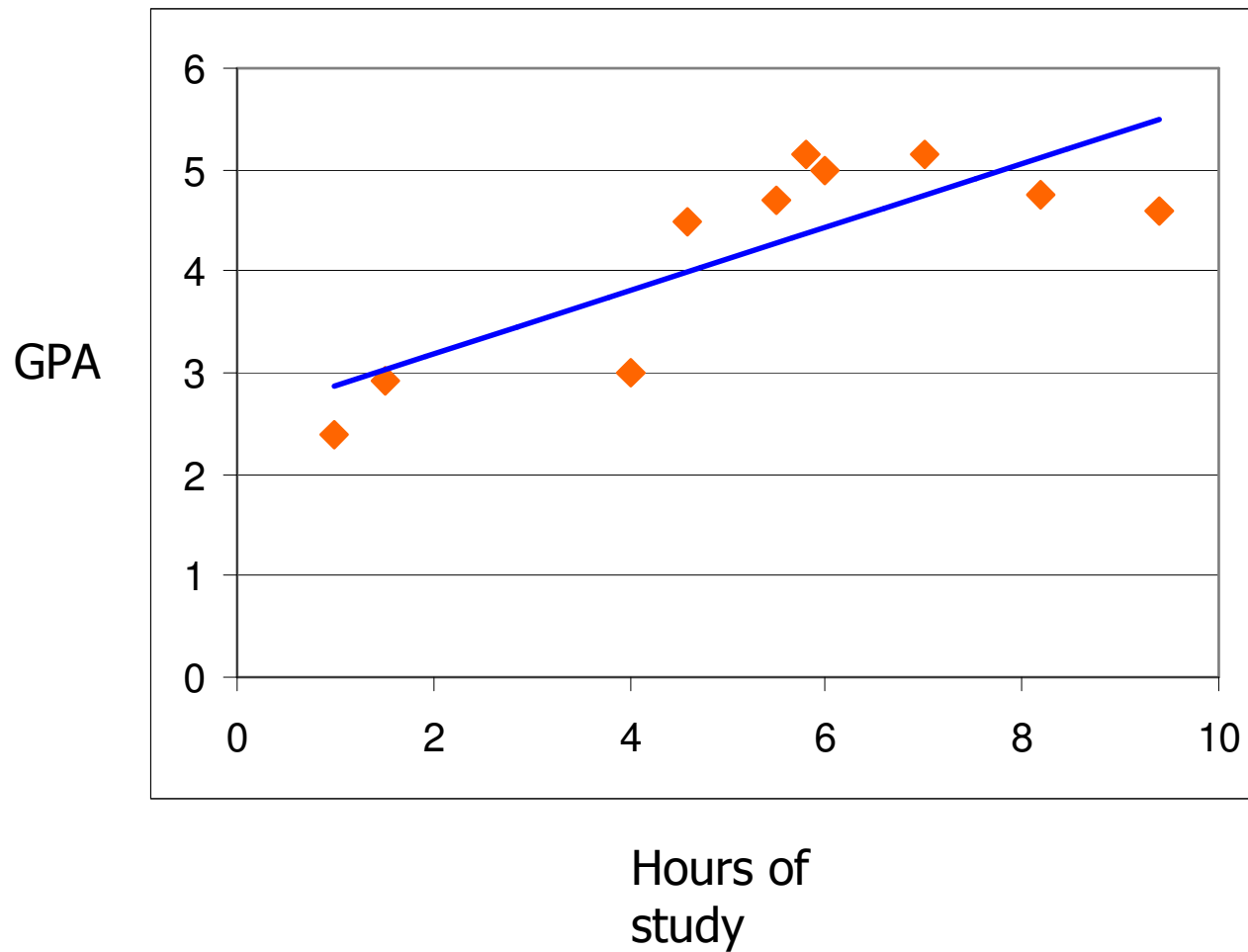
Correlation vs Causation – student performance

Observation	GPA	Male/Female	SATM	Hours of Study (per day)
1	2.4	M	550	1
2	2.91	F	500	1.5
3	4.5	F	650	4.6
4	5.16	M	700	5.8
5	5.14	F	790	7
6	4.75	F	730	8.2
7	4.58	M	720	9.4
8	4.7	M	710	5.5
9	5	M	750	6
10	3	F	600	4

Correlation vs Causation – student performance

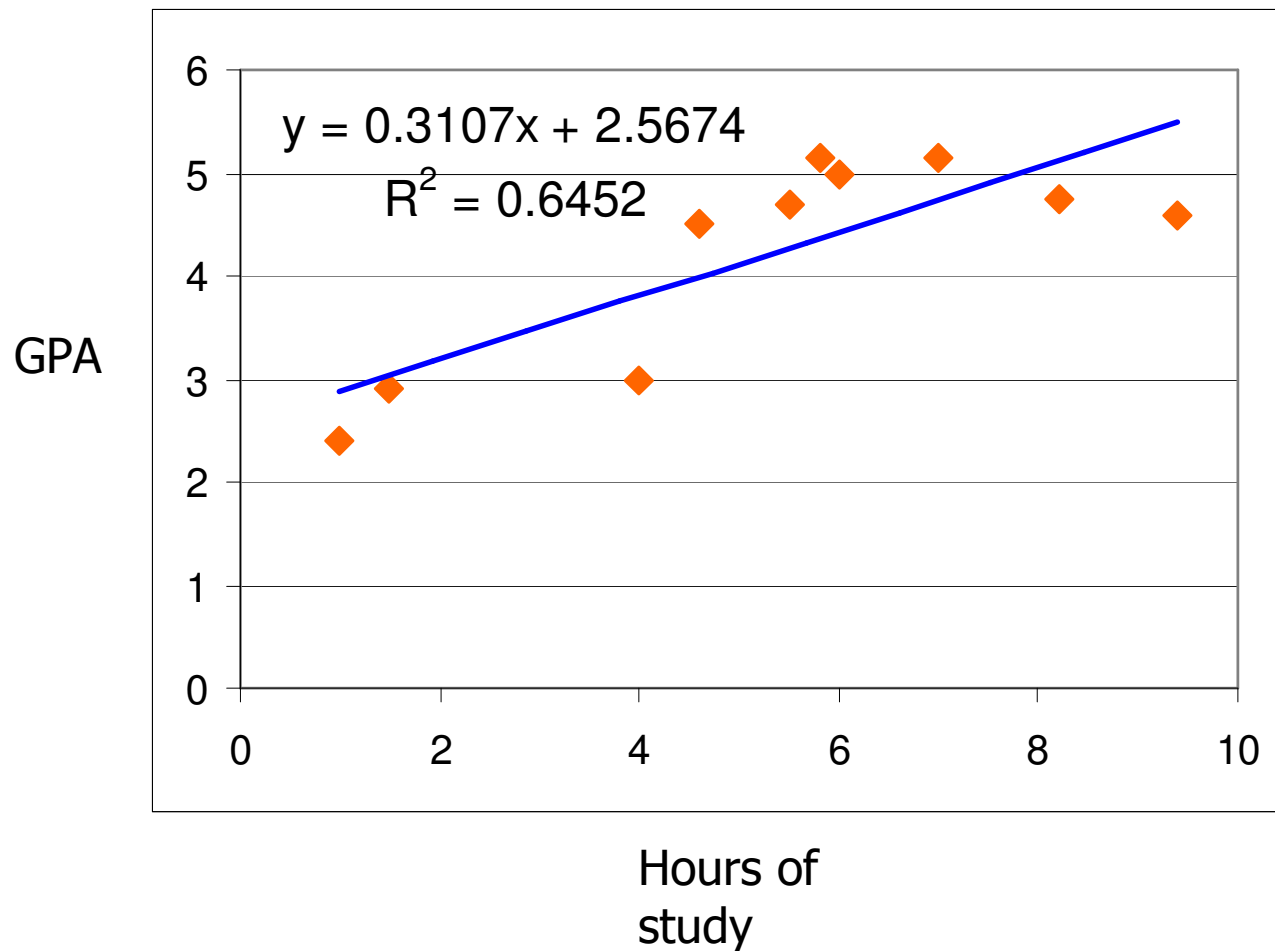


Correlation vs Causation – student performance



Correlation vs Causation – student performance

What can we conclude?





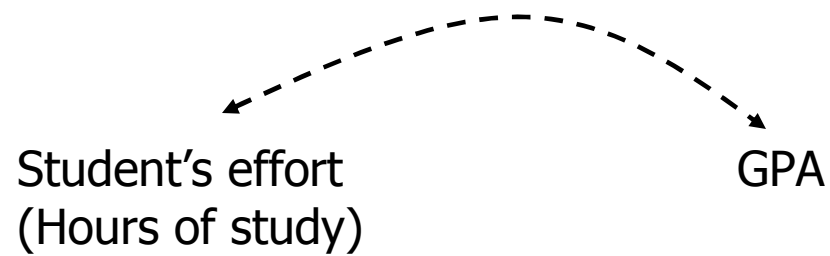
Correlation vs Causation – student performance

two events that occur together are many times claimed to have a cause-and-effect relationship.

- Sometimes, there is NO causal relation between those events (higher SATM and higher GPA).
- They can be both influenced by a lurking variable. (temperature in the ice cream example)
- Sometimes, there is a causal relation (hours of study and higher GPA)
- So, we should be very careful in evaluating results of a scatter plot.
 - Correlation is not causation

Correlation vs Causation

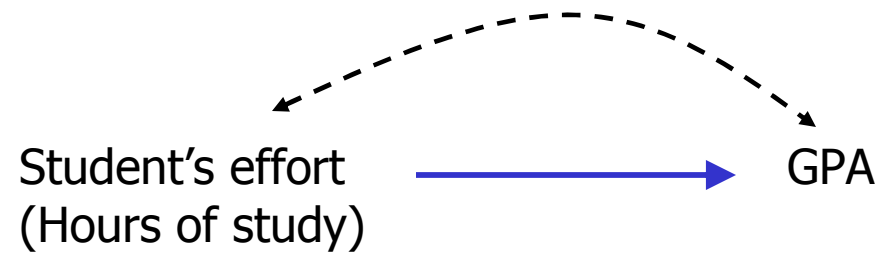
Three types of association:



Correlation vs Causation

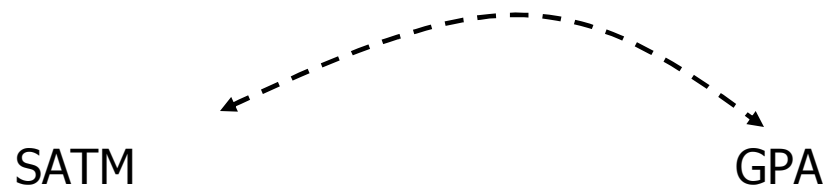
Three types of association:

Causation



Correlation vs Causation

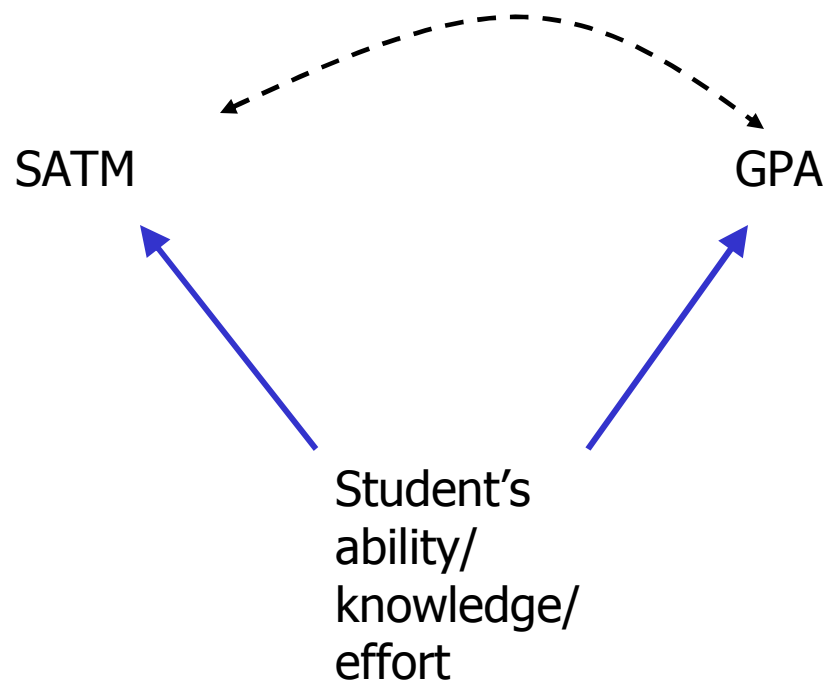
Three types of association:



Correlation vs Causation

Three types of association:

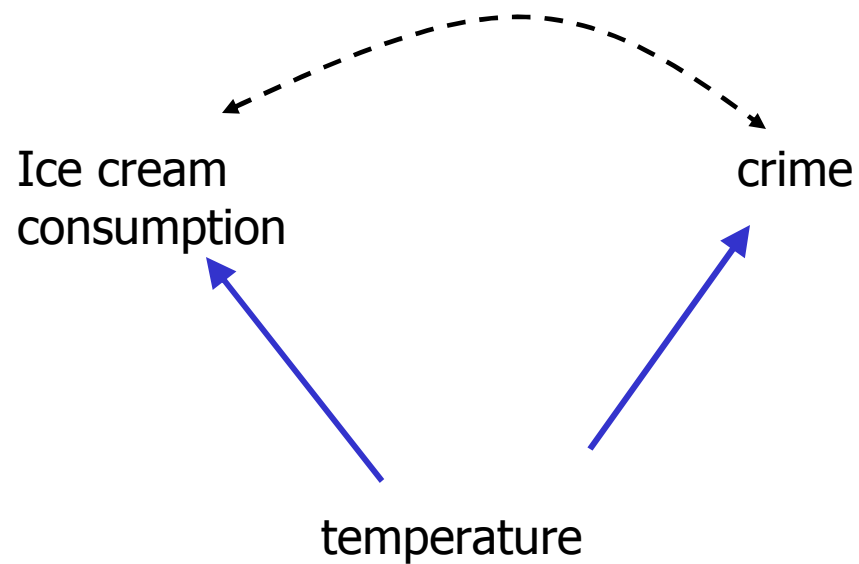
Common
response



Correlation vs Causation

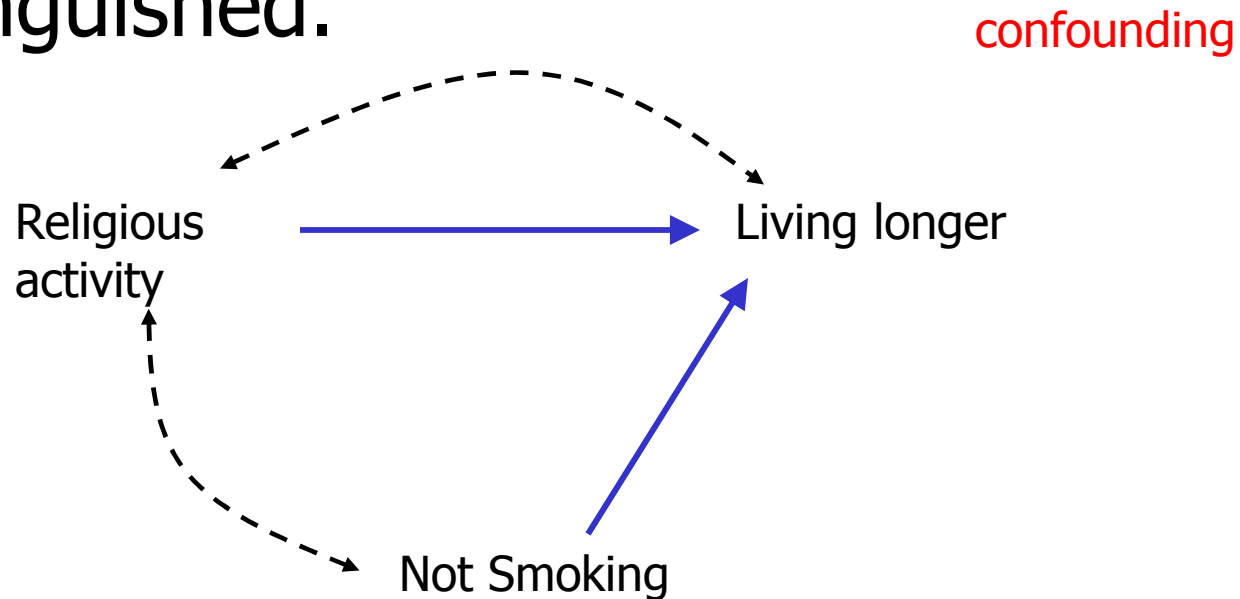
Three types of association:

Common
response



Correlation vs Causation

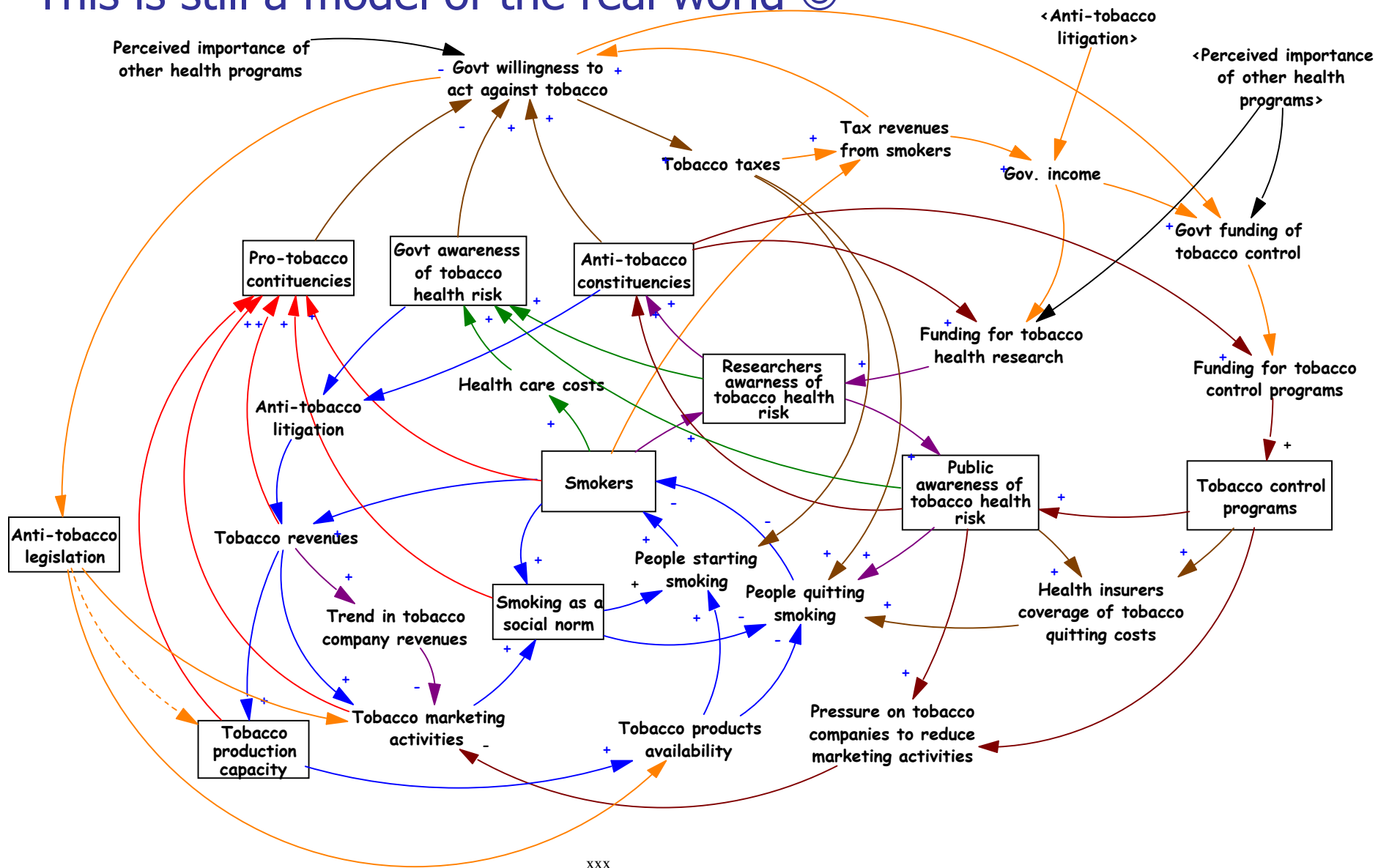
Three types of association: Confounding –
Two variables effect cannot be distinguished.



- How does the real world look like?

The real world!

This is still a model of the real world 😊



xxx

Reference: Richardson (2006), System Dynamics Mapping and Modeling for Tobacco Control



Take home message

That big loopi model is important and we should learn about it in future...

BUT NOT NOW.

Our today take home message:

Correlation is NOT causation

The Ice-cream story



Review

- From Previous section – Two way tables
 - Graph
 - Simpson paradox
- Correlation vs Causation
 - Review of Correlation
 - Two examples