

Phương pháp nghiên cứu trong khoa học liên ngành - Các phương pháp định lượng

Nguyễn Bích Ngọc

Khoa các khoa học liên ngành, ĐHQGHN

Thông tin lớp học

- Lý thuyết: 1/10 (S/C), 8/10 (S)
- Thực hành: 8/10 (S/C)
 - R
 - Laptop cài sẵn R & RStudio
 - Link hướng dẫn cài R & RStudio: <https://rstudio-education.github.io/hopr/starting.html>
- Cuối kỳ 60%: 14/10 (15/10)
 - Phân tích mô tả số liệu, vẽ đồ thị
 - Đọc và phân tích bài báo

Mục tiêu lớp học

- Giới thiệu khái niệm cơ bản của các phương pháp định lượng và thống kê
- Xác định vấn đề và định hướng phương pháp sử dụng (tên phương pháp)

Tài liệu tham khảo

- Cẩm nang nghiên cứu khoa học: từ ý tưởng đến công bố – Nguyễn Văn Tuấn (2nd edition, 2020)
- Từng bước nhập môn nghiên cứu khoa học xã hội – Phạm Hiệp & cộng sự (2022)
- Fundamentals of data visualization – Claus O. Wilke (<https://clauswilke.com/dataviz/index.html>)
- Applied statistics with R – David Dalpiaz (<https://book.stat420.org/>)
- The Scientist's Guide to Writing: How to Write More Easily and Effectively throughout Your Scientific Career – Stephen B. Heard (2nd 2022)
- Understanding research methods – Coursera (<https://www.coursera.org/learn/research-methods/home/info>)

Nội dung

- Giới thiệu chung
- Vấn đề nghiên cứu
- Dữ liệu và nguồn dữ liệu
- Phân tích dữ liệu thăm dò
- Phân tích dữ liệu khẳng định
- Các phương pháp nâng cao

Giới thiệu chung

Khoa học?

Khoa học?

Science

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

For a topical guide, see [Outline of science](#). For other uses, see [Science \(disambiguation\)](#).

Science is a rigorous, systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe.^{[1][2]} Modern science is typically divided into three major

Khoa học?

Science

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

For a topical guide, see [Outline of science](#). For other uses, see [Science \(disambiguation\)](#).

Science is a rigorous, systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe.^{[1][2]} Modern science is typically divided into three major

Nghiên cứu khoa học?

Nghiên cứu khoa học?

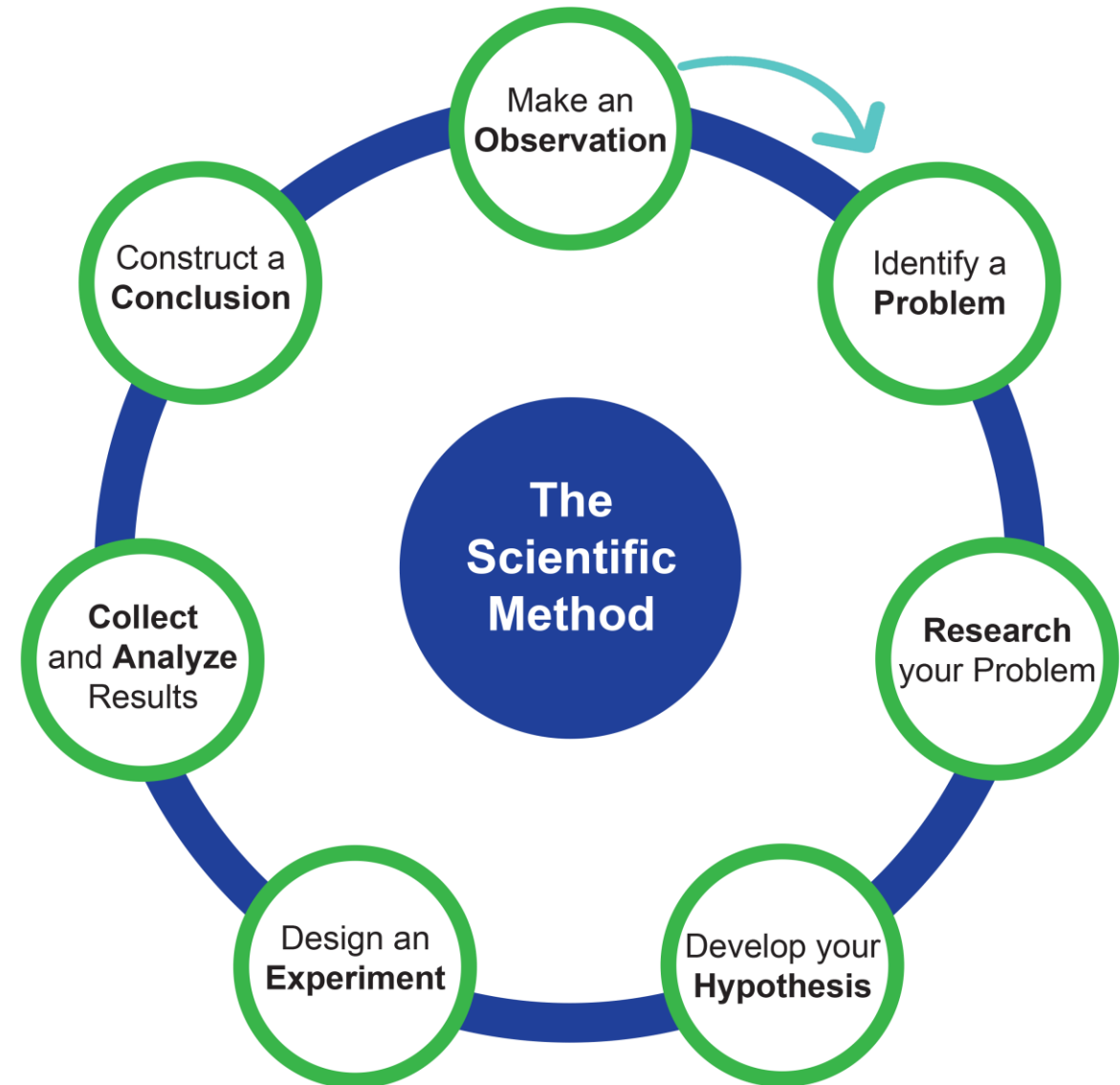
Research is **systematic inquiry** that helps to **make sense of the world** and that helps to make sensible the debates and interpretations that we have of issues of **contemporary significance**.

Professor Sandra Halperin

<https://www.coursera.org/learn/research-methods/home/info>

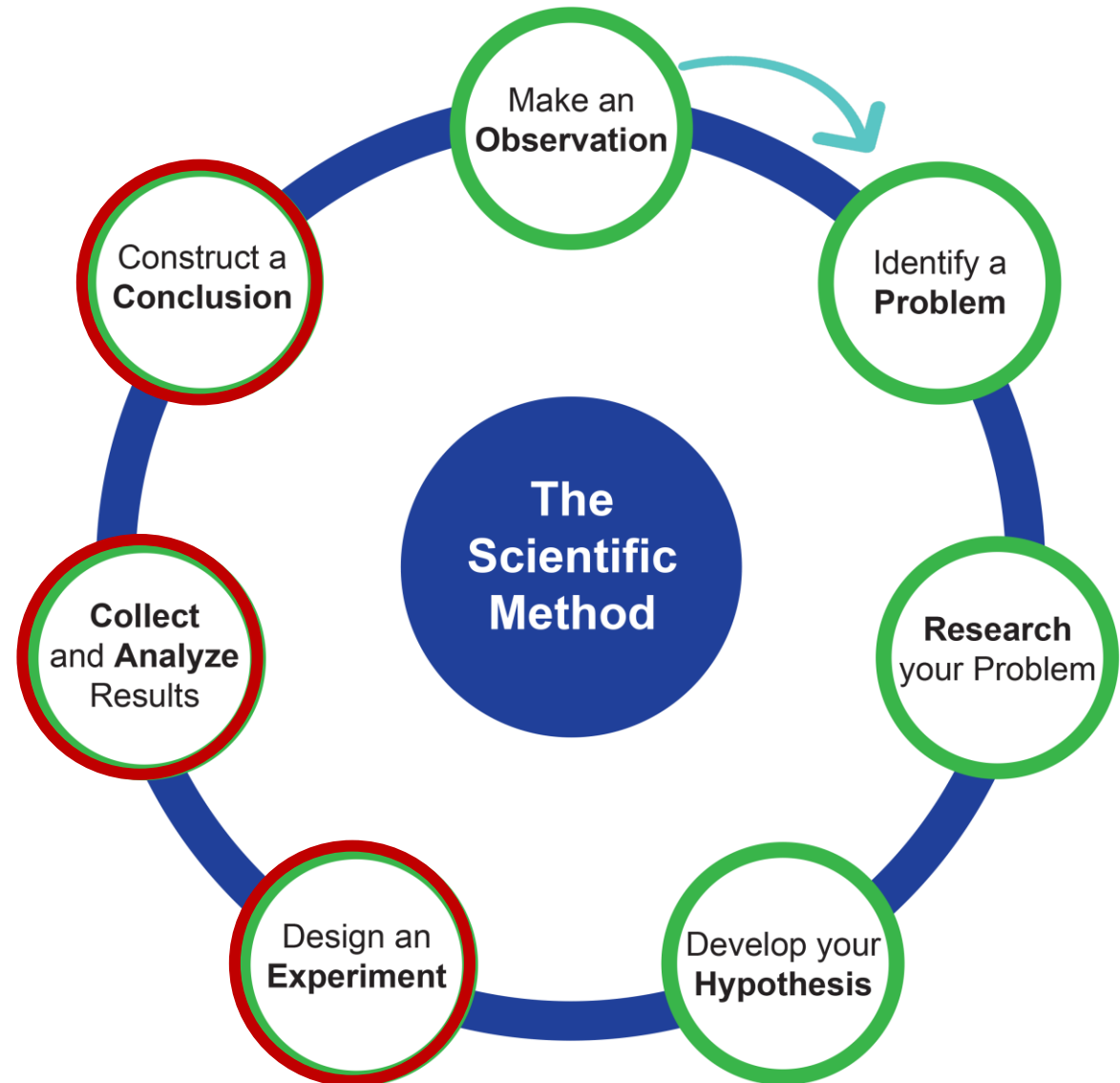
Quá trình nghiên cứu

Quá trình nghiên cứu



Source: <https://www.arfreethinkers.org/>

Phương pháp nghiên cứu?



Source: <https://www.arfreethinkers.org/>

Dữ liệu và nguồn dữ liệu

Bài tập

- Dữ liệu là gì?
- Dự kiến bạn sẽ cần dữ liệu nào cho nghiên cứu của bạn?
- Nguồn của những dữ liệu đó?

Dữ liệu

	Gender	Age.Range	Year	Nationality	Q1	Q2	Q3	Q4	Q5
1	Female	20 - 21 years old	Year 4	Thai	7	5	7	7	7
2	Female	20 - 21 years old	Year 4	Thai	6	5	7	5	6
3	Female	20 - 21 years old	Year 4	Thai	7	7	7	7	7
4	Female	20 - 21 years old	Year 4	Thai	7	2	7	6	7
5	Female	22 - 23 years old	Year 4	Thai	6	6	7	7	7
6	Male	20 - 21 years old	Year 3	Thai	5	4	4	4	4
7	Male	20 - 21 years old	Year 3	Thai	6	4	5	7	6
8	Female	20 - 21 years old	Year 3	Thai	7	4	7	6	7
9	Female	20 - 21 years old	Year 3	Thai	7	5	7	7	7
10	Male	20 - 21 years old	Year 3	Thai	5	5	5	7	6
11	Female	20 - 21 years old	Year 3	Thai	7	5	7	7	7

Question 1	In my opinion, it is important to protect the environment.
Question 2	I actively practice environmental sustainability at home (e.g., energy conservation, recycling).
Question 3	Everyone is responsible for caring for the environment
Question 4	I am concerned about the long-term future of the environment.
Question 5	In my opinion, it is important to conserve natural resources.
Question 6	I think that environmental sustainability is a waste of time and effort.
Question 7	I am a passionate advocate of environmental sustainability.

Dữ liệu

Act 1, Scene 1

[Enter Sampson and Gregory, two high-ranking servants of the Capulet household, carrying swords and shields. Gregory is making fun of Sampson, who sees himself as a fearsome fighter]

Sampson
Gregory, on my word, we'll not carry coals. 1

Gregory
No, for then we should be **colliers**. 2
coal workers

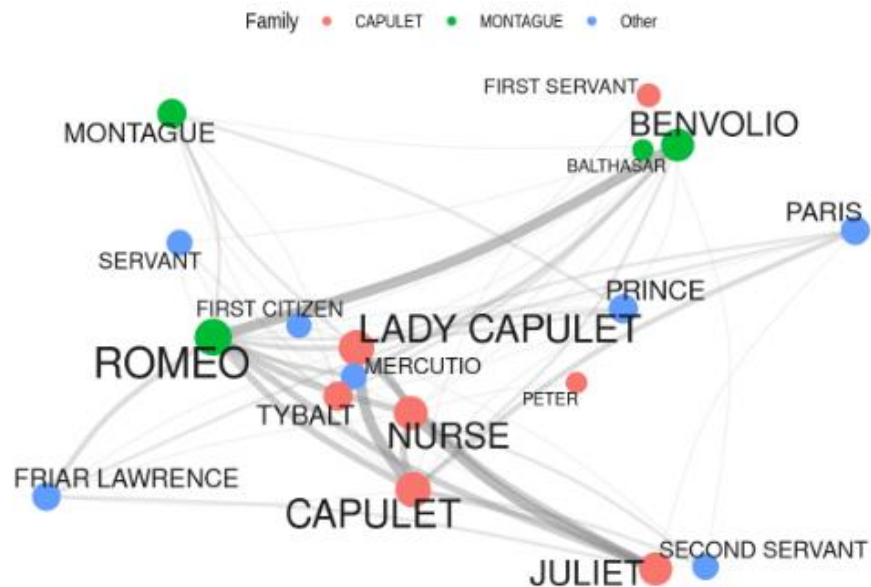
Sampson
I mean, **an** we be **in choler** we'll draw. 3
if angered (our swords)

Gregory
Ay, while you live, draw your neck out of collar. 4

Sampson
I strike quickly, being **moved**. 5
provoked

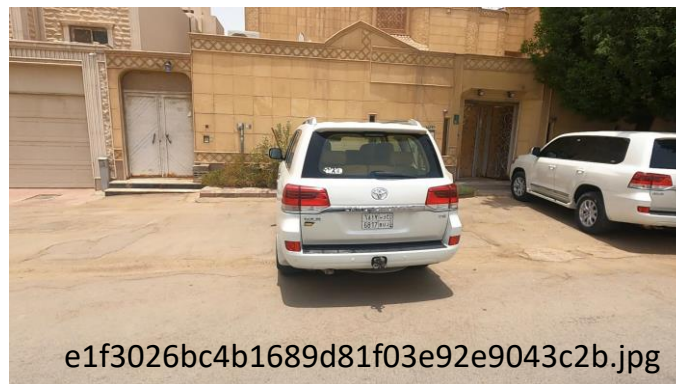


actscene	person	contrib	occurrences
ACT I_SCENE I	BENVOLIO	24	7
ACT I_SCENE I	CAPULET	2	9
ACT I_SCENE I	FIRST CITIZEN	1	2
ACT I_SCENE I	LADY CAPULET	1	10
ACT I_SCENE I	MONTAGUE	6	3
ACT I_SCENE I	PRINCE	1	3
ACT I_SCENE I	ROMEO	16	14
ACT I_SCENE I	TYBALT	2	3
ACT I_SCENE II	BENVOLIO	5	7
ACT I_SCENE II	CAPULET	3	9
ACT I_SCENE II	PARIS	2	5
ACT I_SCENE II	ROMEO	11	14
ACT I_SCENE II	SERVANT	8	3
ACT I_SCENE III	JULIET	5	11
ACT I_SCENE III	LADY CAPULET	11	10



Persona	BALTHASAR	BENVOLIO	CAPULET	FIRST CITIZEN	FIRST SERVANT
BALTHASAR	0	0	1	0	0
BENVOLIO	0	0	3	2	1
CAPULET	1	3	0	1	2
FIRST CITIZEN	0	2	1	0	0
FIRST SERVANT	0	1	2	0	0

Dữ liệu



	A	B	C	
	class	image_path	name	xt
	3	4a48c42c9579ec0399e6c5a3e825e765.jpg	GARBAGE	
	3	4a48c42c9579ec0399e6c5a3e825e765.jpg	GARBAGE	
	3	4a48c42c9579ec0399e6c5a3e825e765.jpg	GARBAGE	
	7	ea906a663da6321bcef78be4b7d1afff.jpg	BAD_BILLBOARD	
	8	1c7d48005a12d1b19261b8e71df7cafe.jpg	SAND_ON_ROAD	
	8	1c7d48005a12d1b19261b8e71df7cafe.jpg	SAND_ON_ROAD	
	8	8ca1b825716ea6755180fde347ac79c1.jpg	SAND_ON_ROAD	
	0	8ca1b825716ea6755180fde347ac79c1.jpg	GRAFFITI	
0	0	8ca1b825716ea6755180fde347ac79c1.jpg	GRAFFITI	
1	2	e1f3026bc4b1689d81f03e92e9043c2b.jpg	POTHOLES	
2	3	c12b006174423ceb3e2e3563a8ca7751.jpg	GARBAGE	
3	3	7fb40d10dde6d5643aa8e197b6b46c2e.jpg	GARBAGE	
4	3	7fb40d10dde6d5643aa8e197b6b46c2e.jpg	GARBAGE	
5	3	f05cd6411a3509a5ddc9d9a52536df01.jpg	GARBAGE	
6	2	f05cd6411a3509a5ddc9d9a52536df01.jpg	POTHOLES	
7	3	b08b7961553eac0b24c7e871836fad9c.jpg	GARBAGE	
8	3	b08b7961553eac0b24c7e871836fad9c.jpg	GARBAGE	
9	3	b08b7961553eac0b24c7e871836fad9c.jpg	GARBAGE	
0	3	b08b7961553eac0b24c7e871836fad9c.jpg	GARBAGE	
1	3	b08b7961553eac0b24c7e871836fad9c.jpg	GARBAGE	

Dữ liệu

Biến

	Gender	Age.Range	Year	Nationality	Q1	Q2	Q3	Q4	Q5
1	Female	20 - 21 years old	Year 4	Thai	7	5	7	7	7
2	Female	20 - 21 years old	Year 4	Thai	6	5	7	5	6
3	Female	20 - 21 years old	Year 4	Thai	7	7	7	7	7
4	Female	20 - 21 years old	Year 4	Thai	7	2	7	6	7
5	Female	22 - 23 years old	Year 4	Thai	6	6	7	7	7
6	Male	20 - 21 years old	Year 3	Thai	5	4	4	4	4
7	Male	20 - 21 years old	Year 3	Thai	6	4	5	7	6
8	Female	20 - 21 years old	Year 3	Thai	7	4	7	6	7
9	Female	20 - 21 years old	Year 3	Thai	7	5	7	7	7
10	Male	20 - 21 years old	Year 3	Thai	5	5	5	7	6
11	Female	20 - 21 years old	Year 3	Thai	7	5	7	7	7

Đối tượng
quan sát/Mẫu

Mẫu, quần thể, cỡ mẫu

- Quần thể? Mẫu?

- Tính đại diện?

https://youtu.be/rxv_sB-wOkY

- Cỡ mẫu?

https://nckh.huph.edu.vn/sites/nckh.huph.edu.vn/files/Ph%C6%B0%C6%A1ng%20ph%C3%A1p%20ch%E1%BB%8Dn%20m%E1%BA%ABu%20v%C3%A0%20t%C3%ADnh%20t%C3%A1n%20c%E1%BB%A1%20m%E1%BA%ABu_revised%20l%E1%BA%A7n%201_5.8.2020_0.pdf

Phương pháp lấy mẫu

- Mẫu ngẫu nhiên (Probability/Random sample)
 - Mẫu ngẫu nhiên đơn giản (Simple random sample)
 - Mẫu ngẫu nhiên hệ thống (Systematic sample)
 - Mẫu ngẫu nhiên phân loại (Stratified sample)
 - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên (Nonprobability sample)
 - Mẫu thuận tiện (Convenience sample)
 - Mẫu hạn ngạch (Quota sample)
 - Mẫu có mục đích (Judgement (or purposive) sample)
 - Mẫu bóng tuyết (Snowball sample)

Biến

- Các loại biến

- Định danh (nominal)

- Thứ bậc (ordinal)

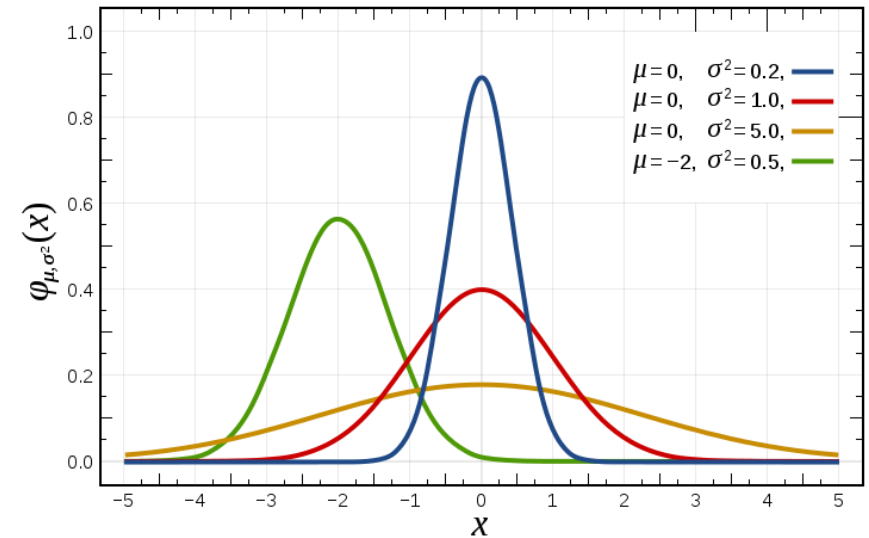
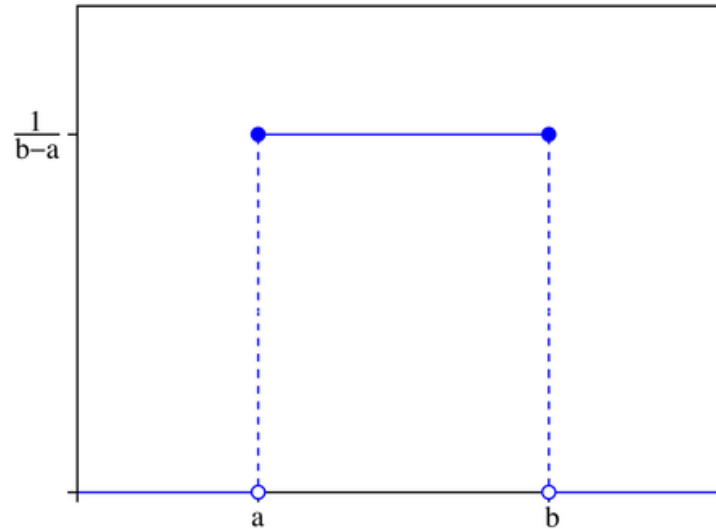
- Liên tục/Định lượng (continuous variables)



Biến gián đoạn/định tính
(discrete variables)

Biến định danh – nominal variables

- Ví dụ:
 - Giới tính
 - Tôn giáo
 - Quốc tịch
- Phân phối
 - Đồng nhất
 - Nhị phân
 - Đa thức

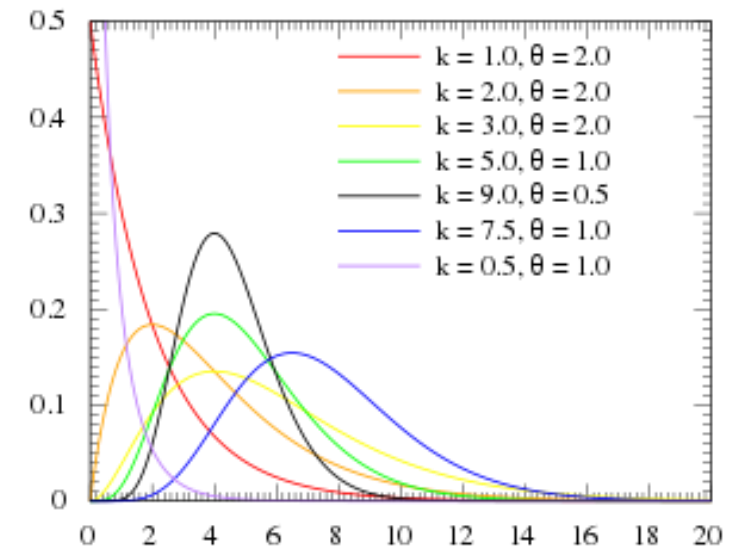
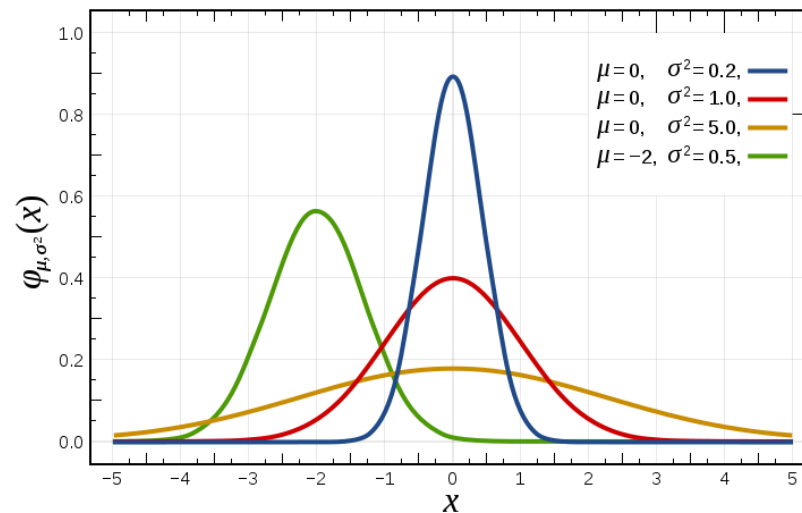


Biến thứ bậc – ordinal variables

- Ví dụ:
 - Thang Likert: hoàn toàn không đồng ý – hoàn toàn đồng ý
 - Trình độ học vấn: THCS, THPT, trung cấp, đại học, sau đại học
 - Điều kiện kinh tế xã hội: thấp, trung bình, cao
 - Đánh giá/chấm điểm: 1 – 5 ★
- Đặc điểm
 - Có tính thứ bậc tự nhiên
 - Không thể khẳng định khoảng cách bằng nhau giữa các giá trị

Biến liên tục/định lượng – continuous variables

- Ví dụ:
 - Tuổi đời (người/di tích)
 - Diện tích
 - Dân số
- Phân phối
 - Đồng nhất
 - Chuẩn (chuẩn tắc)
 - Gamma
 - ...



Bài tập

	A	B	C	D	E	F	G	H
1	id	csmptv	rwtank	iceqac2	dwltyp	livara	nbbdrm	cfdiwq
2	5	74	no	NA	2 facades	58	1	suspicious
3	29	131	yes	modest	2 facades	100	2	neither confident nor suspicious
4	34	102.08	yes	average	4 facades	200	2	neither confident nor suspicious
5	36	71	no	precarious	apartment/studio	70	2	rather confident
6	37	40	no	NA	2 facades	58	1	rather confident
7	39	39.16	no	modest	2 facades	100	3 or more	confident
8	42	115	no	modest	2 facades	200	2	confident
9	44	NA	yes	higher	4 facades	150	3 or more	rather confident
10	45	NA	yes	precarious	4 facades	150	3 or more	confident
11	46	119.57	no	modest	3 facades	150	3 or more	rather confident
12	47	97	no	average	2 facades	90	3 or more	neither confident nor suspicious
13	48	82.6	yes	modest	4 facades	200	3 or more	rather suspicious
14	50	38	yes	modest	4 facades	130	3 or more	confident
15	51	NA	no	NA	apartment/studio	90	1	confident
16	53	162.56	no	modest	4 facades	160	3 or more	confident
17	54	15.27	yes	modest	4 facades	121	2	confident
18	55	71.57	no	precarious	apartment/studio	50	1	no opinion
19	56	42	no	modest	3 facades	102	2	confident
20	57	141.36	no	modest	2 facades	105	3 or more	confident
21	58	77	no	average	apartment/studio	90	2	rather suspicious

Nguồn dữ liệu



Sơ cấp

Thực nghiệm

Khảo sát/Bảng hỏi

Đo đạc ngoài thực địa



Thứ cấp

Cơ sở dữ liệu mở

Báo cáo kỹ thuật của chính phủ

Báo cáo nội bộ

Bảng hỏi

Table 1
Description of the characteristics in the dataset.

Column	Data label	Explanation
Column A	Student Status	Degree student; Exchange student
Column B	Institution	Prince of Songkla University
Column C	Faculty	Faculty of Hospitality and Tourism; College of Computing; Faculty of International Studies
Column D	Gender	Male; Female; I do not wish to say; Other
Column E	Age Range	18–19 years old; 20–21 years old; 22–23 years old; 24 years or above
Column F	Year	Year 1; Year 2; Year 3; Year 4
Column G	Nationality	Thai; Foreign
Column H	Probe	Have you heard about environmental sustainability before? [Answer options: (Yes) and (No)].

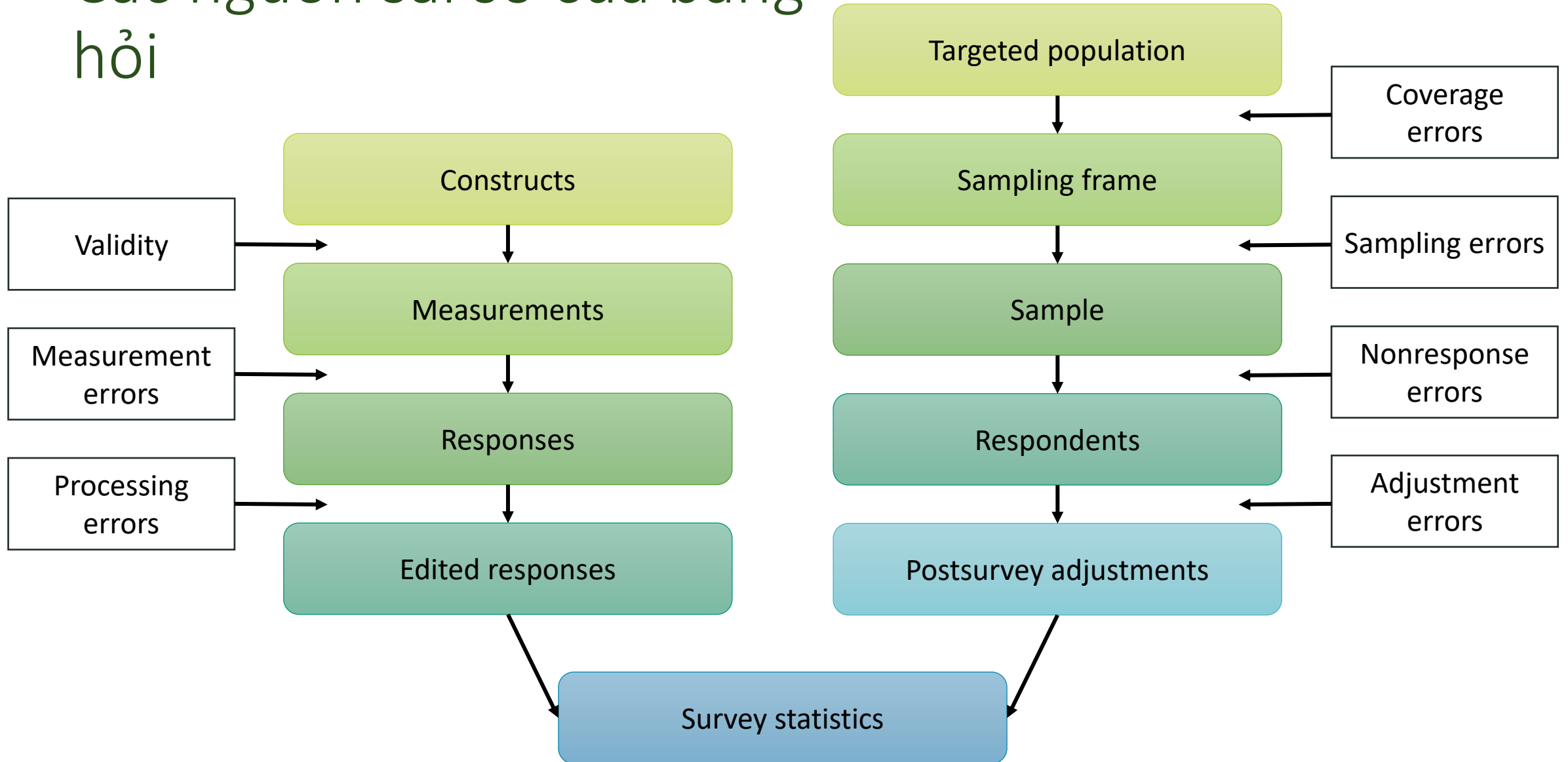
Table 2
Questionnaire organized by their respective factor.

Column	Data label	Explanation
<i>Attitude</i>		
Column I	Question 1	In my opinion, it is important to protect the environment.
Column J	Question 2	I actively practice environmental sustainability at home (e.g., energy conservation, recycling).
Column K	Question 3	Everyone is responsible for caring for the environment
Column L	Question 4	I am concerned about the long-term future of the environment.
Column M	Question 5	In my opinion, it is important to conserve natural resources.
Column N	Question 6	I think that environmental sustainability is a waste of time and effort.
Column O	Question 7	I am a passionate advocate of environmental sustainability.
<i>Perceived behavioral control</i>		
Column P	Question 8	It is easy for me to perform environmentally sustainable activities (e.g., energy conservation, recycling).
Column Q	Question 9	I have control over my actions to support the environment.
Column R	Question 10	It is my decision whether or not to perform environmentally sustainable activities.
Column S	Question 11	I have the ability to carry out environmentally sustainable activities.
Column T	Question 12	I have control over performing environmentally sustainable activities.

Thiết kế bảng hỏi

- Phạm trù (Construct) cần quan tâm
 - Là gì?
 - **Làm sao để đo?**
- Thiết kế bảng hỏi cần chú ý
 - Cách dùng từ
 - Tránh việc sử dụng chỉ một câu hỏi để đo lường cho 1 phạm trù
 - Tâm lý người hỏi và người trả lời
 - **Luôn luôn thử nghiệm trước** bộ câu hỏi

Các nguồn sai số của bảng hỏi

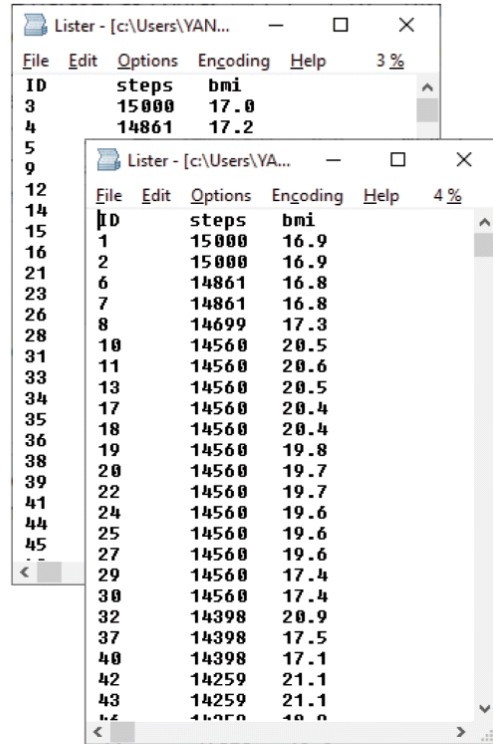


Phân tích dữ liệu thăm dò

Tìm hiểu dữ liệu/Biểu diễn dữ liệu

- Là bước không thể bỏ qua
- Giúp phát hiện những vấn đề trong dữ liệu
- Giúp có hình dung chung về dữ liệu và các mối tương quan giữa các dữ liệu
- Phát triển giả thuyết, và lý thuyết mới

a



The screenshot shows a Notepad window titled 'Lister - [c:\Users\YAN...]' with a menu bar (File, Edit, Options, Encoding, Help) and a status bar (3 %). The text content is a list of data points with three columns: ID, steps, and bmi. The data is as follows:

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
9		
12		
14		
15	1	15000
16	2	15000
21	6	14861
23	7	14861
26	8	14699
28	10	14560
31	11	14560
33	13	14560
34	17	14560
35	18	14560
36	19	14560
38	20	14560
39	22	14560
41	24	14560
44	25	14560
45	27	14560
	29	14560
	30	14560
	32	14398
	37	14398
	40	14398
	42	14259
	43	14259
	44	14259

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

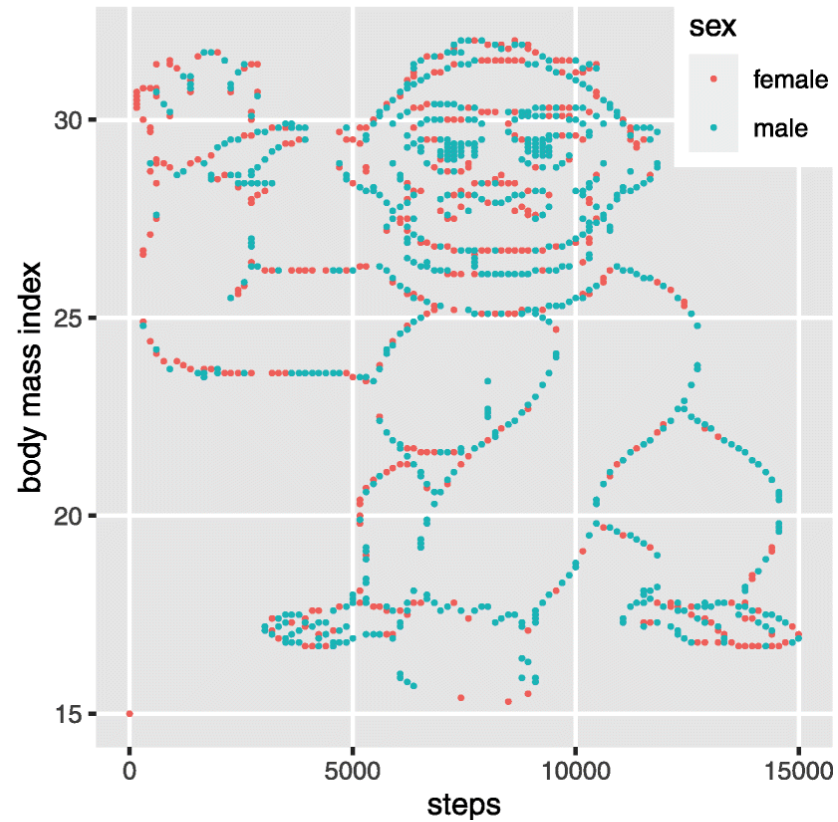
Listner - [c:\Users\YAN...]

ID	steps	bmi
3	15000	17.0
4	14861	17.2

Listner - [c:\Users\YA...]

ID	steps	bmi
1	15000	16.9
2	15000	16.9
6	14861	16.8
7	14861	16.8
8	14699	17.3
10	14560	20.5
11	14560	20.6
13	14560	20.5
17	14560	20.4
18	14560	20.4
19	14560	19.8
20	14560	19.7
22	14560	19.7
24	14560	19.6
25	14560	19.6
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1
44	14259	21.1
45	14259	21.1

b



Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

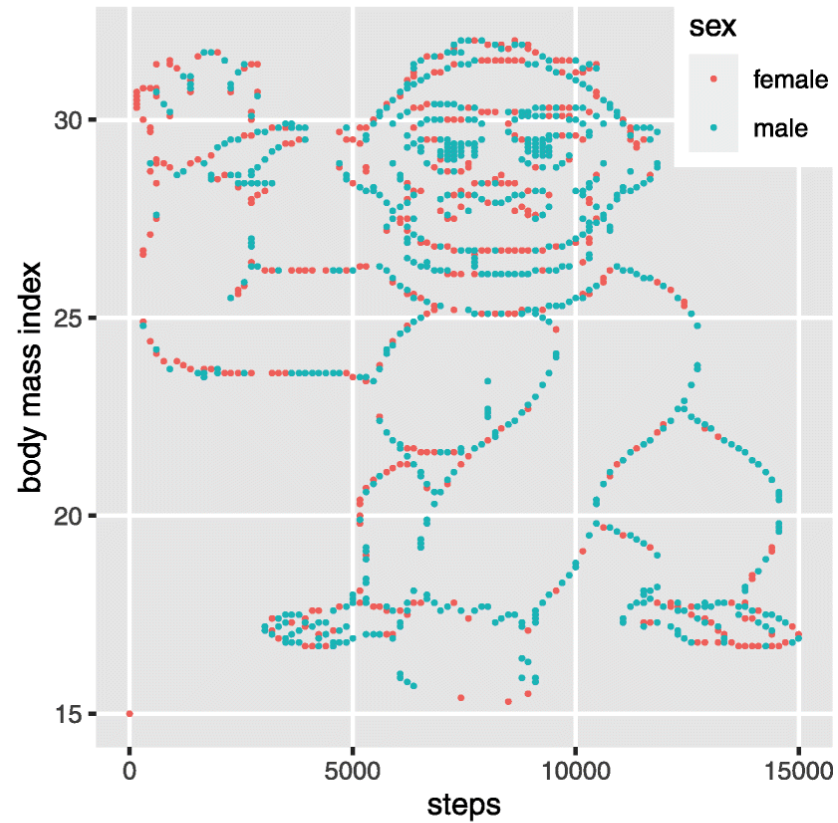
Listser - [c:\Users\YAN...]

ID	steps	bmi
3	15000	17.0
4	14861	17.2

Listser - [c:\Users\YA...]

ID	steps	bmi
1	15000	16.9
2	15000	16.9
6	14861	16.8
7	14861	16.8
8	14699	17.3
10	14560	20.5
11	14560	20.6
13	14560	20.5
17	14560	20.4
18	14560	20.4
19	14560	19.8
20	14560	19.7
22	14560	19.7
24	14560	19.6
25	14560	19.6
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1
44	14259	21.1
45	14259	21.1

b

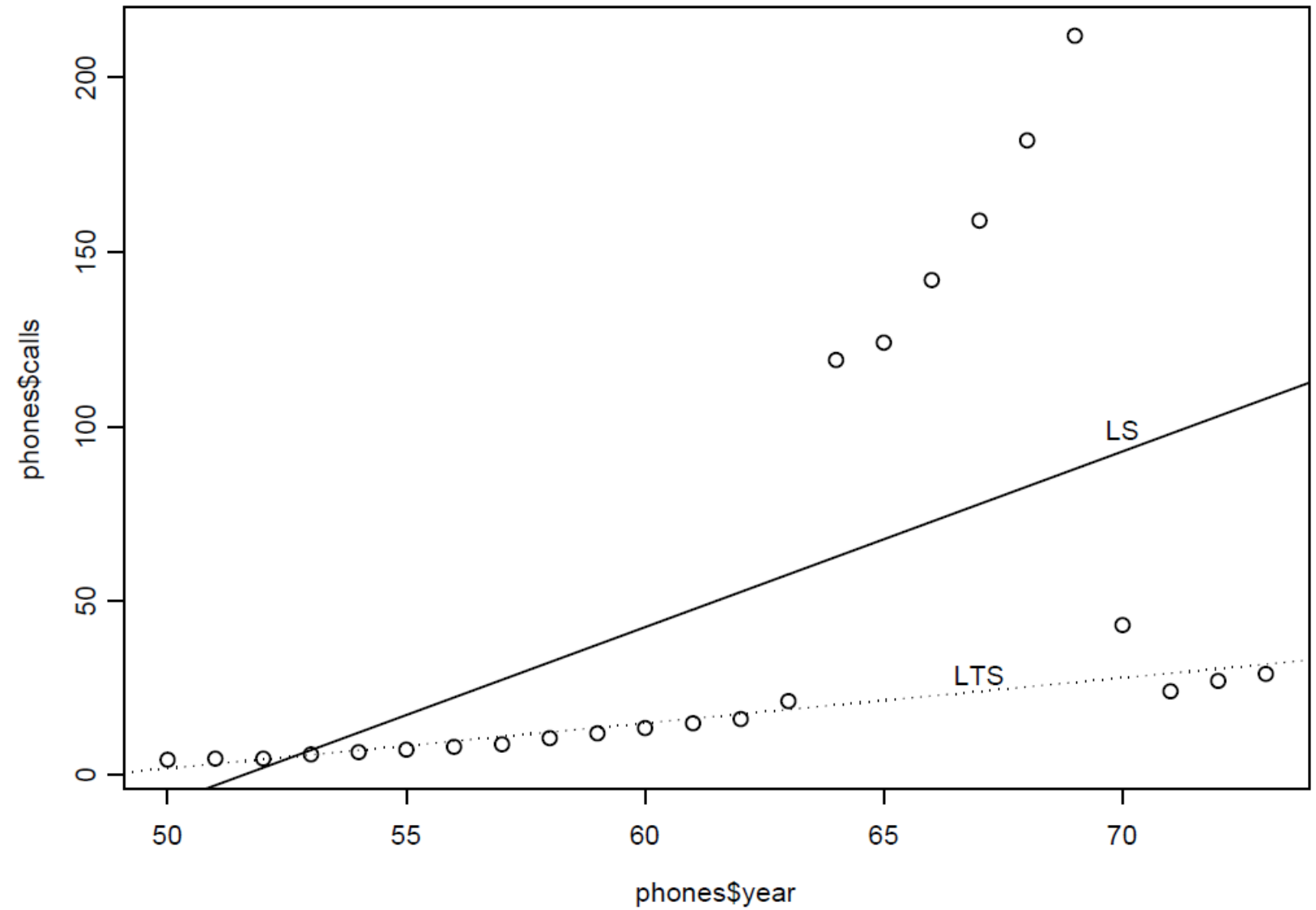


c

	Gorilla <u>not</u> discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

- Dữ liệu điện thoại
- Cuộc gọi (triệu) ra nước ngoài từ Bỉ từ 1950-1973.



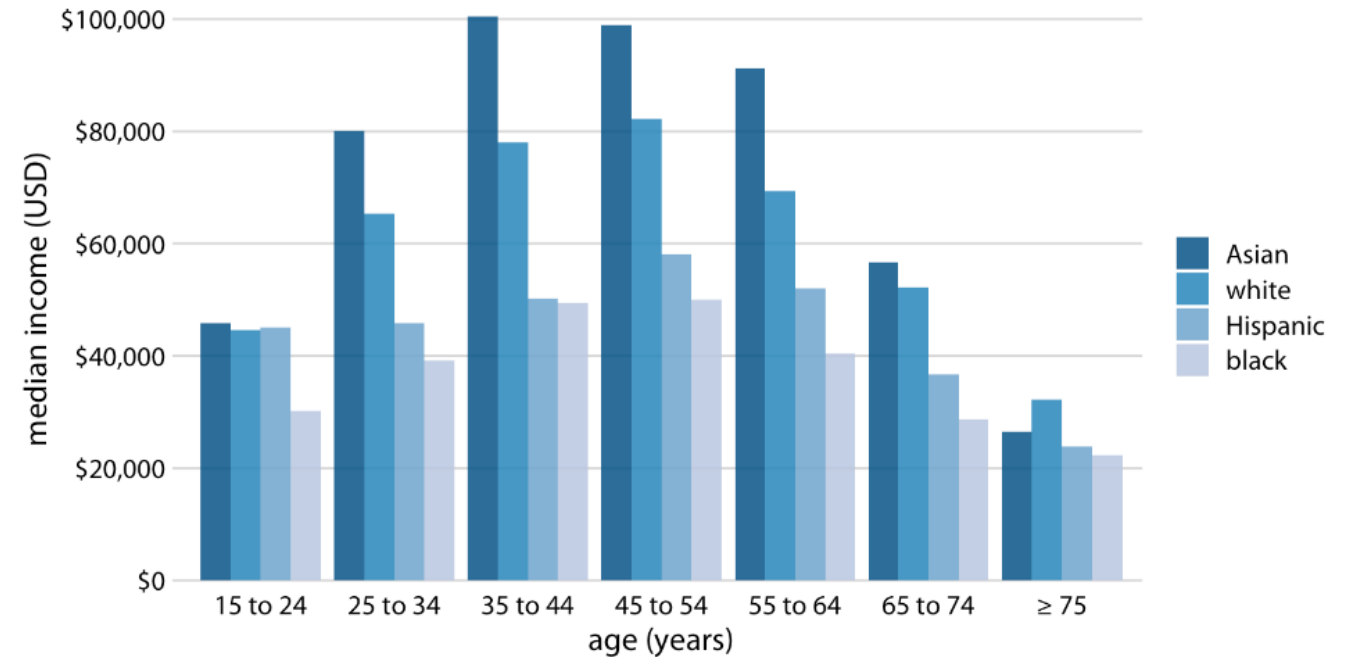
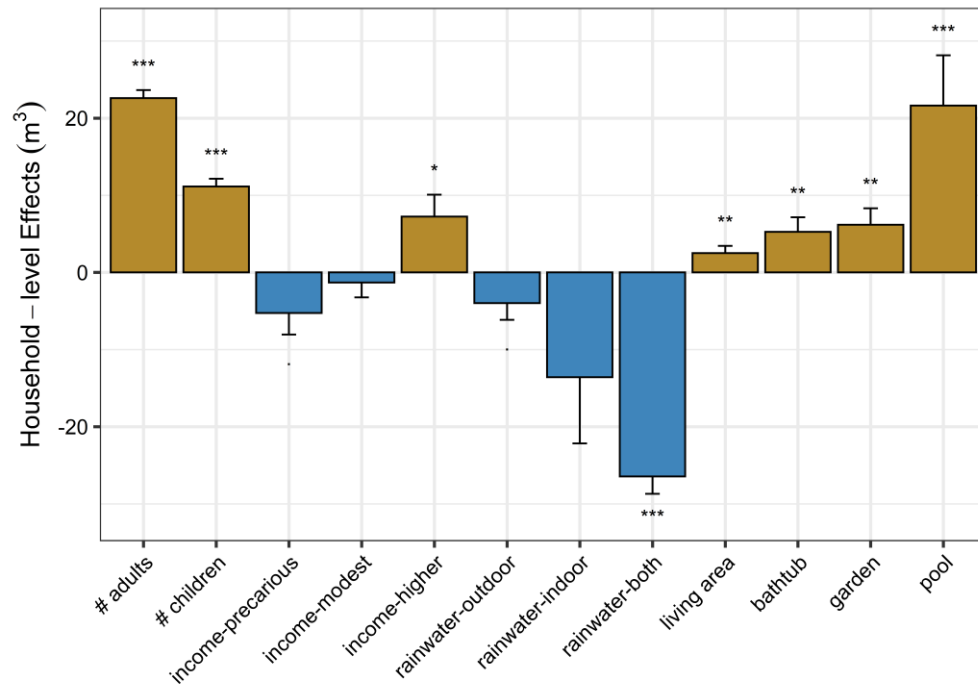
Đồ thị

- Rõ ràng
- Chính xác
- Hiệu quả
- Tối đa thông tin, tối thiểu mực in

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

Một số loại đồ thị thông thường

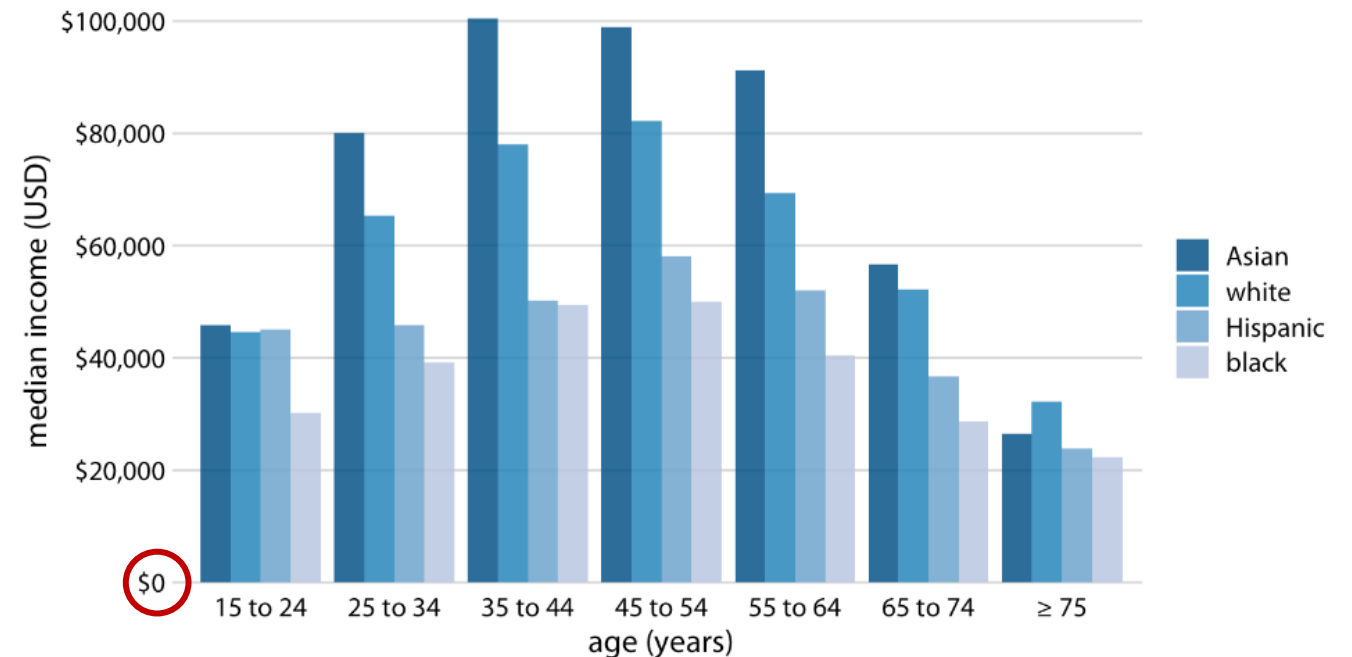
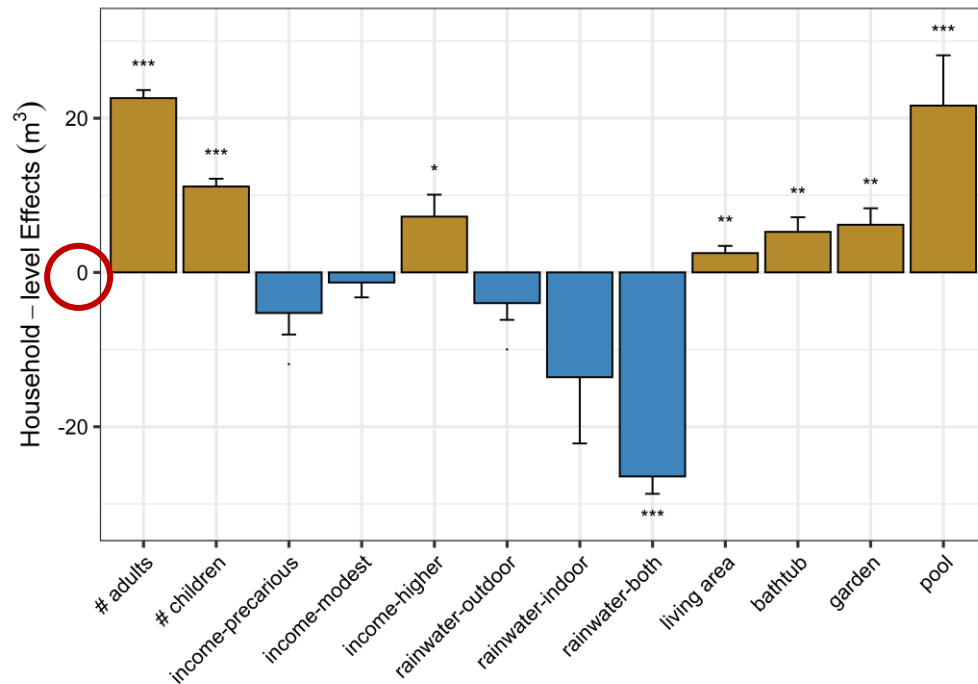
Đồ thị cột – bar charts



Wilke (2018)

Một số loại đồ thị thông thường

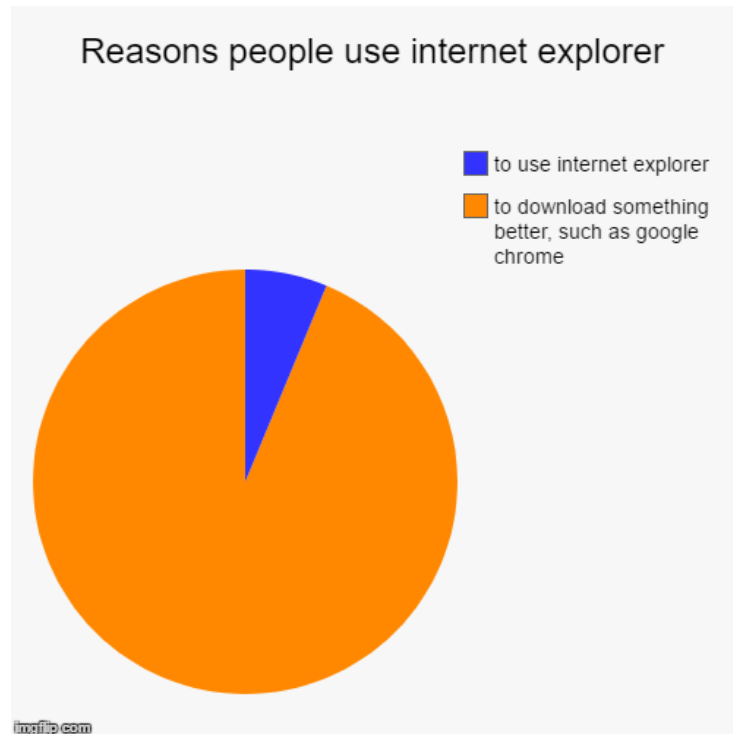
Đồ thị cột – bar charts



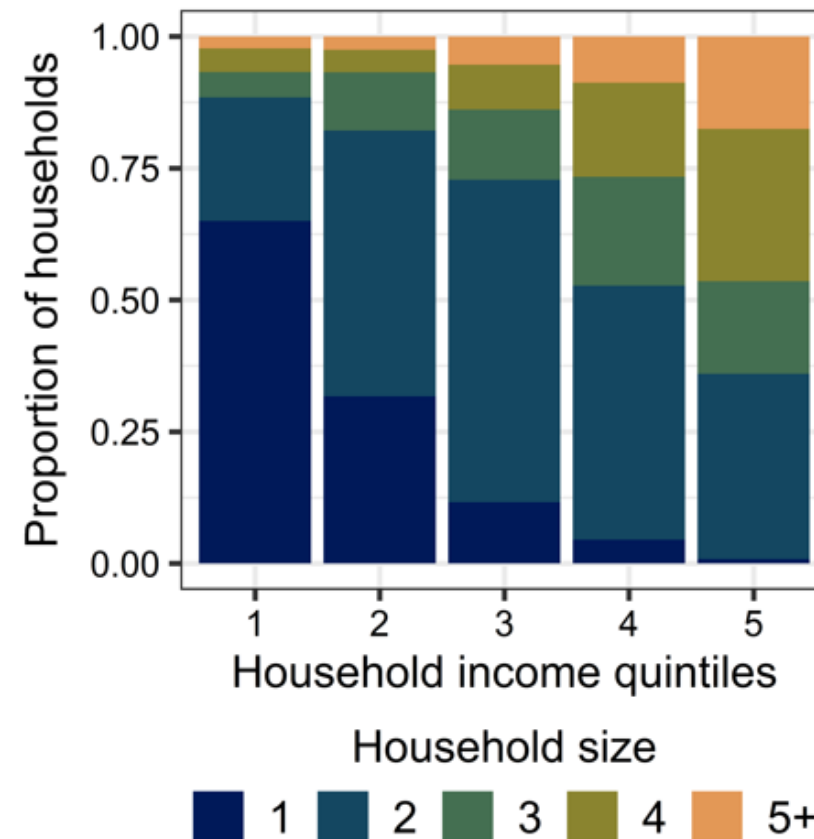
Wilke (2018)

Một số loại đồ thị thông thường

Đồ thị quạt – pie chart

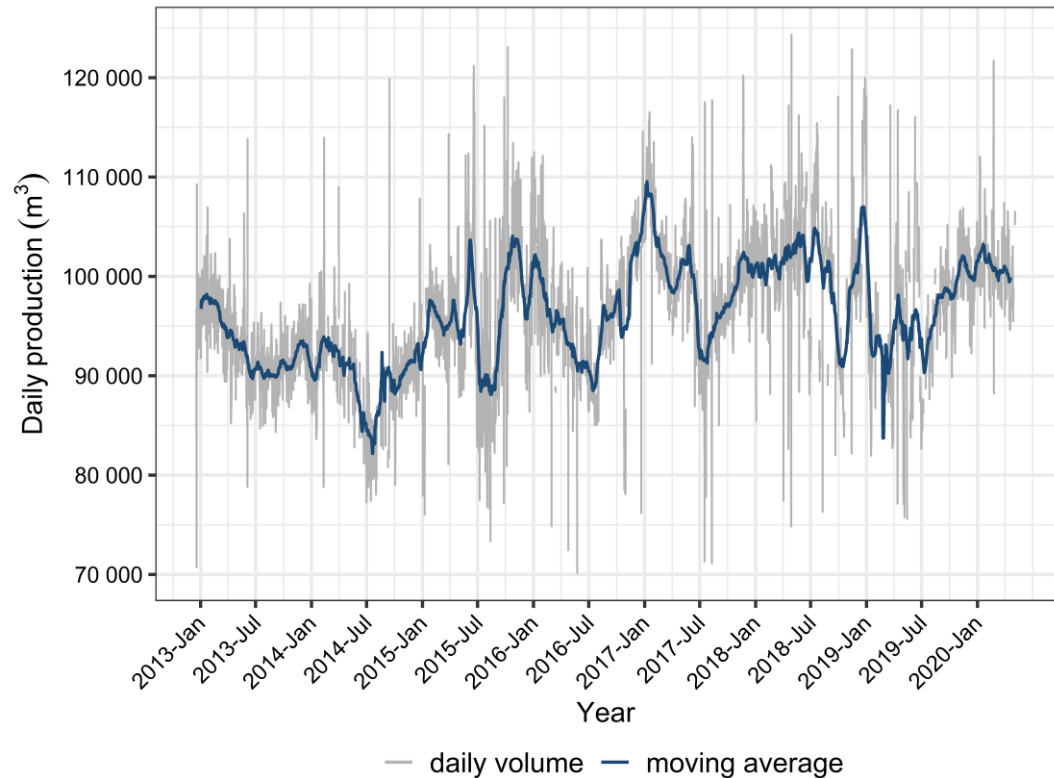
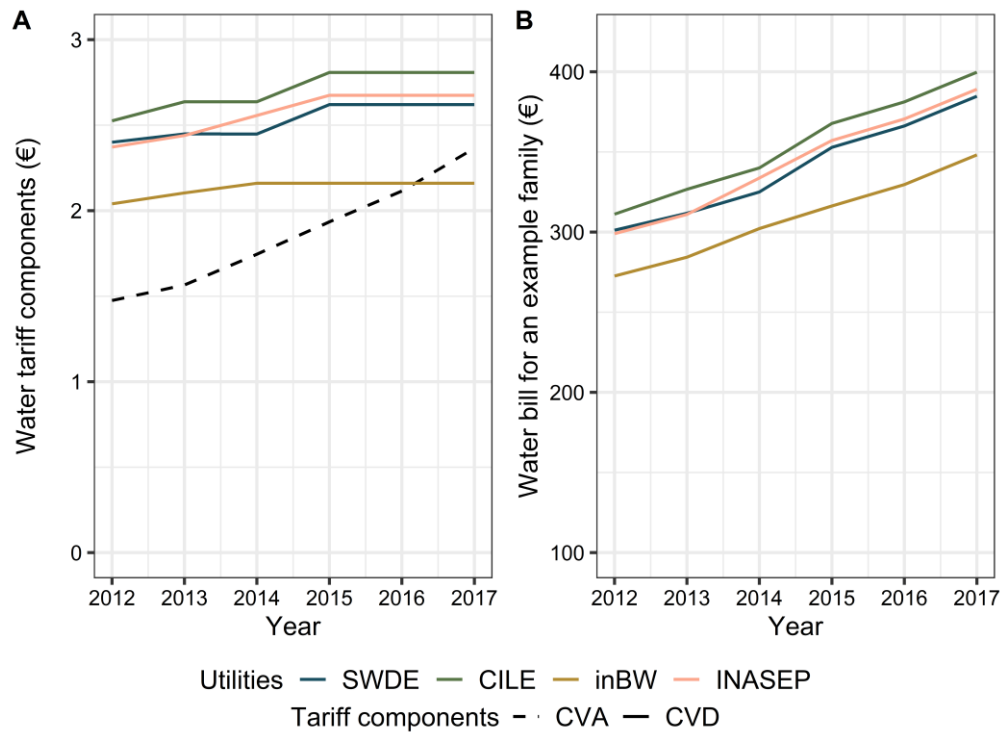


Đồ thị cột xếp chồng – Stacked bar charts



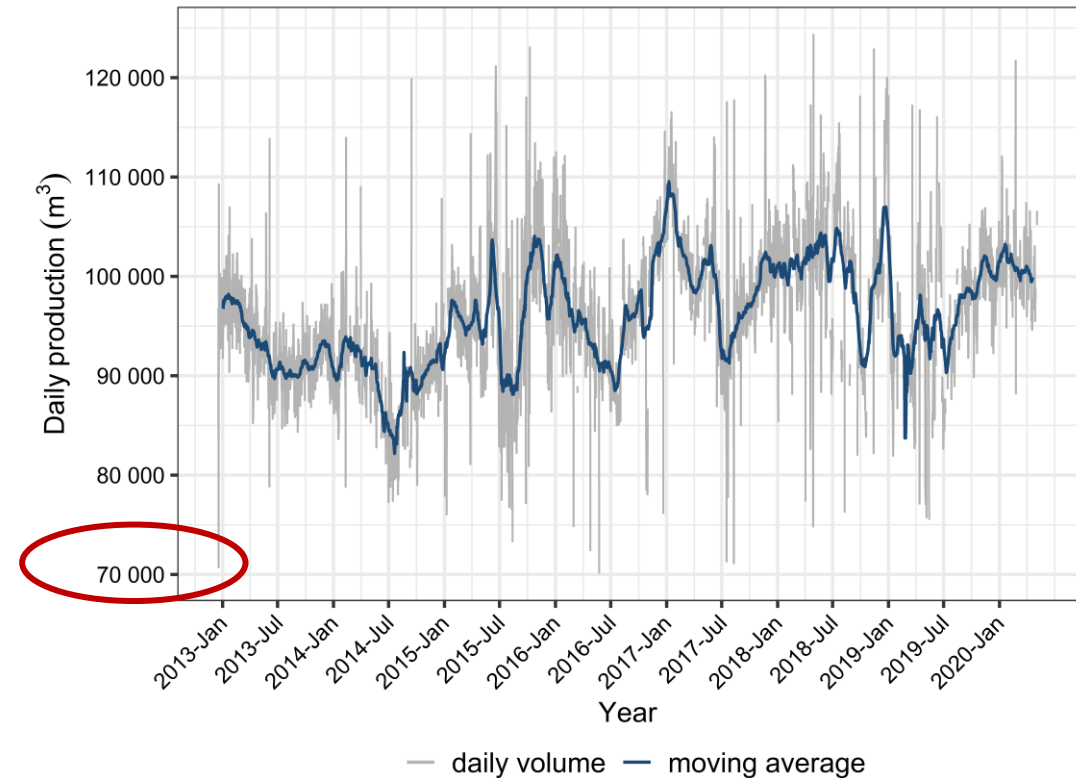
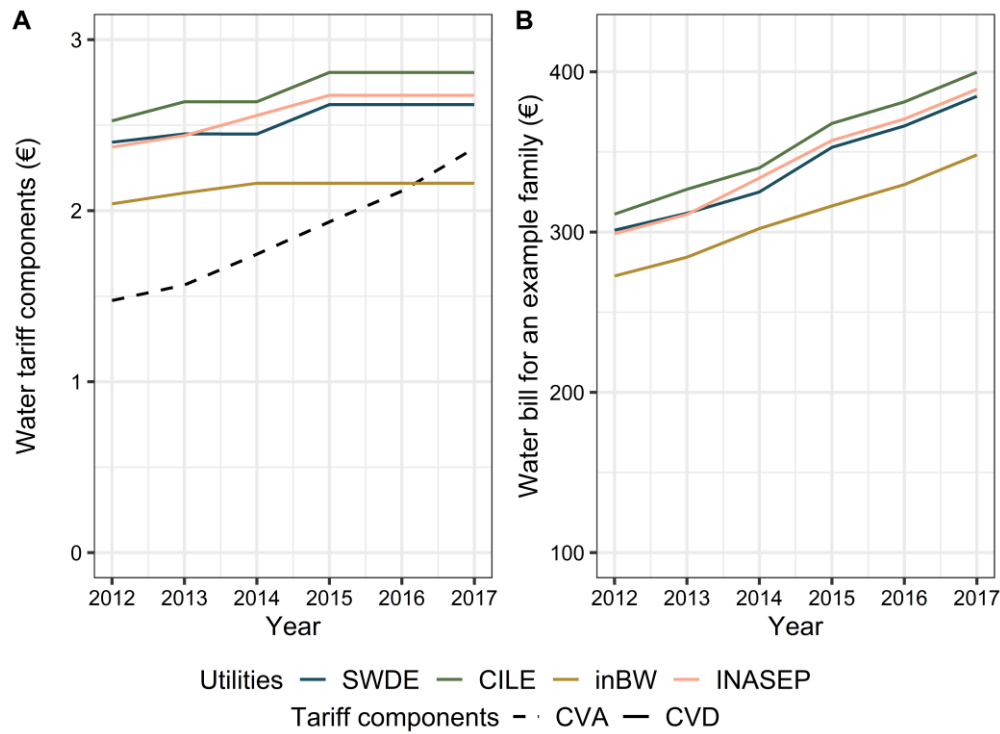
Một số loại đồ thị thông thường

Đồ thị đường – line charts



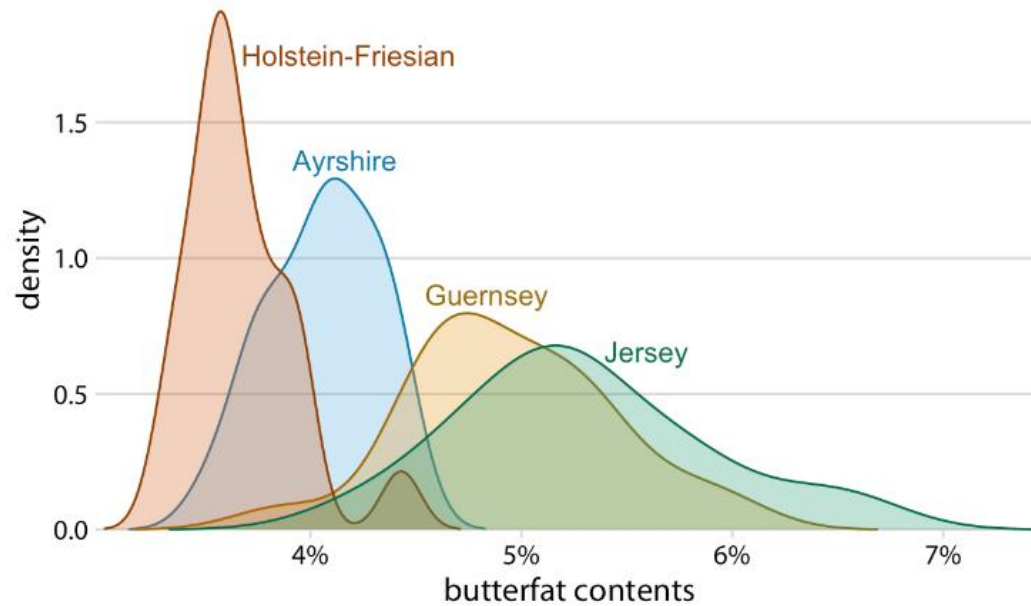
Một số loại đồ thị thông thường

Đồ thị đường – line charts

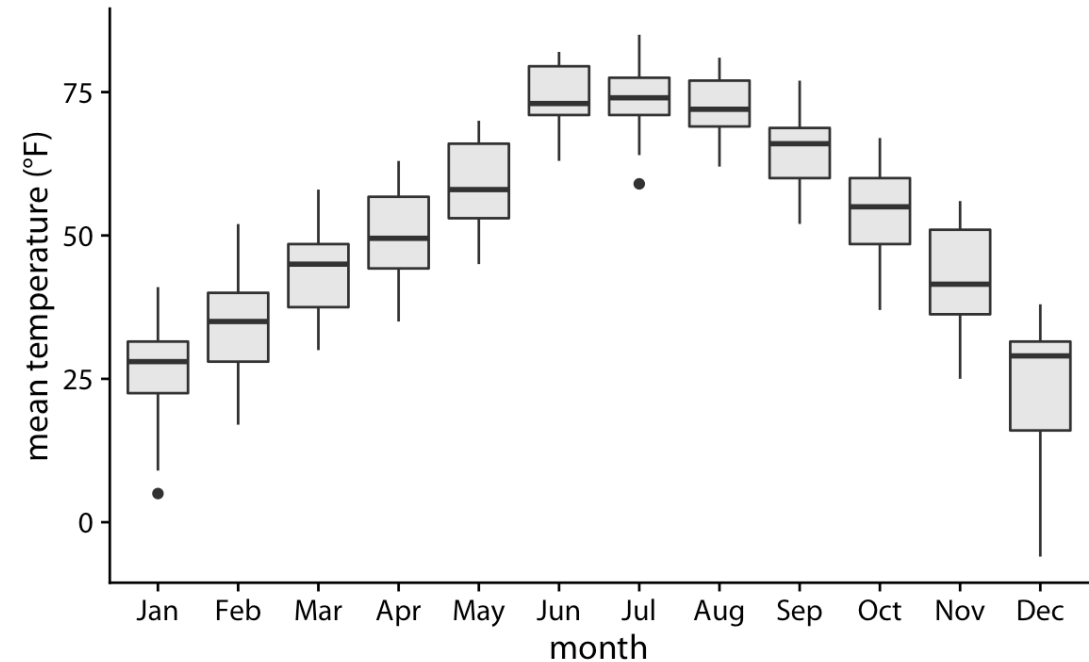


Một số loại đồ thị thông thường

Biểu đồ tần suất - Histogram



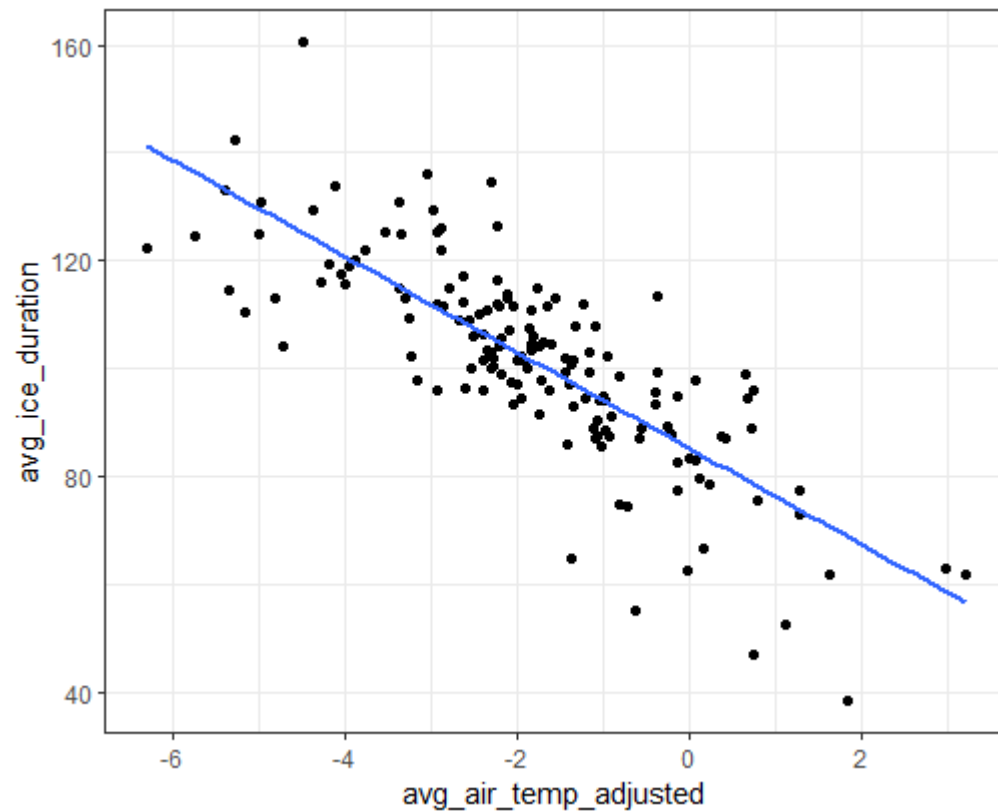
Đồ thị hộp – Box plots



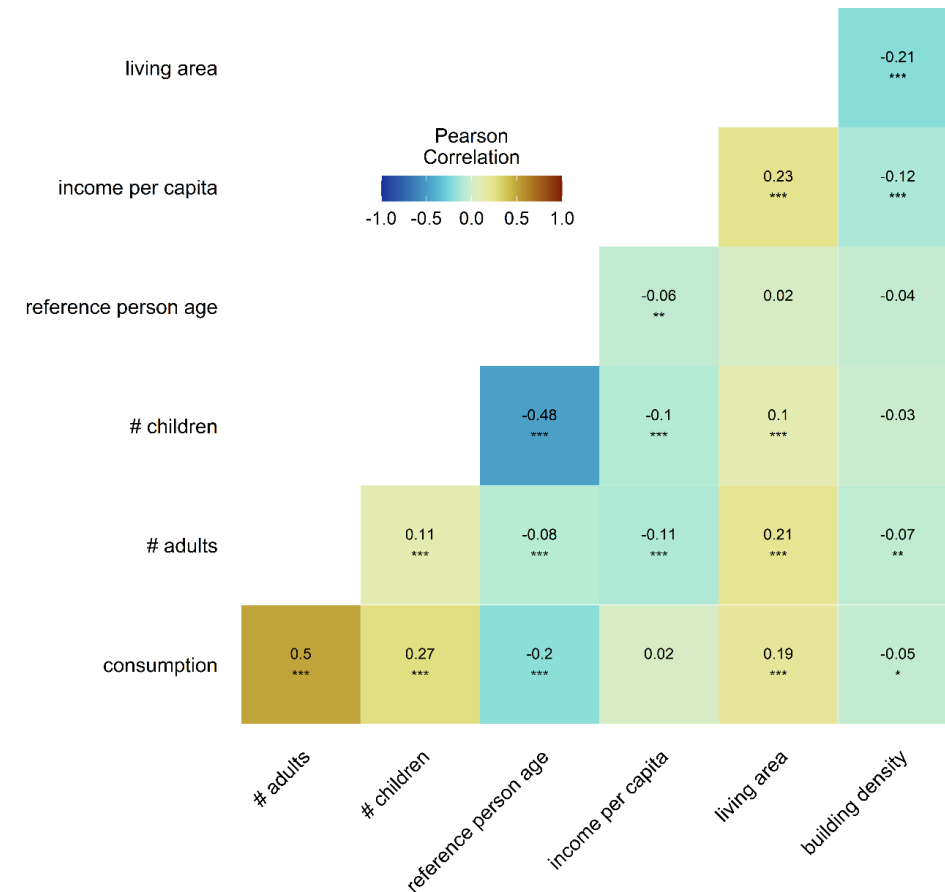
Source Wilke (2018)

Một số loại đồ thị thông thường

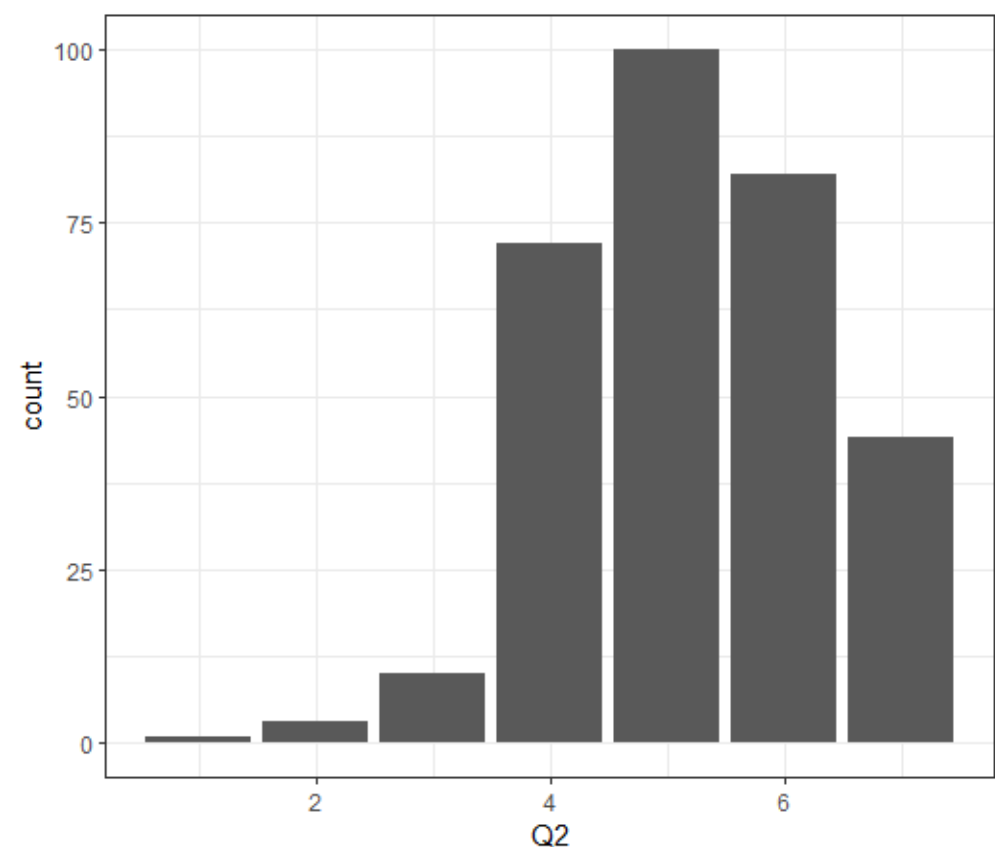
Biểu đồ phân tán - Scatter plots



Bản đồ nhiệt - heatmap



Thống kê mô tả - biến gián đoạn



Question 2

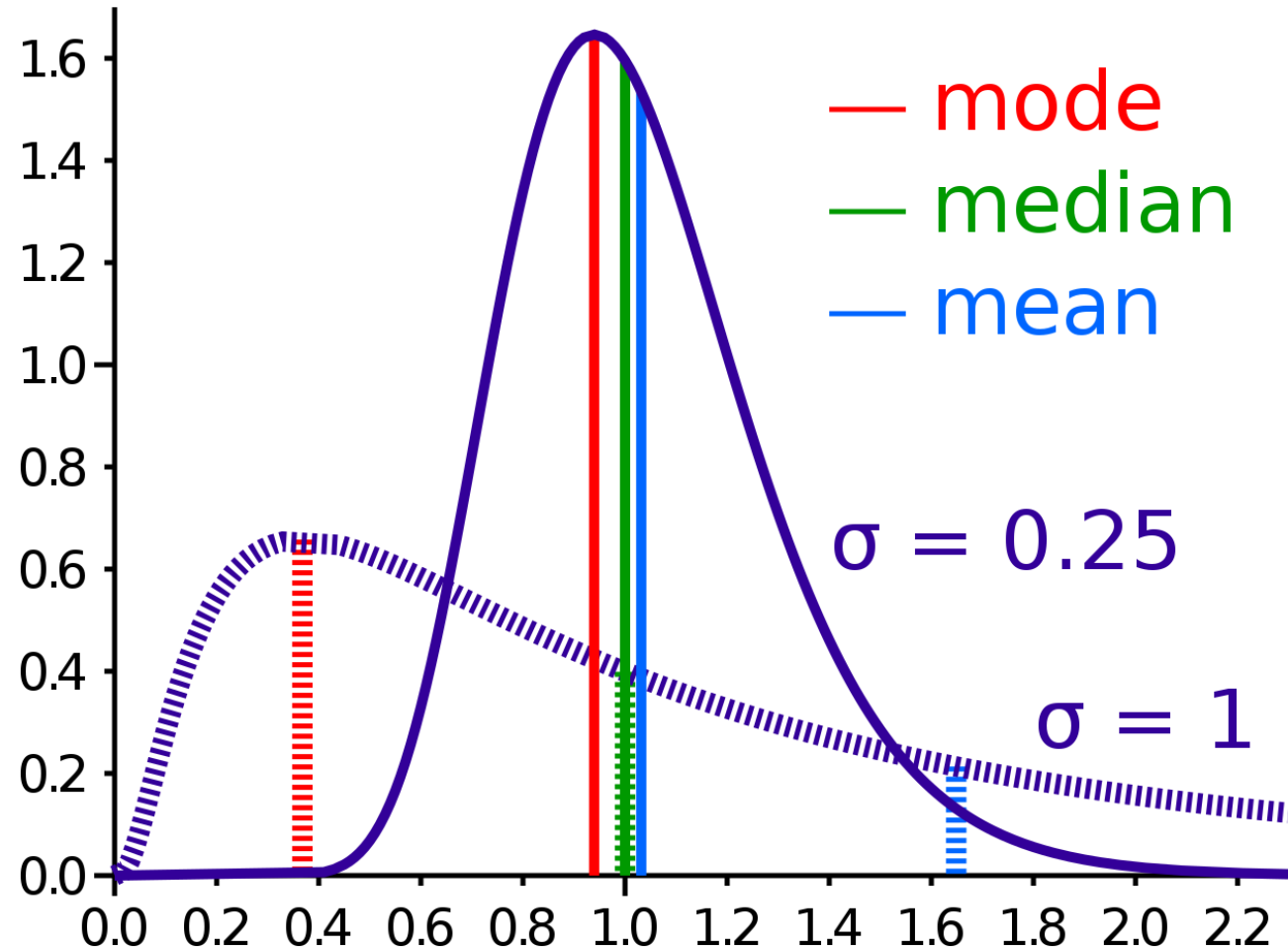
I actively practice environmental sustainability at home (e.g., energy conservation, recycling).

	Q2	count	percent
1	1	1	0.3205128
2	2	3	0.9615385
3	3	10	3.2051282
4	4	72	23.0769231
5	5	100	32.0512821
6	6	82	26.2820513
7	7	44	14.1025641

Thống kê mô tả - Biến liên tục

Đo độ tập trung

Trung bình, trung vị, mode

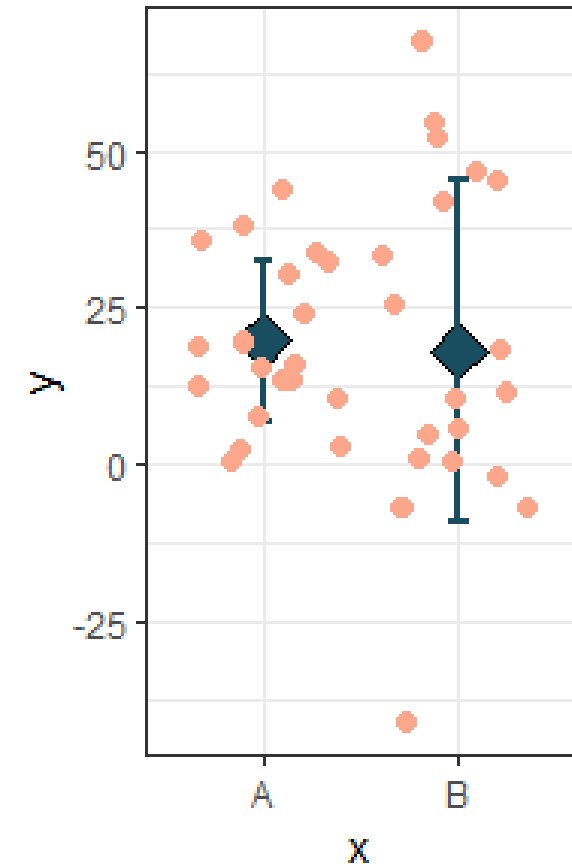
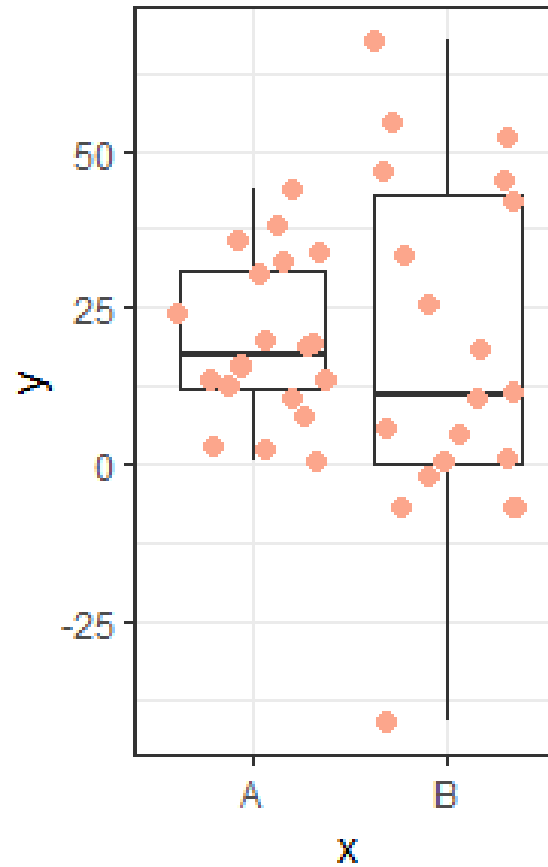


[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Thống kê mô tả - Biến liên tục

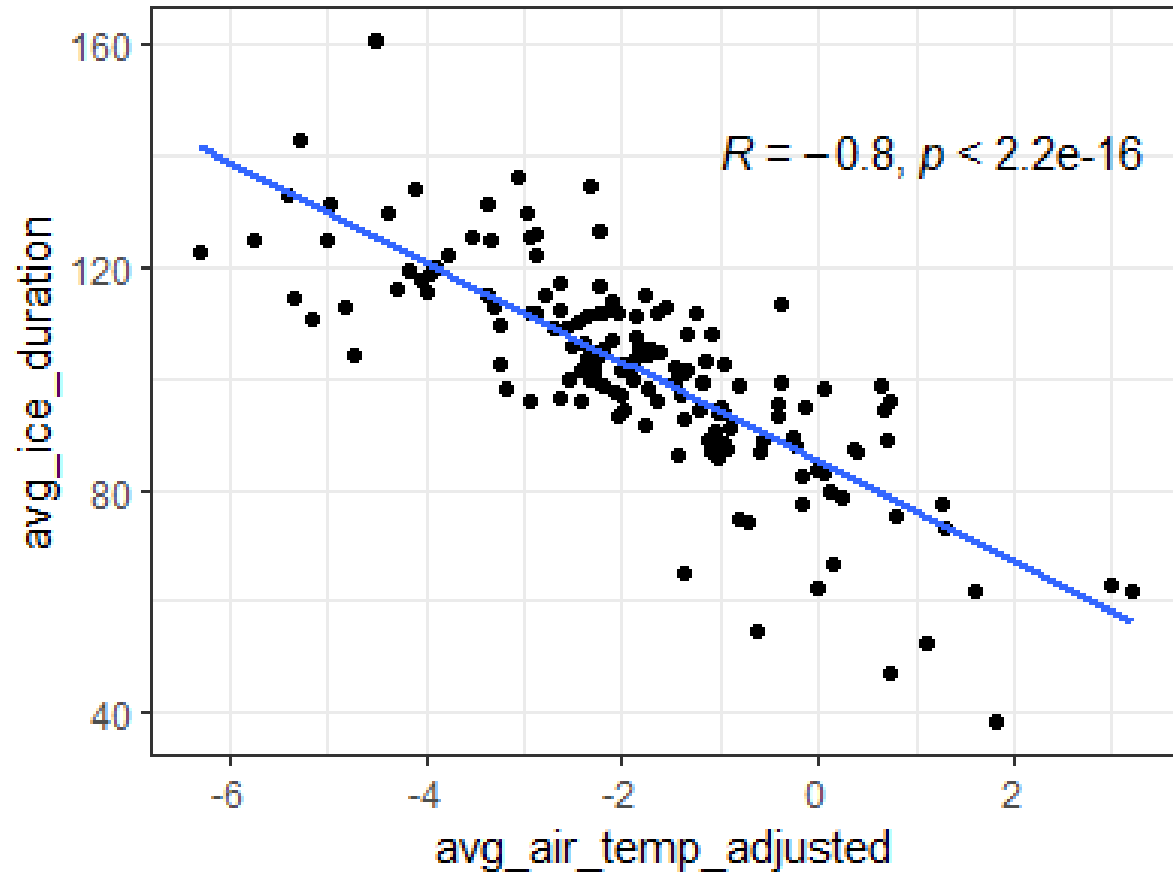
Đo độ phân tán

- Phương sai (Variance)
- Độ lệch chuẩn (Standard deviation)
- Phân vị (Quantile)
- Điểm tứ phân vị (Quartile)
- Giá trị tối thiểu/tối đa (min/max)

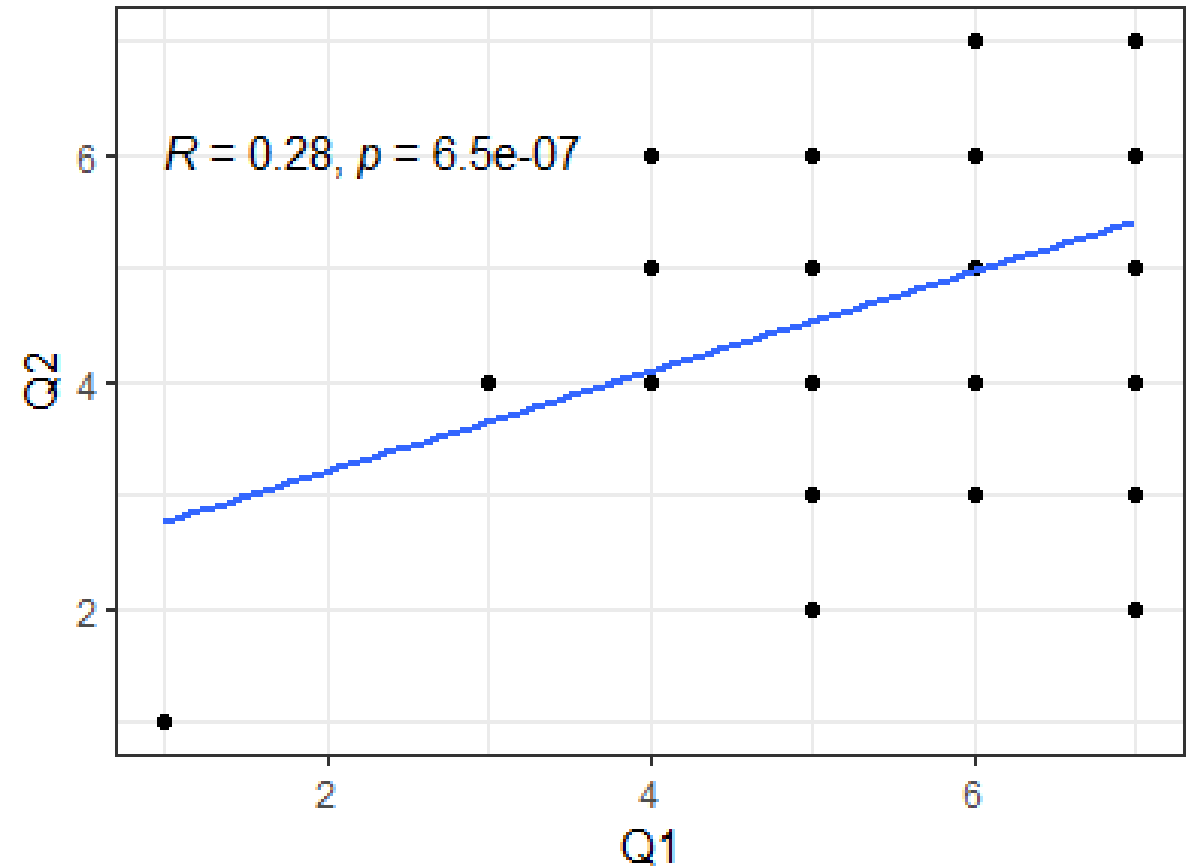


Thống kê mô tả - tương quan

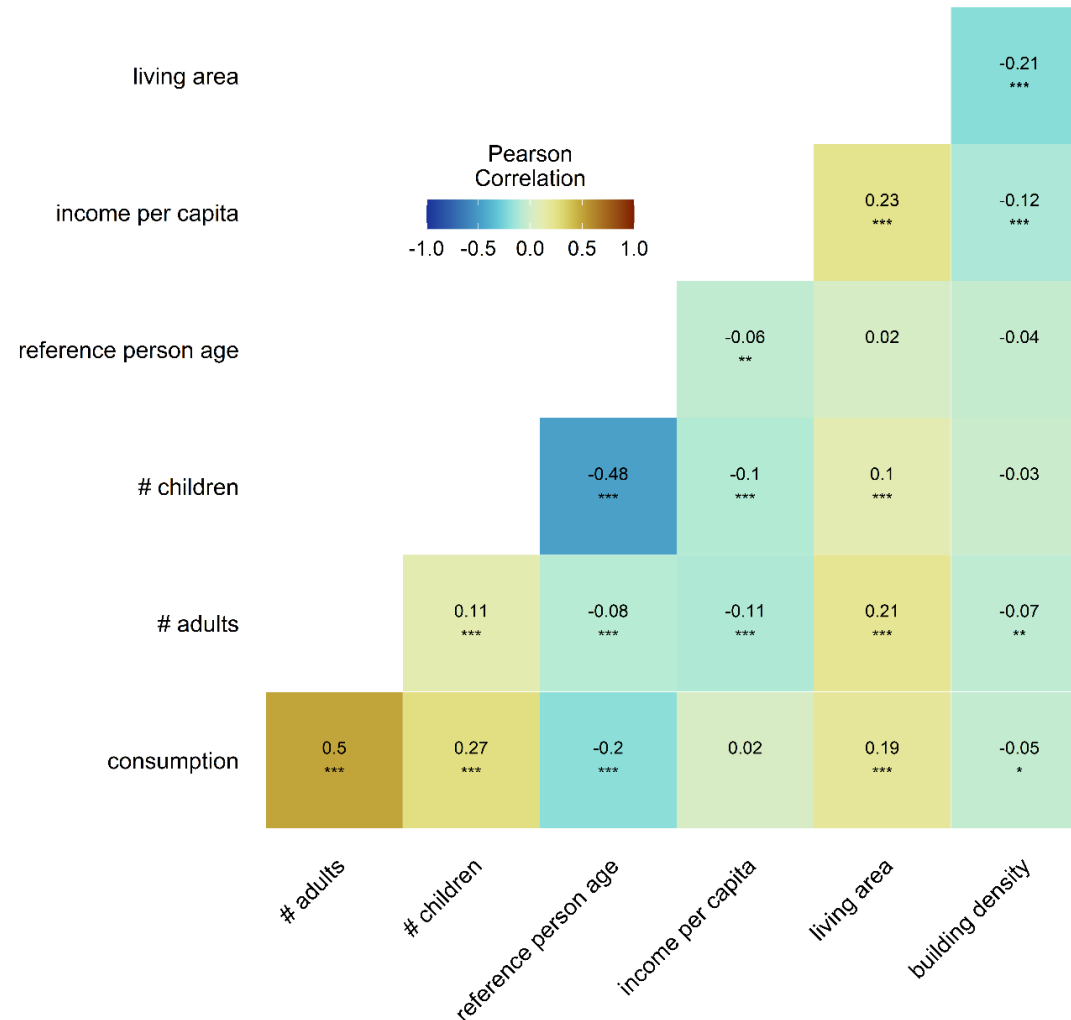
Pearson



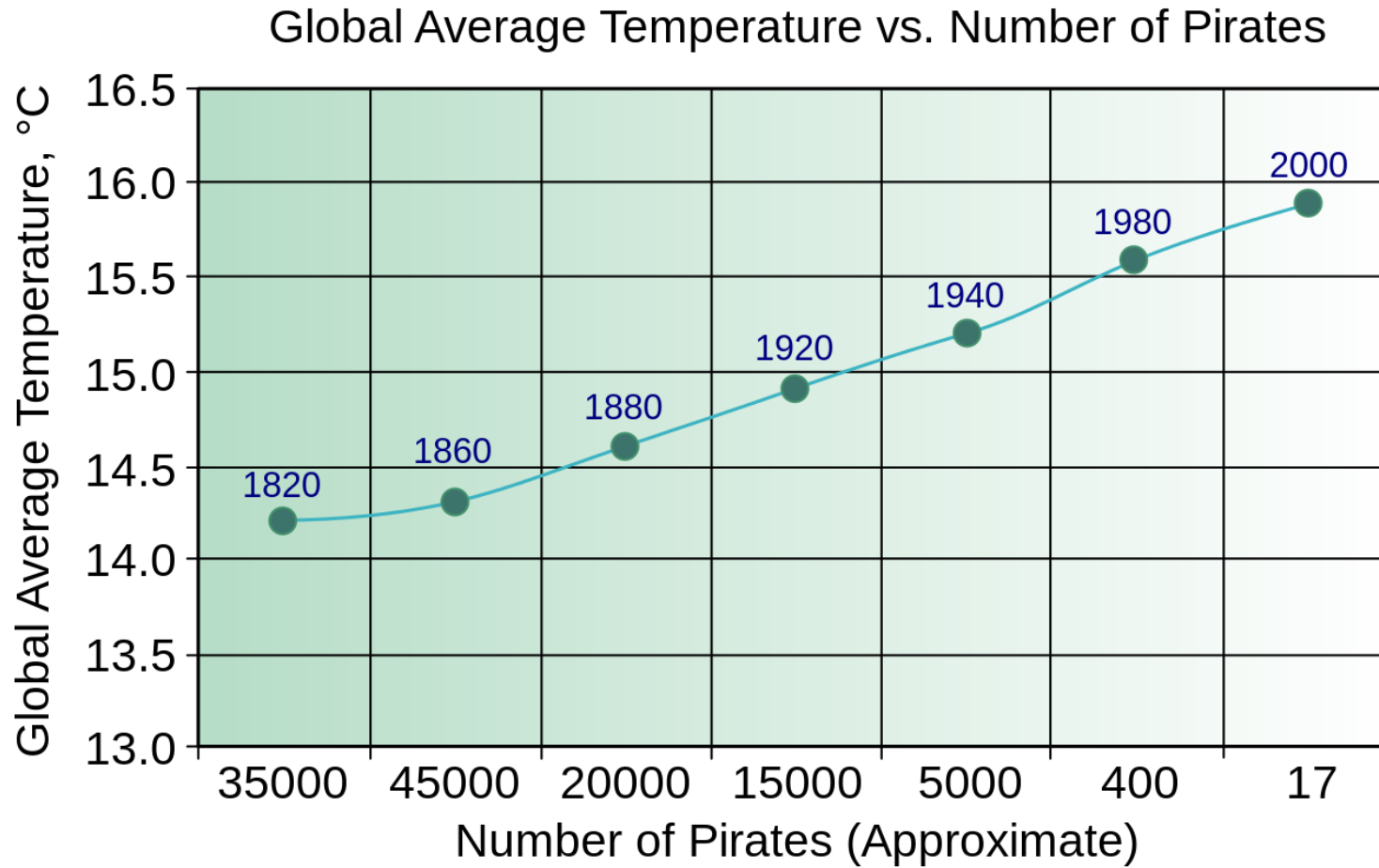
Spearman



Thống kê mô tả - Tương quan



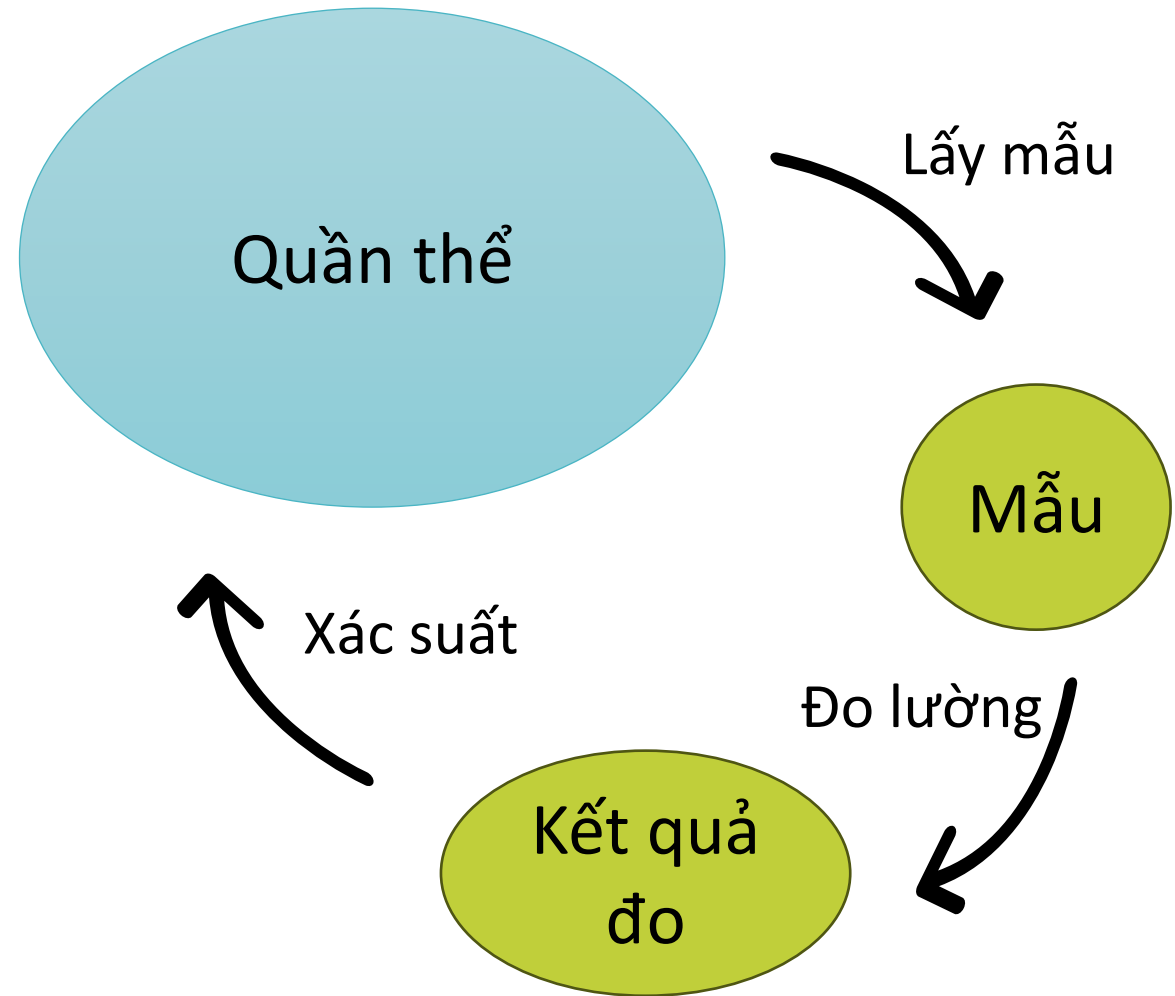
Tương quan và quan hệ nhân quả



Phân tích dữ liệu khẳng định

Thống kê suy luận

- Dùng thông tin về mẫu để suy luận về quần thể
- Kiểm định các giả thuyết (hypothesis) nghiên cứu
- Đưa ra kết luận về mối quan hệ giữa các biến



Kiểm định thống kê (Hypothesis testing)

- Giả thuyết trống/không (Null hypothesis)
- Giả thuyết thay thế (Alternative hypothesis)

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

- Giả thuyết thú vị với nhà nghiên cứu luôn là giả thuyết thay thế
- Giả thuyết được kiểm định luôn là giả thuyết không/trống

Bài tập

- Mức độ tham gia của cộng đồng trong việc bảo tồn các di sản vật thể ảnh hưởng tích cực đến cảm giác về nơi chốn và bản sắc của cư dân địa phương
- Việc thêm các bài giảng về di sản phi vật thể vào trong trường học giúp tăng cường đối thoại và sự tôn trọng đối với các dân tộc khác nhau
- Việc đô thị hóa quá nhanh dẫn đến gia tăng chênh lệch giàu nghèo tại các đô thị

Giá trị p

- Xác suất để thu được kết quả tương tự hoặc cực đoan hơn khi giả thiết rằng giả thuyết trống là đúng
- Giá trị $p < 0.05$: có ý nghĩa về mặt thống kê

	H_0 đúng	H_0 sai
Bác bỏ H_0	Lỗi loại I	✓
Không bác bỏ H_0	✓	Lỗi loại II

- Ý nghĩa về mặt thống kê vs. ý nghĩa thực tế

Kiểm định t (t-test)

- So sánh 2 giá trị trung bình

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

- Giả định
 - Liên tục/định lượng
 - Độc lập
 - Phân phối chuẩn
 - Độ phân tán tương đương
- Trường hợp đặc biệt: Kiểm định t ghép cặp

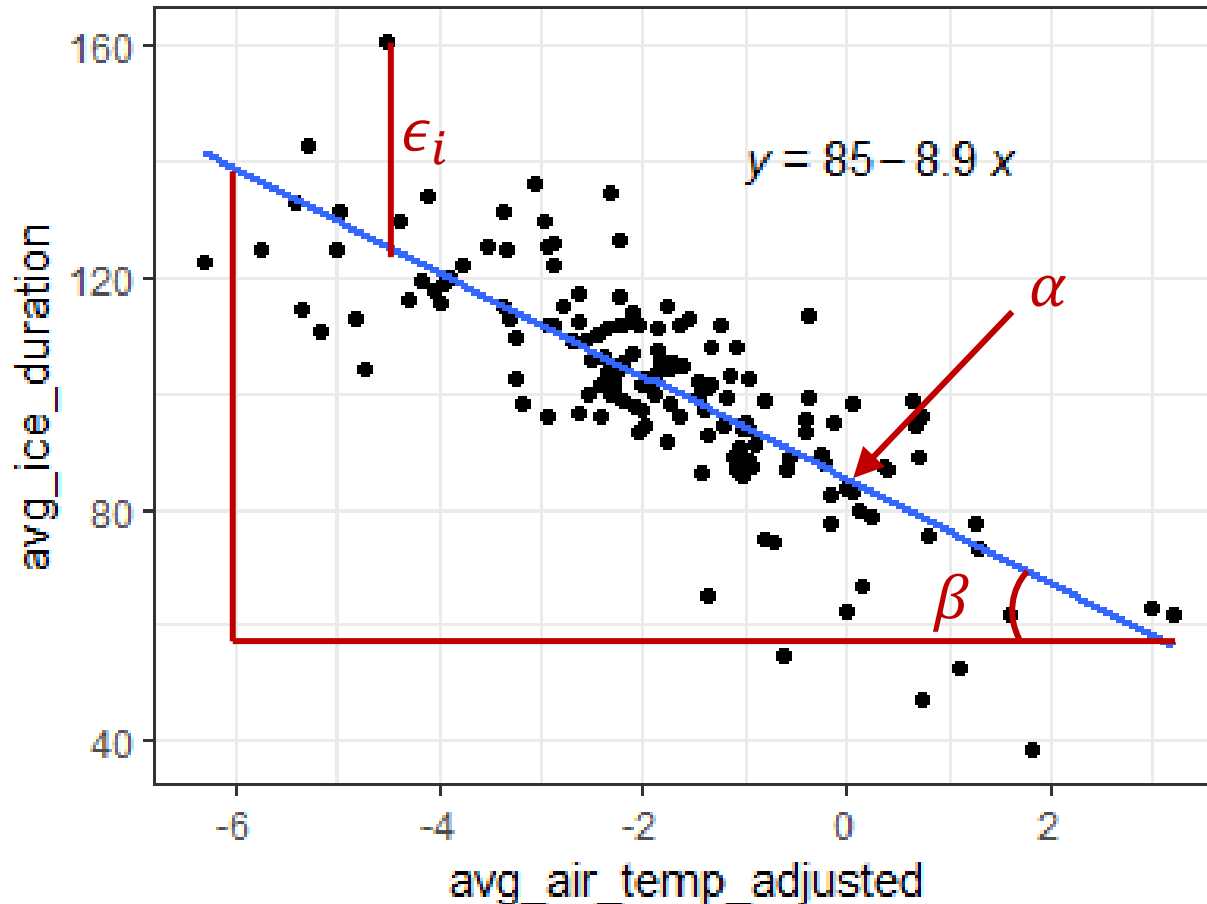
ANOVA

- So sánh giá trị trung bình của nhiều hơn 2 nhóm/điều kiện
 - VD: Chi phí tiền điện của 3 nhóm hộ gia đình thu nhập thấp, trung bình, cao

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_a : ít nhất 1 nhóm có giá trị trung bình khác các nhóm còn lại

Hồi quy tuyến tính đơn giản



$$Y_i = \alpha + \beta X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- Y biến phụ thuộc (dependent/out come variable)
- X biến độc lập (independent/explanatory variable, predictor)
- ϵ phần dư, sai số ngẫu nhiên (residuals, random errors)

Hồi quy tuyến tính đơn giản

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	85.186	1.363	62.50	<2e-16	***
avg_air_temp_adjusted	-8.903	0.552	-16.13	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

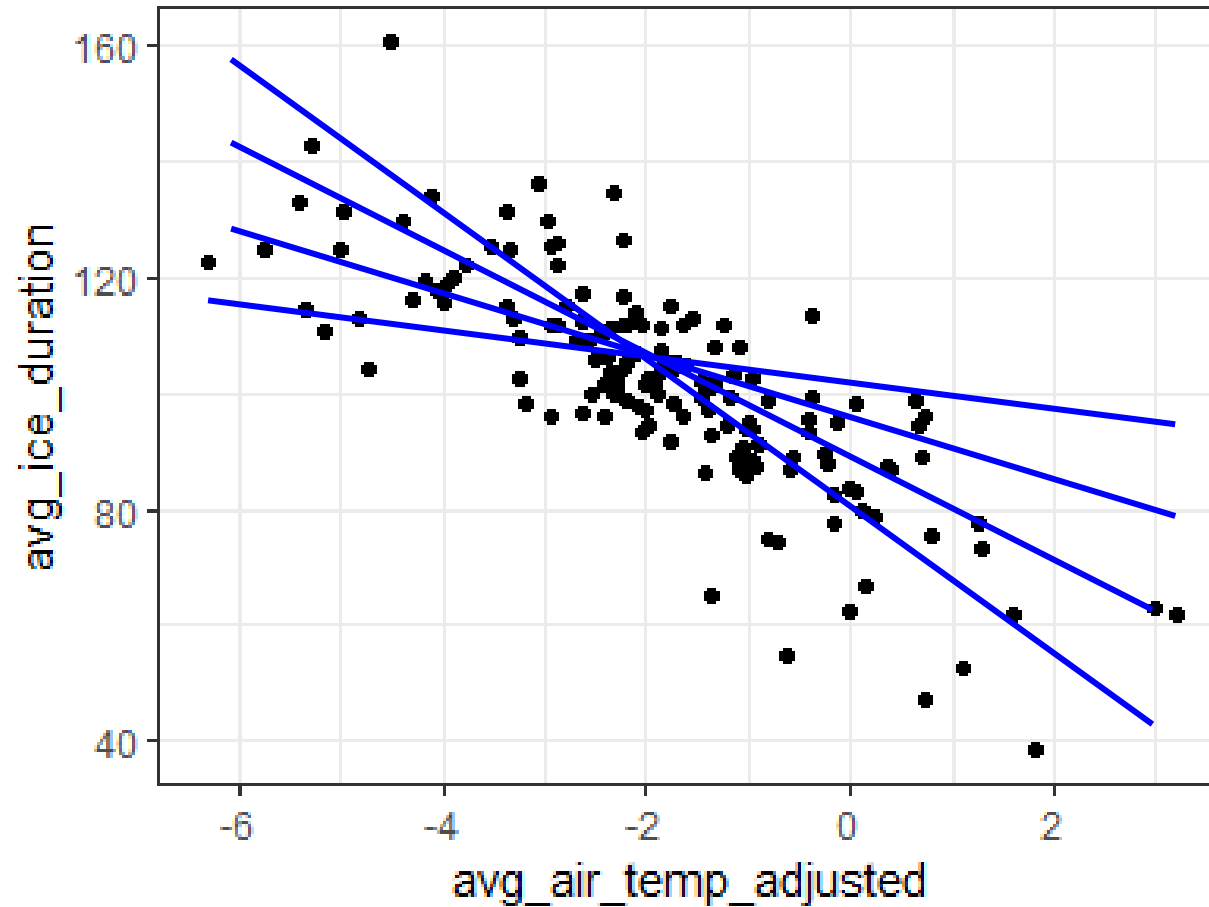
Residual standard error: 11.43 on 150 degrees of freedom

(14 observations deleted due to missingness)

Multiple R-squared: 0.6343, Adjusted R-squared: 0.6319

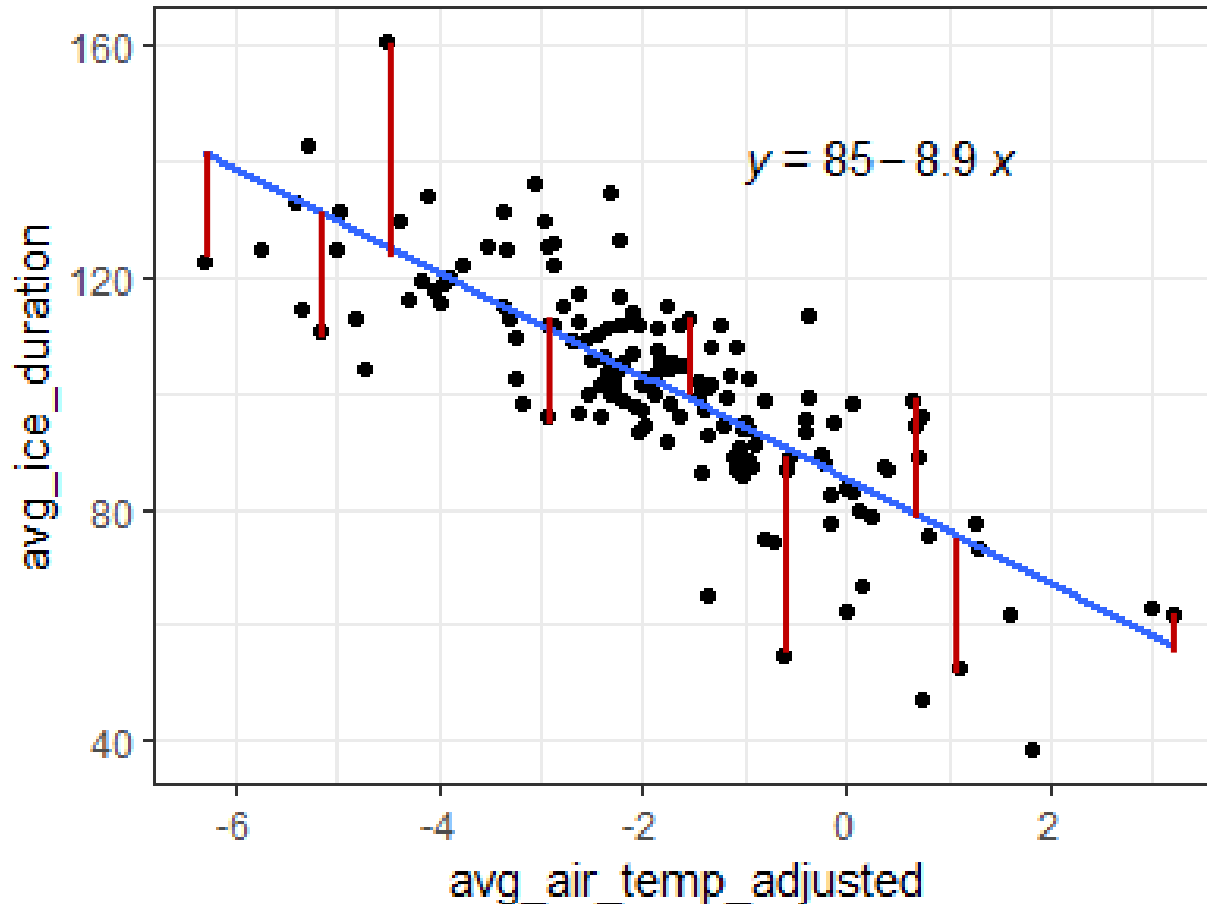
F-statistic: 260.2 on 1 and 150 DF, p-value: < 2.2e-16

Hồi quy tuyến tính đơn giản



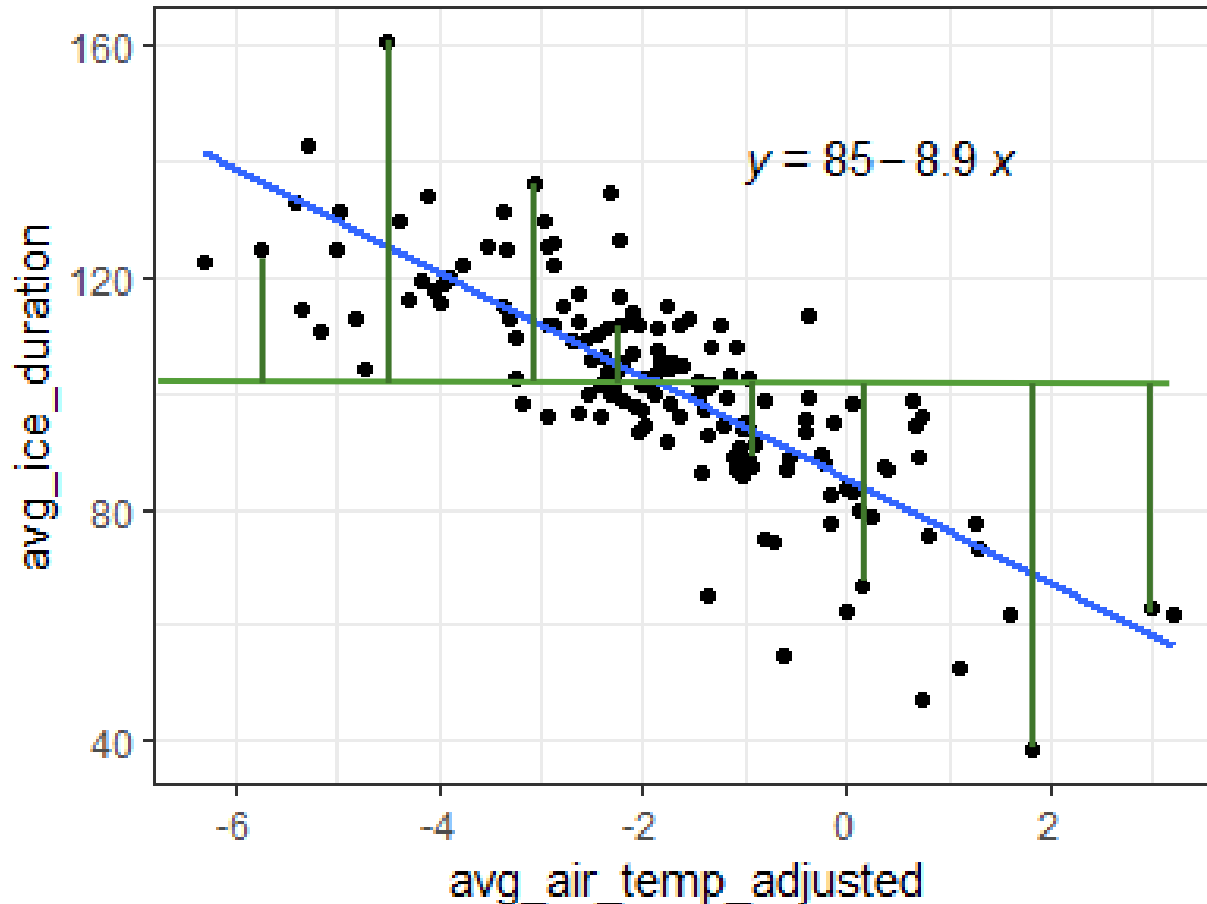
Đường nào?

Hồi quy tuyến tính đơn giản



Tổng bình phương
phần dư - Sum of the
squared residuals
(SSR)

Hồi quy tuyến tính đơn giản



- Độ tương thích (goodness of fit)

$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

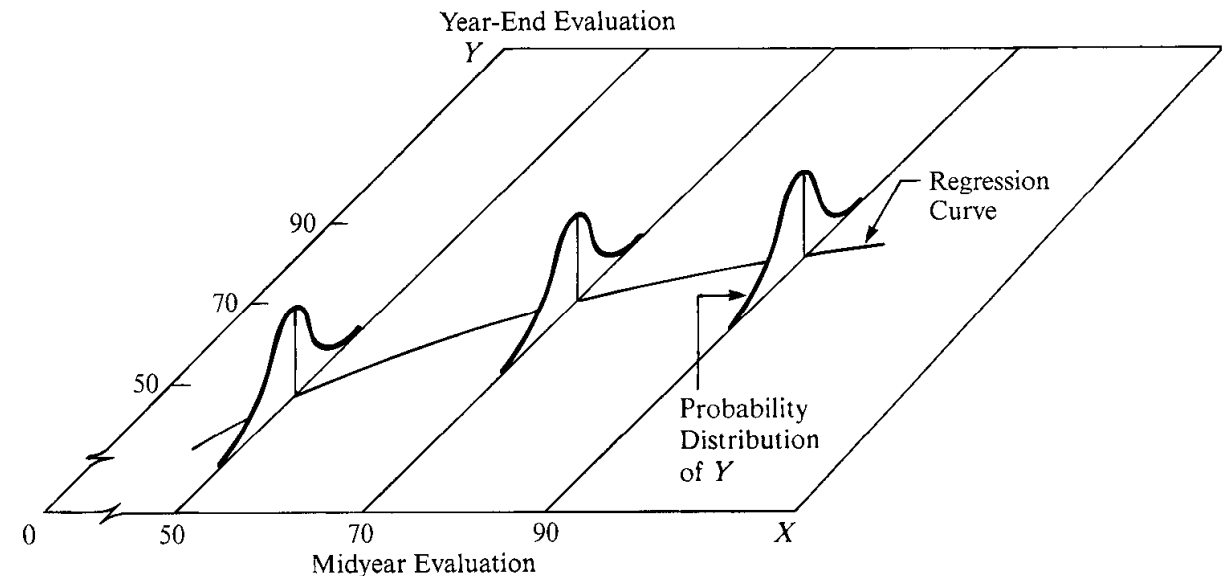
TSS: Total sum of squares

SSR: Sum squared residuals

Hồi quy tuyến tính đơn giản

Các giả định

- Mỗi liên hệ tuyến tính với các tham số khảo sát
- Các giá trị Y độc lập với nhau
- Các sai số ngẫu nhiên tuân theo phân phối chuẩn có cùng phương sai và trung bình = 0



Hồi quy tuyến tính đơn giản

Kiểm tra giả định

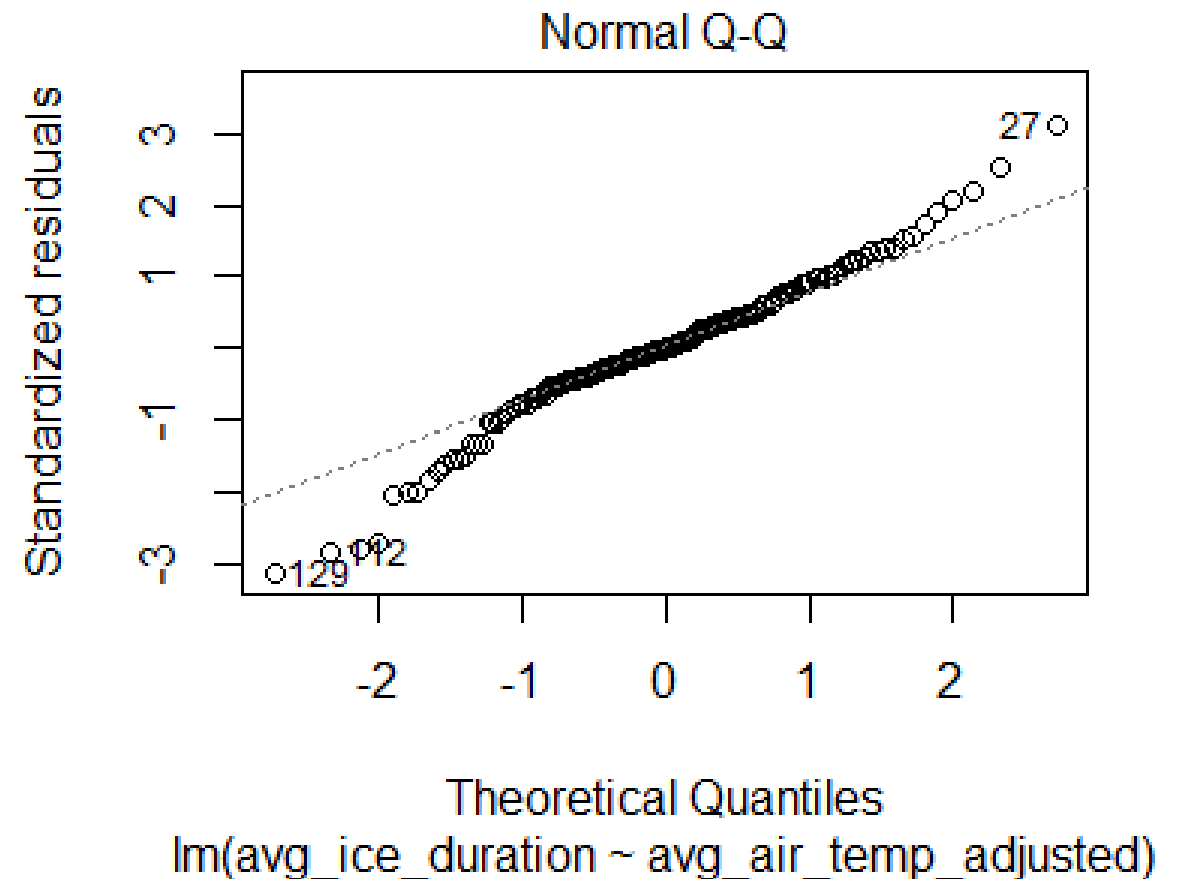
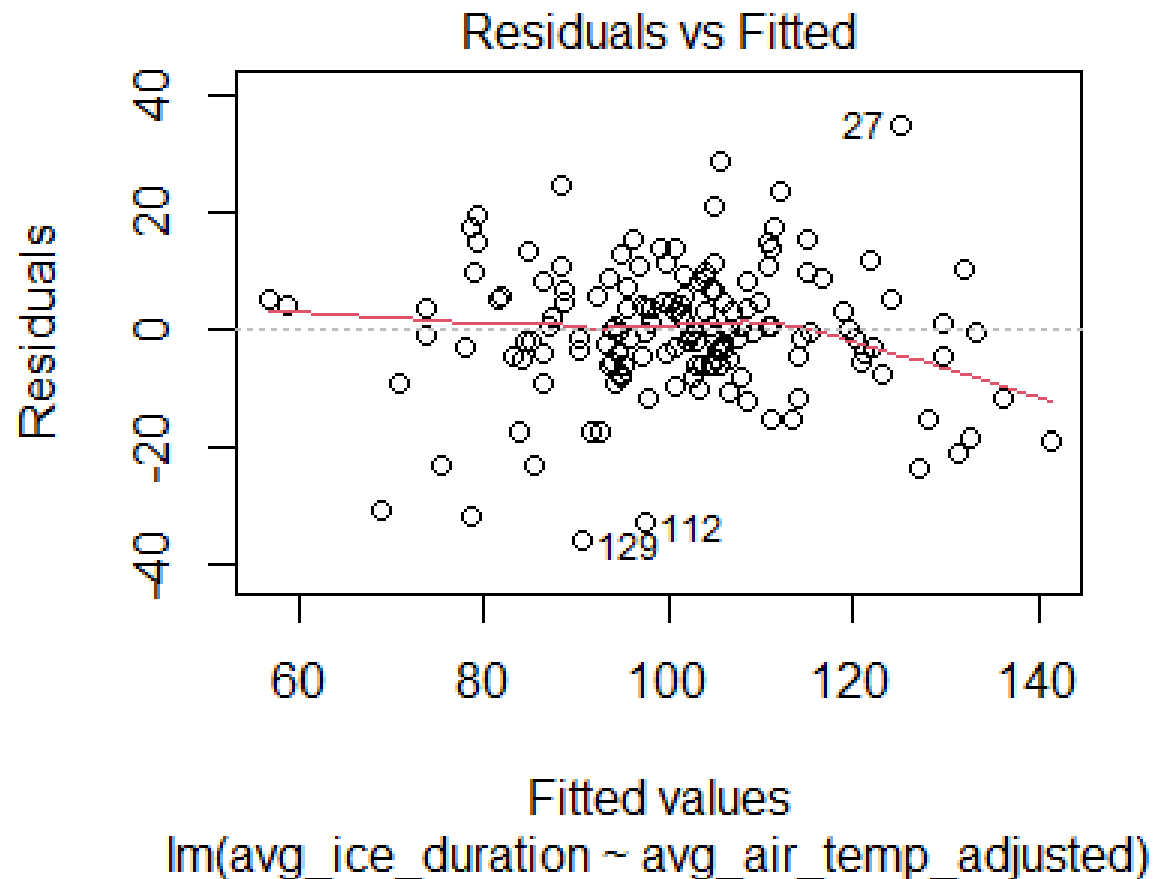
! Giả định phân phối chuẩn là cho phần dư/sai số ngẫu nhiên ϵ không phải biến phụ thuộc Y

! Các kiểm định phân phối chuẩn (vd: Kolmogorov–Smirnov, Shapiro–Wilk) bãi bỏ phân phối chuẩn khi số mẫu lớn

$$H_0: \epsilon \sim N(0, \sigma)$$
$$H_a: \epsilon \text{ không theo phân phối chuẩn}$$

Hồi quy tuyến tính đơn giản

Kiểm tra giả định



Hồi quy tuyến tính với biến nhị phân

Gasoline consumption/Distance ~ Driver Gender

$$Y_i = \alpha + \beta_1 X_{i1} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Biến giả/Dummy variables

$$X_i = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if male} \end{cases}$$

$$\text{Male: } Y_i = \alpha + \epsilon_i$$

$$\text{Female: } Y_i = \alpha + \beta_1 + \epsilon_i$$

Hồi quy tuyến tính với biến nhị phân

Gasoline consumption/Distance ~ Driver Gender

$$Y_i = \alpha + \beta_1 X_{i1} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Biến giả/Dummy variables

$$X_i = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if male} \end{cases}$$

$$\text{Male: } Y_i = \alpha + \epsilon_i$$

$$\text{Female: } Y_i = \alpha + \beta_1 + \epsilon_i$$

t-test

Hồi quy tuyến tính với biến định danh

Gasoline consumption/Distance ~ Car Make

Biến giả - Dummy variables (Toyota, VinFast, Mercedes)

Hồi quy tuyến tính với biến định danh

Gasoline consumption/Distance \sim Car Make

Biến giả - Dummy variables (Toyota, VinFast, Mercedes)

$$X_{i1} = \begin{cases} 1 & \text{if VinFast} \\ 0 & \text{if other} \end{cases} \quad X_{i2} = \begin{cases} 1 & \text{if Mercedes} \\ 0 & \text{if other} \end{cases}$$

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$$\text{Toyota: } Y_i = \alpha + \epsilon_i \quad \text{VinFast: } Y_i = \alpha + \beta_1 + \epsilon_i \quad \text{Mercedes: } Y_i = \alpha + \beta_2 + \epsilon_i$$

Hồi quy tuyến tính với biến định danh

Gasoline consumption/Distance \sim Car Make

Biến giả - Dummy variables (Toyota, VinFast, Mercedes)

$$X_{i1} = \begin{cases} 1 & \text{if VinFast} \\ 0 & \text{if other} \end{cases} \quad X_{i2} = \begin{cases} 1 & \text{if Mercedes} \\ 0 & \text{if other} \end{cases}$$

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

ANOVA

$$\text{Toyota: } Y_i = \alpha + \epsilon_i \quad \text{VinFast: } Y_i = \alpha + \beta_1 + \epsilon_i \quad \text{Mercedes: } Y_i = \alpha + \beta_2 + \epsilon_i$$

Hồi quy tuyến tính bội

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_1 X_{i1}^2 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

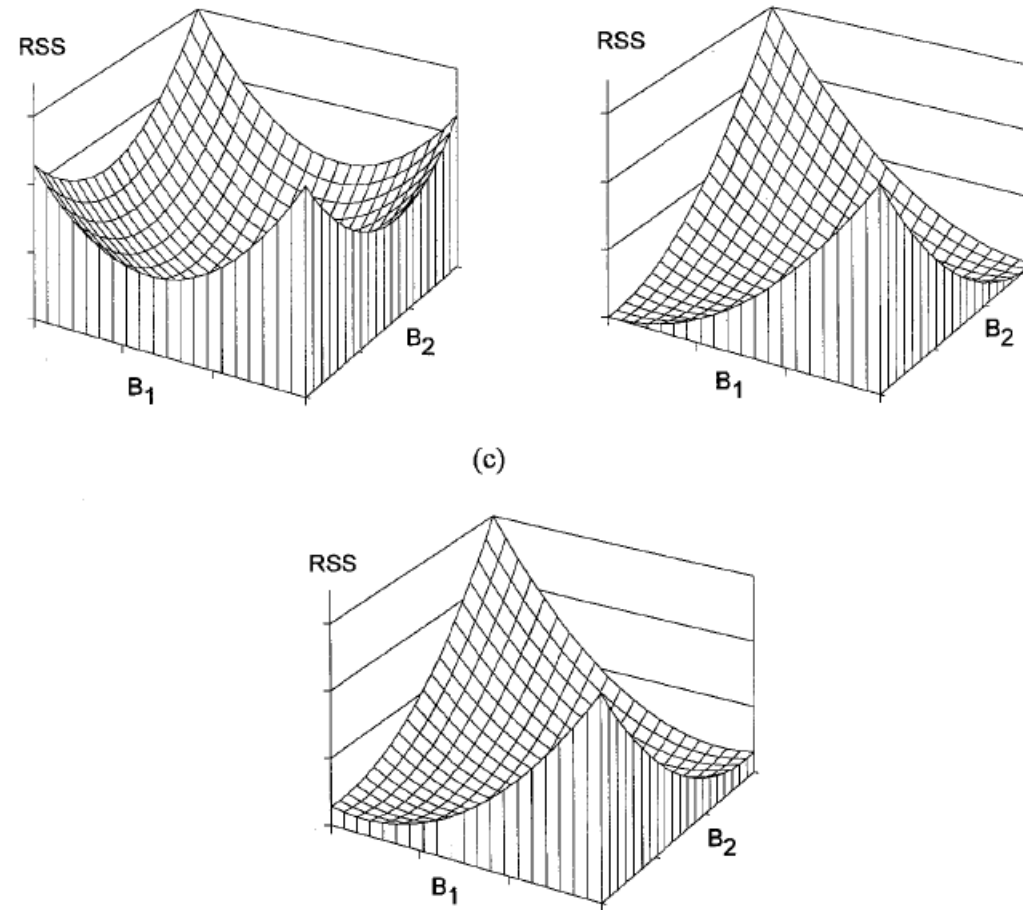
Ví dụ:

Gasoline consumption/Distance ~ Car Age + Driver Age + # of breaks per minute + Driver gender + Car Make +

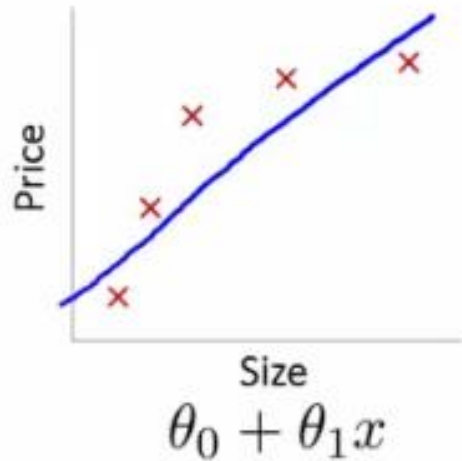
Mô hình càng phức tạp càng tốt?

Tương quan giữa các biến độc lập Multicollinearity

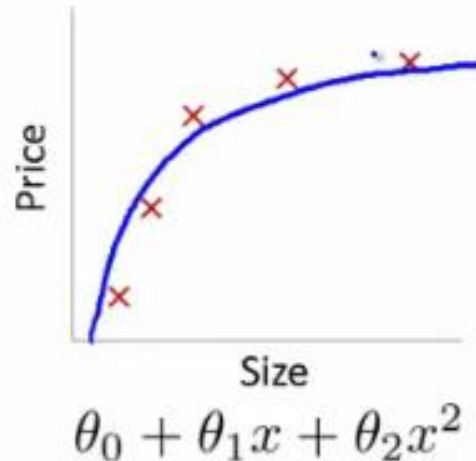
Variance Inflation Factor
(Yếu tố lạm phát phương sai)



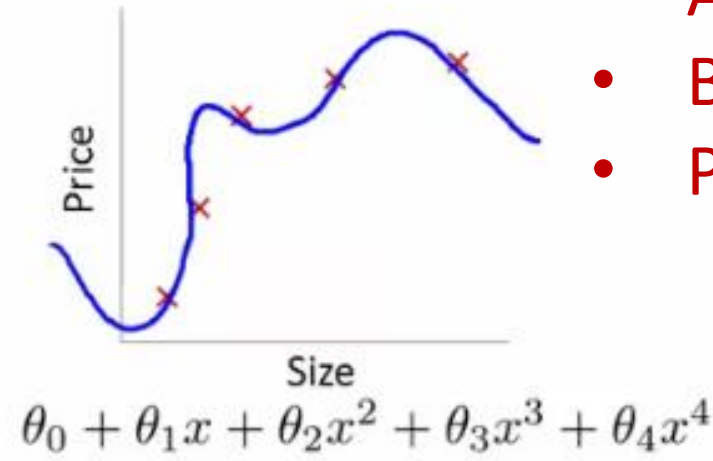
Quá khớp hoặc thiếu khớp với số liệu



High bias
(underfit)



“Just right”



High variance
(overfit)

- Adjusted R^2
- AIC
- BIC
- Predictive power

Các phương pháp nâng cao

Một số kỹ thuật định lượng khác

- Hồi quy tuyến tính suy rộng (Generalized linear regression)
- Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp (Multilevel model/Mixed effects model)
- Phân tích nhân tố (Factor analysis)
- Phân tích thành phần chính (PCA - Principal component analysis)
- Mô hình phương trình cấu trúc (SEM – Structural equation model)
- Thống kê không gian (Spatial statistics)
- ...

Hồi quy tuyến tính suy rộng (Generalized linear regression)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_1 X_{i1}^2 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Hồi quy tuyến tính bội

Y : liên tục/định lượng

X : liên tục/định lượng hoặc gián đoạn

Khi Y là biến gián đoạn?

Hồi quy tuyến tính suy rộng (Generalized linear regression)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_1 X_{i1}^2 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Hồi quy tuyến tính bội

Y : liên tục/định lượng

X : liên tục/định lượng hoặc gián đoạn

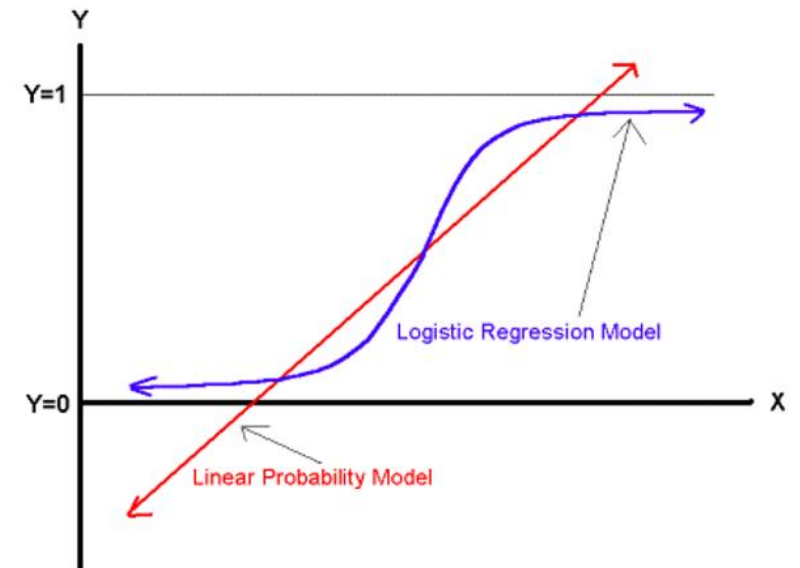
Khi Y là biến gián đoạn?

Nhị phân (Yes/No): Hồi quy Logistic (Logistic regression)

Định danh: Multinomial logistic regression

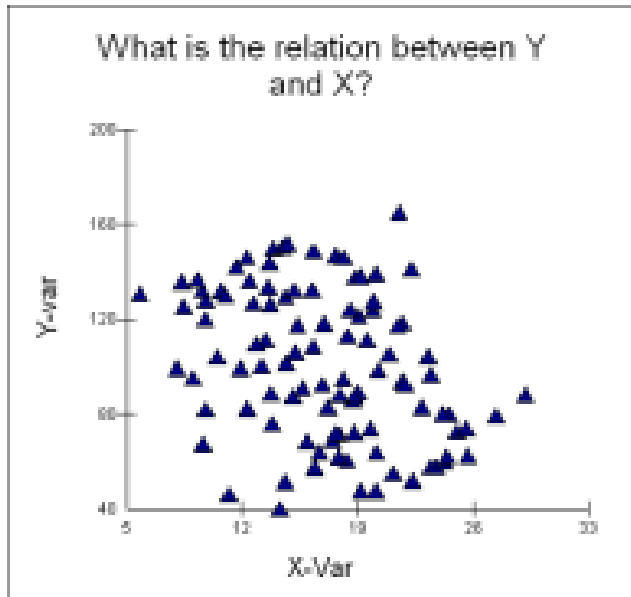
Thứ bậc: Cumulative logistic regression

Biến đếm: Poisson regression



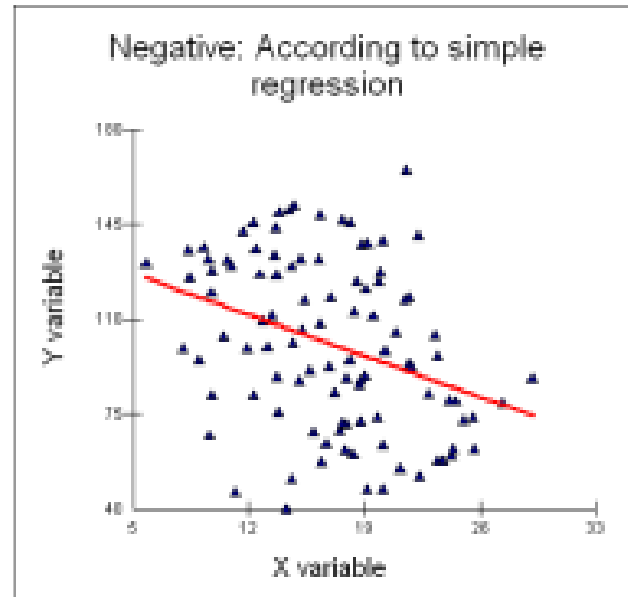
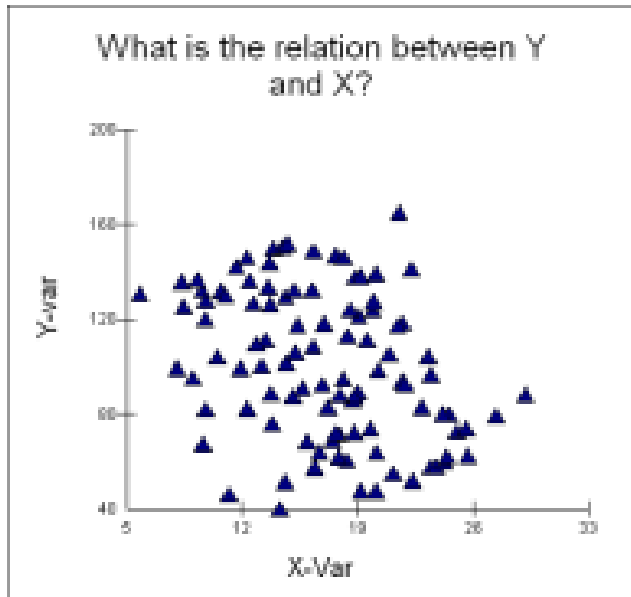
Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model



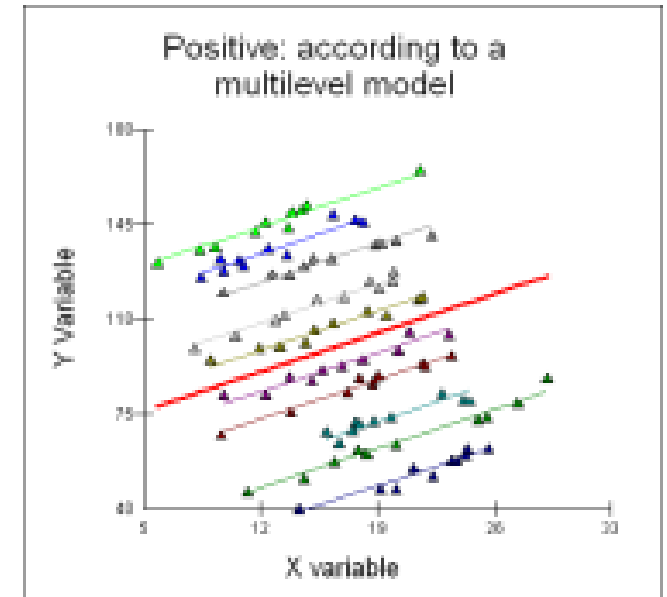
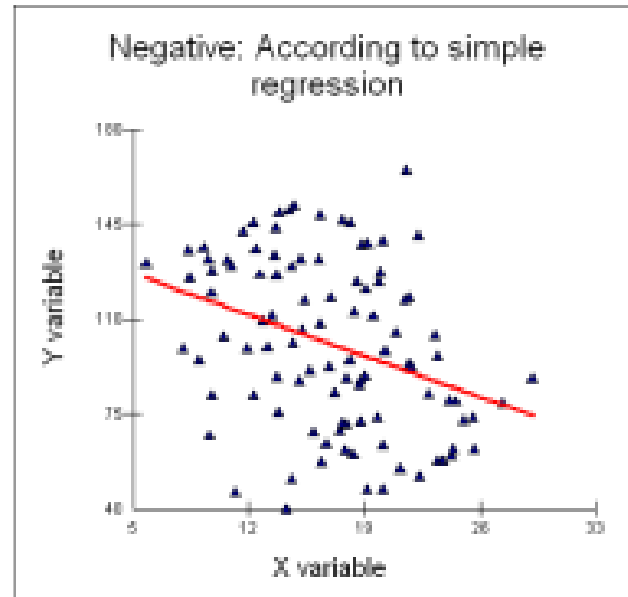
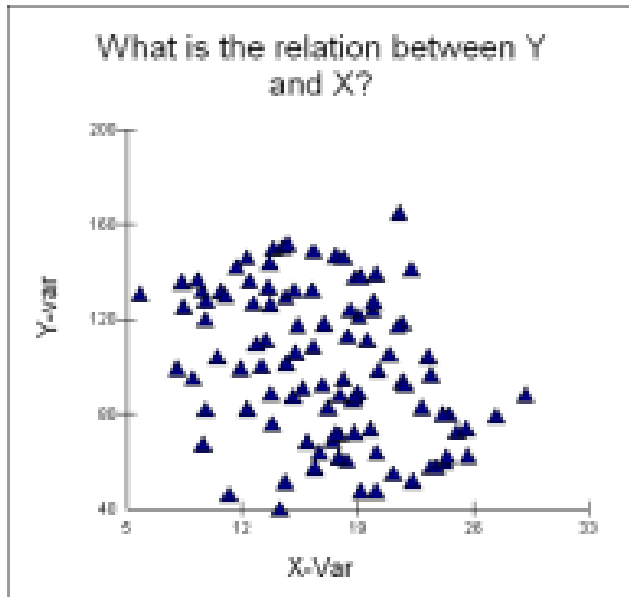
Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model



Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model

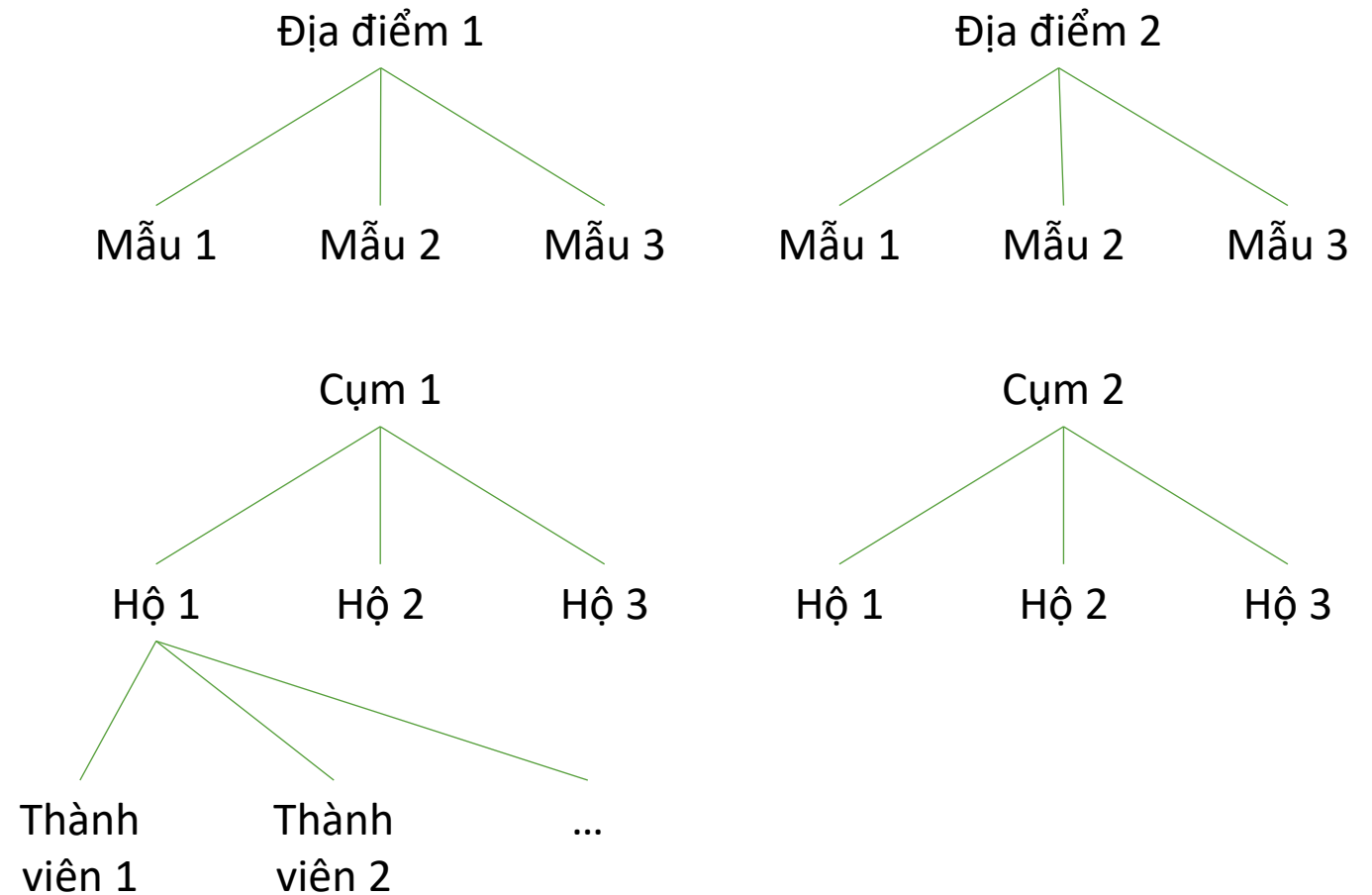


Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model

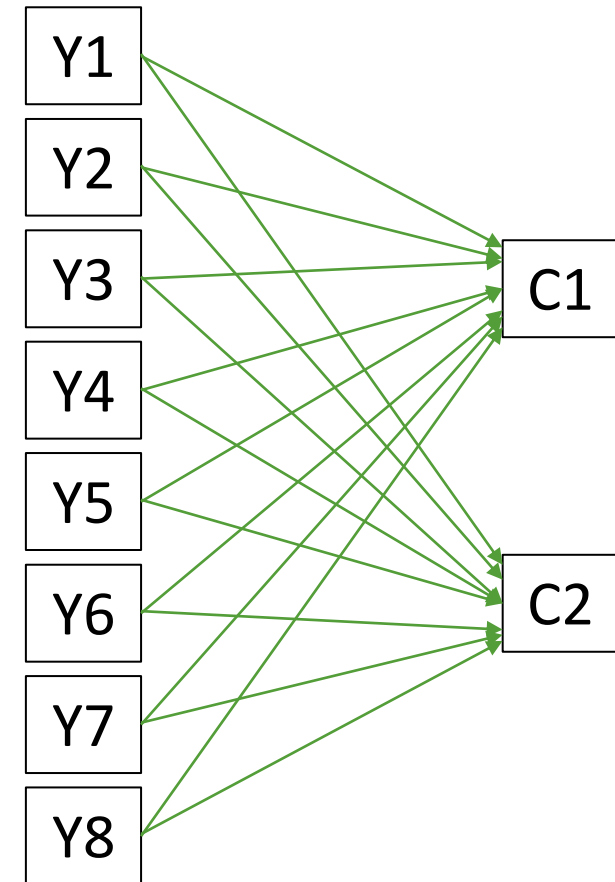


Image by [Chelsea Parlett-Pelleriti](#)



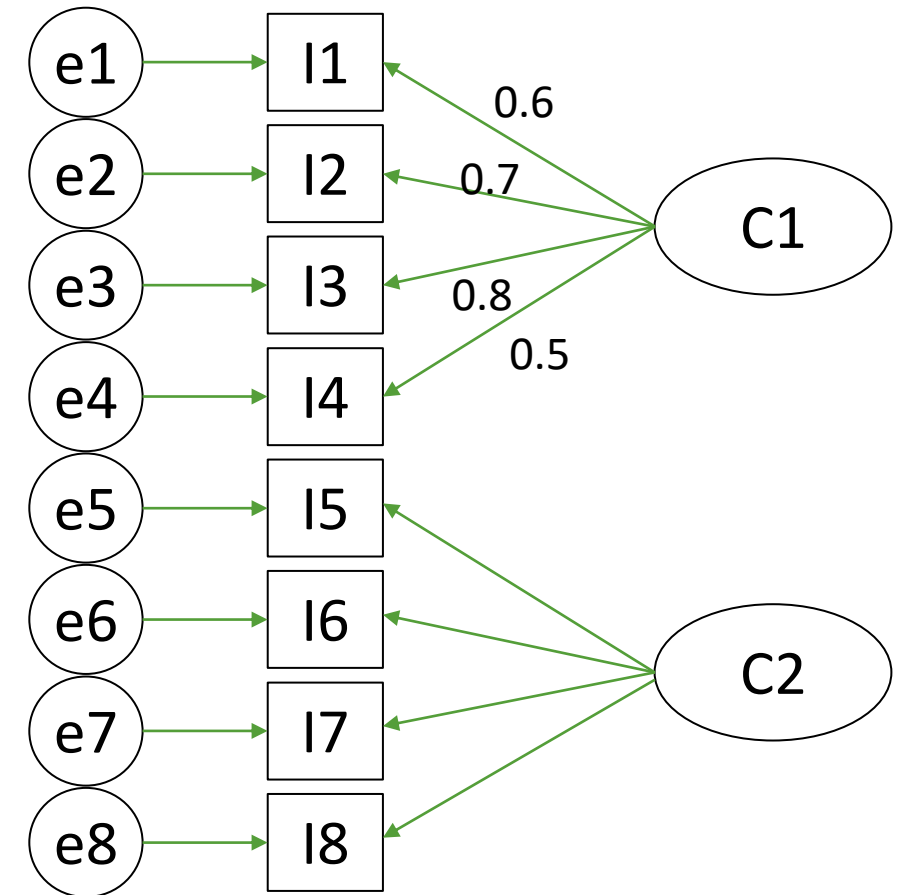
Phân tích thành phần chính (PCA - Principal component analysis)

- Phương pháp giảm chiều dữ liệu
- Không phân biệt biến độc lập hay phụ thuộc
- Phương pháp khảo sát (không phải phương pháp suy luận)
- Bước trước cho hồi quy tuyến tính để giảm đa cộng tuyến (multicollinearity)



Phân tích nhân tố (Factor analysis)

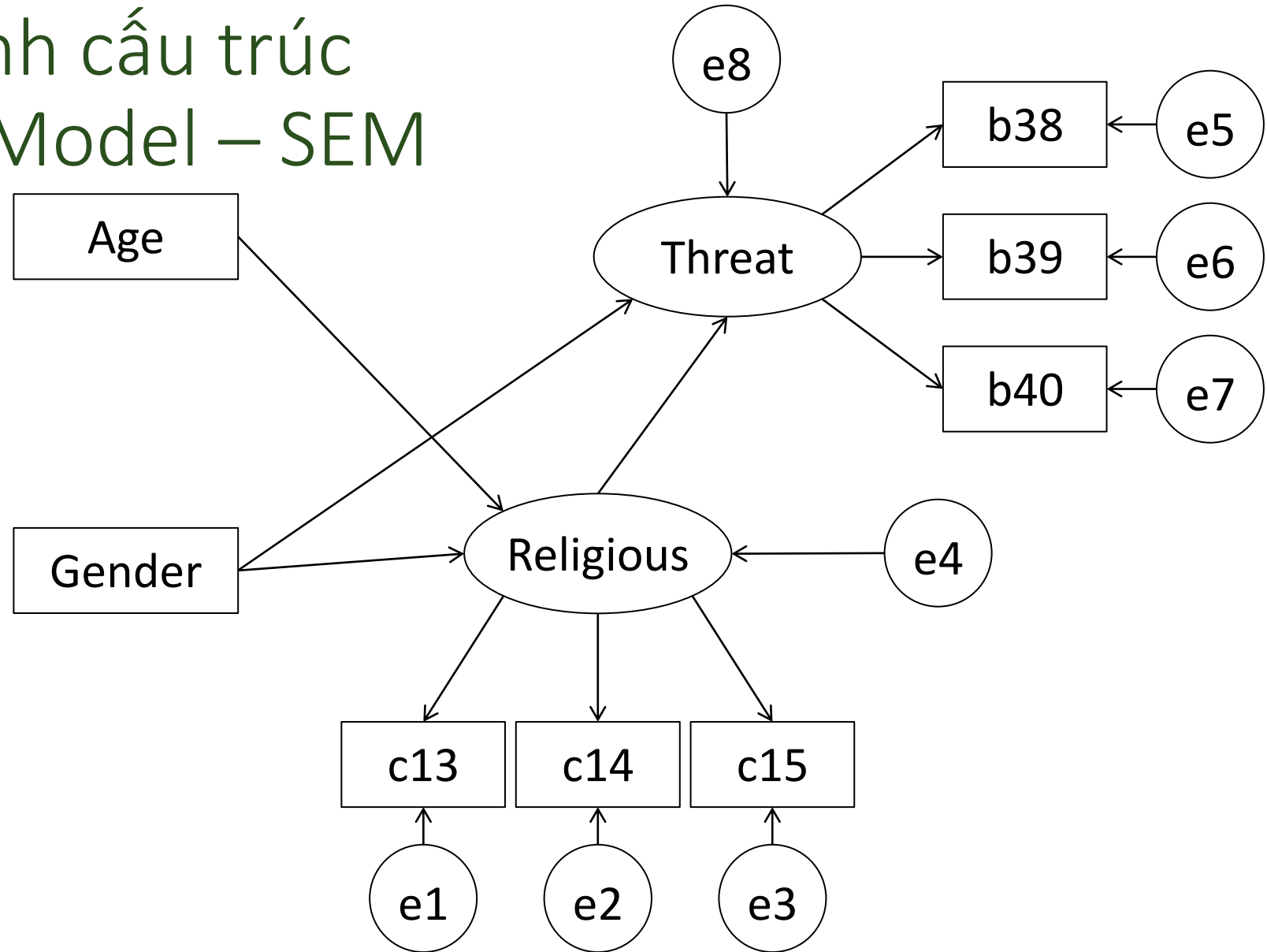
- Thường áp dụng cho dữ liệu bảng hỏi
- Đo phạm trù tiềm ẩn (Latent construct)
- Phân tích nhân tố khám phá/khẳng định (Exploratory/Confirmatory Factor Analysis)



Mô hình phương trình cấu trúc

Structural Equation Model – SEM

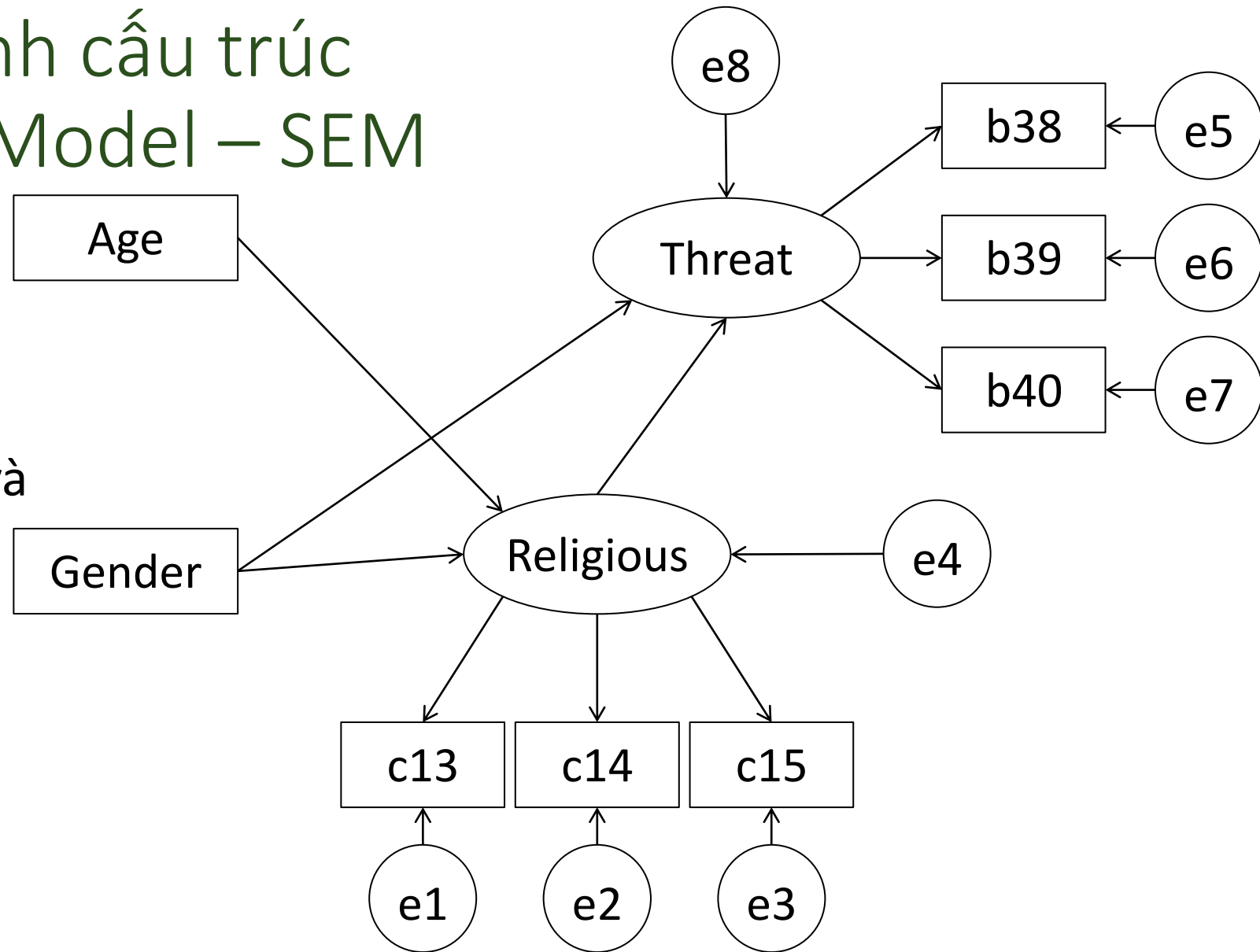
- Tiềm ẩn (latent) và biểu hiện (manifest)



Mô hình phương trình cấu trúc

Structural Equation Model – SEM

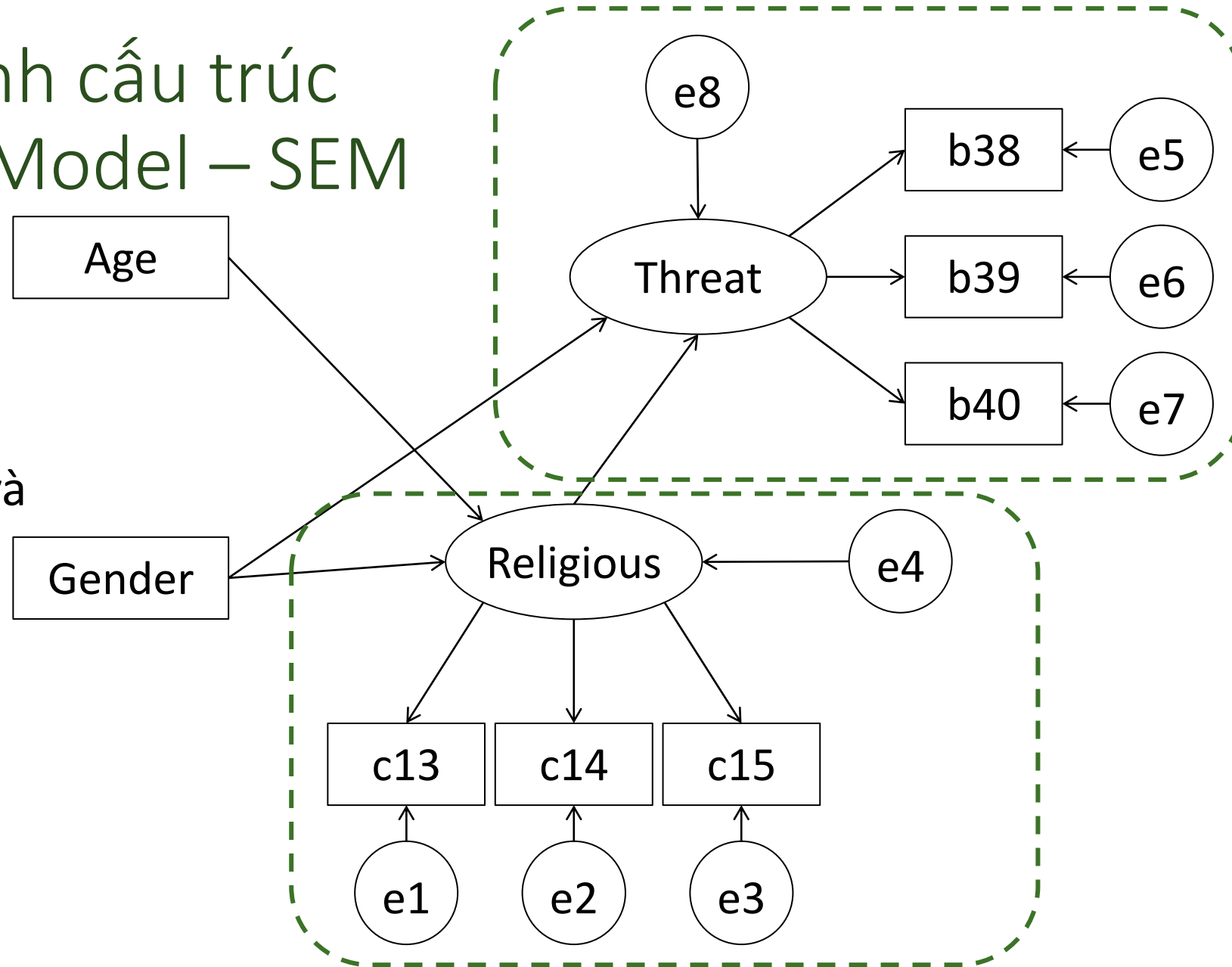
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)



Mô hình phương trình cấu trúc

Structural Equation Model – SEM

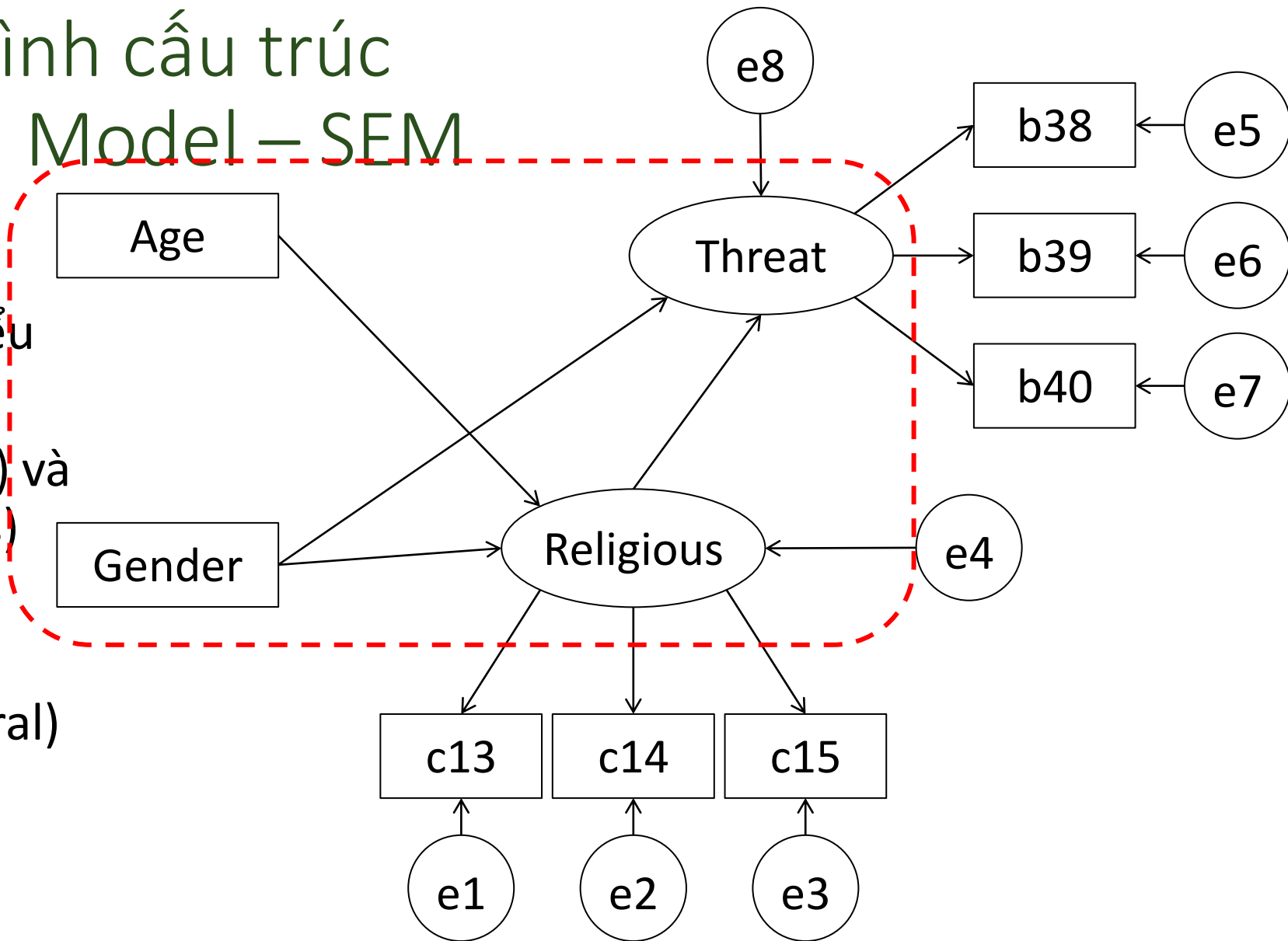
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement)



Mô hình phương trình cấu trúc

Structural Equation Model – SEM

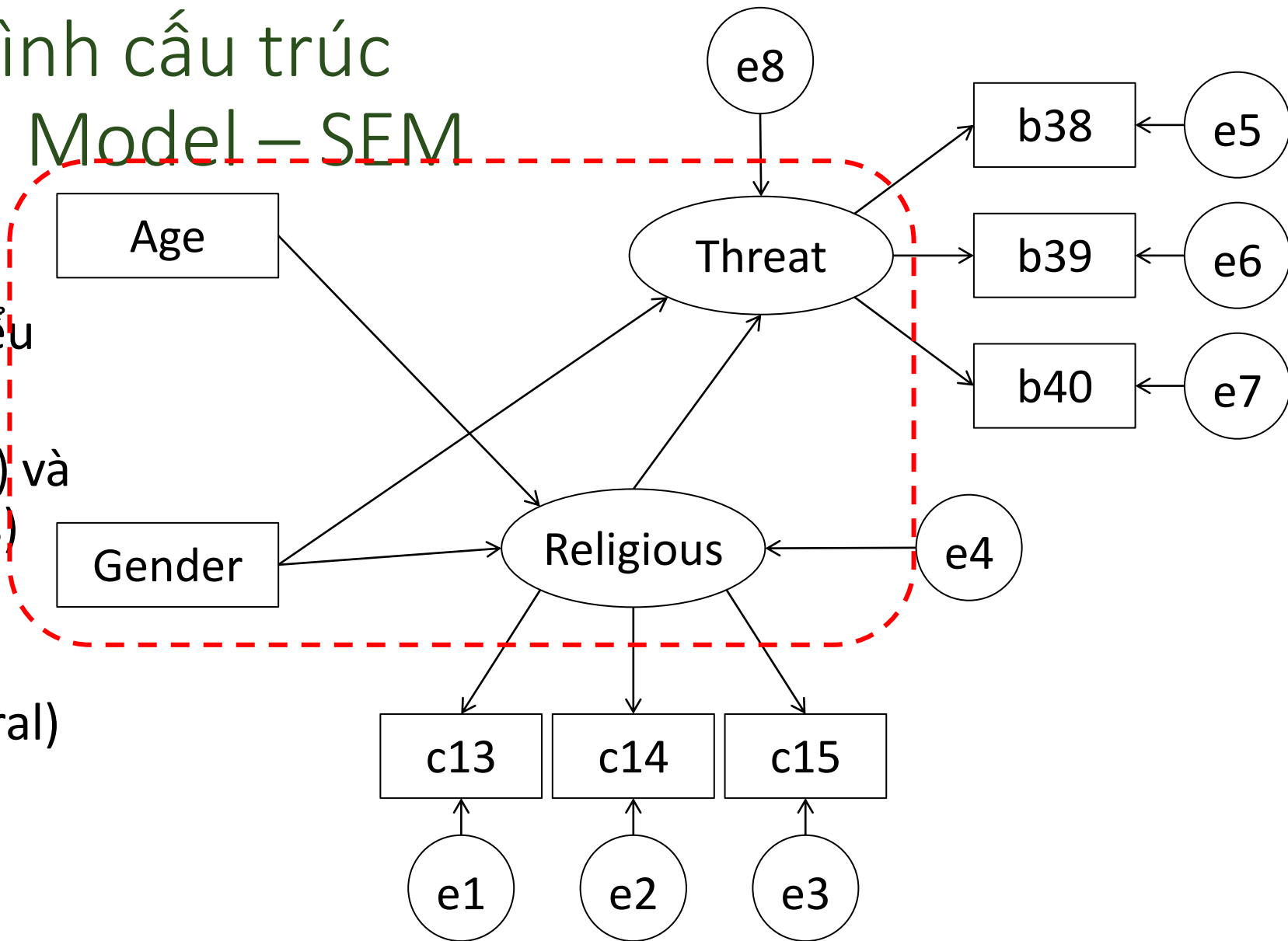
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement) và mô hình cấu trúc (structural)



Mô hình phương trình cấu trúc

Structural Equation Model – SEM

- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement) và mô hình cấu trúc (structural)
- Tác động trực tiếp và gián tiếp (Direct vs indirect effects)



Thống kê không gian Spatial Statistics

