

Phương pháp nghiên cứu trong khoa học liên ngành

Nguyễn Bích Ngọc

Khoa các khoa học liên ngành, ĐHQGHN

Phân tích định lượng

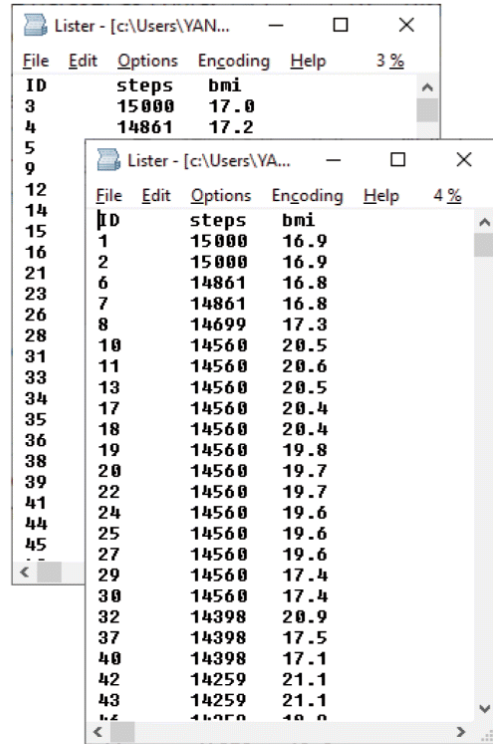
Các bước phân tích định lượng

- Làm sạch dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

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

a



The image 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
	45	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

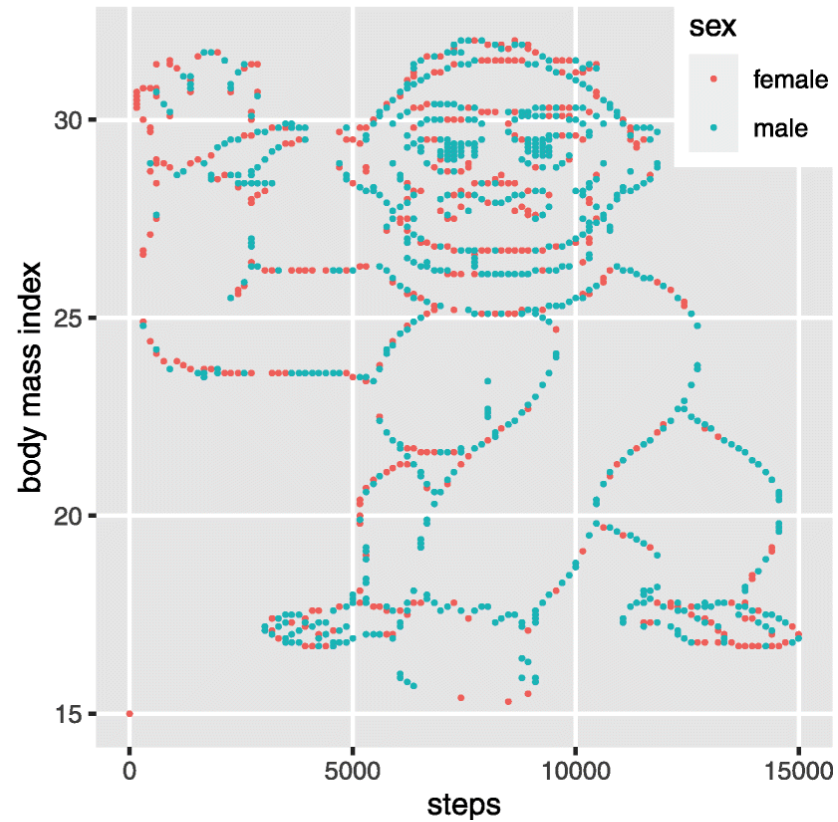
List - [c:\Users\YAN...]

ID	steps	bmi
3	15000	17.0
4	14861	17.2

List - [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



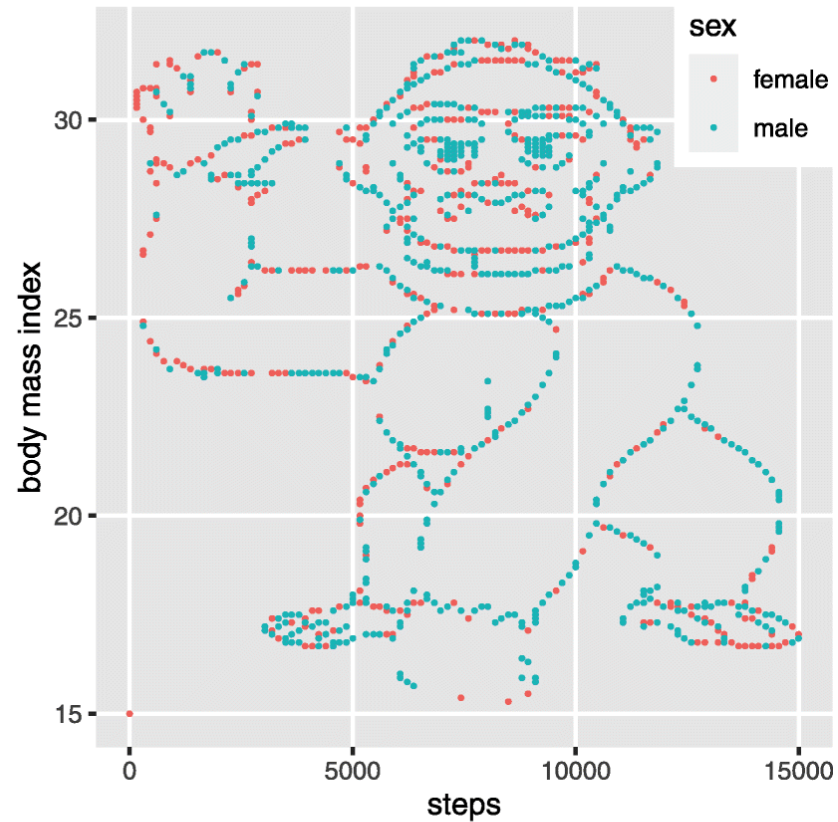
Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

ID	steps	bmi
3	15000	17.0
4	14861	17.2

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

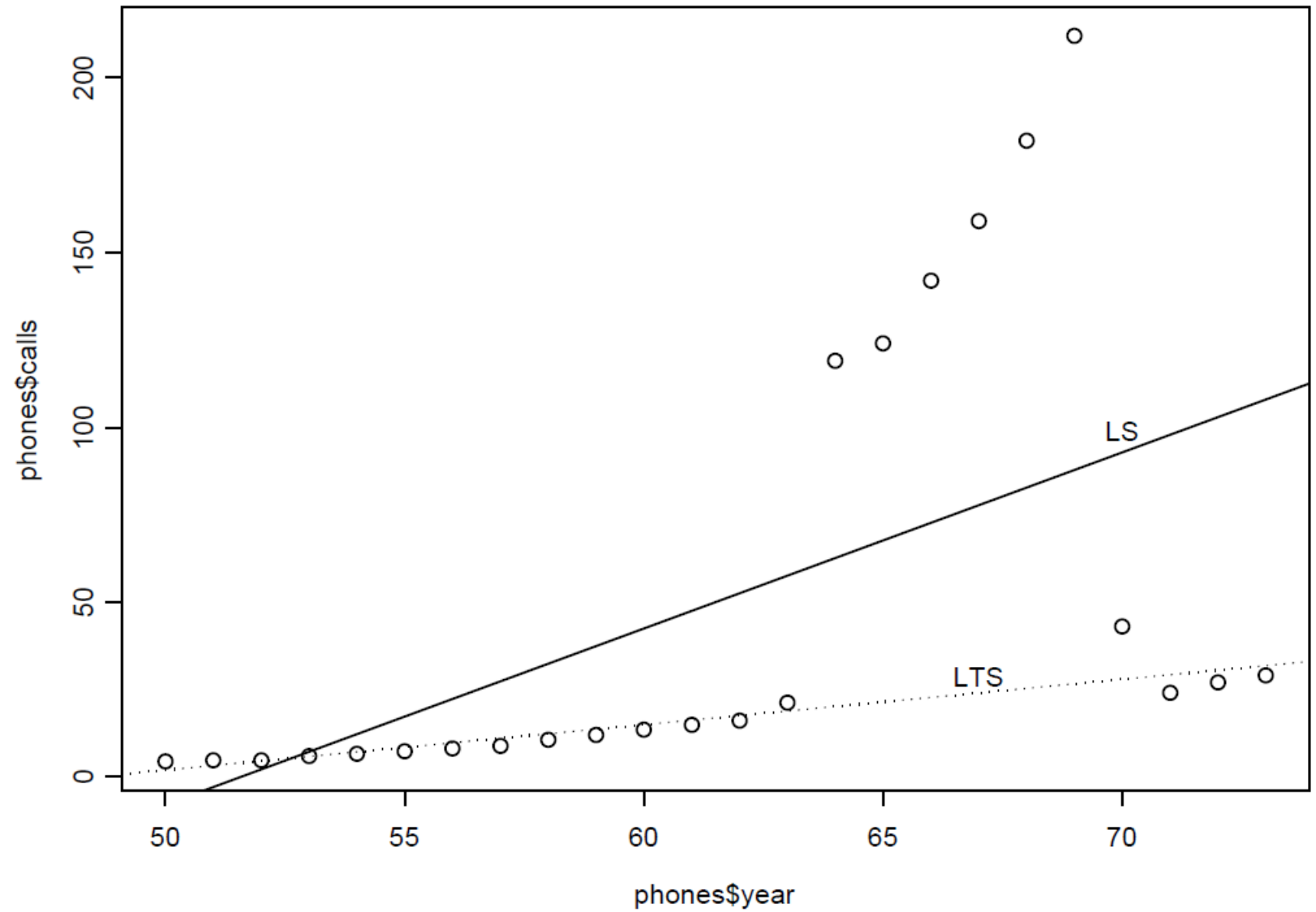


c

	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.



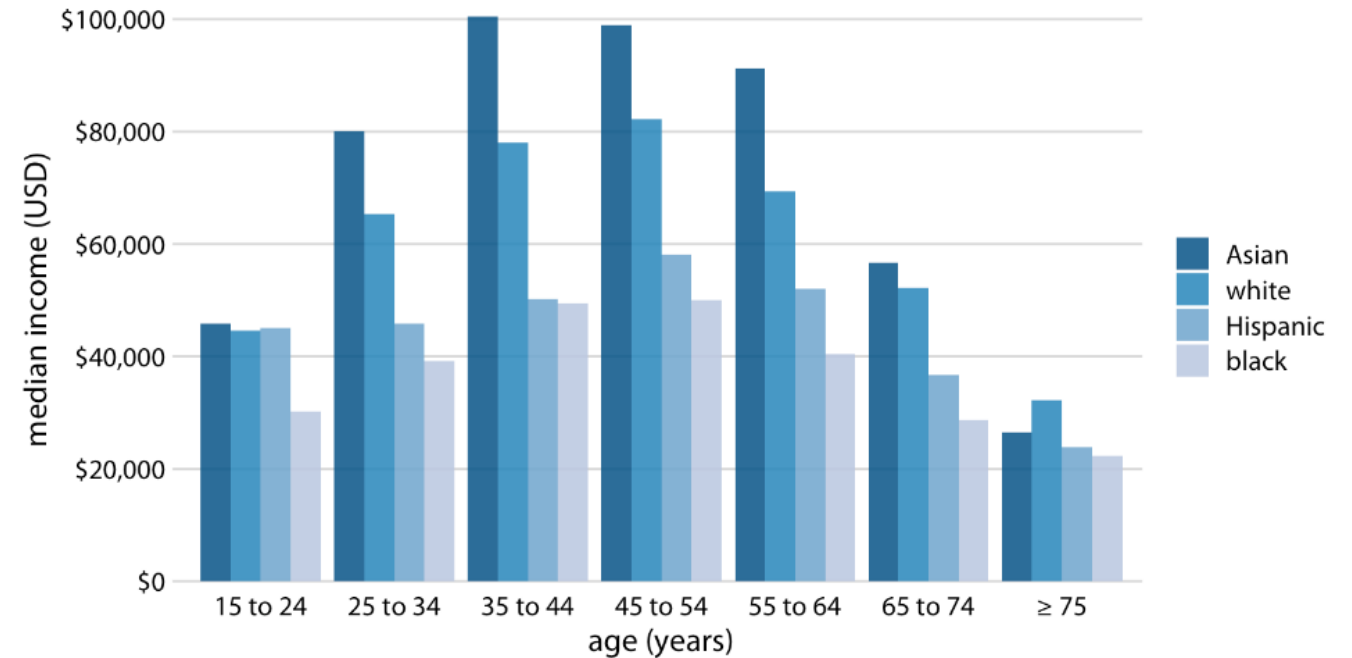
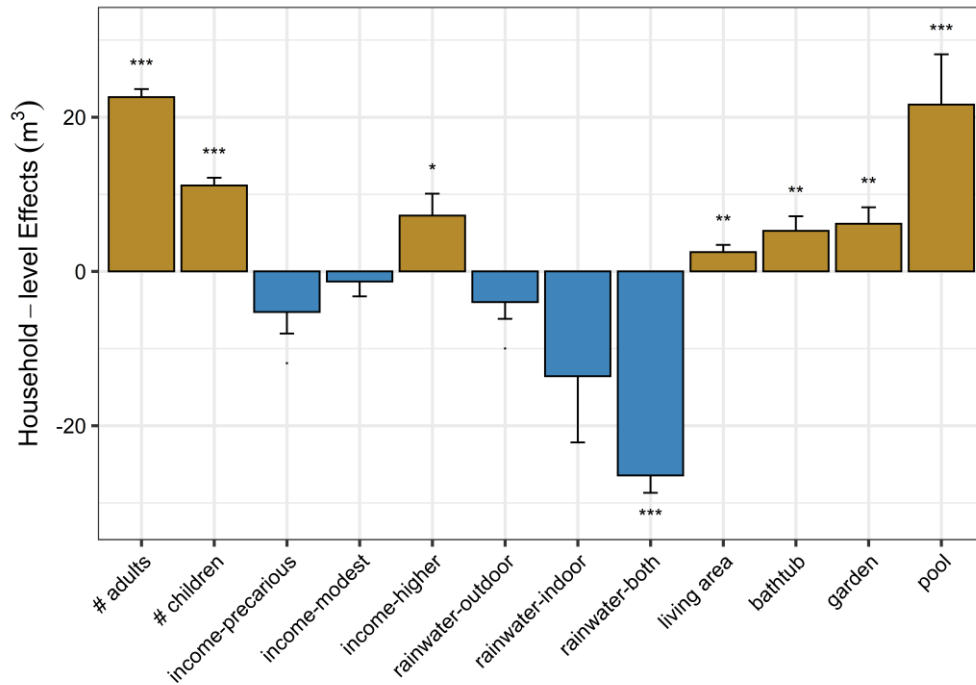
Đồ thị

- 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

Đồ thị thông thường cho biến liên tục

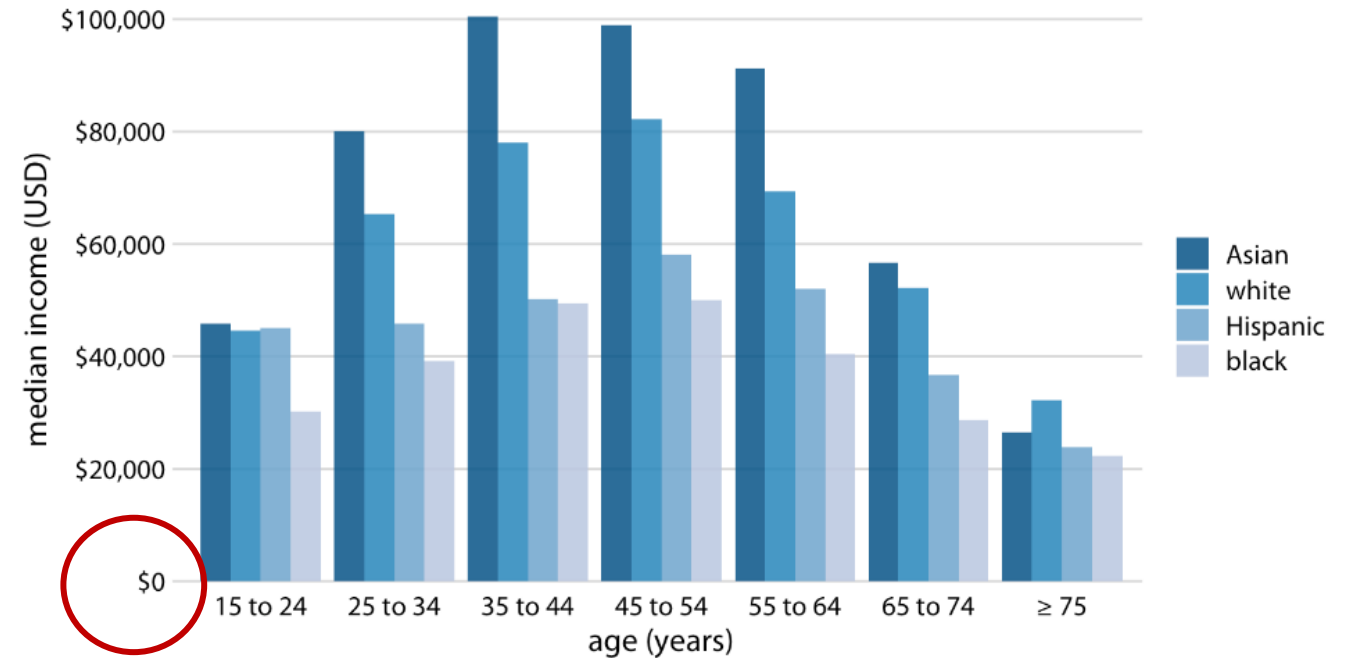
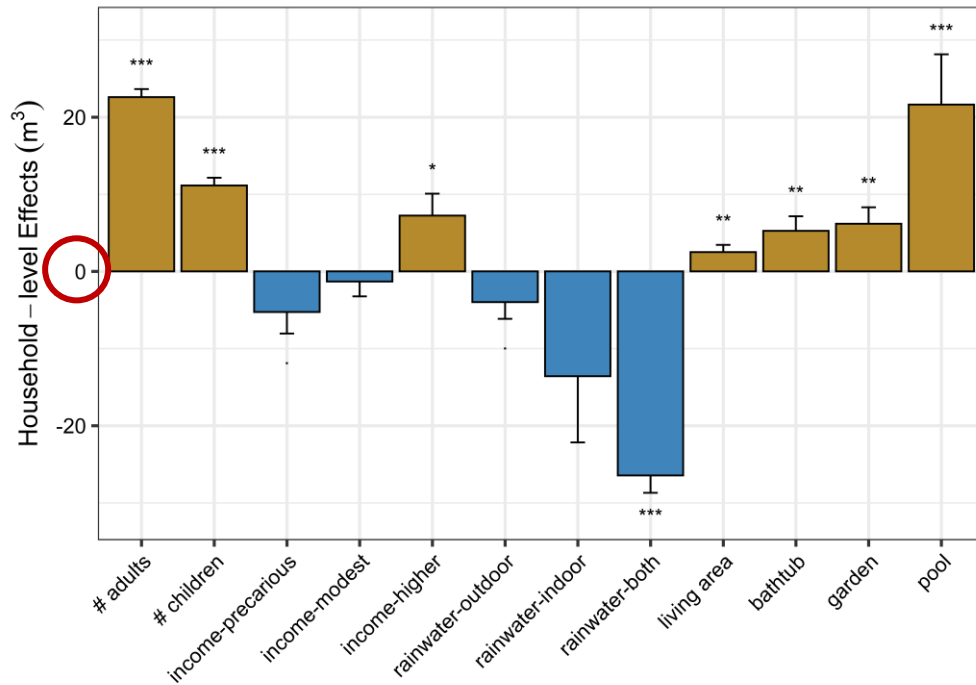
Đồ thị cột – bar chart



Wilke (2018)

Đồ thị thông thường cho biến liên tục

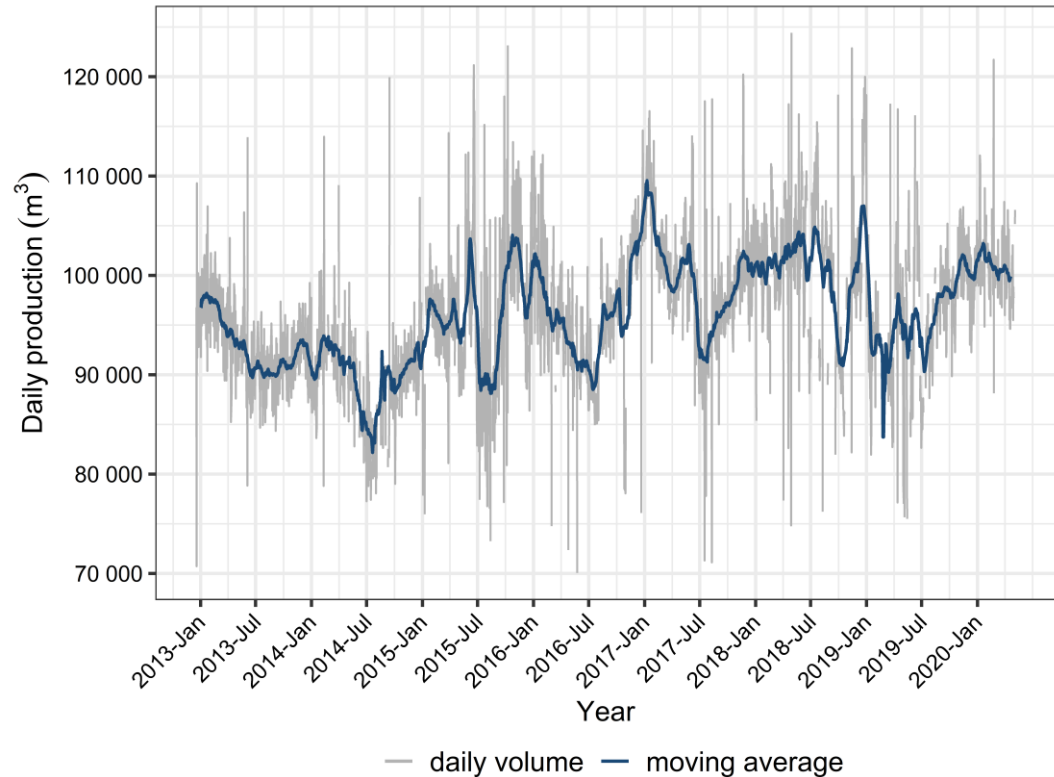
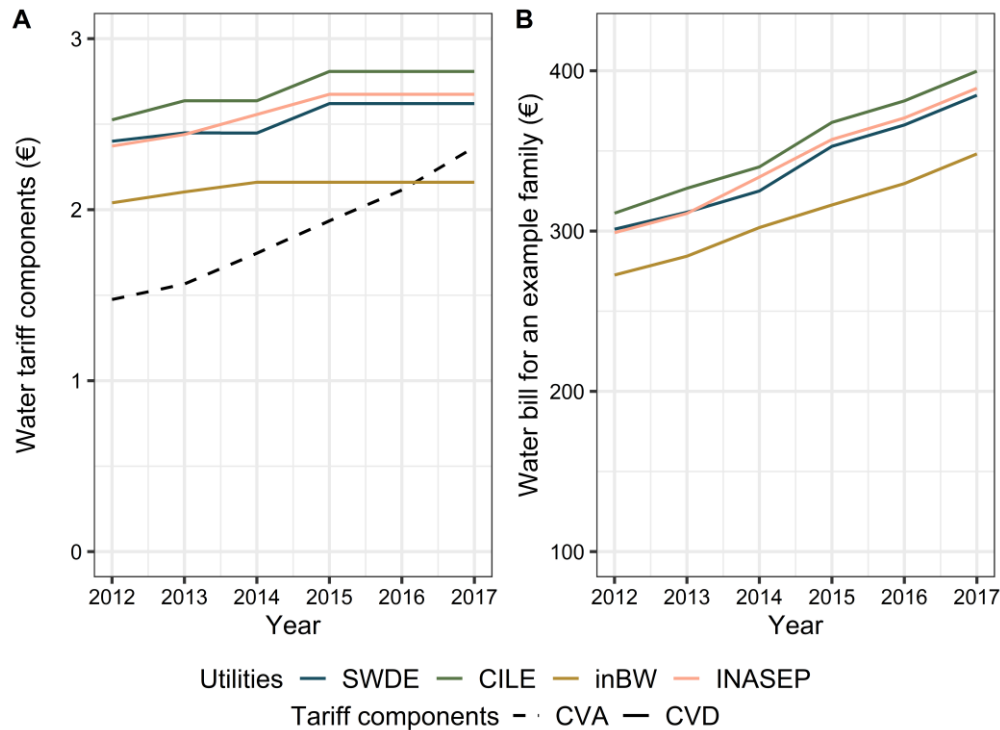
Đồ thị cột – bar chart



Source: Wilke (2018)

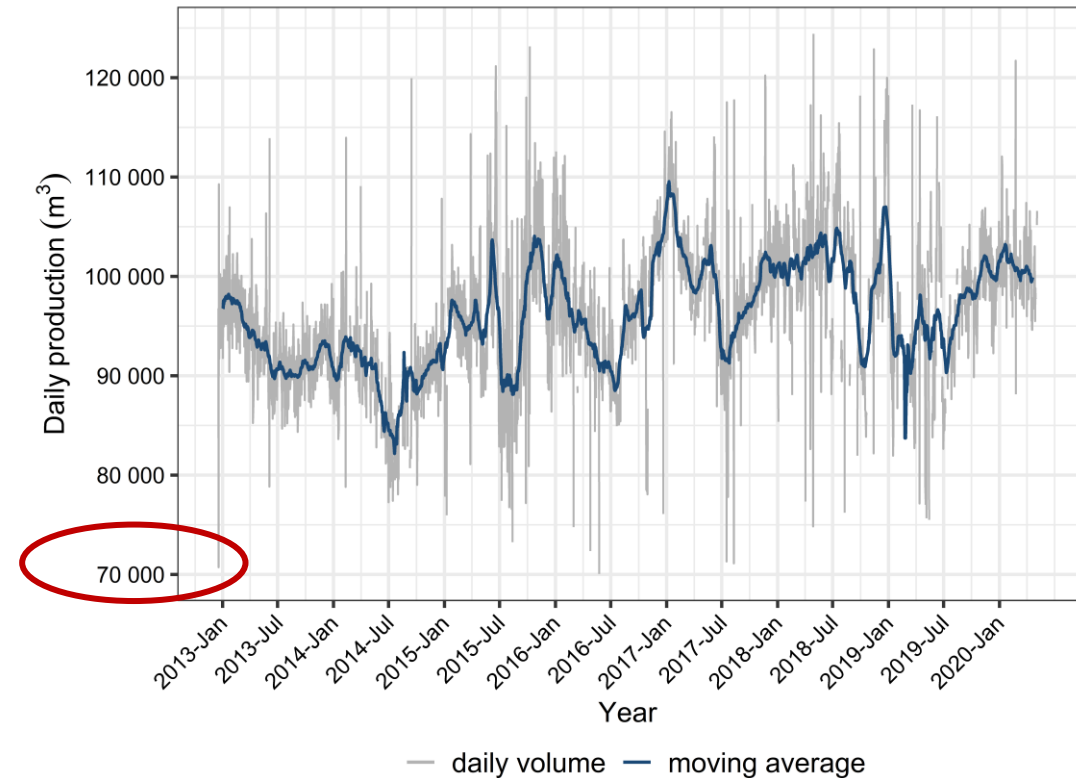
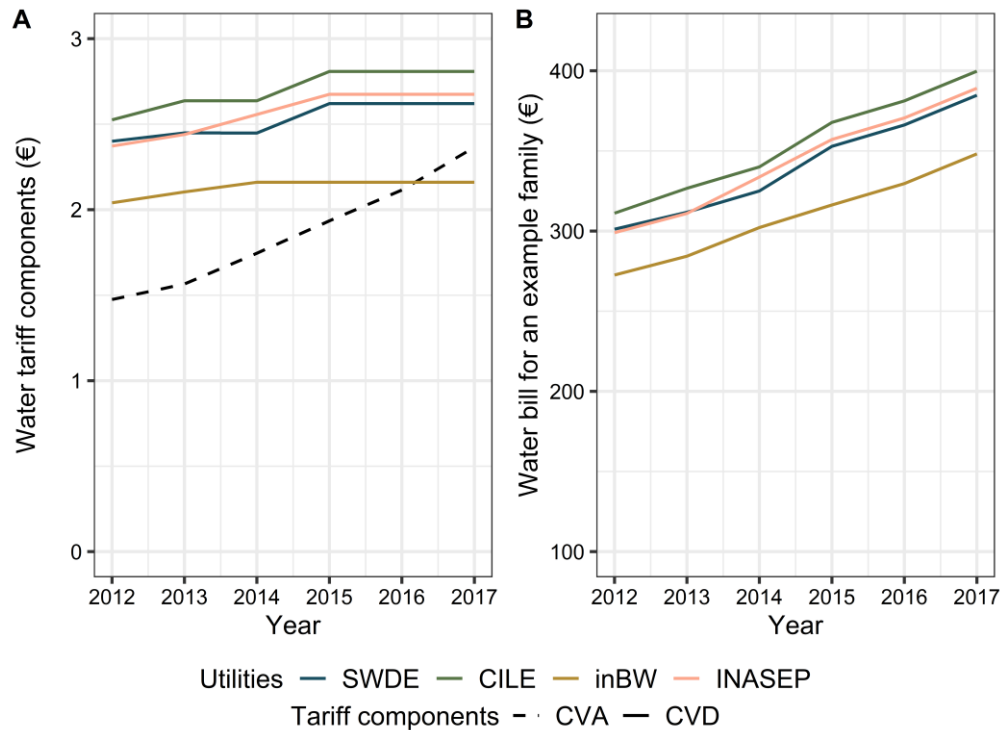
Đồ thị thông thường cho biến liên tục

Đồ thị đường – line chart



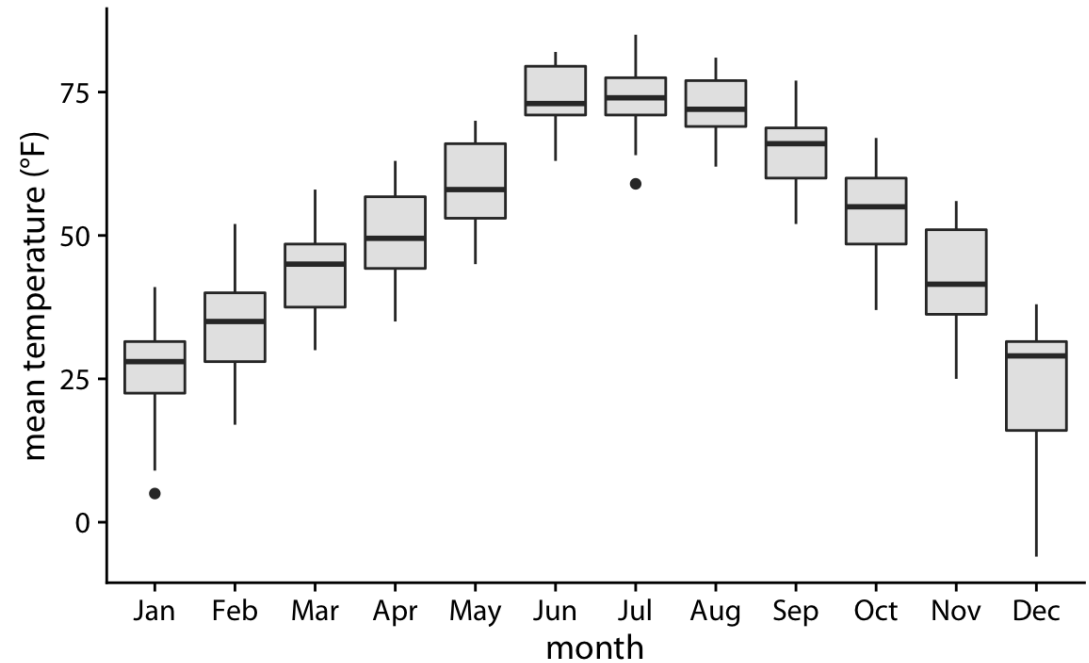
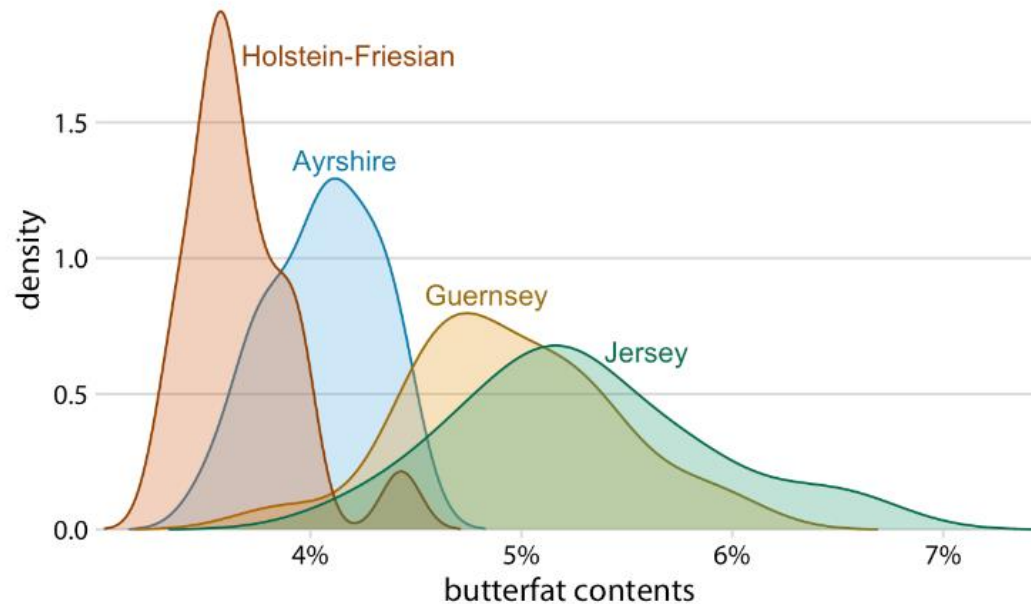
Đồ thị thông thường cho biến liên tục

Đồ thị đường – line chart



Đồ thị thông thường cho biến liên tục

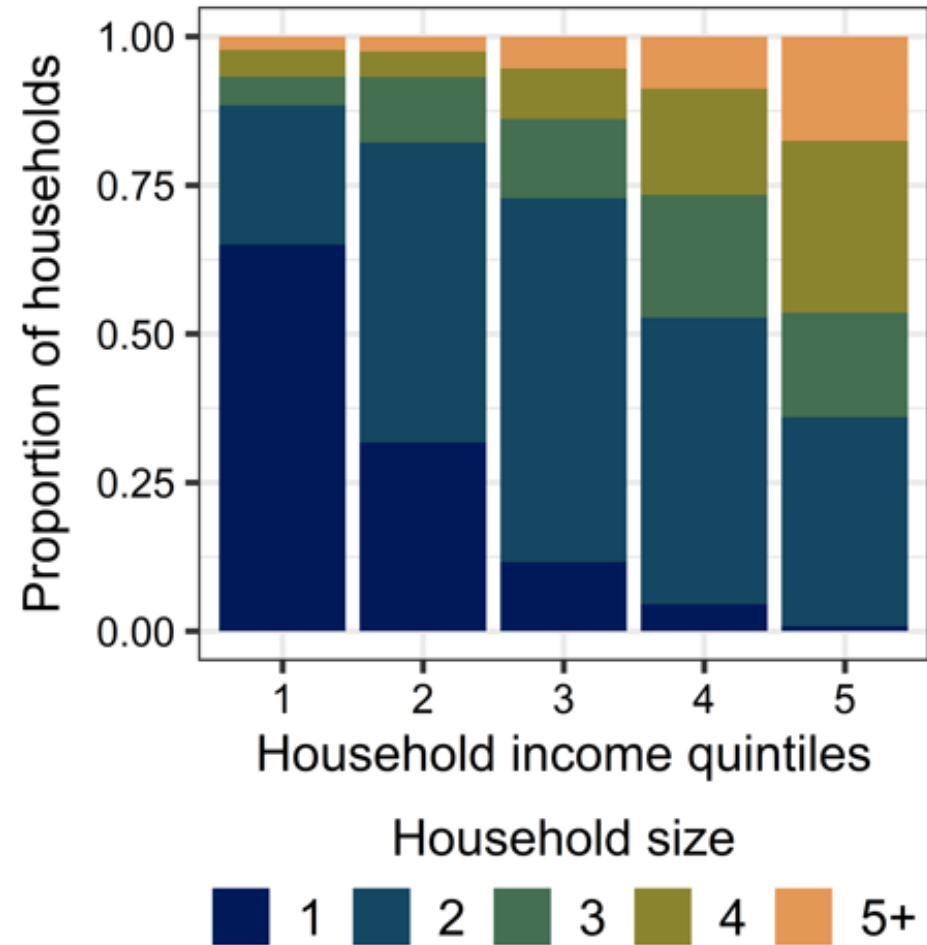
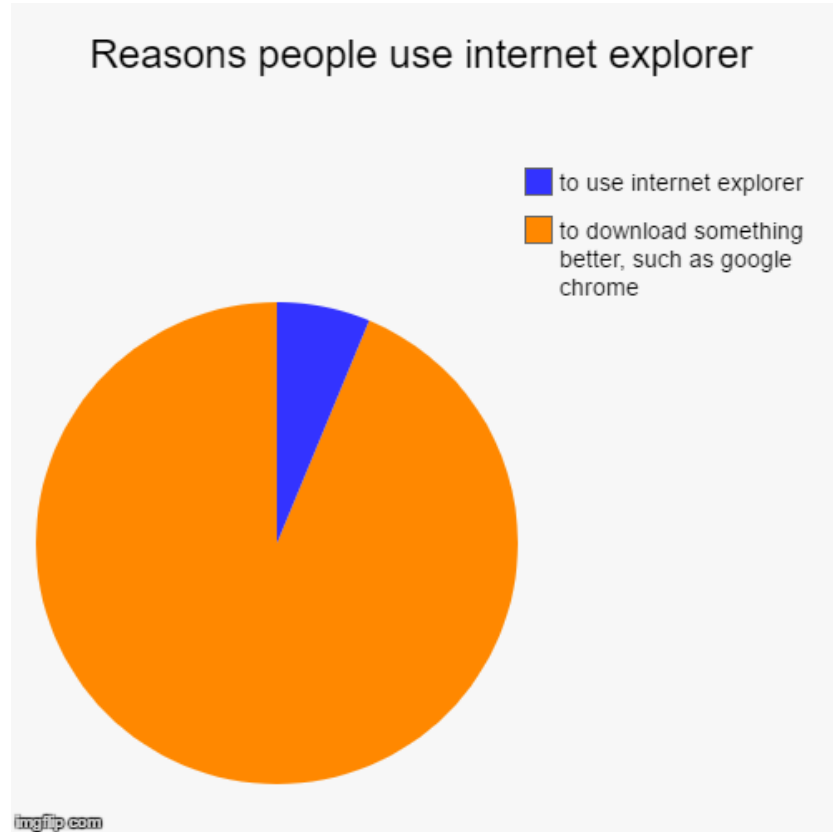
Biểu đồ tần suất - Histogram



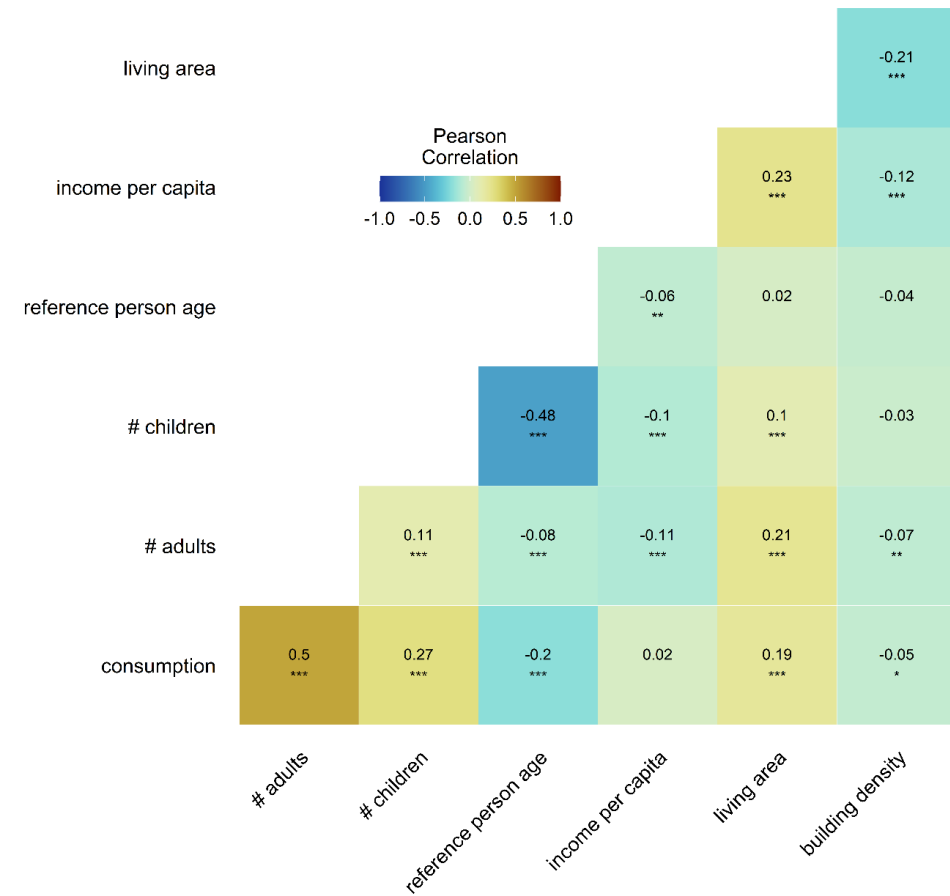
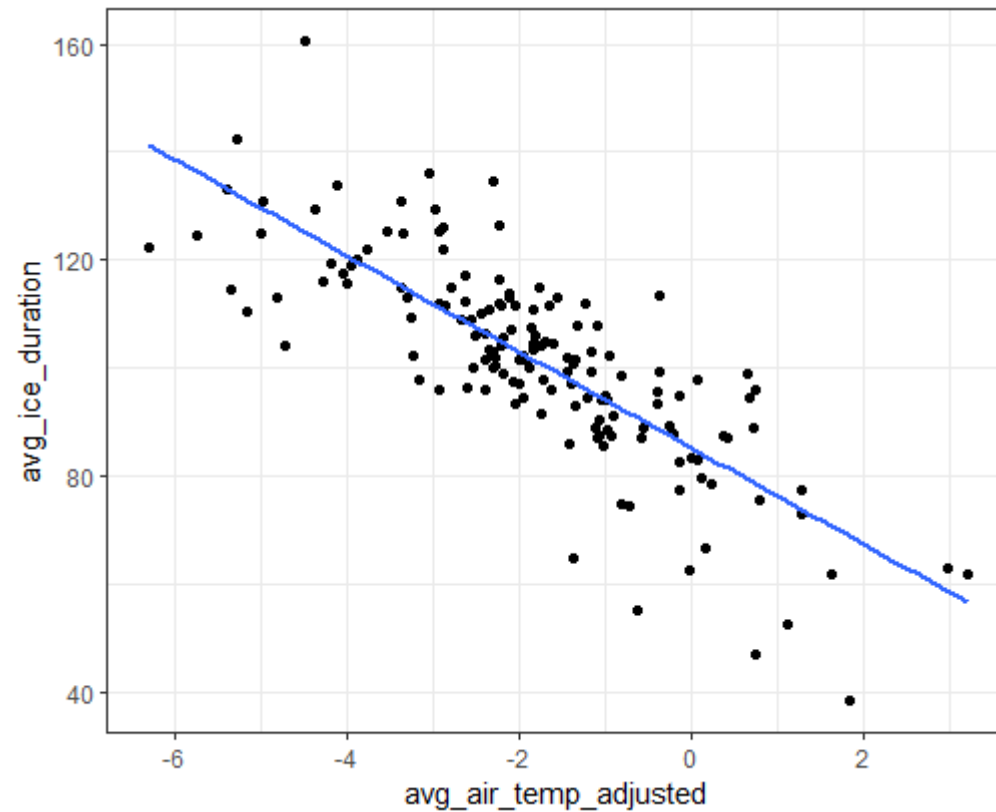
Đồ thị hộp – Box plots

Source Wilke (2018)

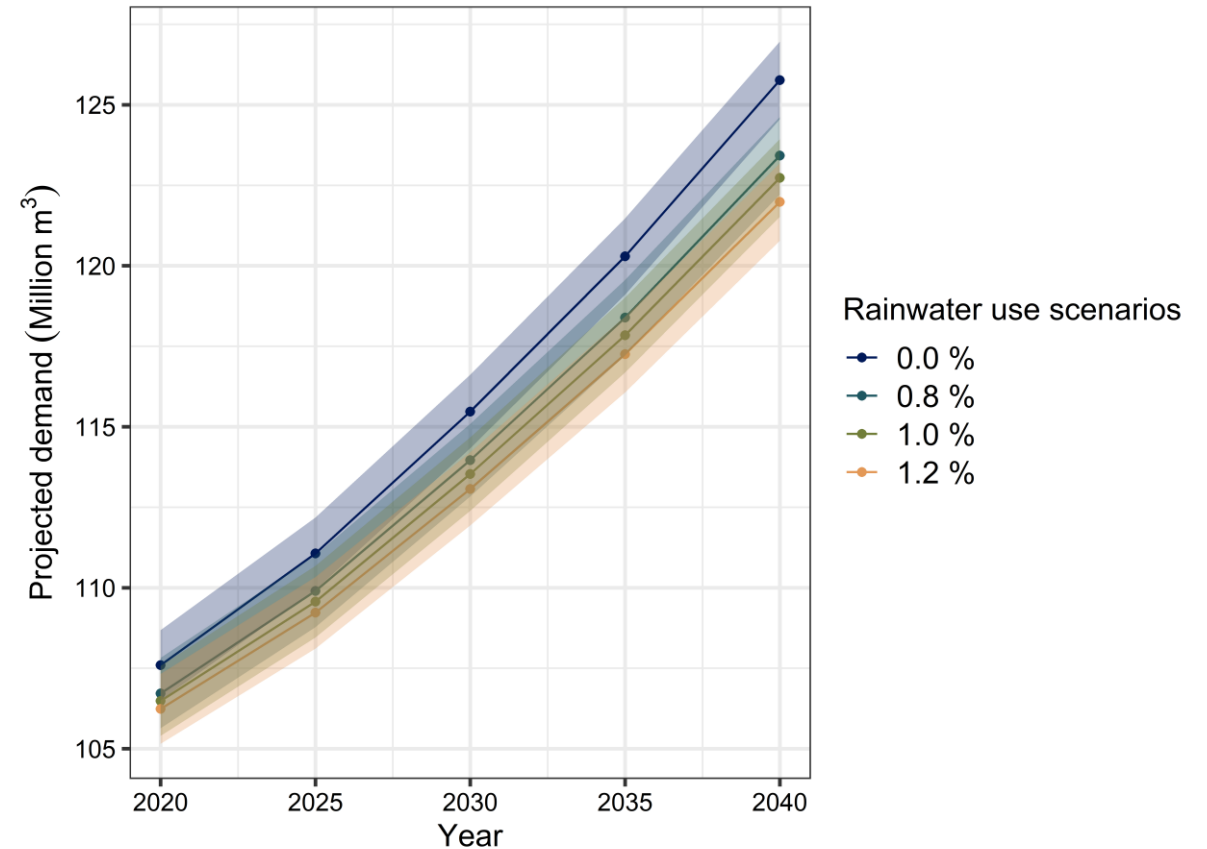
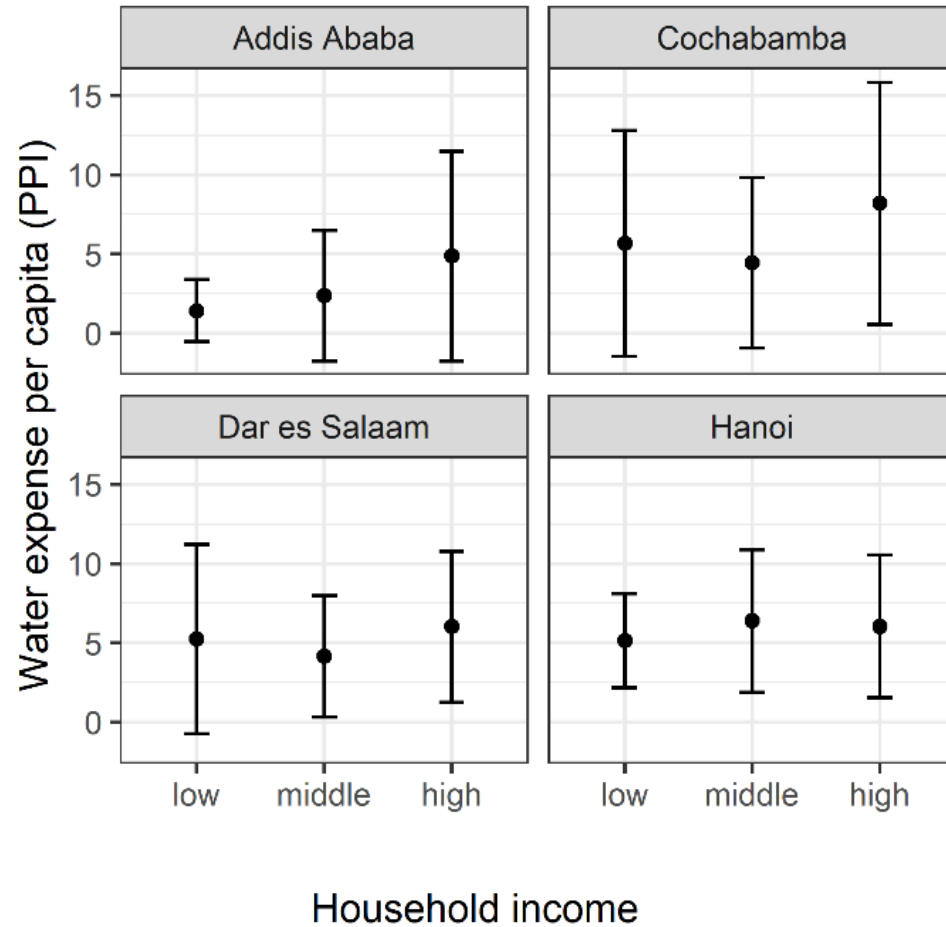
Đồ thị thông thường cho biến gián đoạn



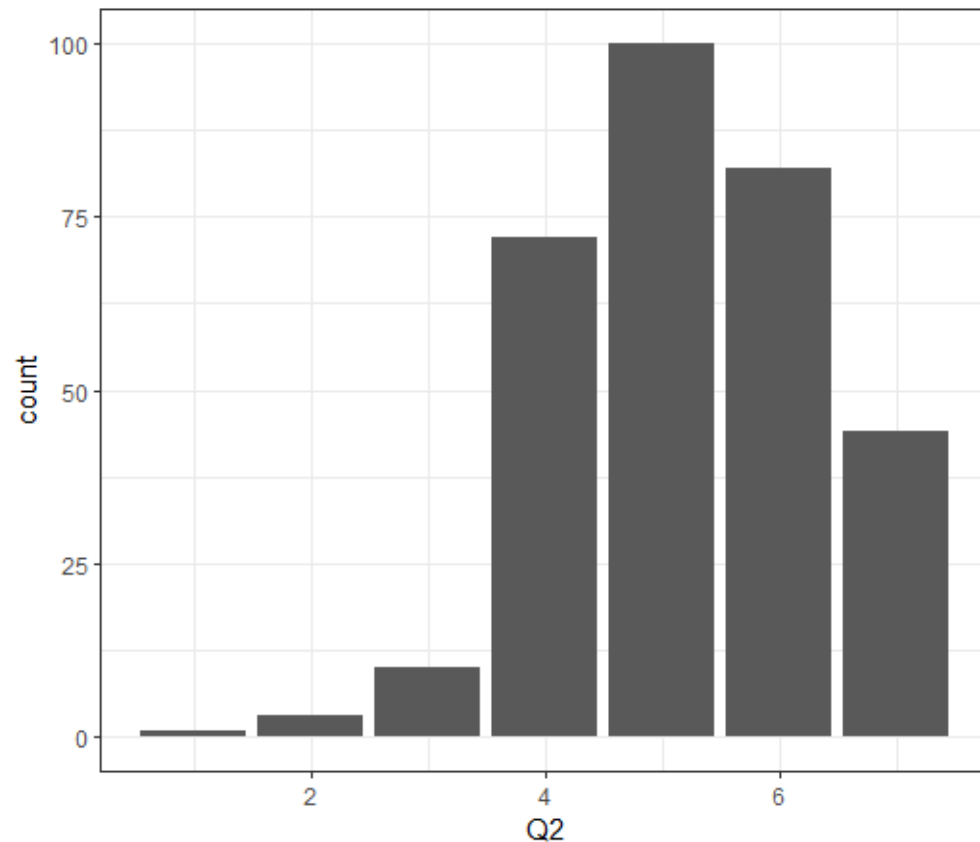
Biểu diễn quan hệ giữa các biến liên tục



Biểu diễn quan hệ giữa biến liên tục và gián đoạn



Thống kê mô tả - biến gián đoạn

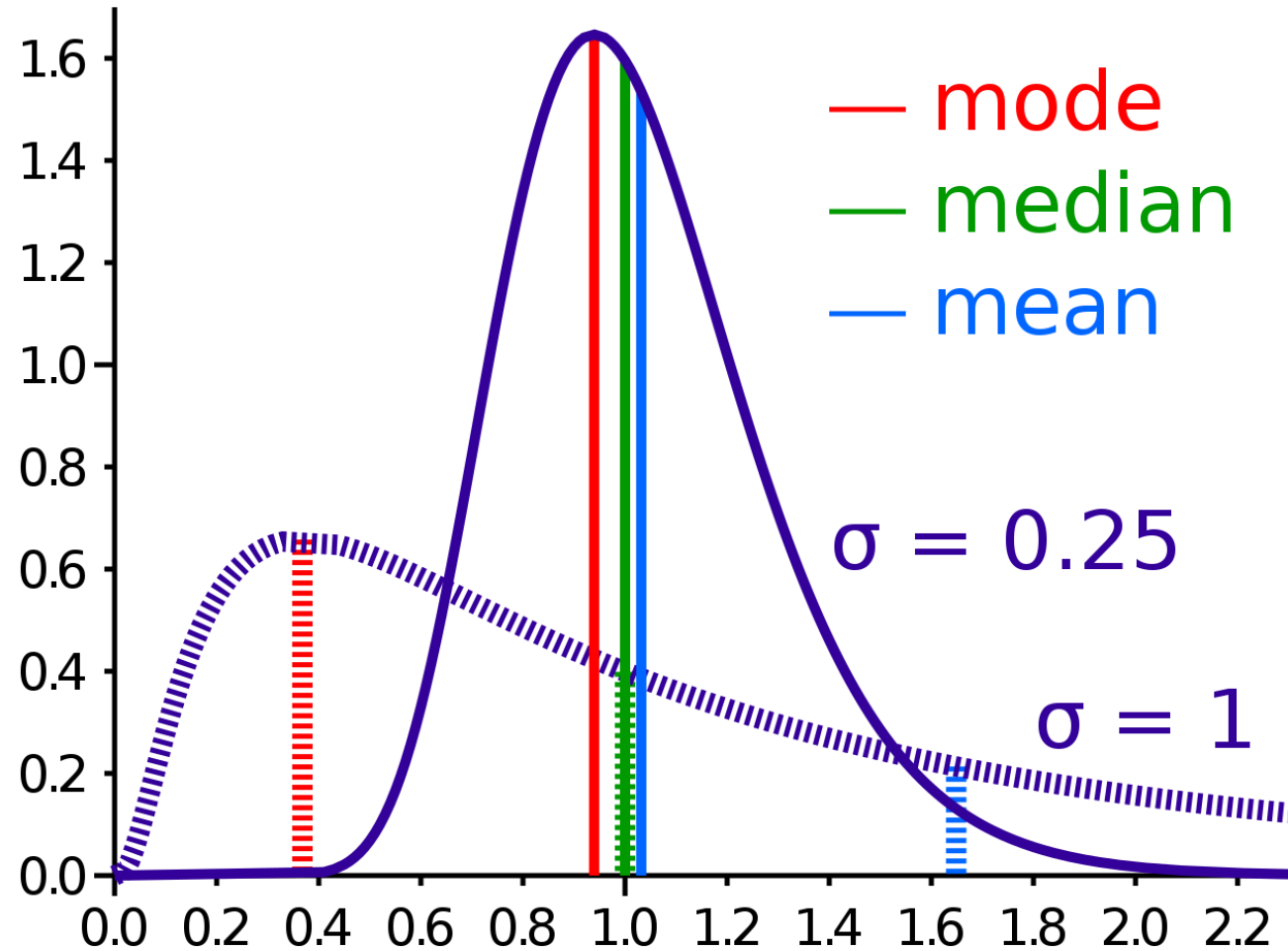


	Q2	count	percent
1	1	1	0.3205128
2	2	3	0.9615385
3	3	10	3.2051282
4	4	72	23.0769231
5	5	100	32.0512821
6	6	82	26.2820513
7	7	44	14.1025641

Thống kê mô tả - Biến liên tục

Đo độ tập trung

Trung bình, trung vị, mode

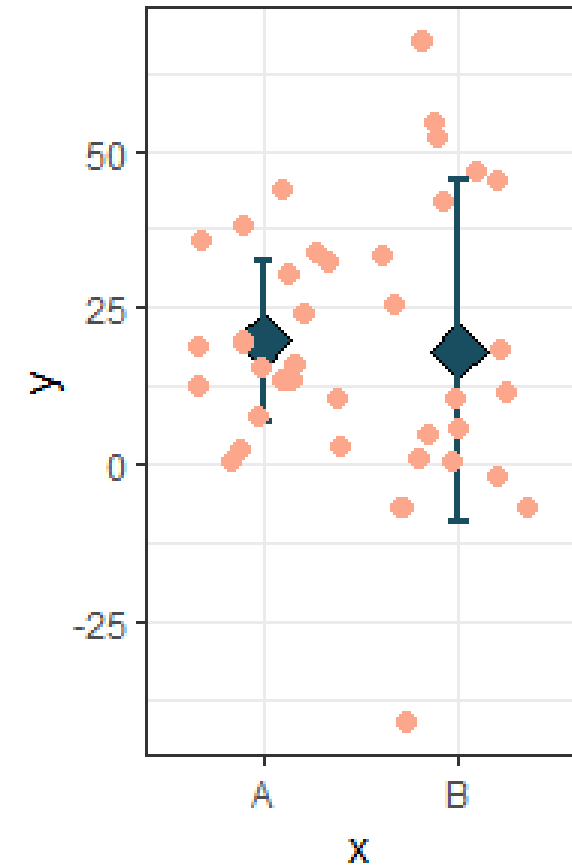
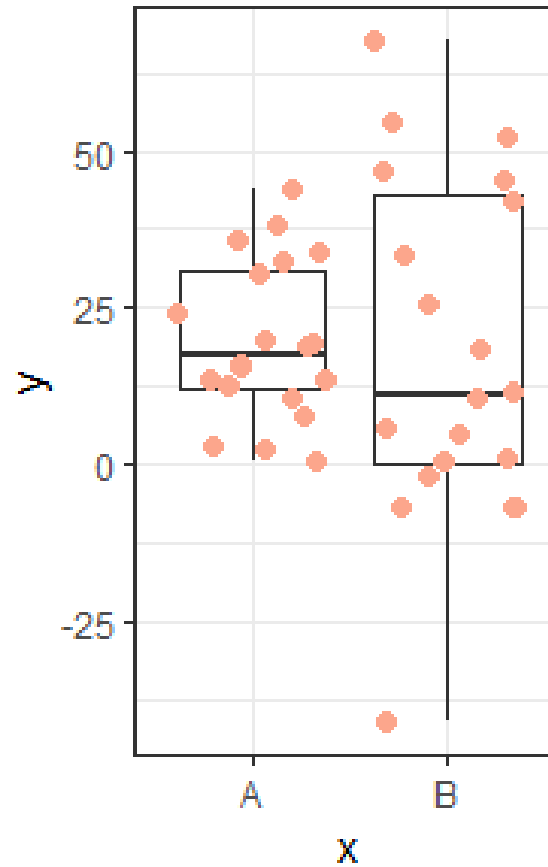


[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Thống kê mô tả - Biến liên tục

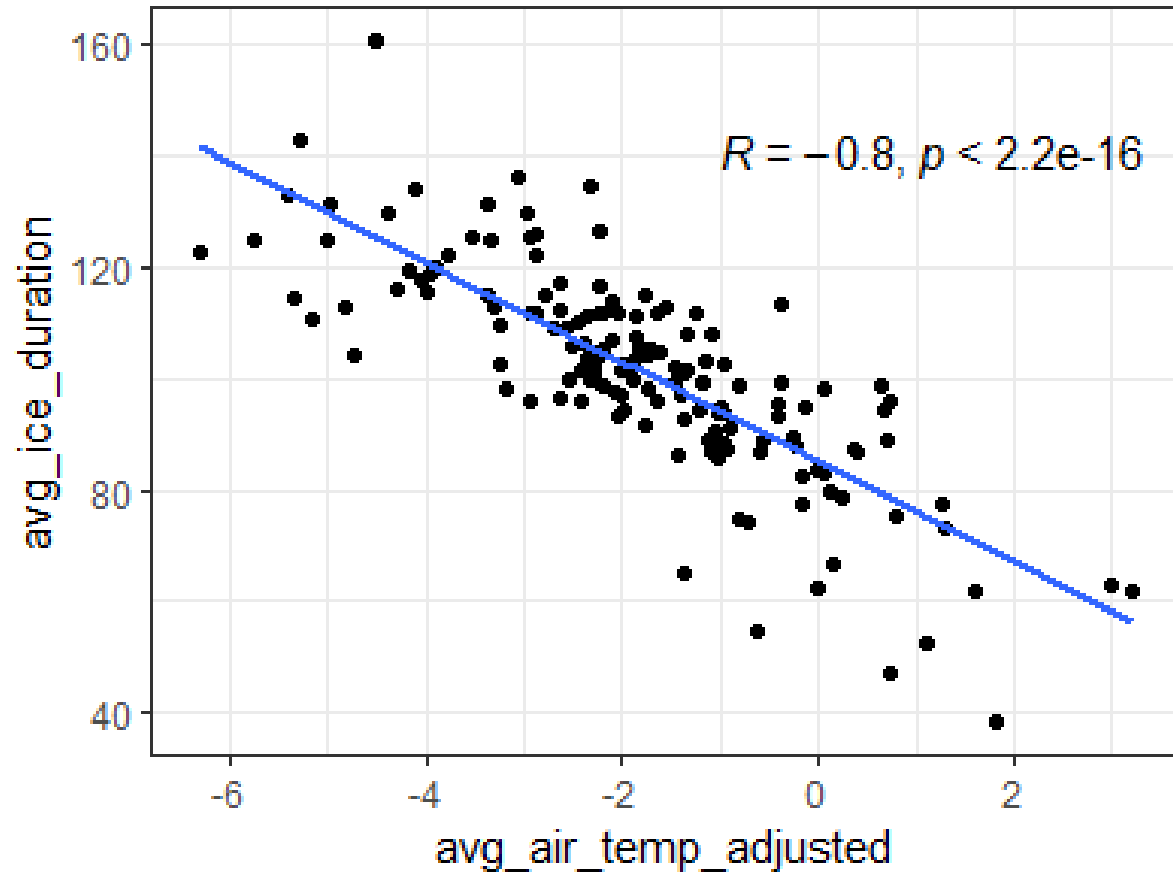
Đo độ phân tán

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

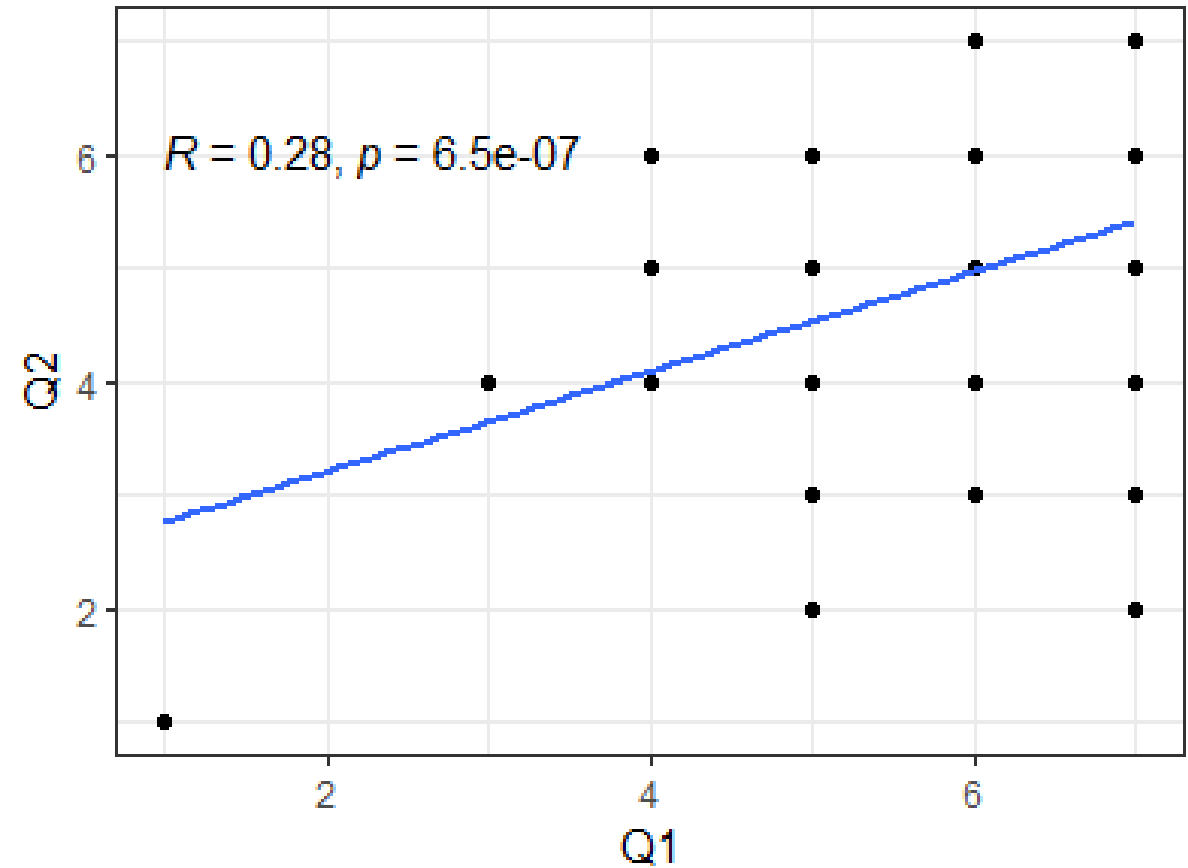


Thống kê mô tả - tương quan

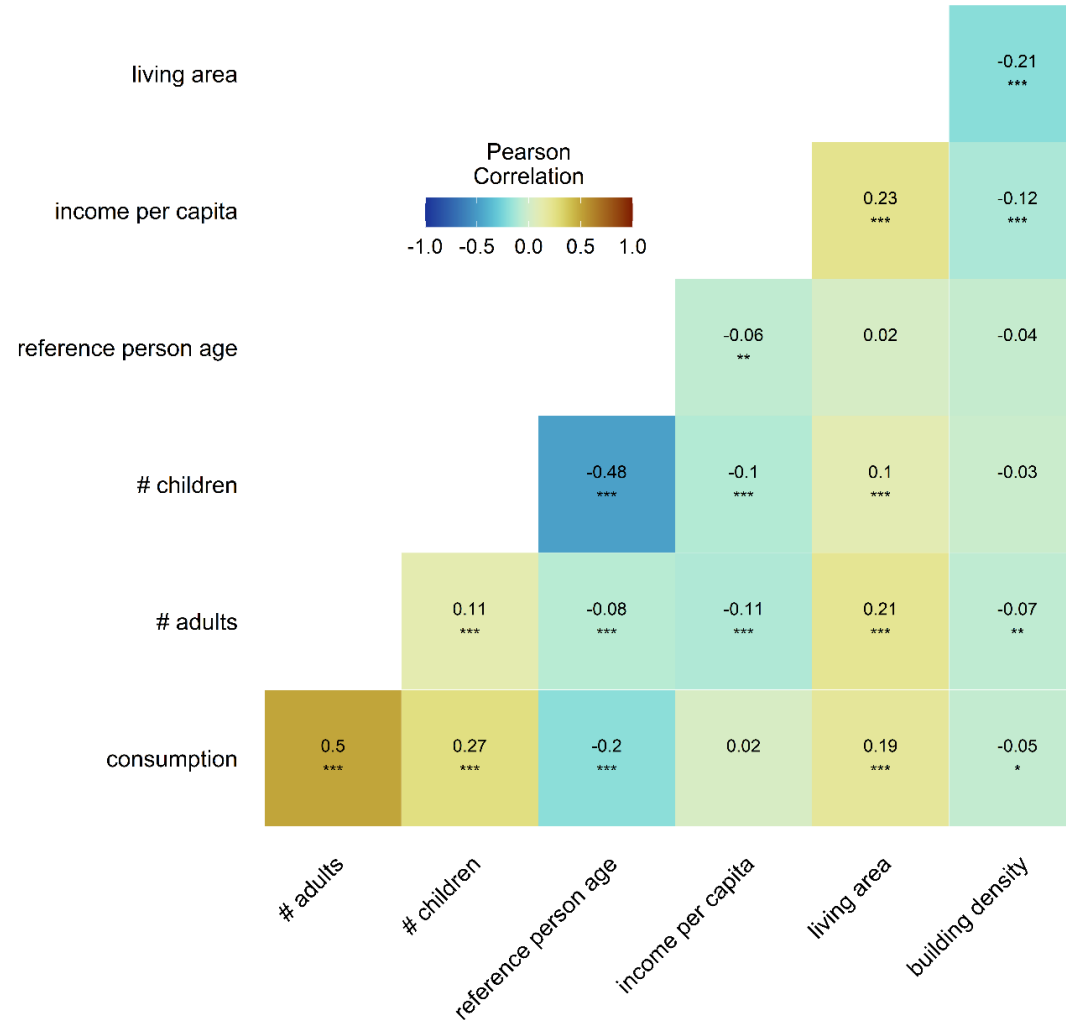
Pearson



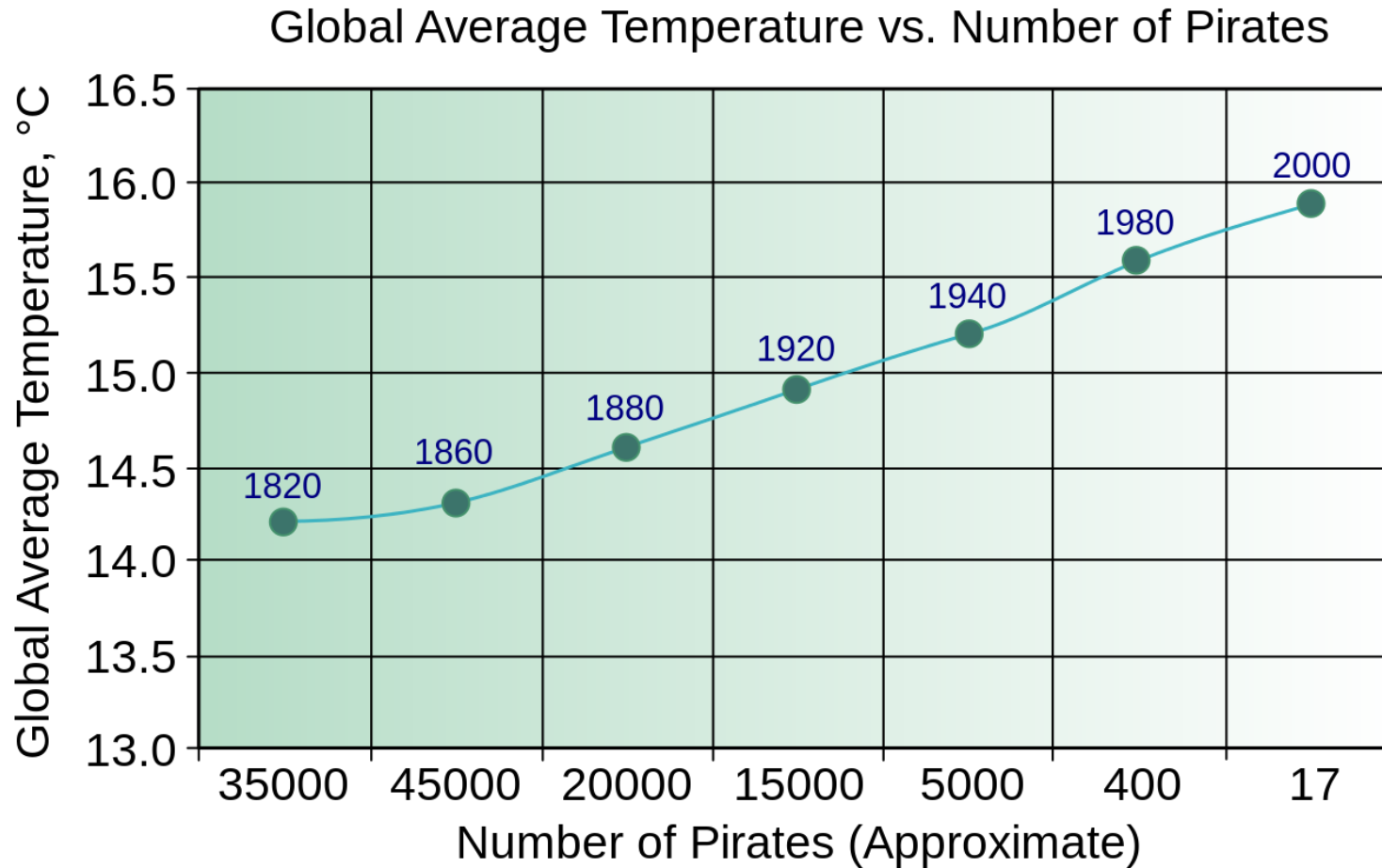
Spearman



Thống kê mô tả - Tương quan



Tương quan và quan hệ nhân quả

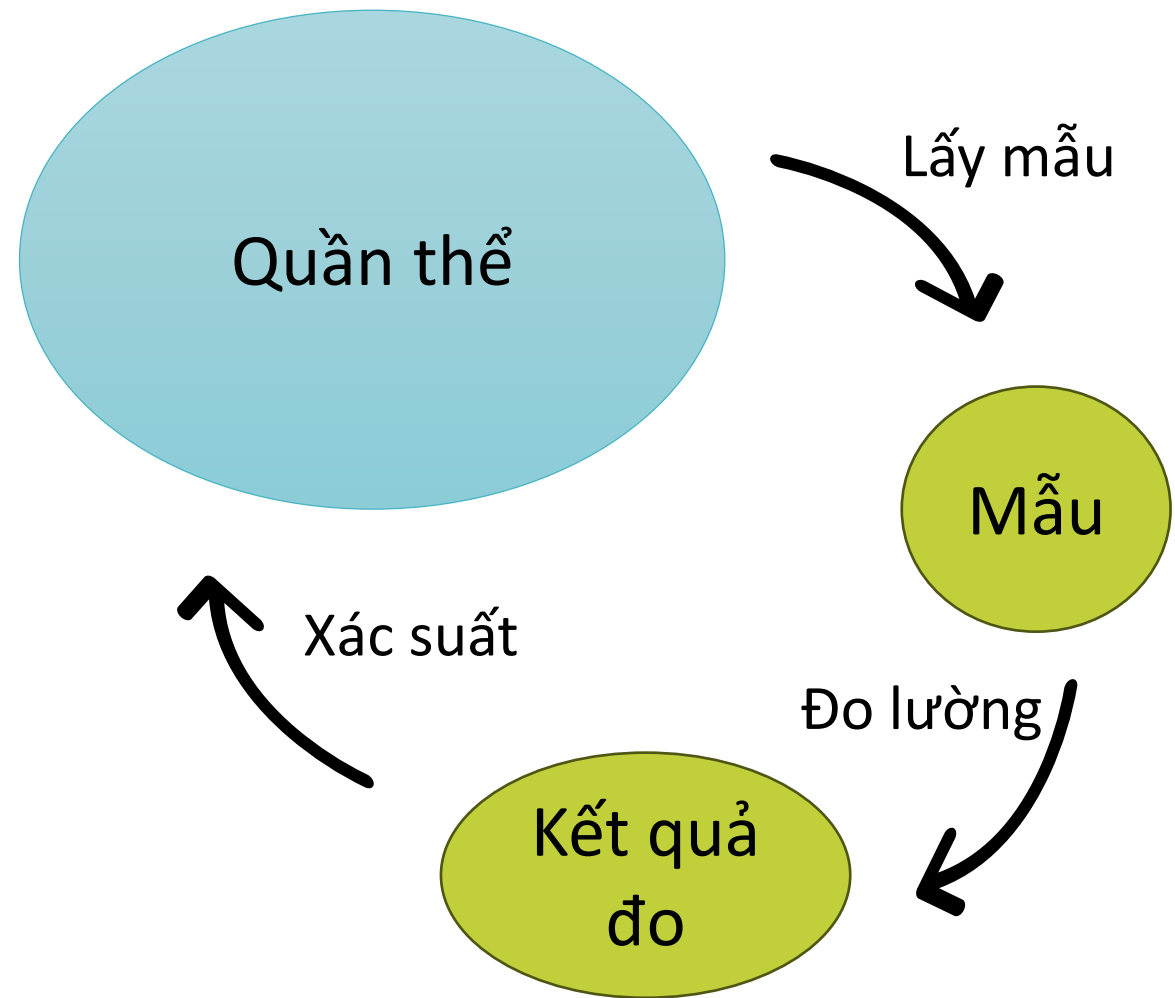


Bài tập

- Tìm hiểu và biểu diễn dữ liệu nước sạch tiêu thụ
- Tính toán một vài đại lượng thống kê miêu tả của dữ liệu nước sạch tiêu thụ

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

- Việc tăng nhiệt độ không khí có làm tăng phong hóa đá vôi?
- Thu nhập của các hộ gia đình có làm du lịch cộng đồng khác hay không khác thu nhập hoàn toàn từ nông nghiệp trước đây?

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

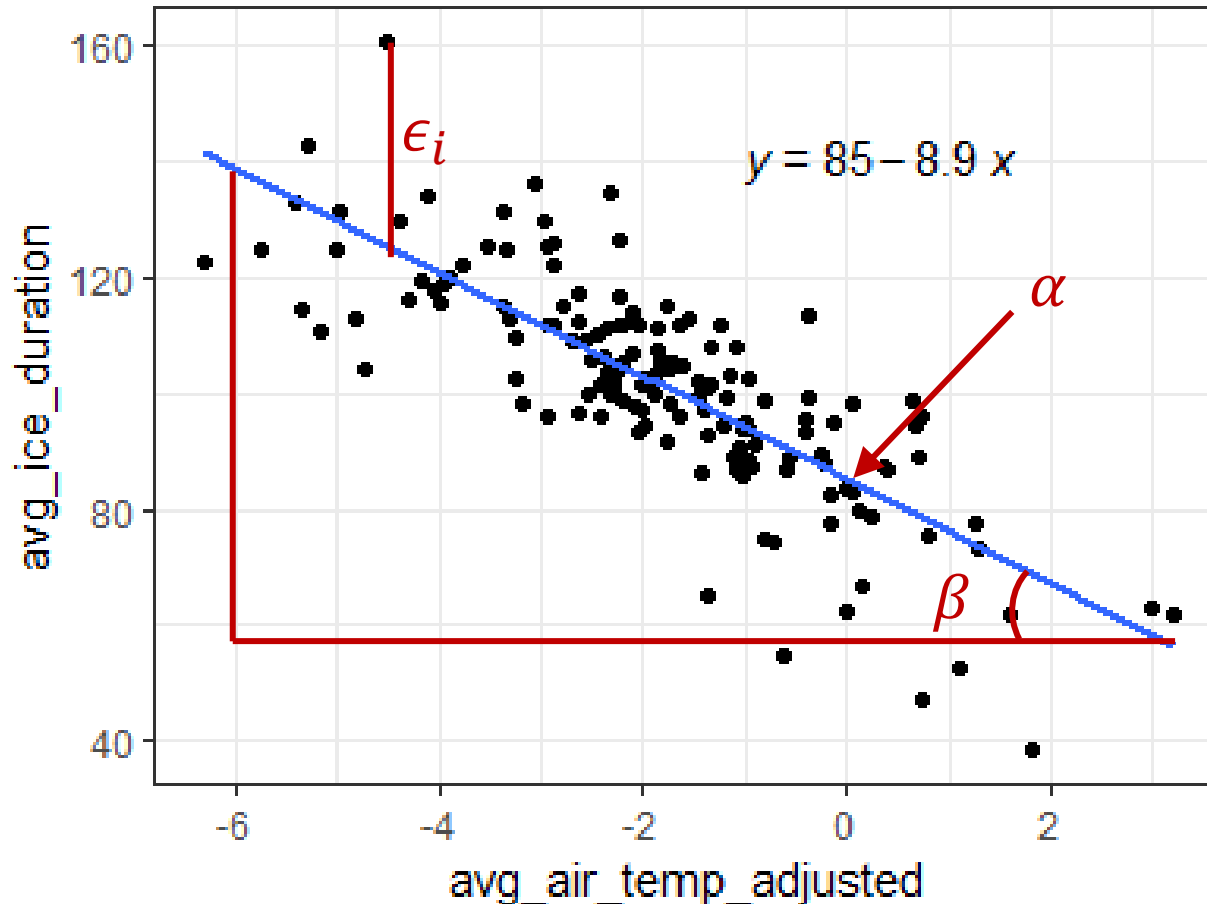
ANOVA

- 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

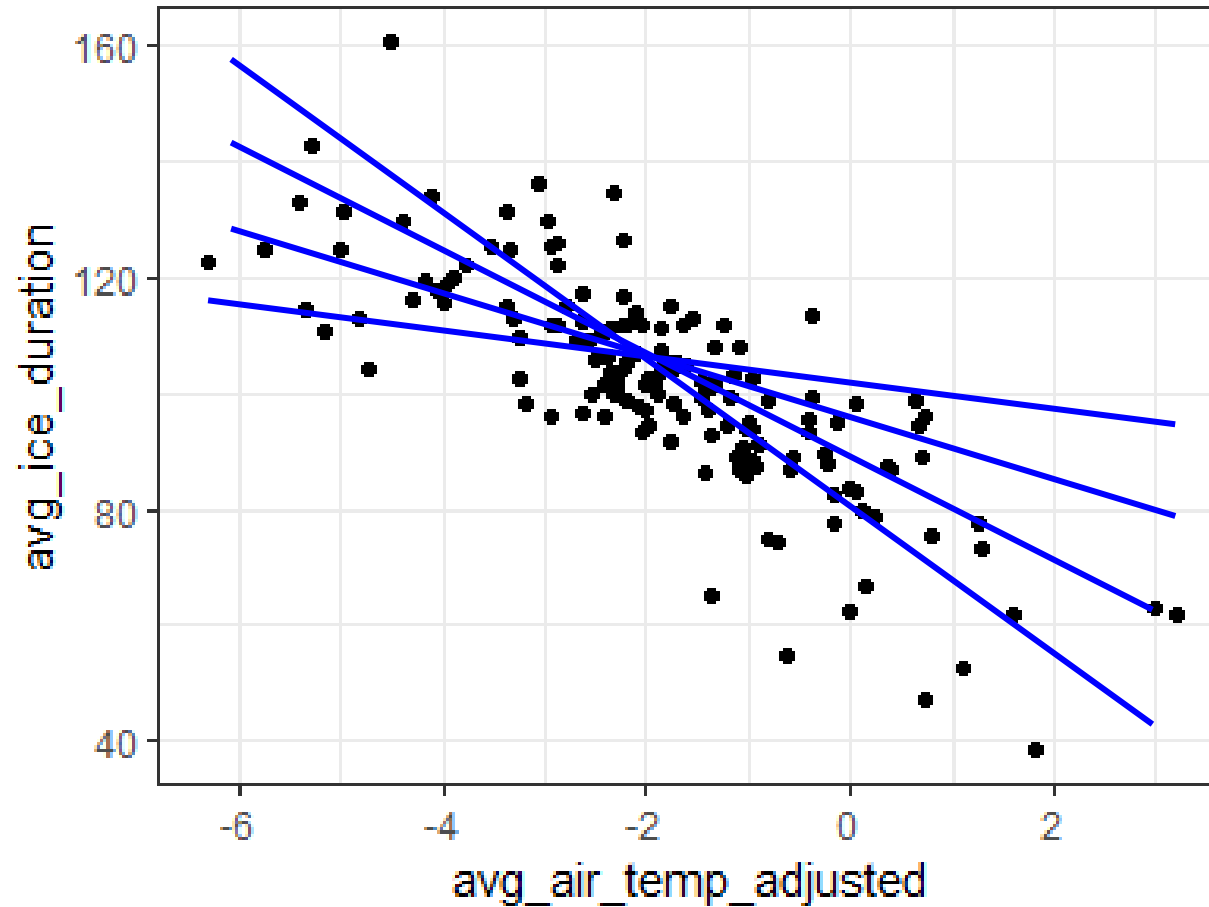
Residual standard error: 11.43 on 150 degrees of freedom

(14 observations deleted due to missingness)

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

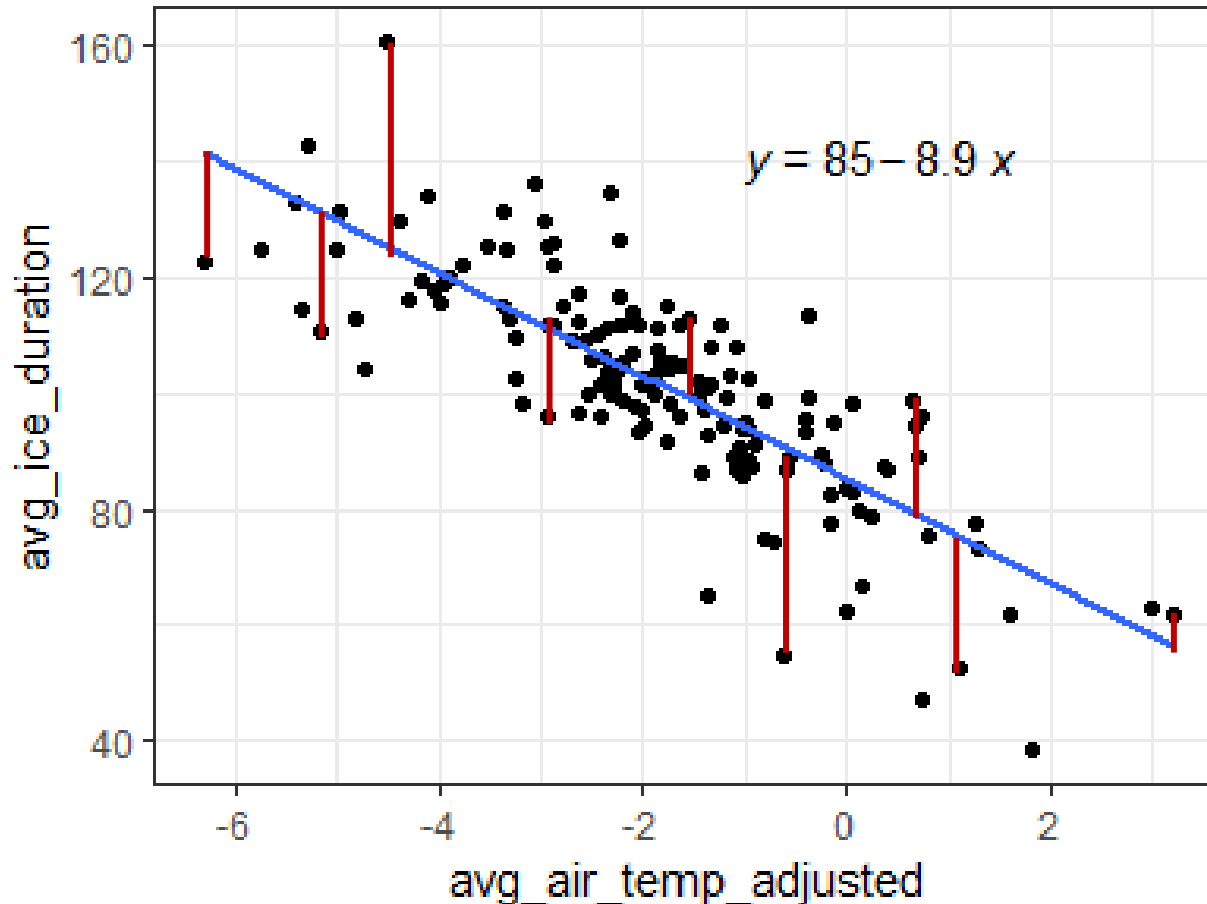
F-statistic: 260.2 on 1 and 150 DF, p-value: < 2.2e-16

Hồi quy tuyến tính đơn giản



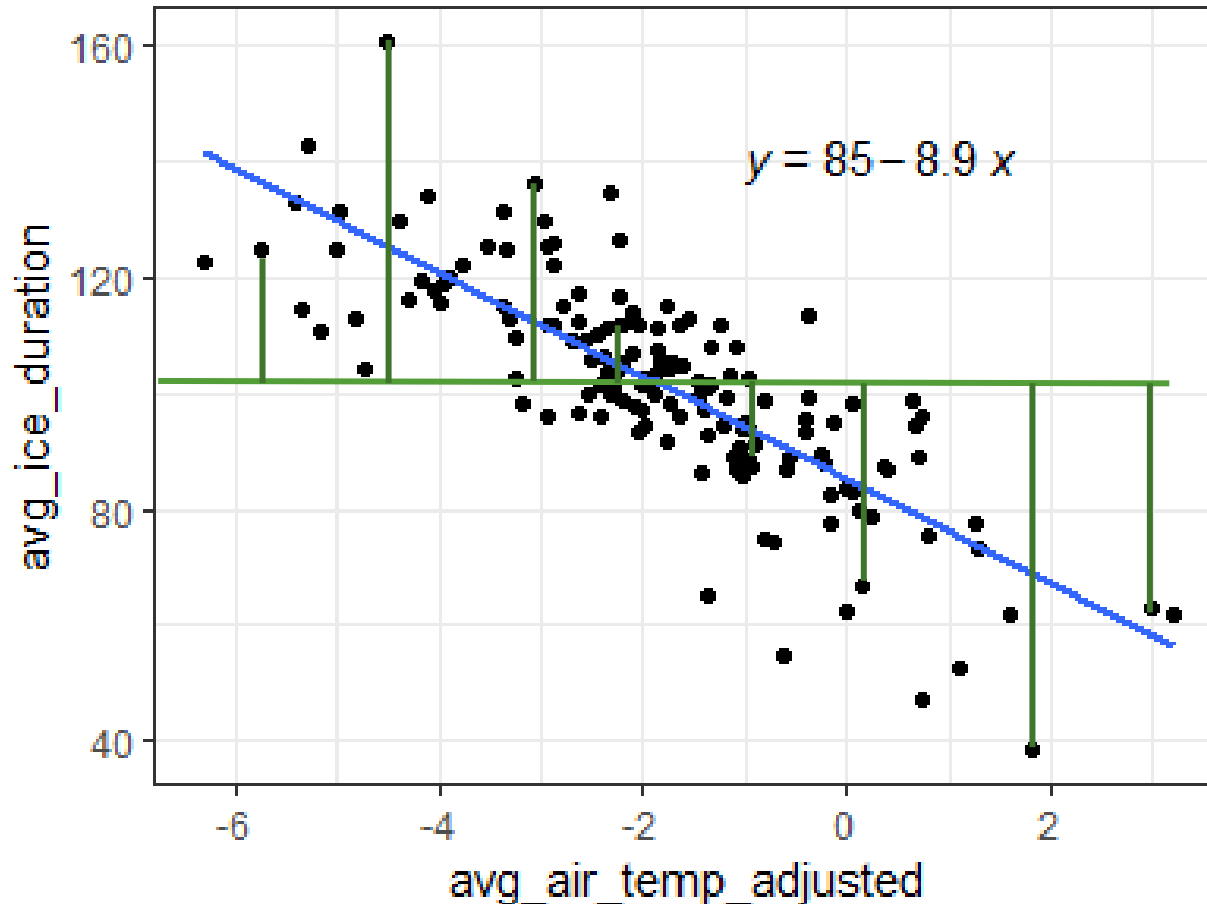
Đường nào?

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)

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

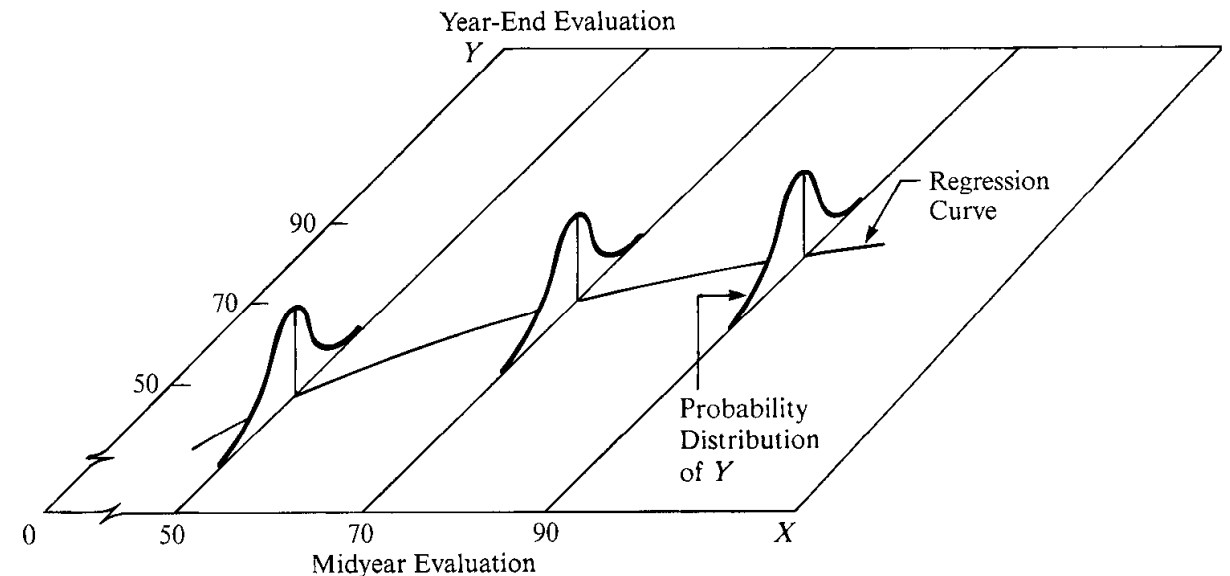
TSS: Total sum of squares

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

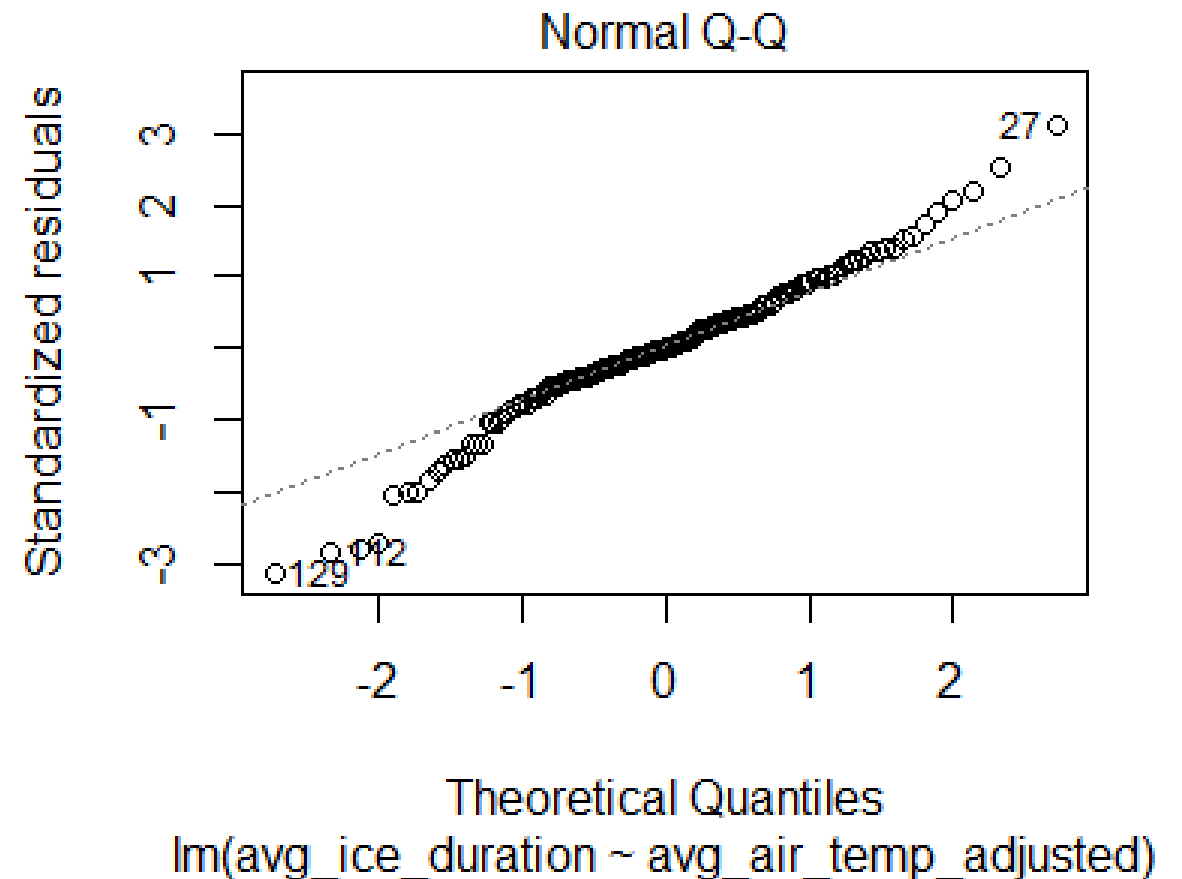
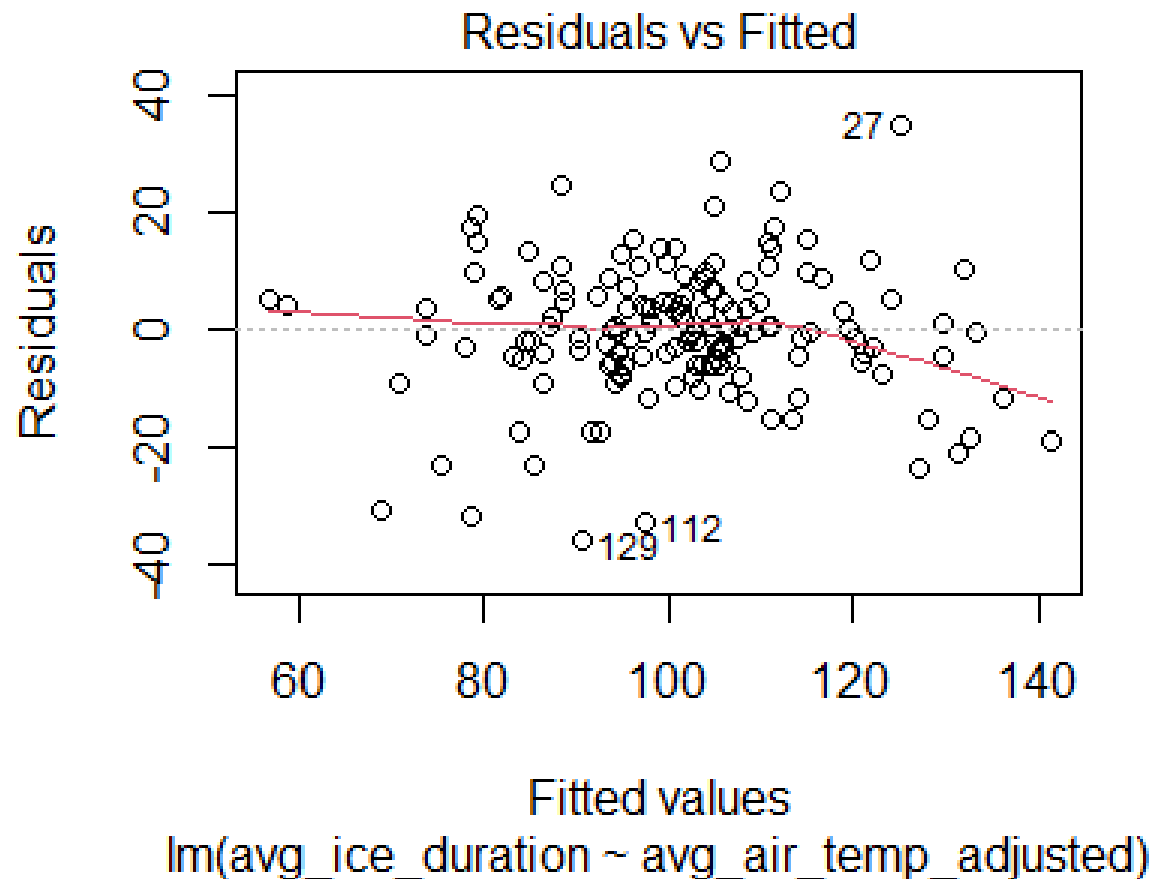
! Giả định phân phối chuẩn là cho phần dư/sai số ngẫu nhiên ϵ không phải biến phụ thuộc Y

! 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

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

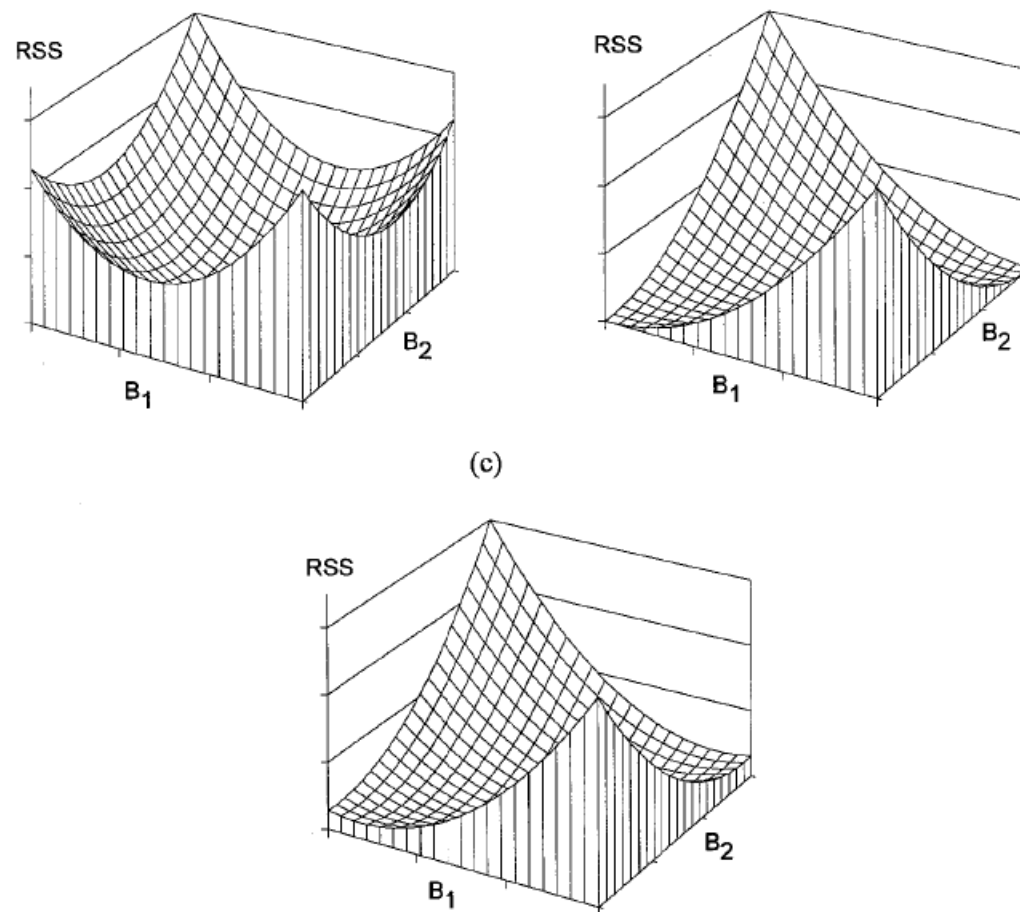
Ví dụ:

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

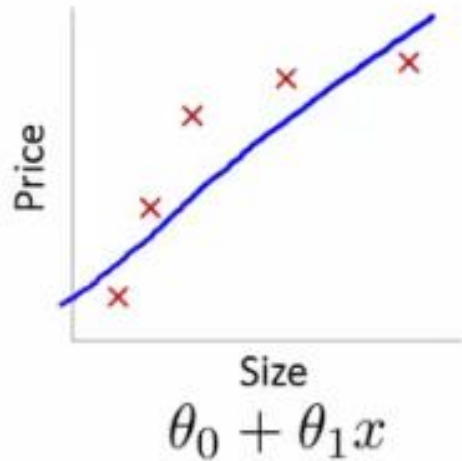
Mô hình càng phức tạp càng tốt?

Tương quan giữa các biến độc lập Multicollinearity

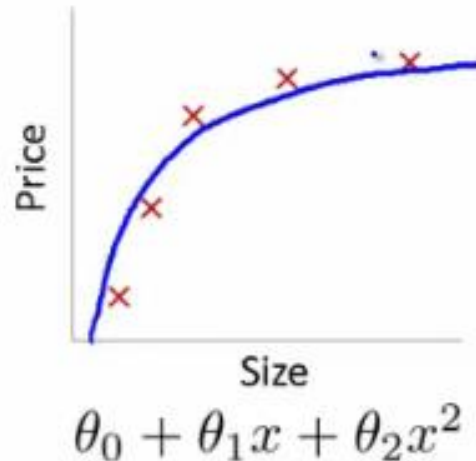
Variance Inflation Factor
(Yếu tố lạm phát phương sai?)



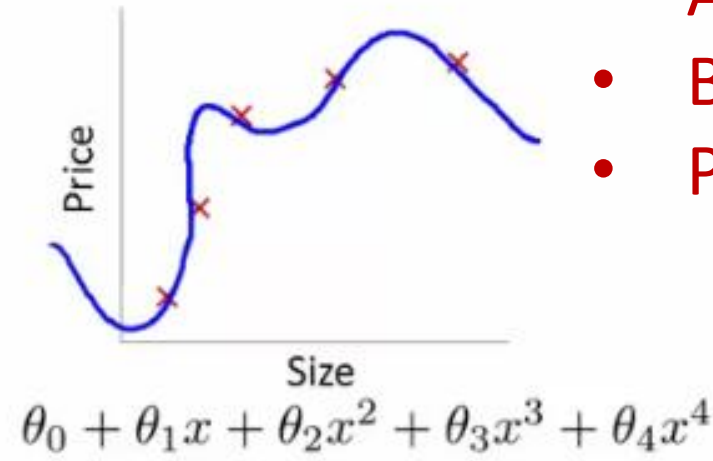
Quá khớp hoặc thiếu khớp với số liệu



High bias
(underfit)



"Just right"



High variance
(overfit)

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

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

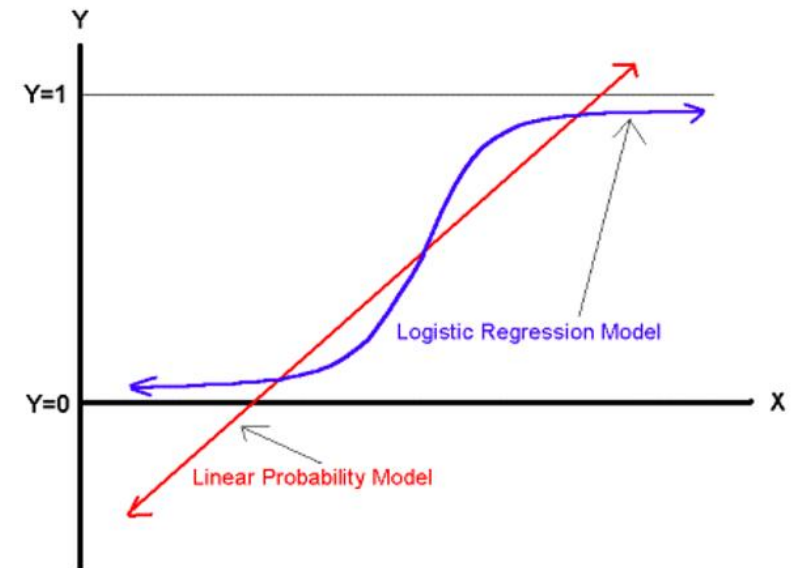
Khi Y là biến gián đoạn?

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

Định danh: Multinomial logistic regression

Thứ bậc: Cumulative logistic regression

Biến đếm: Poisson regression

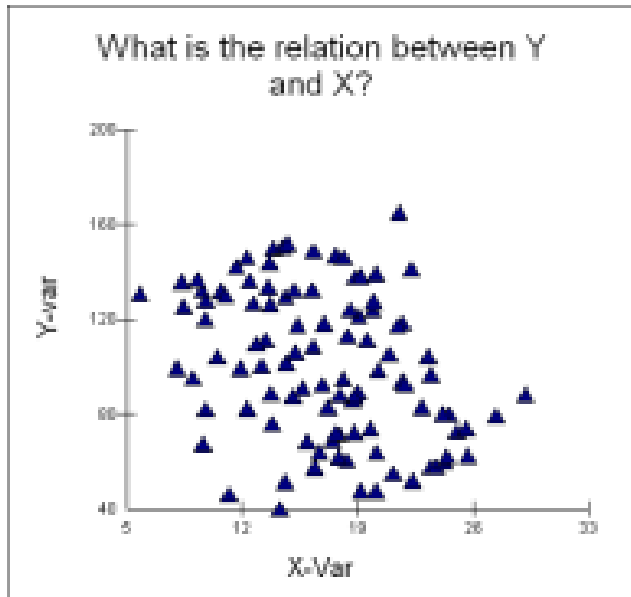


Một số kỹ thuật định lượng khác

- 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)
- ...

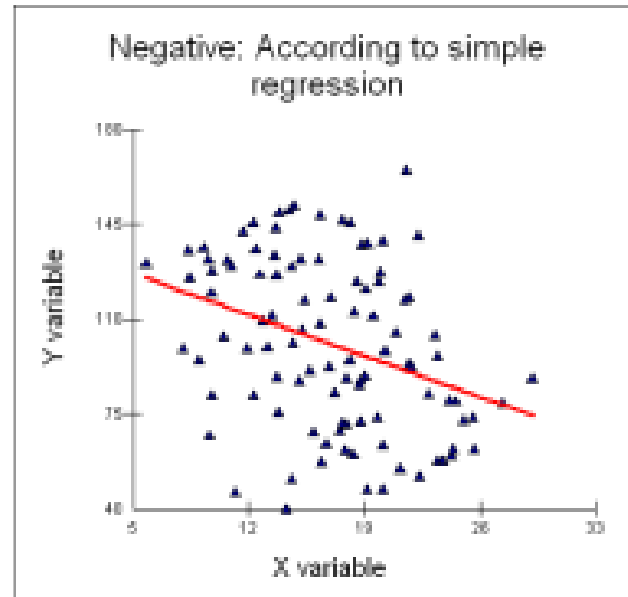
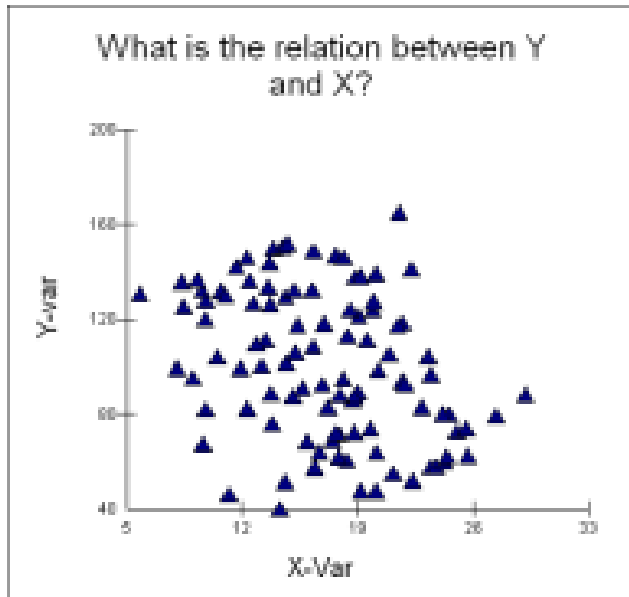
Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model



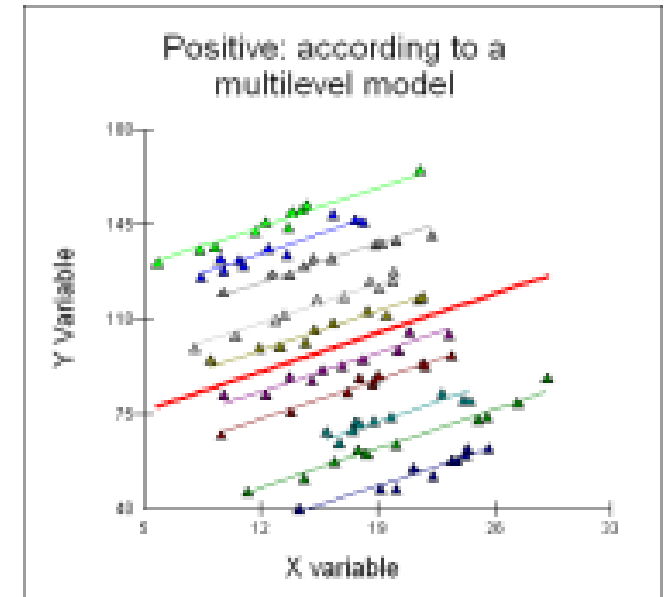
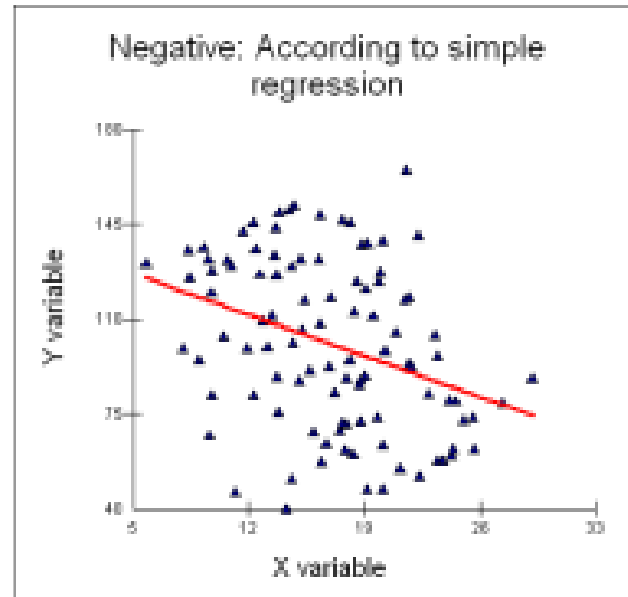
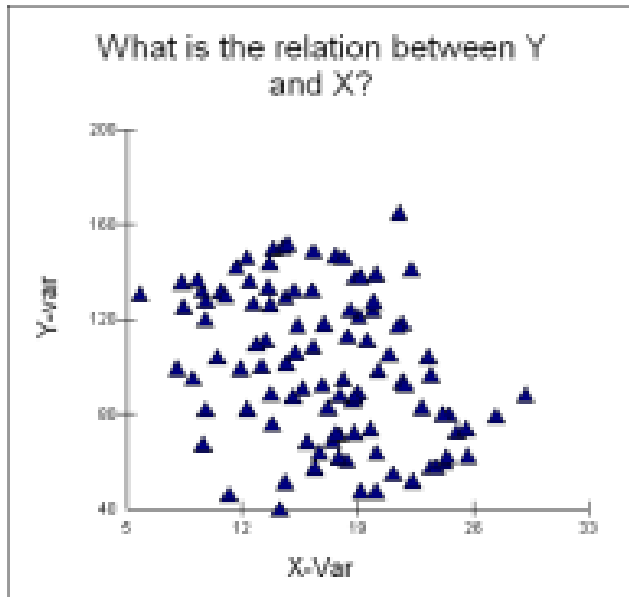
Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model



Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model

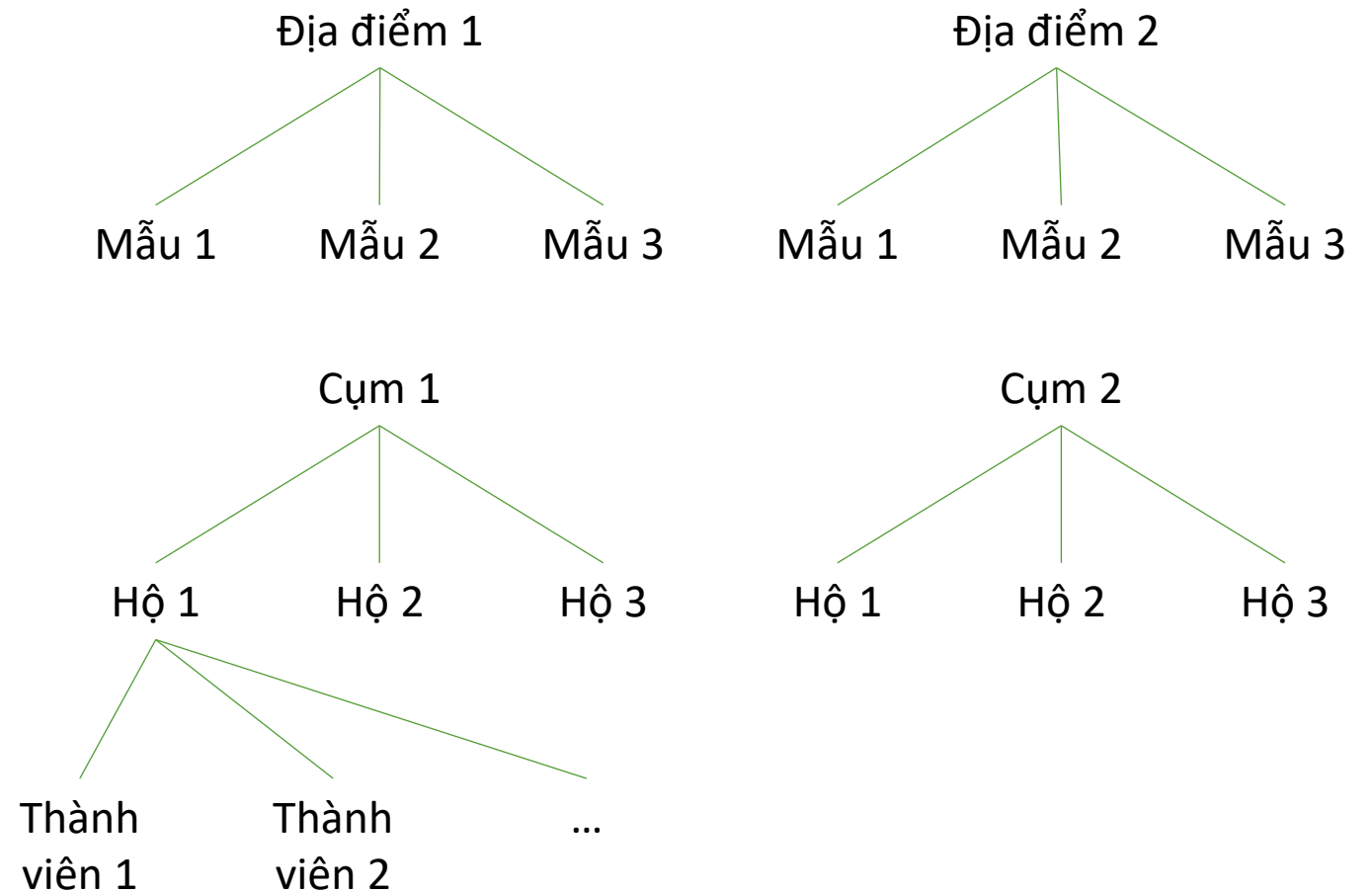


Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

Multilevel model/Mixed effect model

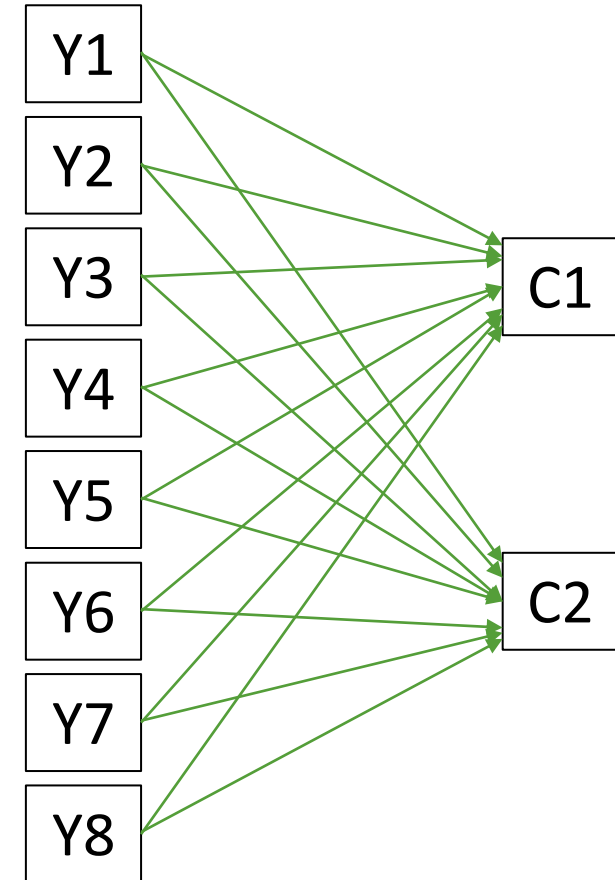


Image by [Chelsea Parlett-Pelleriti](#)



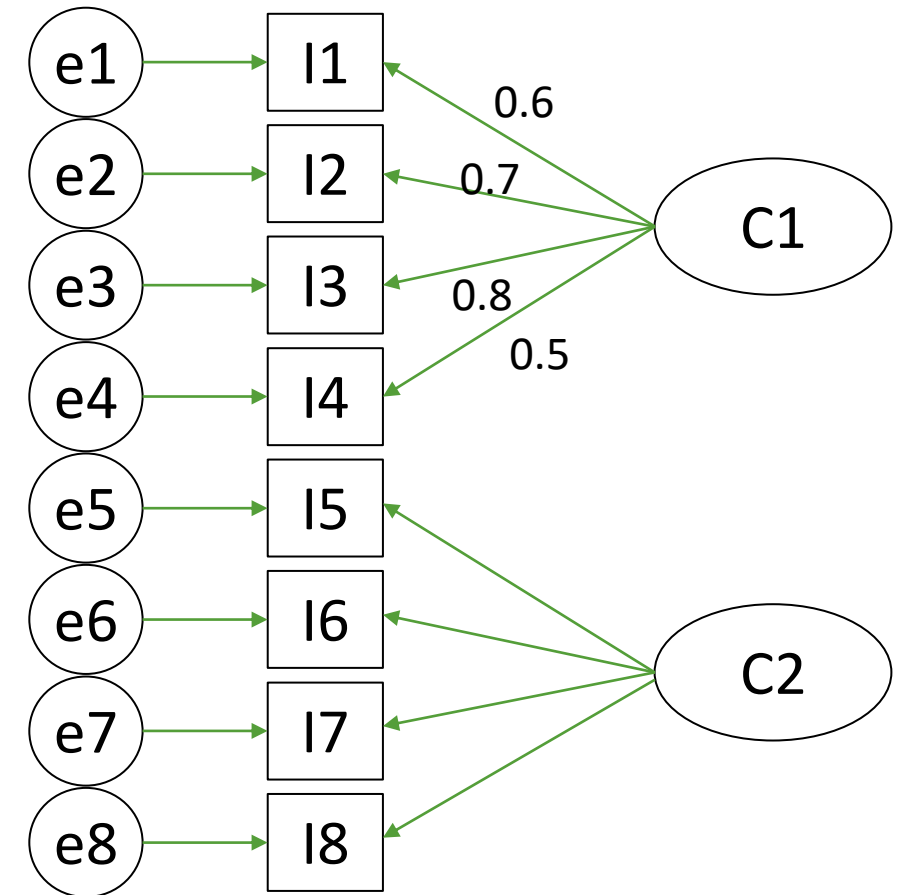
Phân tích thành phần chính (PCA - Principal component analysis)

- 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)

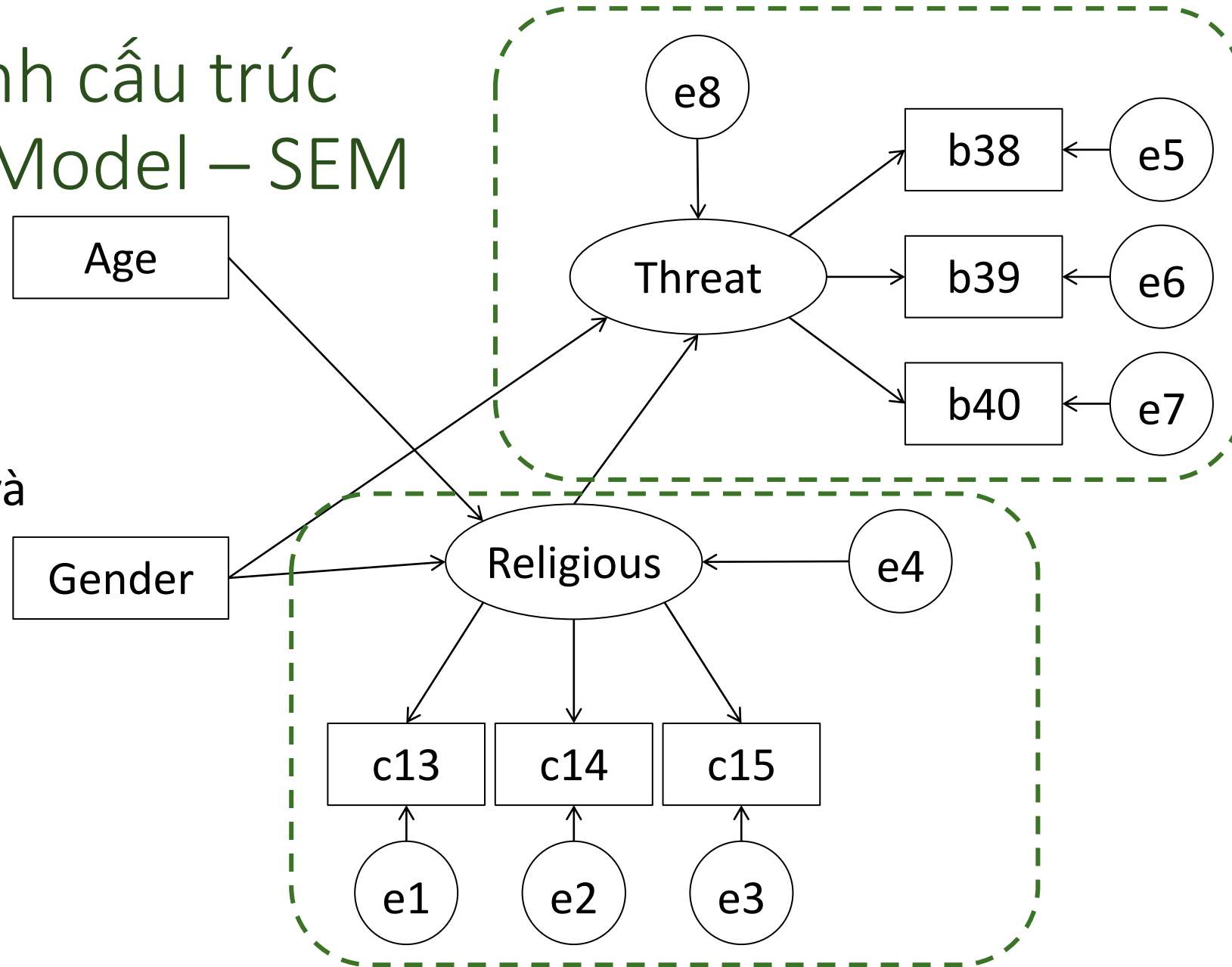
- Thường áp dụng cho dữ liệu bảng hỏi
- Đo phạm trù tiềm ẩn (Latent construct)
- Phân tích nhân tố khám phá/khẳng định (Exploratory/Confirmatory Factor Analysis)



Mô hình phương trình cấu trúc

Structural Equation Model – SEM

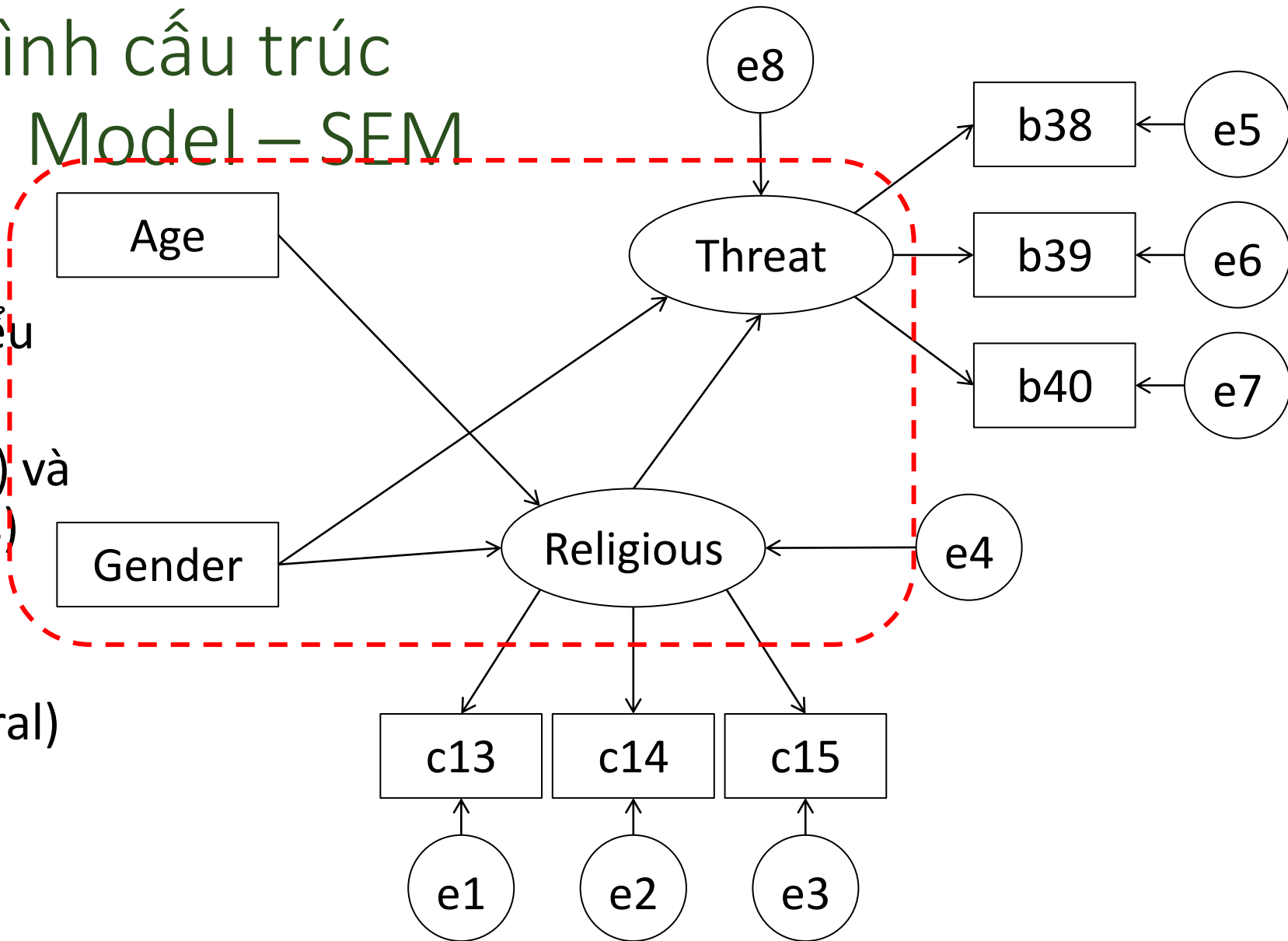
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement)



Mô hình phương trình cấu trúc

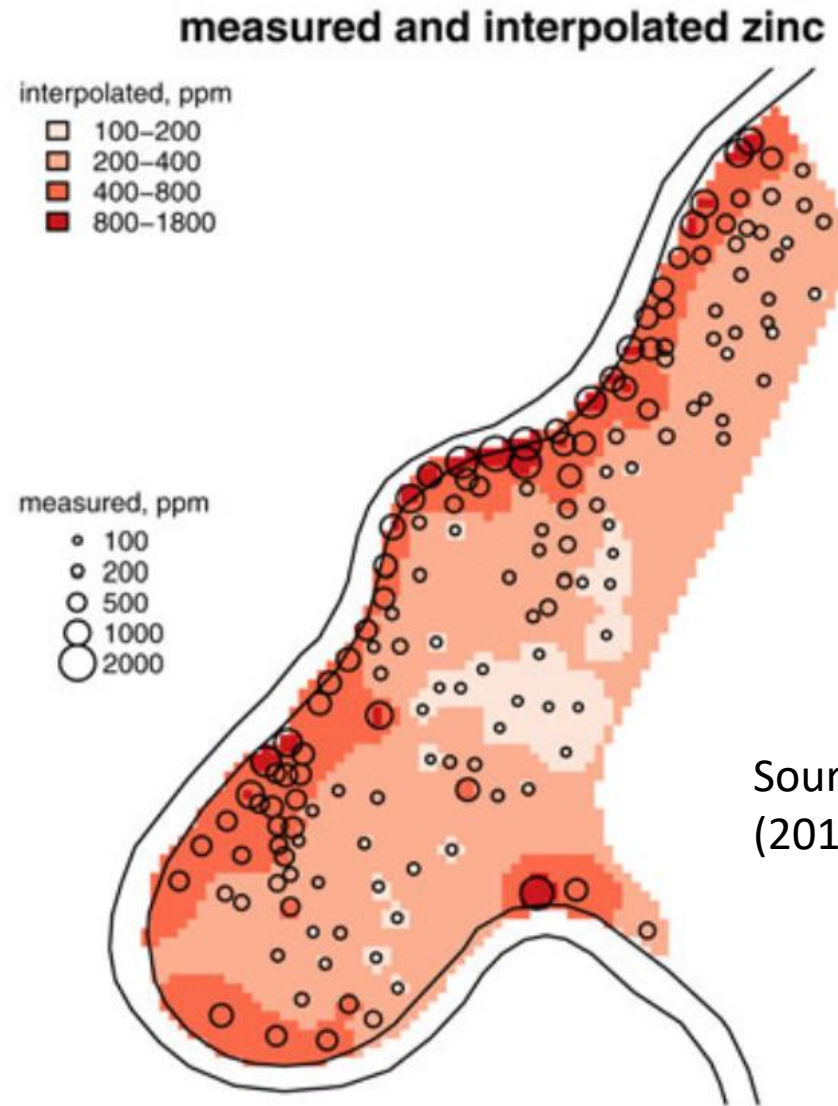
Structural Equation Model – SEM

- 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)
- Tác động trực tiếp và gián tiếp (Direct vs indirect effects)



Thống kê không gian Spatial Statistics

- Tương quan không gian
- GIS



Source: Bivand
(2013)

Phân tích định tính

Các phương pháp phân tích định tính

- Dân tộc học/Quan sát khách thể (Ethnographic)
- Phân tích trường hợp (Case study)
- Phỏng vấn sâu (One-on-one interview)
- Thảo luận nhóm tập trung (Focus group)
- Phân tích tài liệu (Records analysis)

Phỏng vấn sâu

- Phỏng vấn không cấu trúc (unstructured) và bán cấu trúc (semi-structured)
- Tập trung vào các câu hỏi mở, thăm dò
- Đòi hỏi kỹ năng phỏng vấn và ý thức rõ ràng về mục đích nghiên cứu
- Ưu điểm: mang lại nhiều thông tin, mang tính khám phá
- Nhược điểm: đòi hỏi nhiều thời gian và kỹ năng, có những nghi ngại về kết quả nghiên cứu