Bài giảng 6: Xử lý Khuyết dữ liệu trong hồi quy

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

Vấn đề khuyết dữ liệu

Khuyết dữ liệu là vấn đề xảy ra khá phổ biến trong các dữ liệu được thu thập thực tế:

- không phản hồi trong mẫu khảo sát;
- thoát ra khỏi nghiên cứu dài;
- thí nghiệm bị lỗi;
- người dùng không cung cấp thông tin.

Việc thiếu hụt thông tin sẽ khiến cho việc ước lượng và phân tích dữ liệu trở nên thiếu chính xác.

Vấn đề khuyết dữ liệu

	IQ	wbeing	jobperf
1	78	13	*NA*
2	84	9	*NA*
3	84	10	*NA*
4	85	10	*NA*
5	87	*NA*	*NA*
6	91	3	*NA*
7	92	12	*NA*
8	94	3	*NA*
9	94	13	*NA*
10	96	*NA*	*NA*
11	99	6	7
12	105	12	10

- `IQ`: điểm số IQ của người dự tuyển;
- `wbeing`: điểm số hạnh phúc của người dự tuyển;
- `jobperf`: điểm số hiệu suất công việc của người dự tuyển (sau thời gian thử việc 6 tháng).

Ví dụ minh họa việc bỏ qua dữ liệu khuyết

Ta xét ví dụ minh họa sau về ảnh hưởng của việc bỏ qua dữ liệu khuyết trong phân tích dữ liệu.

- $X_1 \sim \mathcal{N}(0, 1)$;
- $X_2 \sim \mathcal{B}(0.4)$;
- Mô hình hồi quy tuyến tính (chính xác)

$$Y = 1 + 2X_1 + X_2$$

Ta mong muốn ước lượng các hệ số của mô hình

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

với hai kịch bản:

- dữ liệu đầy đủ;
- dữ liệu khuyết X_1.

Ví dụ minh họa việc bỏ qua dữ liệu khuyết

Trong kịch bản dữ liệu bị khuyết X_1:

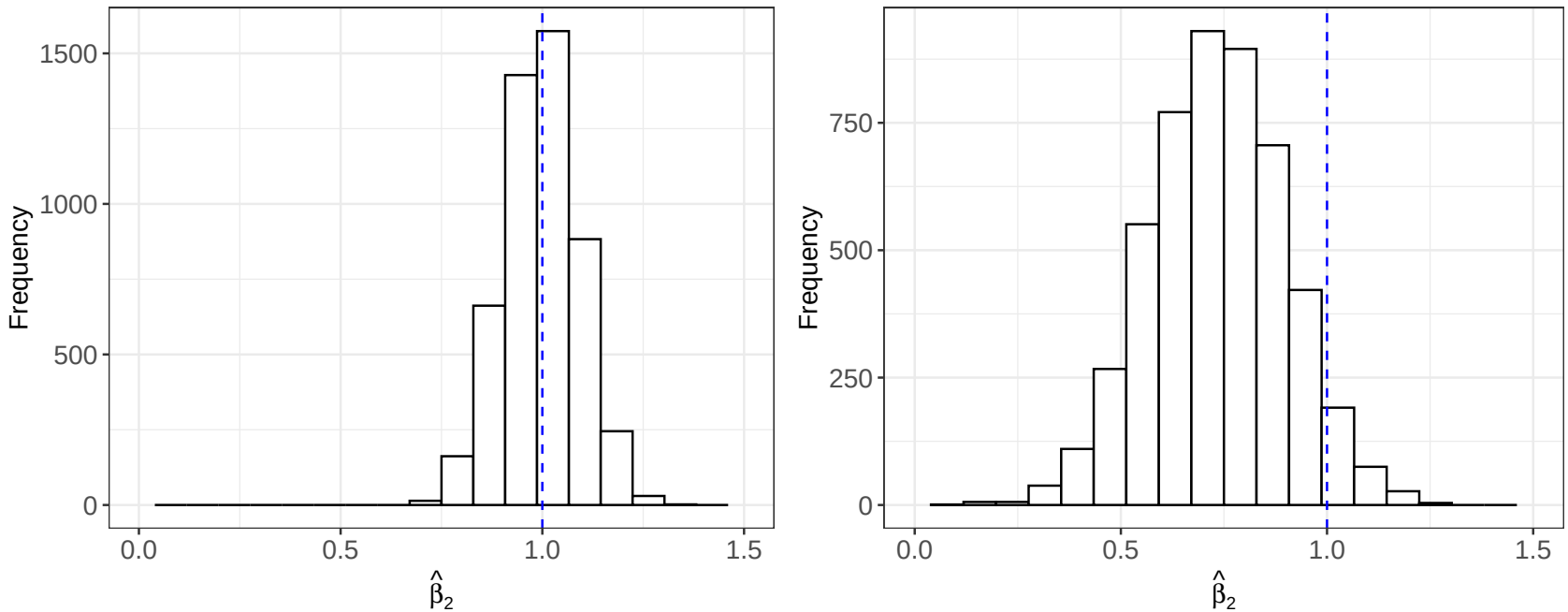
- Y và X_2 được quan sát đủ;
- X bị xóa (khuyết) tại một số quan sát với các xác suất:

$$p_{miss} = \begin{cases} 0.6 & \text{nếu } Y \leq 0, X_2 = 0 \\ 0.4 & \text{nếu } Y \leq 0, X_2 = 1 \\ 0.3 & \text{nếu } Y > 0, X_2 = 0 \\ 0.8 & \text{nếu } Y > 0, X_2 = 1 \end{cases}$$

Trong mọi kịch bản

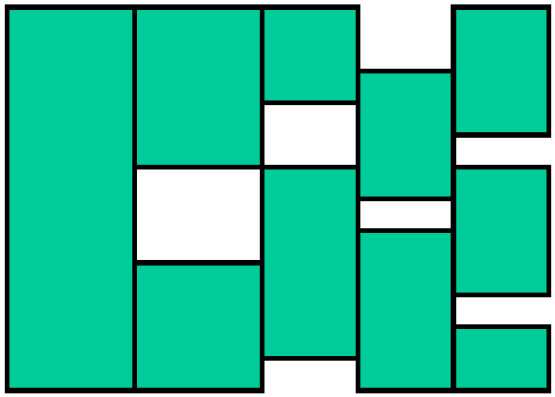
- cỡ mẫu $n = 500$;
- lặp lại mô phỏng trong 5000 lần;

Ví dụ minh họa việc bỏ qua dữ liệu khuyết

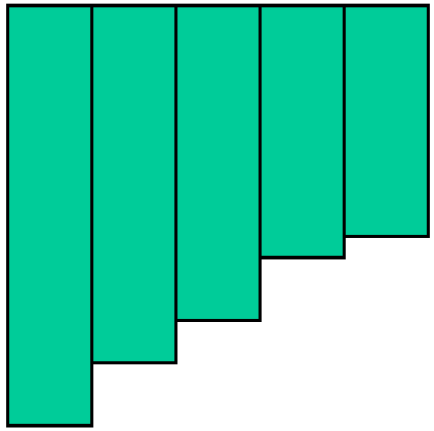


- Giá trị chính xác của β_2 là 1 (đường nét đứt màu xanh dương).
- Ước lượng của hệ số $\widehat{\beta_2}$ bị lệch rất lớn khi loại bỏ dữ liệu khuyết của X_1 (hình bên phải).

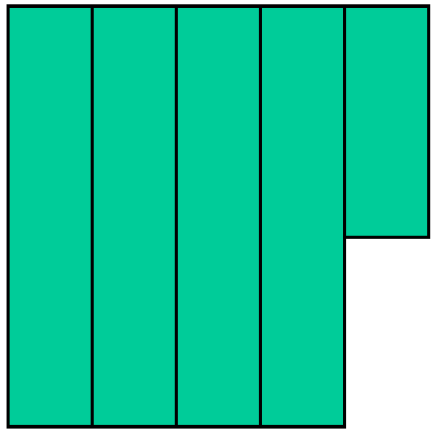
Các dạng dữ liệu bị khuyết



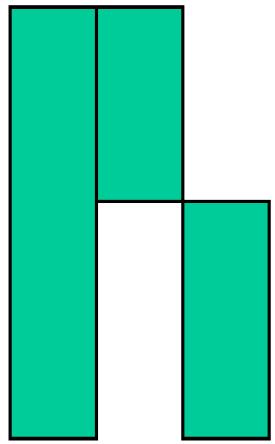
(a) tổng quát



(b) đơn điệu



(c) một biến khuyết



(b) khớp tệp tin

Cơ chế khuyết dữ liệu

- **Khuyết hoàn toàn ngẫu nhiên - Missing completely at random (MCAR):** Cơ chế khá đơn giản. Không có mối quan hệ nào giữa việc một điểm dữ liệu bị thiếu và bất kỳ giá trị nào trong tập dữ liệu, bị thiếu hoặc được quan sát. Hoặc dữ liệu bị thiếu chỉ là một tập hợp con ngẫu nhiên của dữ liệu.

  Ví dụ: Bệnh nhận bỏ buổi khám đã lên lịch do thời tiết xấu hoặc do bị hỏng xe.

- **Khuyết ngẫu nhiên - Missing at random (MAR)**: Cơ chế này có nghĩa là sự khuyết dữ liệu không có liên hệ đến các biến bị thiếu mà liên quan đến một số biến được quan sát hoàn toàn.

  Ví dụ: nếu hầu hết mọi người trong cuộc khảo sát không trả lời một câu hỏi nhất định thì tại sao họ lại làm vậy? Có phải là câu hỏi không rõ ràng?

- **Khuyết không ngẫu nhiên - Missing not at random (MNAR)**: cơ chế khuyết dữ liệu không chỉ có liên hệ tới một số biến được quan sát mà còn liên hệ tới những biến bị thiếu dữ liệu.

  đối tượng nghiên cứu bị loại bỏ bởi vì họ có kết quả trị liệu kém hoặc bị chết.

Cơ chế khuyết dữ liệu

Ảnh hưởng của khuyết dữ liệu tới kết quả phân tích tùy thuộc vào loại cơ chế bị khuyết:

- MCAR: không ảnh hưởng tới kết quả phân tích;

  do phân phối của biến với dữ liệu quan sát được, và với dữ liệu bị khuyết là như nhau;
- MAR và MNAR: nếu phân tích chỉ dựa trên phần dữ liệu được quan sát, sai số sẽ rất cao.

Nhận diện các dạng khuyết dữ liệu

Ta xét các tiêu chí sau:

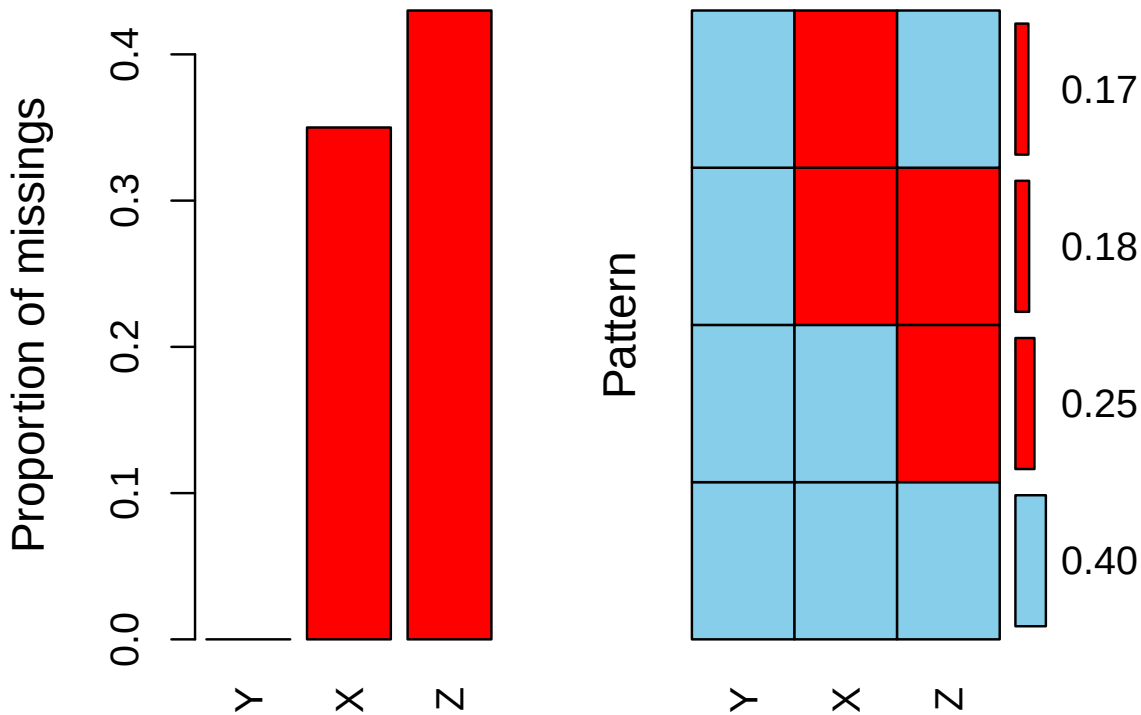
- phần trăm dữ liệu bị khuyết theo từng biến được thu thập trong nghiên cứu

  nếu tối đa $\leq 10\%$, ta có thể bỏ qua các dữ liệu bị khuyết;
- phân phối của một biến theo thành phần bị thiếu/quan sát được của biến khác

  nếu phân phối của hai phần là tương đồng nhau, cơ chế khuyết là MCAR.

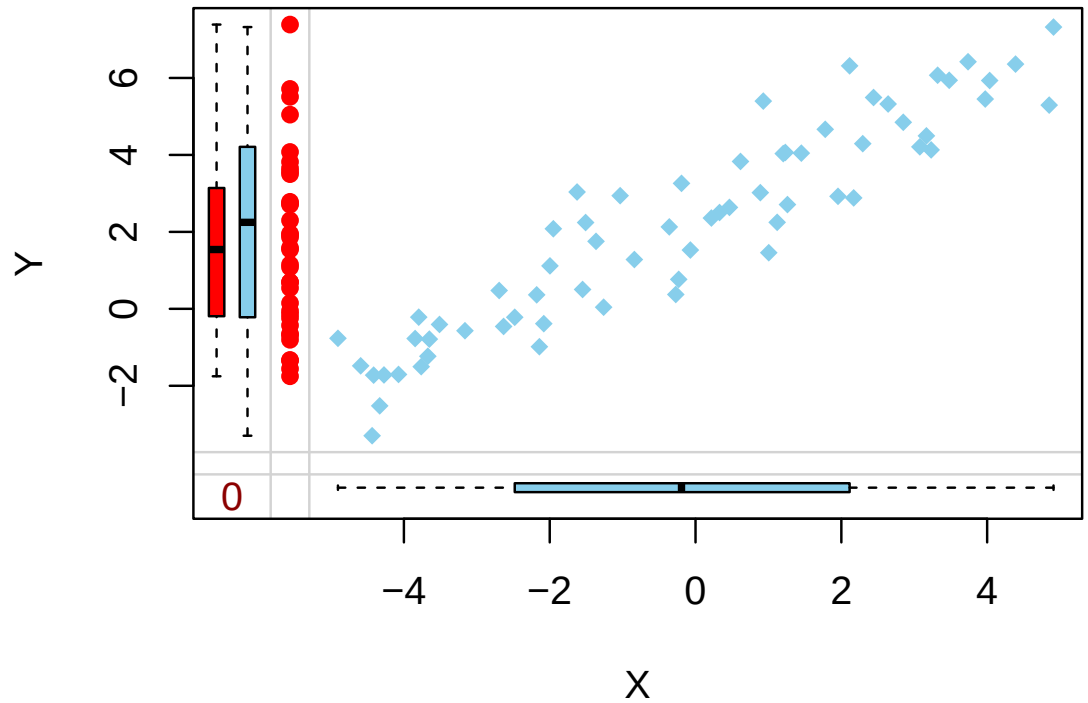
Nhận diện các dạng khuyết dữ liệu

Ví dụ 1:



- Hai X và Z biến bị khuyết dữ liệu.
- Tối đa 43% dữ liệu bị khuyết.

Nhận diện các dạng khuyết dữ liệu



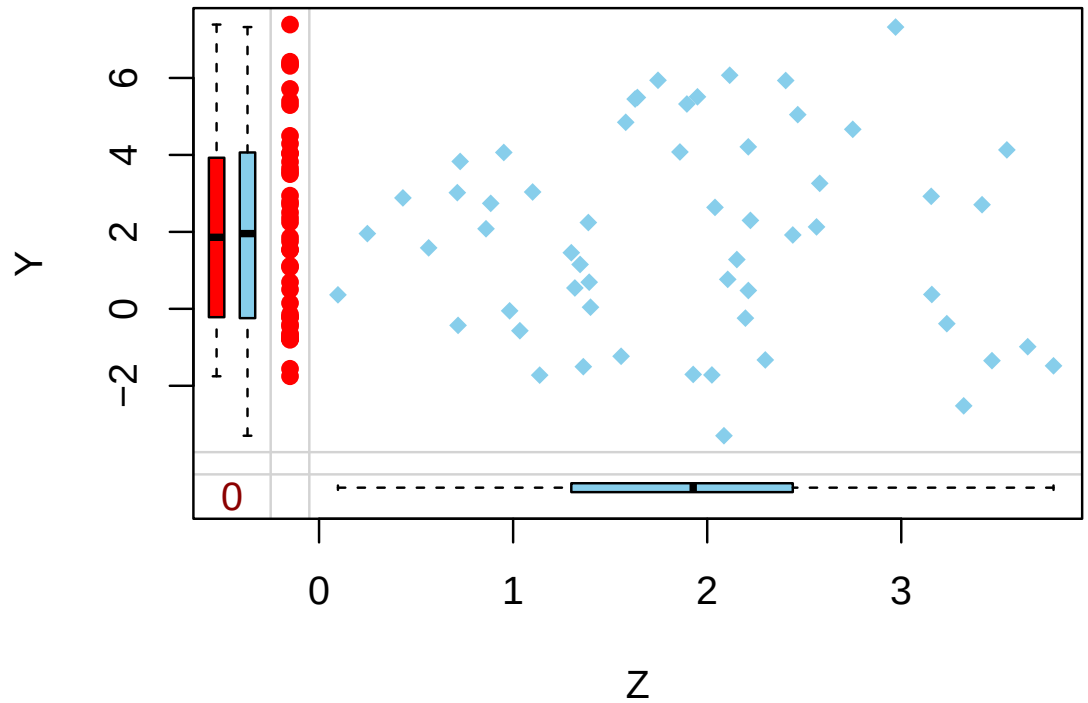
Phối phối của Y trên

- phần quan sát được của X (biểu đồ boxplot màu xám); và
- phần khuyết của X (biểu đồ boxplot màu đỏ),

là khá tương đồng nhau.

$\hookrightarrow$ cơ chế khuyết của X có thể là MCAR.

Nhận diện các dạng khuyết dữ liệu



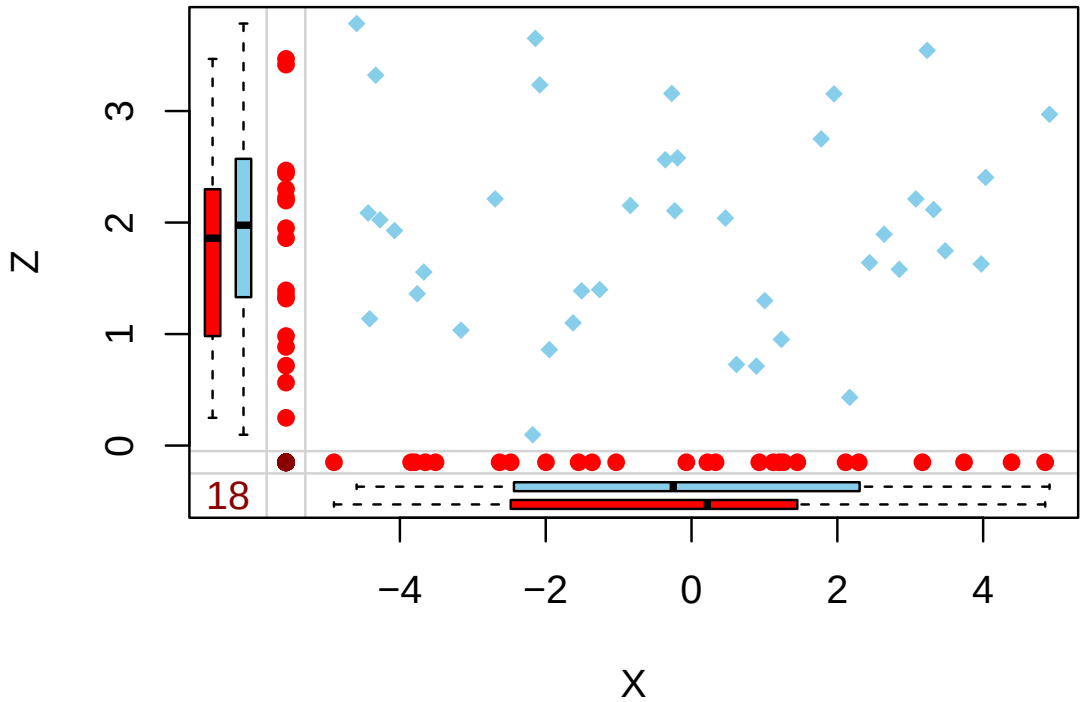
Phối phối của Y trên

- phần quan sát được của Z (biểu đồ boxplot màu xám); và
- phần khuyết của Z (biểu đồ boxplot màu đỏ),

là khá tương đồng nhau.

$\hookrightarrow$ cơ chế khuyết của Z có thể là MCAR.

Nhận diện các dạng khuyết dữ liệu



Phối phối của X trên

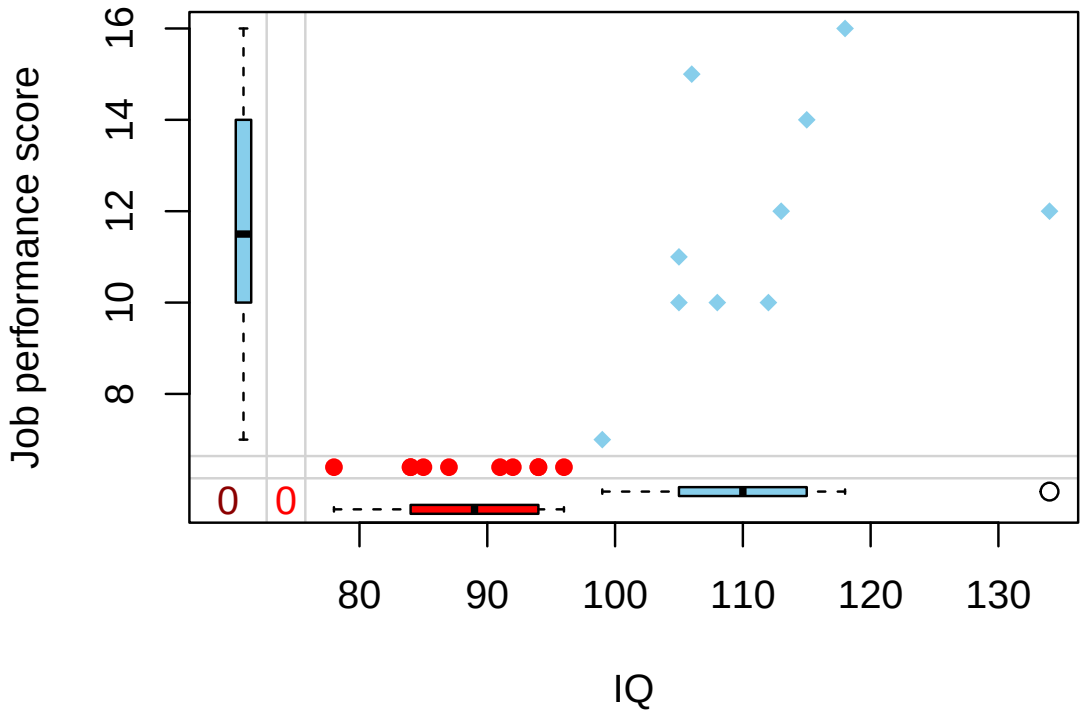
- phần quan sát được của Z (biểu đồ boxplot màu xám); và
- phần khuyết của Z (biểu đồ boxplot màu đỏ),

là khá tương đồng nhau. Tương tự với Z.

$\hookrightarrow$ cơ chế khuyết có thể là MCAR.

Nhận diện các dạng khuyết dữ liệu

Ví dụ 2:



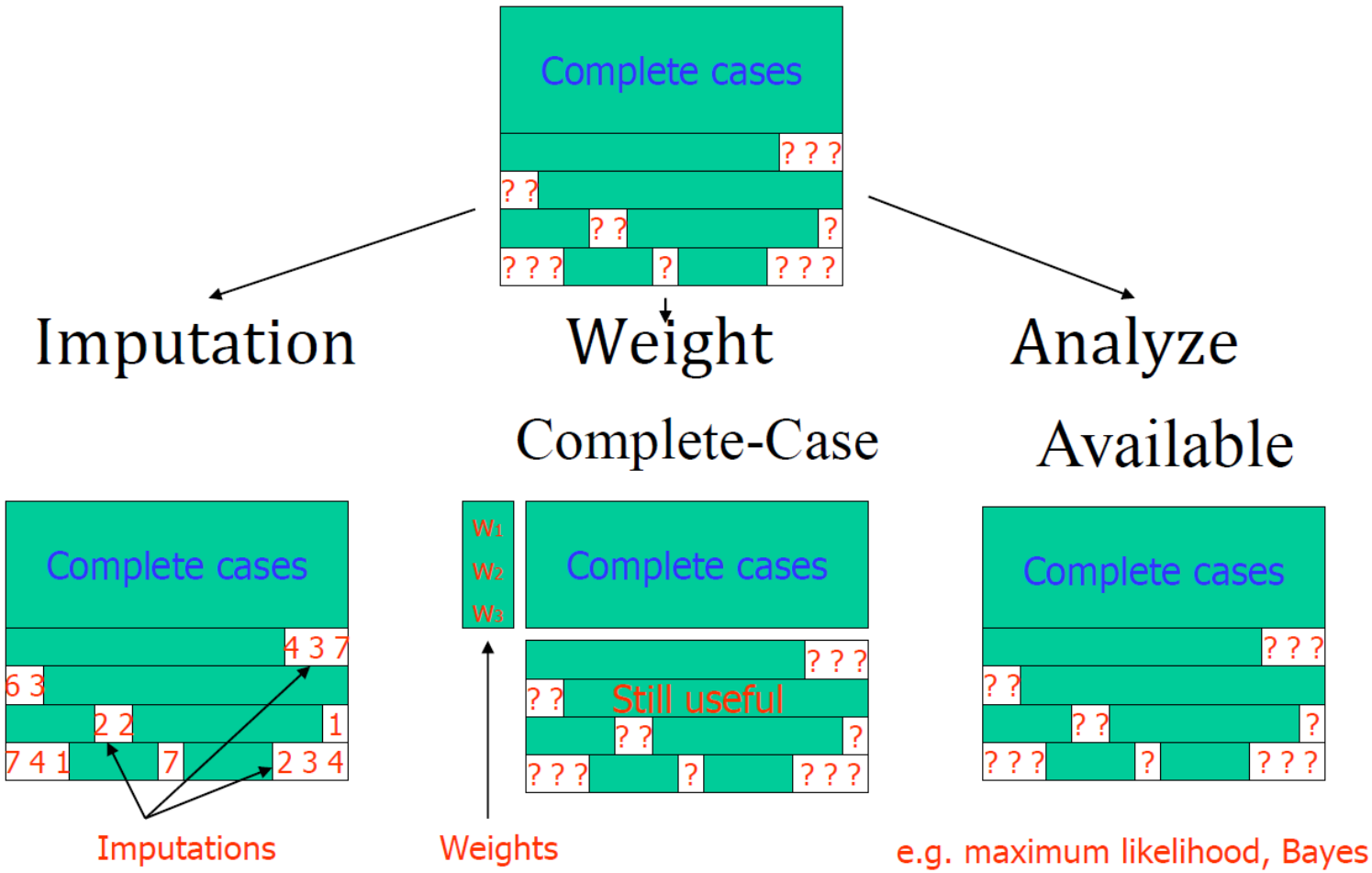
Phối phối của IQ trên

- phần quan sát được của Job performance score (biểu đồ boxplot màu xám); và
- phần khuyết của Job performance score (biểu đồ boxplot màu đỏ),

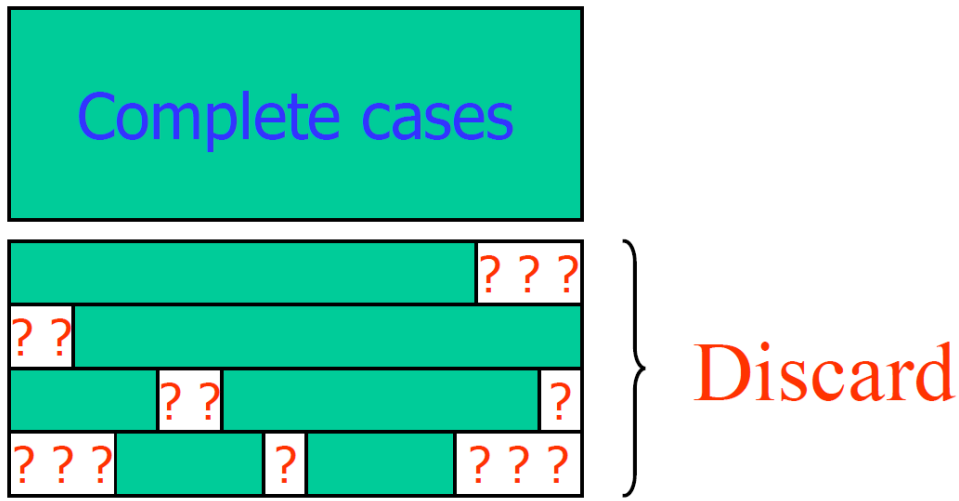
là khác nhau (những người có IQ thấp không được đánh giá khả năng làm việc).

↪ cơ chế khuyết của Job performance score có thể là MAR.

Một số chiến lược cơ bản



Complete-case Analysis



Chỉ sử dụng vùng dữ liệu được quan sát đối với tất cả các biến.

Complete-case Analysis

Ưu điểm

- dễ sử dụng;
- được cài đặt mặc định trong các phần mềm thống kê.

Nhược điểm

- chỉ đúng trong trường hợp MCAR;
- loại bỏ đi một lượng đáng kể thông tin của dữ liệu;

  $\hookrightarrow$ chỉ nên dùng khi % khuyết dữ liệu nhỏ;
- phương sai của ước lượng bị tăng lên so với khi phân tích dữ liệu đầy đủ;
- kết quả phân tích sẽ gặp sai lầm nghiêm trọng khi cơ chế khuyết không phải MCAR.

Single imputation

Single imputation là một trong số những kỹ thuật để "gán" hoặc suy đoán dữ liệu thiếu.

Các dữ liệu bị khuyết của một biến được thay thế bằng một giá trị ước tính duy nhất:

- trung bình (mean imputation);
- ước đoán theo hồi quy (regression imputation);
- ước đoán theo hồi quy ngẫu nhiên (stochastic regression imputation).

Hạn chế

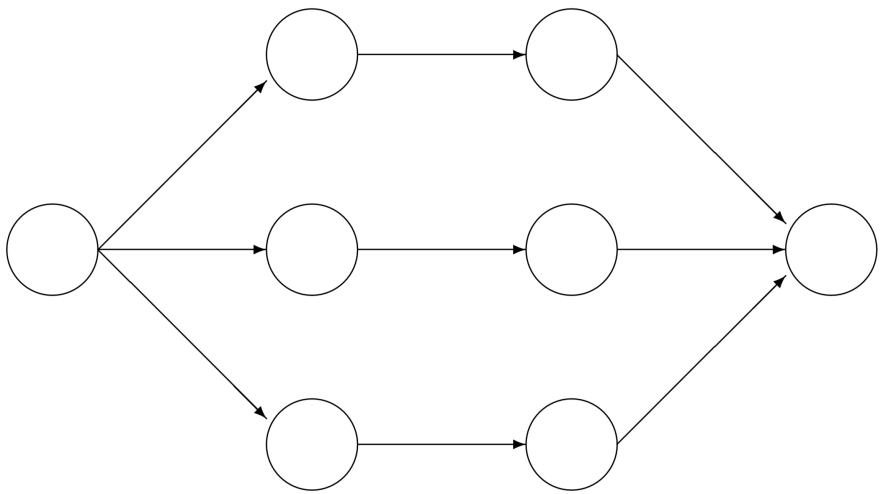
- đánh giá thấp độ không chắc chắn do dữ liệu khuyết gây ra;
- giả định rằng dữ liệu được ước đoán là giá trị thực có thể quan sát được khi dữ liệu đã hoàn chỉnh $\Rightarrow$ thiếu tính thực tế;
- dữ liệu luôn là ngẫu nhiên $\Rightarrow$ chỉ gán dữ liệu khuyết bằng 1 giá trị đơn lẻ là không phù hợp.

Multiple imputation

Multiple imputation "gán" hoặc suy đoán dữ liệu thiếu bằng các giá trị ước đoán khác nhau, thực hiện trong nhiều lần.

Sự không chắc chắn về dữ liệu khuyết được giải quyết thông qua việc tạo ra nhiều tập dữ liệu khác nhau

$\hookrightarrow$ thực hiện phân tích trên các tập dữ liệu mới và đưa ra nhận xét chung.



Multiple imputation

Phương pháp tiếp cận hiểu quả cho multiple imputation là Chained Equation.

Các thuật toán thường được sử dụng

- predictive mean matching - pmm (phù hợp cho mọi dạng dữ liệu);
- Bayesian linear regression (phù hợp cho dữ liệu định lượng);
- logistic regression (phù hợp cho dữ liệu định tính với 2 nhóm);
- Bayesian polytomous regression (phù hợp cho dữ liệu định tính với nhiều nhóm);
- proportional odds model (phù hợp với dữ liệu định tính có thứ tự với nhiều nhóm).

Quy trình hồi quy với dữ liệu khuyết

Quy trình xử lý dữ liệu cho bài toán hồi quy với dữ liệu khuyết:

1. Chuẩn bị dữ liệu:

   kiểm tra % dữ liệu bị khuyết;

   điều tra dạng dữ liệu bị khuyết và cơ chế khuyết có thể: MCAR, MAR hoặc MNAR;

   áp dụng multiple imputation để tạo ra nhiều tập dữ liệu ước đoán khác nhau.
2. Xây dựng và huấn luyện mô hình: sử dụng các thuật toán xây dựng mô hình hồi quy trên các tập dữ liệu ước đoán.
4. Tổng hợp kết quả ước lượng trên các tập dữ liệu để tạo thành một kết quả chung.
5. Đánh giá độ chính xác của mô hình.
6. Áp dụng mô hình và giám sát.