

Dữ liệu và Ứng dụng trong Phân tích Chính sách

Lê Việt Phú

Trường Chính sách Công và Quản lý Fulbright

30/05/2023

Too much data – too little information!

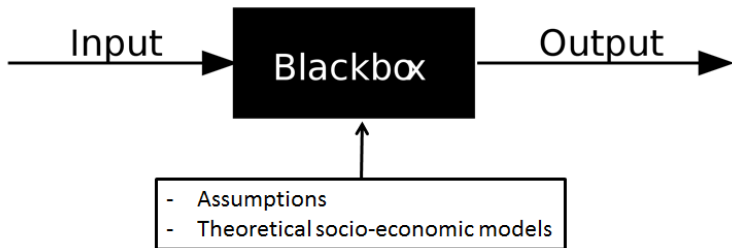
Cần thiết phải phân tích, phân nhóm, giảm chiều dữ liệu, khám phá mối quan hệ tiềm ẩn giữa các nhóm biến số – nhằm mục đích diễn tả thực tại và khuyến nghị hoặc dự báo!

Một số ứng dụng khác của phân tích dữ liệu trong kinh tế và kinh doanh

- ▶ Cảnh báo giao dịch gian lận trong thanh toán online (fraudulent credit card transaction)
- ▶ Giới thiệu mua hàng (recommender system)
- ▶ Quảng cáo bán hàng
- ▶ Giá cả hàng hóa, bất động sản, Uber ridesharing
- ▶ Cho vay tín dụng
- ▶ Dự báo doanh nghiệp phá sản
- ▶ Đầu tư chứng khoán
- ▶ Y tế, chuẩn đoán bệnh tật theo triệu chứng
- ▶ Các chính sách kinh tế xã hội
- ▶ Cách mạng công nghiệp 4.0

Mục đích của môn học phân tích dữ liệu

- ▶ Nhận diện bản chất của các mô hình sử dụng trong phân tích dữ liệu.
- ▶ Sử dụng Excel để tiến hành các phân tích định lượng căn bản.
- ▶ Diễn giải, phân tích, và phê phán các kết quả nghiên cứu.



Tổng quan phần Phân tích Dữ liệu

- ▶ Thu thập, phân loại, tổng hợp, tổ chức, phân tích và diễn giải kết quả:
 - Nhận dạng dữ liệu và phân tích thống kê mô tả
 - Sử dụng các kiến thức thống kê để làm rõ đặc điểm và mối quan hệ giữa các biến số
- ▶ Các bước trong xây dựng mô hình phân tích dữ liệu ứng dụng cơ bản:
 - Xây dựng mô hình dựa trên các lý thuyết kinh tế để thiết lập quan hệ giữa các biến số;
 - Lựa chọn phương pháp ước lượng dựa trên thiết kế nghiên cứu và khả năng cho phép của dữ liệu;
 - Diễn giải, kiểm định và giải thích kết quả; Chuẩn đoán và xử lý các vấn đề có liên quan;
 - Kiểm tra tính vững của kết quả trong các điều kiện khác nhau;
- ▶ Ứng dụng phân tích khám phá dữ liệu trong phân tích kinh tế và kinh doanh.

Thu thập dữ liệu

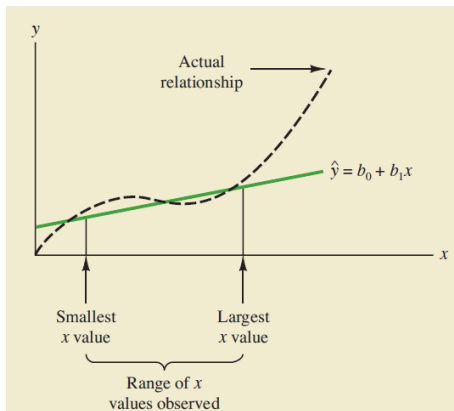
- ▶ Dữ liệu thử nghiệm/thí nghiệm (experimental data): thu được trong nghiên cứu khoa học cơ bản dựa trên các điều kiện có kiểm soát chặt chẽ trong phòng thí nghiệm.
 - Ví dụ thử nghiệm một loại thuốc chữa bệnh nào đó. Hai nhóm bệnh nhân có tình trạng bệnh tật như nhau. Một nhóm cho điều trị theo phác đồ (gọi là nhóm điều trị), một nhóm không (gọi là nhóm đối chứng). Sau một thời gian, so sánh mức độ tiến triển của bệnh thì có thể xác định tác động của phác đồ điều trị đối với mỗi nhóm.
- ▶ Dữ liệu phi thử nghiệm: là các dữ liệu quan sát/điều tra (observational/surveyed data). Hầu hết các dữ liệu kinh tế xã hội đều là dữ liệu phi thử nghiệm.
 - Không đảm bảo được điều kiện các nhóm tương đồng nên khi phân tích chính sách hay quyết định kinh doanh sẽ gặp rất nhiều khó khăn để **thiết lập quan hệ nhân quả** giữa các biến số kinh tế.

Nhân quả (Causation) hay Tương quan (Correlation)?

- ▶ Nhân quả: A gây ra B. Ví dụ: tăng giá điện là nguyên nhân dẫn đến giảm tiêu thụ.
- ▶ Tương quan: A và B biến động cùng hoặc ngược chiều. Ví dụ: giá điện và thu nhập người dân trong 10 năm qua đều tăng. Nhưng không có quan hệ nhân quả, giá điện tăng không phải do thu nhập tăng hay ngược lại.
- ▶ Cảnh báo với quan hệ nhân quả: giá điện tăng chưa chắc là nguyên nhân dẫn đến việc giảm tiêu thụ điện. Có thể do các nguyên nhân khác như sự xuất hiện của các thiết bị tiết kiệm điện năng, hay thu nhập bị giảm, hay hành vi tiêu dùng thay đổi dẫn đến tiêu thụ điện giảm. Để xác lập quan hệ nhân quả phải kiểm soát và loại trừ các nguyên nhân khác có thể gây ra kết luận sai, dẫn đến đề xuất quyết định sai.

Về vấn đề tương quan hay nhân quả, và độ vững của nghiên cứu

Nhận diện được vấn đề overfitting, bias, prediction vs causal inference; và các mô hình, thuật toán dùng để hạn chế sai sót.



→ Bài giảng về mô hình định lượng, thiết lập quan hệ nhân quả và ứng dụng trong phân tích chính sách.

Nhận dạng dữ liệu và phân tích thống kê mô tả

1. Nhận dạng các loại dữ liệu
2. Xử lý và làm sạch dữ liệu
3. Trình bày dữ liệu bằng bảng biểu và đồ thị thống kê
4. Khai thác và truy vấn dữ liệu từ các bộ dữ liệu thông dụng
5. Các công cụ phân tích dữ liệu

1. Nhận dạng dữ liệu

- ▶ Dữ liệu điều tra thông thường (dữ liệu chéo, dữ liệu bảng, dữ liệu chuỗi thời gian)
- ▶ Phân loại dữ liệu điều tra theo cách thu thập:
 - Lấy mẫu theo xác suất: điều tra ngẫu nhiên hóa đơn giản/hệ thống, điều tra ngẫu nhiên phân tầng/nhóm, quyền số điều tra
 - Lấy mẫu điều tra phi ngẫu nhiên: lấy mẫu thuận tiện, lấy mẫu hệ thống
- ▶ Dữ liệu phi cấu trúc: dữ liệu mạng xã hội, dữ liệu lớn, dữ liệu text

Loại hình thu thập dữ liệu và đặc tính dữ liệu quyết định phương pháp và mục đích sử dụng!

Các cấu trúc dữ liệu phổ biến

- ▶ Dữ liệu chéo (cross-sectional data): thông tin của cá nhân, hộ gia đình, công ty, đơn vị địa lý tại một thời điểm.
- ▶ Dữ liệu chuỗi thời gian (time-series data): các quan sát được theo dõi lặp lại qua nhiều kỳ, tập trung vào đặc tính tần suất và độ trễ của tác động.
- ▶ Dữ liệu chéo gộp (chéo kết hợp - pooled cross-sectional data): quan sát qua nhiều kỳ nhưng của các cá thể khác nhau.
- ▶ Dữ liệu bảng (panel/longitudinal data): các cá thể kinh tế được theo dõi lặp lại ít nhất là 2 kỳ, tập trung vào đặc tính thay đổi giữa các kỳ quan sát.

Ví dụ nhận dạng dữ liệu

- ▶ Đọc các bộ dữ liệu trong thư mục **Data1**
- ▶ Cho biết loại hình dữ liệu, đơn vị quan sát, và giải thích các biến số
- ▶ Có thể sử dụng các loại dữ liệu trên cho phân tích gì, các rủi ro gặp phải đối với từng loại phân tích đó [trên góc độ dùng dữ liệu và các công cụ để phân tích]?

Bộ dữ liệu chéo WAGE1 về tiền lương bình quân theo giờ và đặc điểm cá nhân

Household	Var1	Var2	Var3	...
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...

- ▶ Mỗi dòng dữ liệu là một quan sát (cá thể). Thứ tự dữ liệu không ảnh hưởng đến kết quả phân tích.
- ▶ Dễ phân tích nhất nhưng cũng gặp nhiều vấn đề nhất.
- ▶ *Câu hỏi: So sánh tiền lương giữa phụ nữ và nam giới, giữa người có gia đình và độc thân, người có trình độ giáo dục và số năm kinh nghiệm khác nhau.*

Dữ liệu chuỗi thời gian PHILLIPS về tỷ lệ lạm phát và thất nghiệp ở Mỹ

Year	Var1	Var2	Var3	...
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...

- ▶ Thứ tự của dữ liệu rất quan trọng. Cần lưu ý đến tần suất thu dữ liệu và tính chu kỳ.
- ▶ Khó phân tích nhất trong các loại dữ liệu.
- ▶ *Câu hỏi: vẽ đồ thị xu hướng biến động của tỷ lệ thất nghiệp và lạm phát theo thời gian. Dữ liệu trên có phù hợp với lý thuyết về đường Phillips không?*

Dữ liệu chéo gộp HPRICE3

Household	Year	Var2	Var3	...	...
1	1	...	...	...	...
2	1	...	...	...	...
3	1	...	...	...	...
4	2	...	...	...	...
5	2	...	...	...	...
6	2	...	...	...	...

- ▶ Hai hoặc nhiều bộ dữ liệu chéo gộp lại thành một bộ dữ liệu lớn, có thêm biến thời gian để xác định thời điểm thu thập dữ liệu. Các quan sát không nhất thiết phải trùng lặp lại giữa các thời kỳ.
- ▶ Có thể cung cấp nhiều thông tin hơn dữ liệu chéo thuần túy, nhưng không mạnh như dữ liệu bảng.
- ▶ *Câu hỏi: các nhân tố nào ảnh hưởng đến giá nhà? và các nhân tố đó thay đổi theo thời gian như thế nào?*

Dữ liệu bảng CRIME2

Household	Year	Var2	Var3	...	...
1	1	...	...	...	...
2	1	...	...	...	...
3	1	...	...	...	...
1	2	...	...	...	...
2	2	...	...	...	...
3	2	...	...	...	...

- ▶ Mỗi cá thể được thu dữ liệu trong ít nhất 2 kỳ, kỳ trước và kỳ sau.
- ▶ Cho phép thực hiện các phân tích tốt nhất trong các loại dữ liệu. Tuy nhiên yêu cầu cao về kỹ thuật.

2. Xử lý làm sạch dữ liệu

- ▶ Đọc dữ liệu từ MS Excel
- ▶ Xuất dữ liệu từ các phần mềm thống kê thông dụng
- ▶ Cách mã hóa các biến số
- ▶ Xử lý các vấn đề liên quan đến dữ liệu:
 - Dữ liệu bị thiếu/mất (missing values)
 - Dữ liệu ngoại vi (outliers)
 - Tính chính xác/nhiều của dữ liệu (signal/noise)

3. Sử dụng bảng biểu thống kê và đồ thị để mô tả dữ liệu

- ▶ Các đại lượng về giá trị trung bình, trung vị, mode
- ▶ Các đại lượng về phân phối: Khoảng phân phối, các phân vị, độ lệch chuẩn, hệ số CV
- ▶ Độ lệch (skewness)
- ▶ Đồng phương sai, hệ số tương quan
- ▶ Đồ thị phân phối, đồ thị điểm, đồ thị tương quan, đồ thị hộp (boxplot)...

Câu hỏi: Chỉ sử dụng các thống kê mô tả và đồ thị bảng biểu liệu bạn có thể nhận định gì về nguyên nhân gây ra tác động trong các bộ dữ liệu đã cung cấp?

4. Khai thác và truy vấn dữ liệu từ các bộ dữ liệu thông dụng

Phân loại theo cấp độ điều tra của dữ liệu và khả năng ứng dụng:

- ▶ Phân tích cấp độ cá nhân, hộ gia đình, doanh nghiệp: Dữ liệu microdata xuất phát từ các điều tra thống kê lấy mẫu với quyền số.
- ▶ Phân tích theo cấp độ phạm vi địa giới hành chính: Tỉnh/thành phố, quốc gia: Dữ liệu tổng hợp thành các chỉ số thông dụng như GDP, chất lượng môi trường kinh doanh, chính sách can thiệp...

Các bộ dữ liệu và cơ quan cung cấp dữ liệu thông dụng

- ▶ Các cơ sở dữ liệu thông dụng: VHLSS, SME, Enterprise Census, VARHS, Population Census (IPUMS), Agricultural Census...
- ▶ Các cơ quan tổ chức: GSO, WB, IMF, ADB, các tổ chức quốc tế

5. Các Công cụ Phân tích Dữ liệu

- ▶ MS Excel
- ▶ Các phần mềm thống kê phân tích chuyên dụng như Minitab, R, Stata, Eview, SPSS, SAS...

Bài giảng 2: Sử dụng mô hình định lượng trong phân tích chính sách.