疑似相関にだまされない因果分析の方法

データから因果分析を行う際には、下表のようなガイドラインがあります。データを分析した結果、原因とされるもの（“原因”）と結果とされるもの（“結果”）が因果関係にあると判断するためのものです。このガイドラインは、主に生物学・医学の研究で使われているものですが、機械学習や統計を利用したデータ分析に役立つ部分も多いでしょう。

相関関係から因果関係を見抜くのにもっとも確実な方法は、実験です。主な方法として、**ランダム化比較試験**があります。医療分野以外では、A/Bテストと呼ばれることが多いかもしれません。たとえば、あるアンケートの結果、「朝ごはんを食べていること」と「成績」の間に強い相関が見られたとします。「朝ごはんを食べると成績が上がる」という結論を導くためには、ここで朝ごはんを食べるグループ（介入グループ）と朝ごはんを食べないグループ（比較グループ）をランダムに分けた上で、成績の差があるかどうかの実験を行う必要があります。また、朝ごはんを食べるか否か以外は、各グループの特徴を同じにする必要もあります。このように、“原因”の有無だけを変化させて“結果”を観察する実験がランダム化試験なのです。

■ 因果分析におけるガイドライン

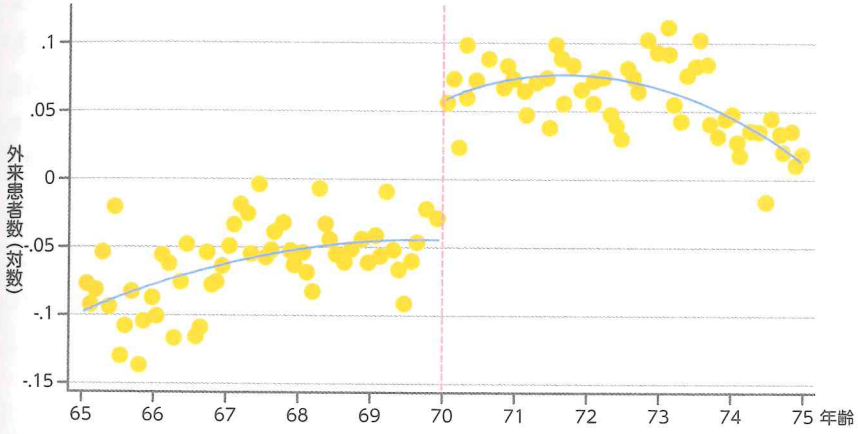
1	強固性	“原因”と“結果”の間に強い関連があると数値（統計）からわかること
2	一致性	観察対象、実証する手法などの条件を変えても結果が一致すること
3	特異性	“原因”以外の要素と“結果”の相関や、“結果”以外の要素と“原因”の相関が強くないこと。“原因”と“結果”の相関だけが際立って強いこと
4	時間的先行性	“原因”の後に“結果”が起こること
5	量-反応関係	“原因”の値が大きくなると、“結果”の値も単調に大きくなること
6	妥当性	各分野（例えば生物学・医学）の常識にもとづいてもっともらしいこと
7	整合性	過去の知見と矛盾しないこと
8	実験	観察された関連性を支持する実験的研究（例えば動物実験）が存在すること
9	類似性	すでに確立している別の因果関係と類似した関係があること

出典：Hill, Austin Bradford (1965). "The Environment and Disease: Association or Causation?". Proceedings of the Royal Society of Medicine. 58 (5): 295–300. PMC 1898525. PMID 14283879.

実験を行うことが困難な場合は、今持っているデータを使って、実験に近い分析（疑似実験）を行います。その分析手法の1つが、**回帰分断デザイン**です。この手法では、境界線の前後では“原因”以外の要素がほぼ変わらないことを利用します。“原因”だけを変化させて“結果”を観察するランダム化比較実験と同じような状況が再現できるためです。横軸に年齢を、縦軸に外来患者数の対数を取ると、70歳を境に病院の外来患者数が増加していることが見て取れます。また、医療費の自己負担比率は、70歳を超えると3割から2割になります。これ以外の要素は70歳の直前直後ではほぼ同じと考えられるので、自己負担比率（“原因”）だけを変化させて外来患者数（”結果”）の変化を観察できます。

似た手法として、中断時系列デザインがあります。この手法では、時系列データを利用して横軸を時間にとります。ある時刻を境に“原因”が変動したときの変化（たとえば、消費増税による消費行動への影響）を観察するときに有効です。

■ 回帰分断デザイン



参照：Shigeoka, Hitoshi. 2014. "The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection." American Economic Review, 104 (7): 2152-84.

まとめ ▶ 相関と因果を分けて考え、適切な手法を選ぶ