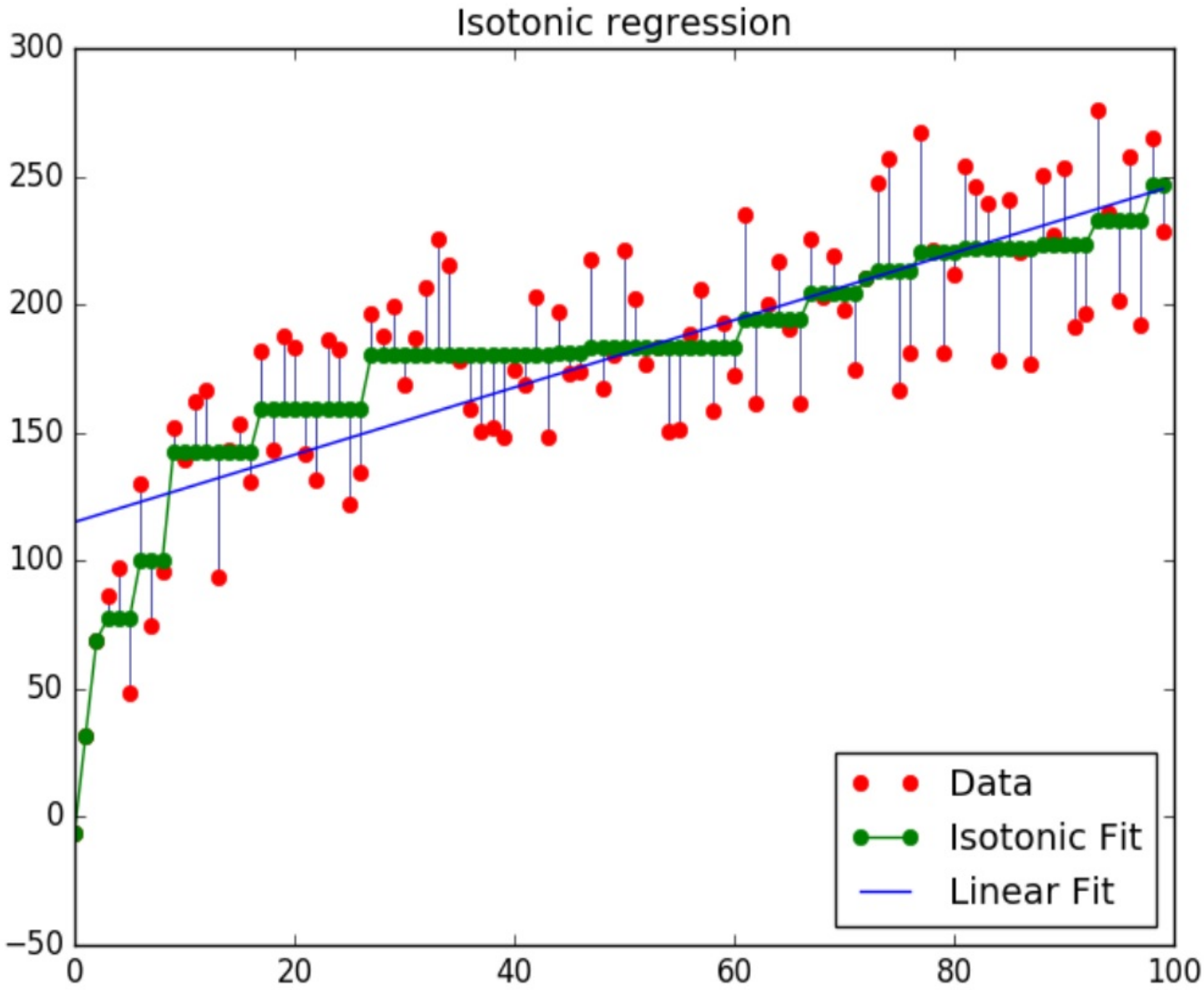
SF 2930 Regression Analysis
LECTURE 13.2
Regression and Causality

Timo Koski

KTH Royal Institute of Technology

2023

Isotonic Regression

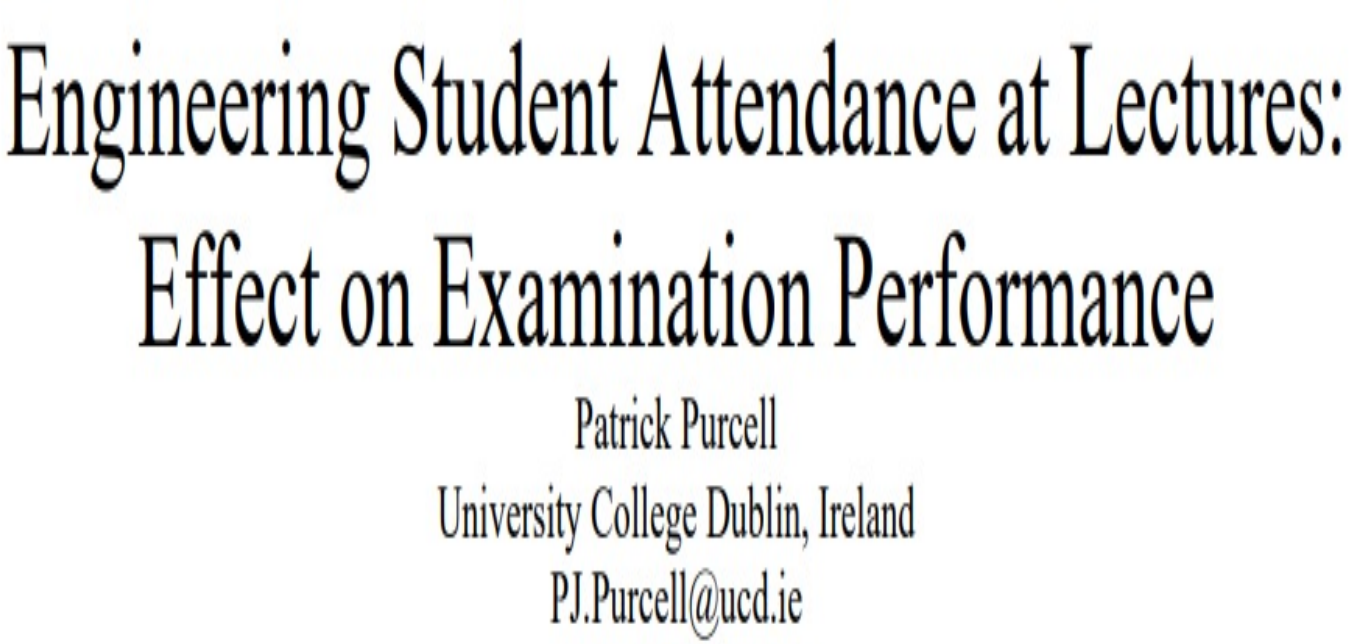


Part 1: Back to Lecture 1

Simple Linear Regression model: An Observational Study

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*. International Conference on Engineering Education – ICEE 2007 Coimbra, Portugal September 3 – 7, 2007

Linear Regression: Prediction of Success at an Exam by Attendance at Lectures



Engineering Student Attendance at Lectures:
Effect on Examination Performance

Patrick Purcell
University College Dublin, Ireland
PJ.Purcell@ucd.ie

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

- A linear regression analysis of the data showed a strong correlation between lecture attendance and examination performance.

evident between class attendance and examination performance. Examination of these figures also shows that the pass mark of 40% can be attained at relatively low attendance levels (< 20% attendance).

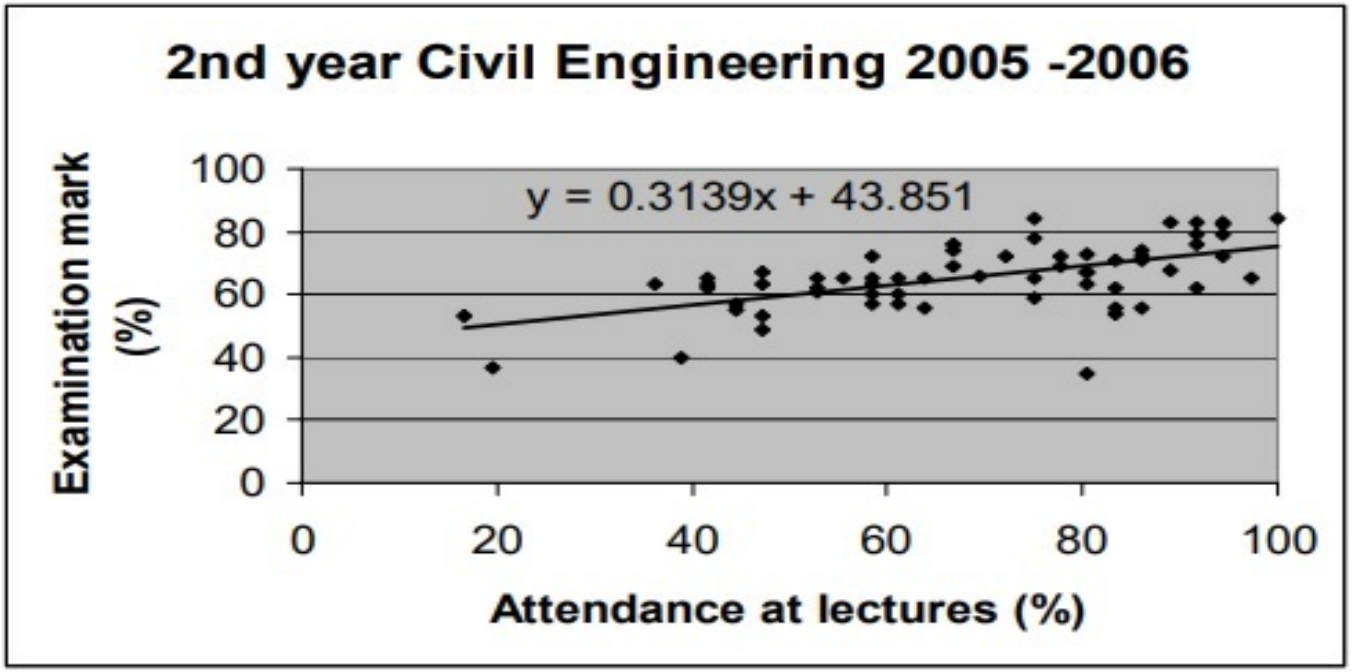


FIGURE I
PERFORMANCE OF SECOND-YEAR CIVIL ENGINEERING STUDENTS

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

towards continuous assessment.

Examination of Figures 1 and 2 clearly shows that students who have chosen to attend lectures regularly perform significantly better in their examinations than students that have chosen not to attend lectures. The best-fit equations ($y = 0.31x + 43.90$ and $y = 0.32x + 33.91$) indicate that each 10% increase in lecture attendance results in an approximate 3% improvement in examination performance. These correlation equations compare favourably to other studies, for example, Lockwood et al. [9].

Question: A Policy Recommendation?

There are other studies confirming the findings of Purcell.

- Question: Are we now in our right to request an audience with Rektor Anders Söderholm at Rektor's office in Brinellvägen 8 to advocate the following policy reform:

  Statistical research has shown that exam results will be substantially improved, were KTH to introduce mandatory attendance at lectures!

Question: A Policy Recommendation?

There are other studies confirming the findings of Purcell.

- Question: Are we now in our right to request an audience with Rektor Anders Söderholm at Rektor's office in Brinellvägen 8 to advocate the following policy reform:

  Statistical research has shown that exam results will be substantially improved, were KTH to introduce mandatory attendance at lectures!

- A decree of mandatory attendance by the Rektor would be an **intervention**.

Rolf Sandahl, Gustav Jakob Petersson: *Kausalitet i filosofi, politik och utvärdering*, Studentlitteratur 2016

Hur kan vi bäst ta reda på om offentliga insatser ger avsedda resultat? Denna fråga har under senare år diskuterats livligt både inom forskarvärlden och inom olika myndigheter och förvaltningar, inte minst i den så kallade evidensdebatten.

Svaret på denna fråga beror på vad vi menar med kausalitet. Detta har filosoferna diskuterat under århundraden och beroende på hur kunskapen ska användas kan vissa aspekter av kausalitet vara mer relevanta att spåra än andra.

Prediction and causation are very different. Typical questions are:

- **Prediction**: Predict Y after observing $X = x$
- **Causation**: Predict Y after setting $X \mapsto x$.

Causation involves predicting the effect of an intervention. For example:

- Prediction: Predict examination performance of person after observing attendance at lectures.
- Causation: Predict examination performance of a person after having decreed attendance at x % of lectures.

The difference between passively observing $X = x$ and actively intervening and setting $X \mapsto x$ is significant and requires different techniques and, typically, much stronger assumptions. This is the area known as **causal inference**.

There are two types of causal questions.

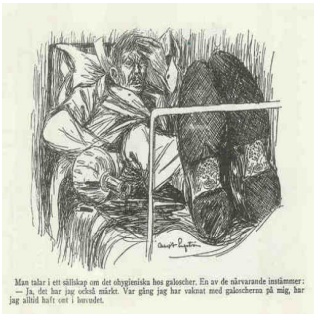
- The first deals with questions like this: do cell phones cause brain cancer? In this case, there are variables X and Y and we want to know the causal effect of X on Y . The challenges are: find a parameter θ that characterizes the causal influence of X on Y and find a way to estimate θ. This is usually what we mean when we refer to causal inference.
- The second question is: given a set of variables, determine the causal relationship between the variables. This is called causal discovery. This problem is statistically impossible.

Part 2: Causality & Correlation

Albert Engström, 1869 – 1940, Swedish cartoonist

Man talar i ett sällskap om det ohygieniska hos galoscher. En av de närvarande instämmer: — Ja, det har jag också märkt. Var gång jag har vaknat med galoscherna på mig, har jag alltid haft ont i huvudet.

Common Cause



There is a common cause for headache in the morning and having fallen asleep with galoshes on: excessive consumption of alcoholic beverages during the evening/night before.

Confounding: Z is a Confounder

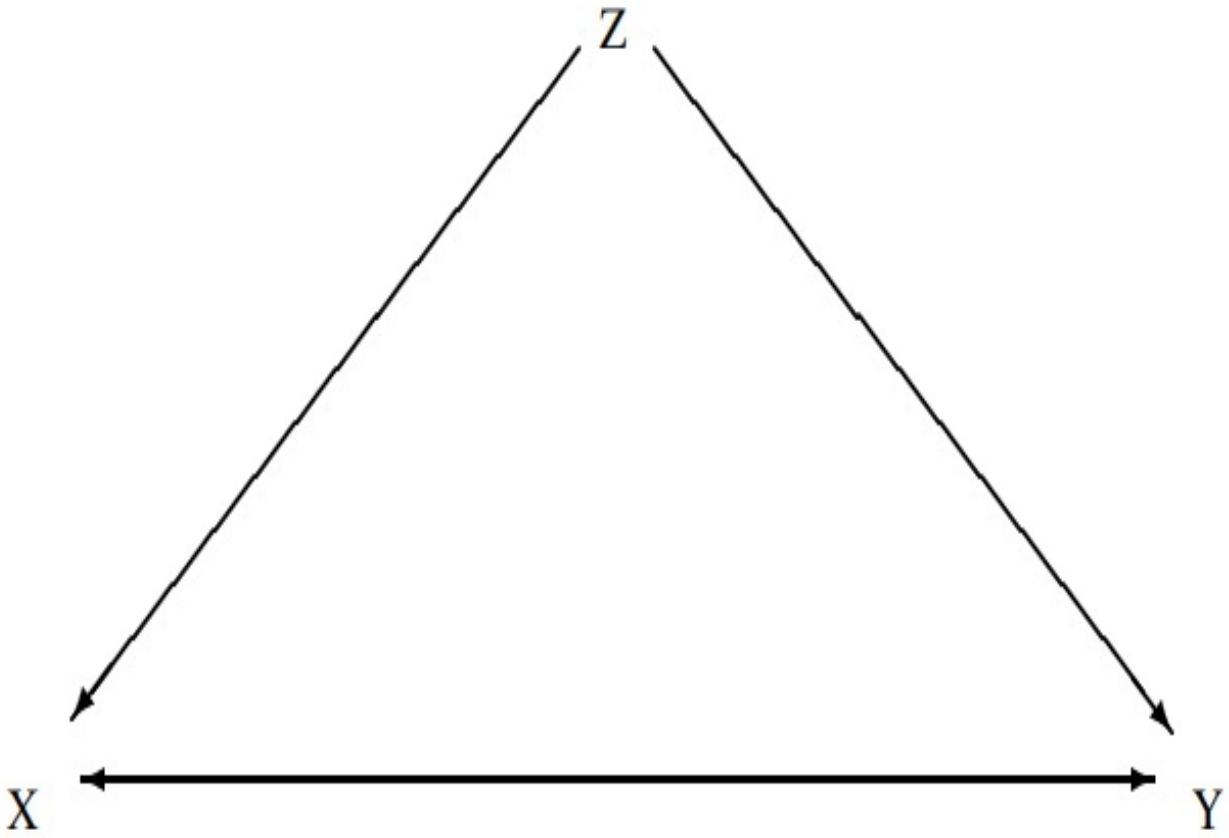
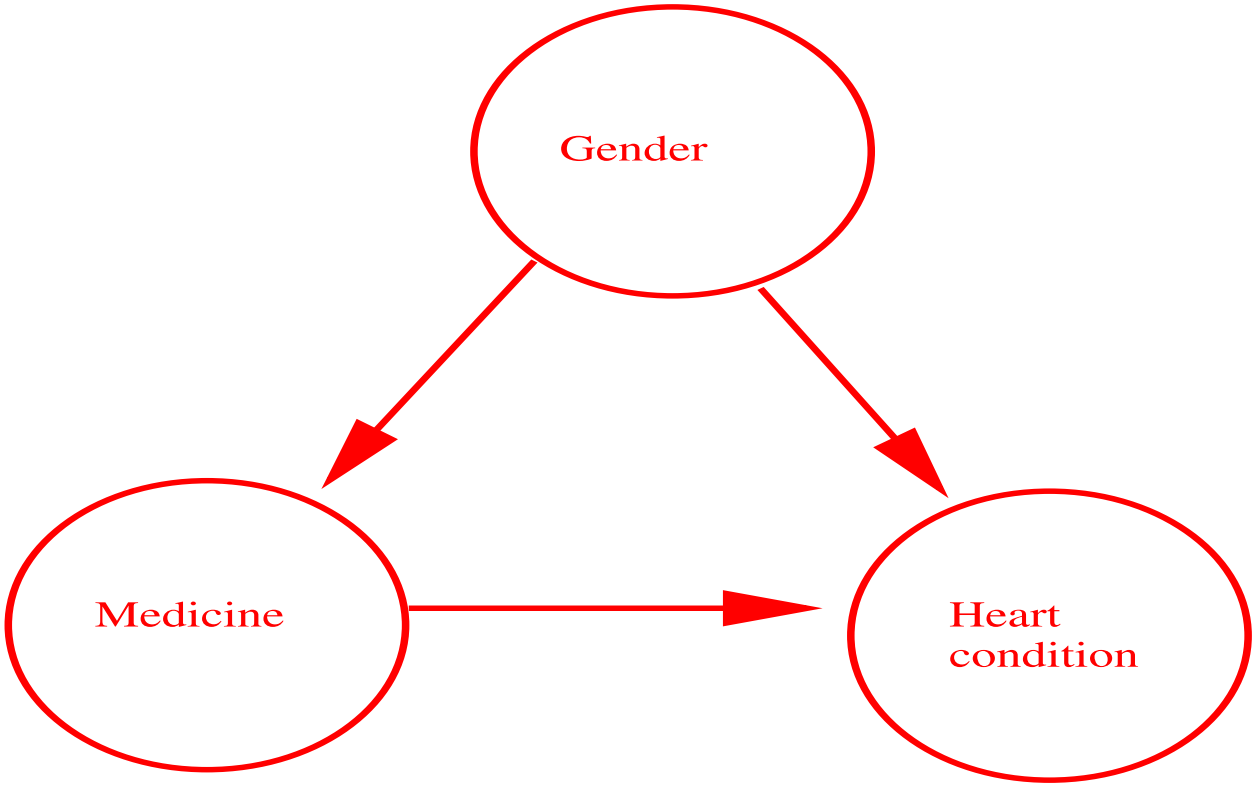


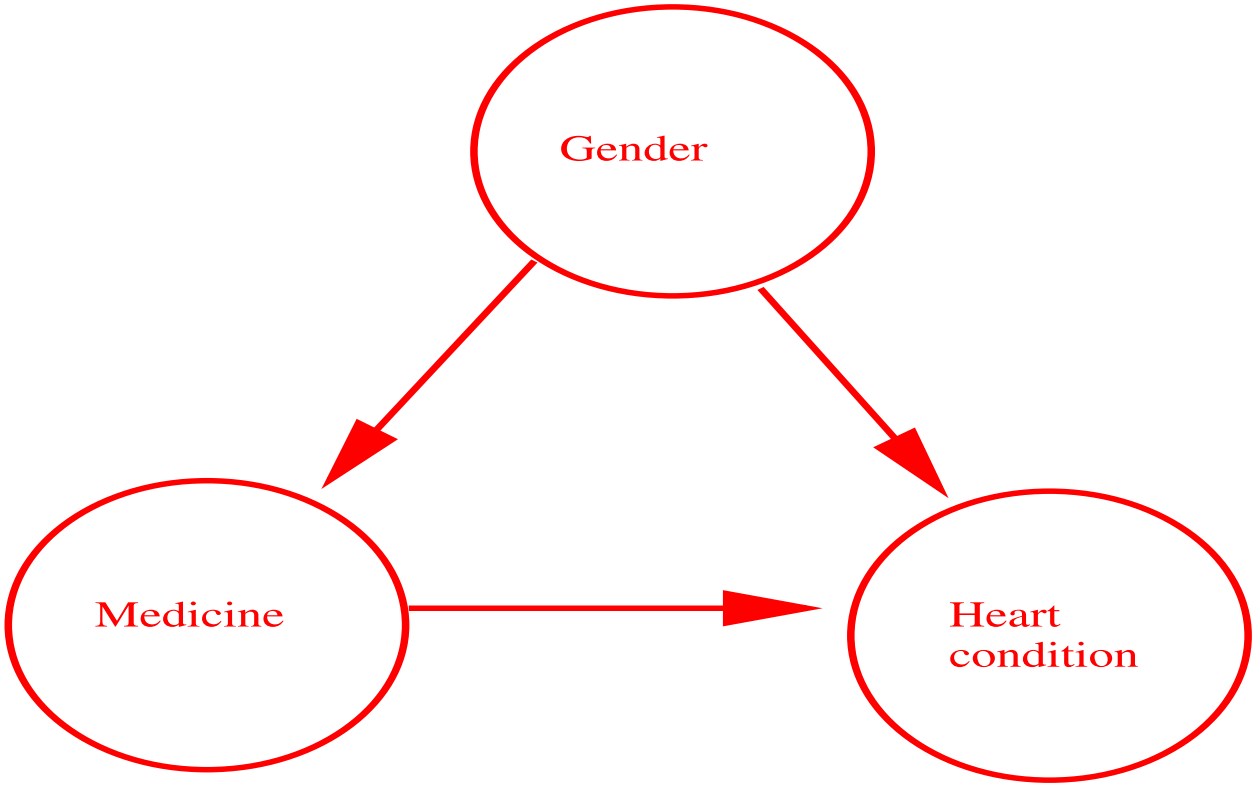
Figure 3.4: Is the relationship between x and y, really caused by z?

Confounder



- Unaccounted common causes: confounders.
- Example: gender may cause both selfselection of treatment and heart condition.

Confounder



- Unaccounted common causes: confounders.
- Example: gender may cause both selfselection of treatment and heart condition.

Counterfactuals/Potential Outcomes

Suppose now that X is a binary variable that represents some exposure. So $X = 1$ means the subject was exposed and $X = 0$ means the subject was not exposed. We can address the problem of predicting Y from X by estimating $E(Y|X = x)$.

Counterfactuals/Potential Outcomes

To address causal questions, we introduce counterfactuals or potential outcomes. Let Y_1 denote the response if the subject is exposed. Let Y_0 denote the response if the subject is not exposed.

> Once a subject is exposed, we observe Y_1. We cannot observe Y_0, what i.e., would have happened, if a subject had not been exposed.

Counterfactuals

Confounding variables. are variables that affect both X and Y. These variables explain why two groups of people are different in taking a medicine. In other words, these variables account for the dependence between X and Y_0, Y_1. By definition, there are no such variables in a randomized experiment (more of that later).

Causal questions involve the the distribution $p(y_0, y_1)$ of the potential outcomes. We can interpret $p(y_1)$ as $p(y|X \mapsto 1)$ and we can interpret $p(y_0)$ as $p(y|X \mapsto 0)$. The mean treatment effect or mean causal effect is defined by

$$\theta := E[Y \mid X \mapsto 1] - E[Y \mid X \mapsto 0]. \tag{1}$$

The parameter θ has the following interpretation:θ is the mean response if we expose everyone minus the mean response if we exposed no-one. In general,

$$E[Y_1] \neq E[Y|X \mapsto 1], \quad E[Y_0] \neq E[Y|X \mapsto 0].$$

We say that there is no unmeasured confounding, or that ignorability holds, if Z is a confounding variable, iff

$$E[Y_1] = E[E[Y_1 \mid X = 1, Z]] = E[E[Y_1 \mid Z]]$$

$$E[Y_0] = E[E[Y_1 \mid X = 0, Z]] = E[E[Y_0 \mid Z]]$$

and $E[Y_1 \mid Z]$ and $E[Y_0 \mid Z]$ can be estiamated from data.

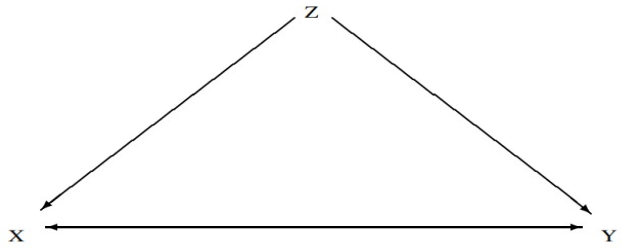


Figure 3.4: Is the relationship between x and y, really caused by z?

$$X = \beta_{xz} Z$$

$$Y = \beta_0 + \beta_{yx} X + \beta_{yz} Z + \epsilon$$

$$E[Y \mid X = x, Z = z] = \beta_0 + (\beta_{yx} + \beta_{yz})\, z$$

$X \mapsto 0$.

$$E[Y \mid X \mapsto 0, Z = z] = \beta_0 + \beta_{yz} z$$

$X \mapsto 1$

$$E[Y \mid X \mapsto 1, Z = z] = \beta_0 + \beta_{yx} + \beta_{yz} z$$

Suppose $\beta_{yz} < 0$ and $(\beta_{yx} + \beta_{yz}) > 0$. If we ignore Z,

$$E[Y \mid X = x] = E[Y \mid X \mapsto x] = \beta_0 + \beta_{yx} x$$

,

Causal Questions (Dawid 2007)

Causal questions come in two principal flavours: questions about the effects of applied causes (*EoC*), and questions about the causes of observed effects (*CoE*). Examples are:

- Effects of causes : I have a headache. Will taking aspirin help?
- Causes of effects : My headache has gone. Is it because I took aspirin?

Much of classical statistical design and analysis - for example randomized agricultural or medical experiments —has been crafted to address EoC-type questions. CoE-type questions are gen- erally more of intellectual than practical interest. They are also much more problematic, both philosophically and methodologically.

Part 3: Causality and Statistics

Early prominent statisticians have argued that **causation can only be inferred from randomized experiments** (R.A. Fisher) or that there is no need for a notion of causality at all (K. Pearson: The Grammar of Science, 1892 1st Ed.).
For Pearson **the only proper goal of scientific investigation is to provide descriptions of experience (perceptions) in a mathematical form (e.g., a coefficient of correlation). An effort to advance beyond this description means in this view to evoke hidden or metaphysical entities such as causes, and is not scientific**. Following in Pearson's footsteps a majority of statistical studies claim 'correlation` or 'association', unless randomized experimental trials are performed.

Causality and Statistics

Terry Speed is quoted as saying : *Considerations of causality should be treated as they have always been treated in statistics : preferably not at all (but if necessary, then with great care)*.

Causality in Statistics

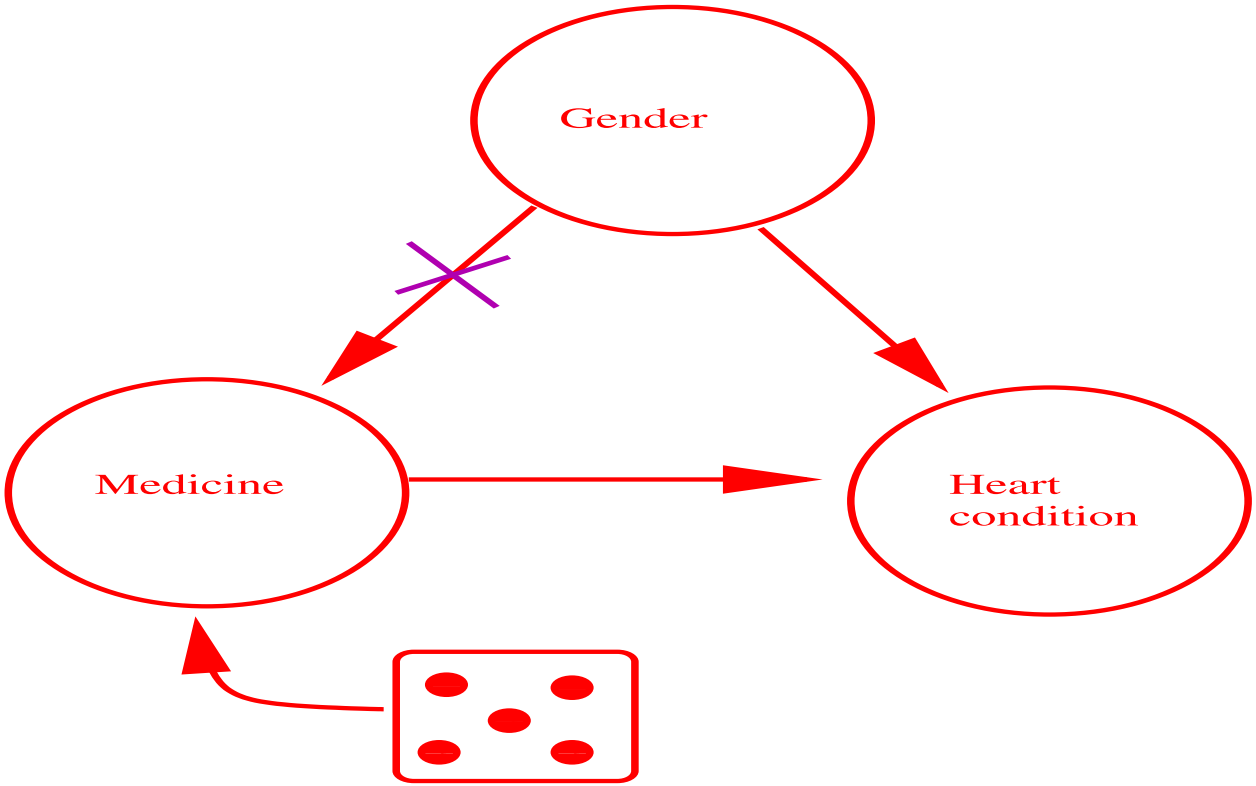
Probabilistic reasoning and statistical studies claim 'correlation' or 'association', unless randomized experimental trials are performed. Yet, statistical methods are routinely used to justify causal inference from data not obtained by randomized experiments.
D. Freedman (1999): From Association to Causation: Some Remarks on the History of Statistics. Statistical Science*, vol. 14, 3, 243–258.*

Traditionally, Statistics has been concerned with uncovering and describing associations, and statisticians have been wary of causal interpretations of their findings. **But users of Statistics have rarely had such qualms. For otherwise what is it all for?**
(A.P. Dawid, 2007)

The ability to infer causal relationships forms the basis for learning and acting in an intelligent manner in the external world. Knowledge of causal relationships as opposed to mere statistical associations gives a sense of genuine understanding as well as a sense of potential control emanating from the capability to *predict the consequences of actions that have not been performed*. (Judea Pearl).

- Experimental studies: Sir R. Fisher introduced randomized controlled trials (RCT) as prerequisite for estimation of causal effects.

The role of randomization (2)



- The randomized assignment overrides the original causal mechanisms.

The randomized controlled trial (RCT) is generally taken as a gold-standard for the assessment of causal effects. In particular, randomization helps guard against many of the difficulties of interpreting observational evidence displayed by the above examples.

- By ensuring that assignment of treatment is entirely unrelated to other variables, randomization can eliminate confounding and the problem of the common cause. When we compare outcomes in different treatment groups, we are "comparing like with like", and can therefore ascribe any observed differences in effect to the only real difference between the groups: the treatment applied.
- External intervention to apply a treatment ensures that any observed associations between treatment and response can indeed be given a causal interpretation.
- Because application of treatment necessarily precedes measurement of outcome, the direction of causality is clear.

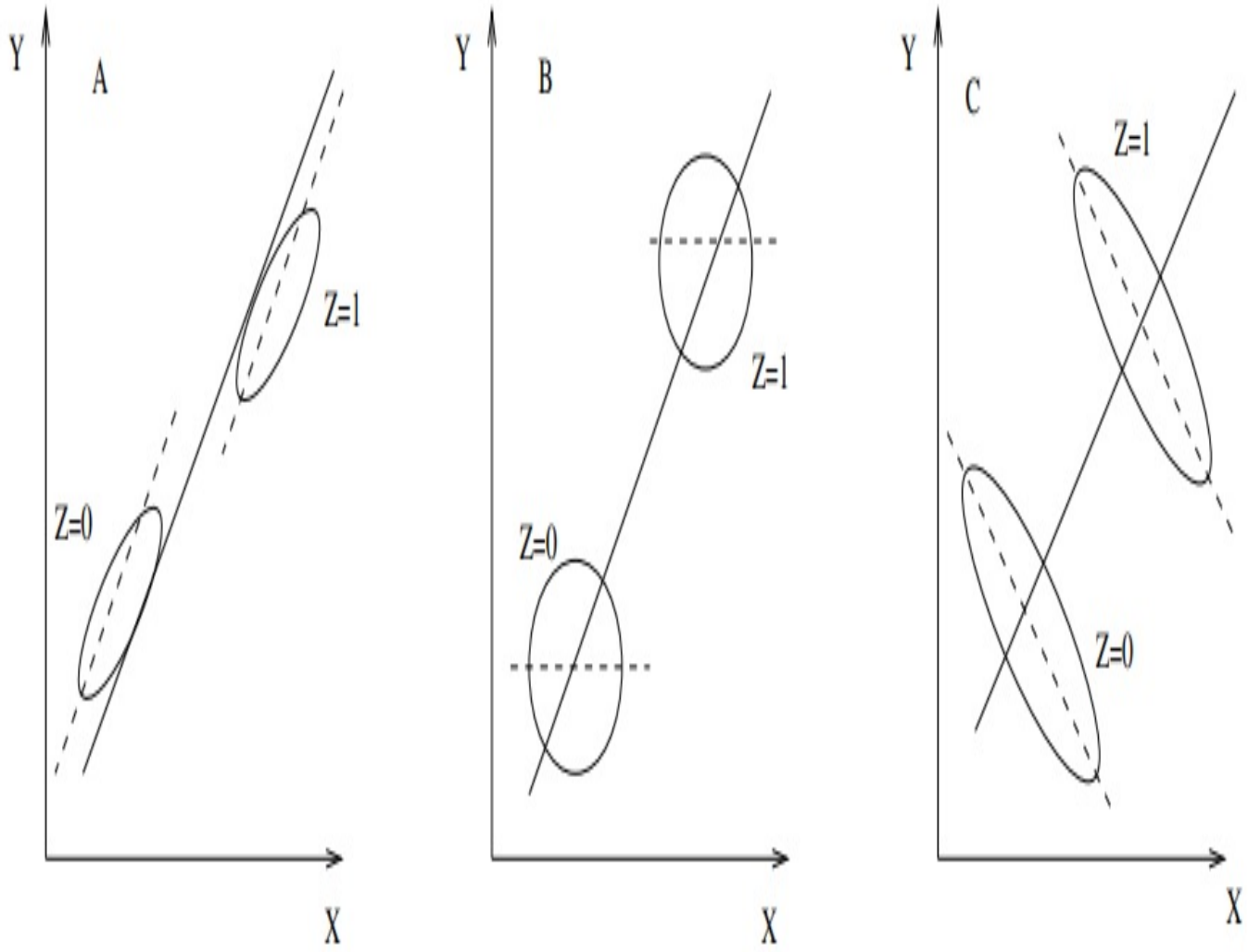
Observational Studies

However, it is often not possible, for pragmatic or ethical reasons, to conduct a RCT; or we may want to interpret observational data collected by others.

- Observational studies: we may have to rely on observational data, where the putative "causal factors" are not under the control of the investigator.
    A) interested in comparing what would happen to a patient (or plot of soil) under various treatments that we might apply: that is, we are really concerned with assessing and comparing possible treatment **interventions**, applied to and compared on one and the same subject - having, necessarily, constant characteristics.

B) The variables measured, and the conditions under which they are produced and recorded, may not be identical with those in the new situations we wish to understand. In the presence of such *confounding*, we will not know whether to ascribe observed differences between the responses in different treatments groups to the treatments themselves, or to other differences between the groups, whose effects would persist even if all units were treated identically.

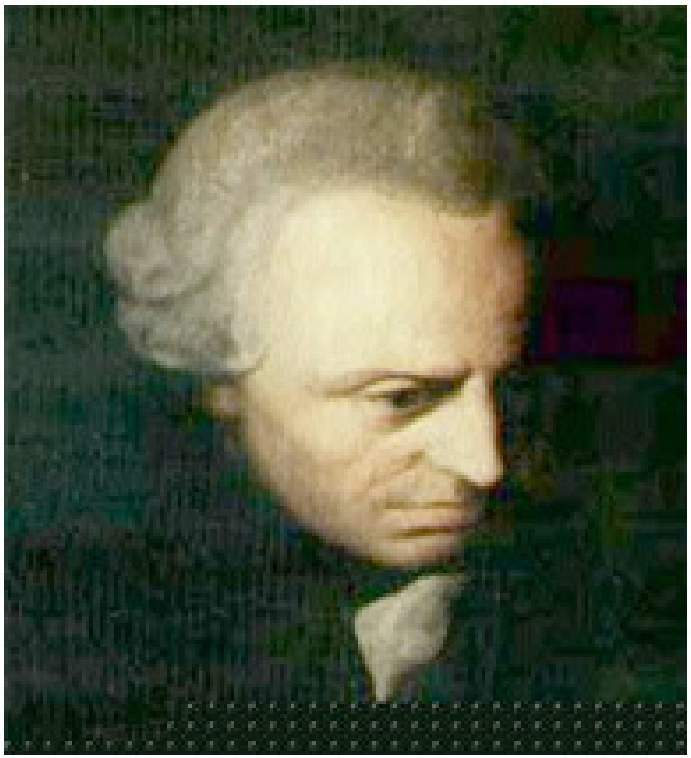
Observational research on postmenopausal hormone therapy suggested a 40-50 % reduction in coronary heart disease incidence among women using these preparations. In contrast, the Women's Health Initiative clinical trial of estrogen plus progestin found an elevated incidence.
Even after age adjustment, estrogen-plus-progestin hazard ratio estimates for coronary heart dis- ease, stroke, and venous thromboembolism in the observational study were 39-48 % lower than those in the clinical trial.


Y
A
Z=1
Z=0
X
Y
B
Z=1
Z=0
X
Y
C
Z=1
Z=0
X

Selection

An important warning of the perils of interpreting observational evidence: The underlying problem is that those choosing or chosen to receive the treatment may well not be typical of the population at large. People who take HRT or supplements tend to be healthier, so that the observational studies may be measuring the influence of other factors than that under investigation.

- Humean view that causal relations are not directly observable vs. the Kantian view that people hold prior beliefs about unobserved powers of causes.

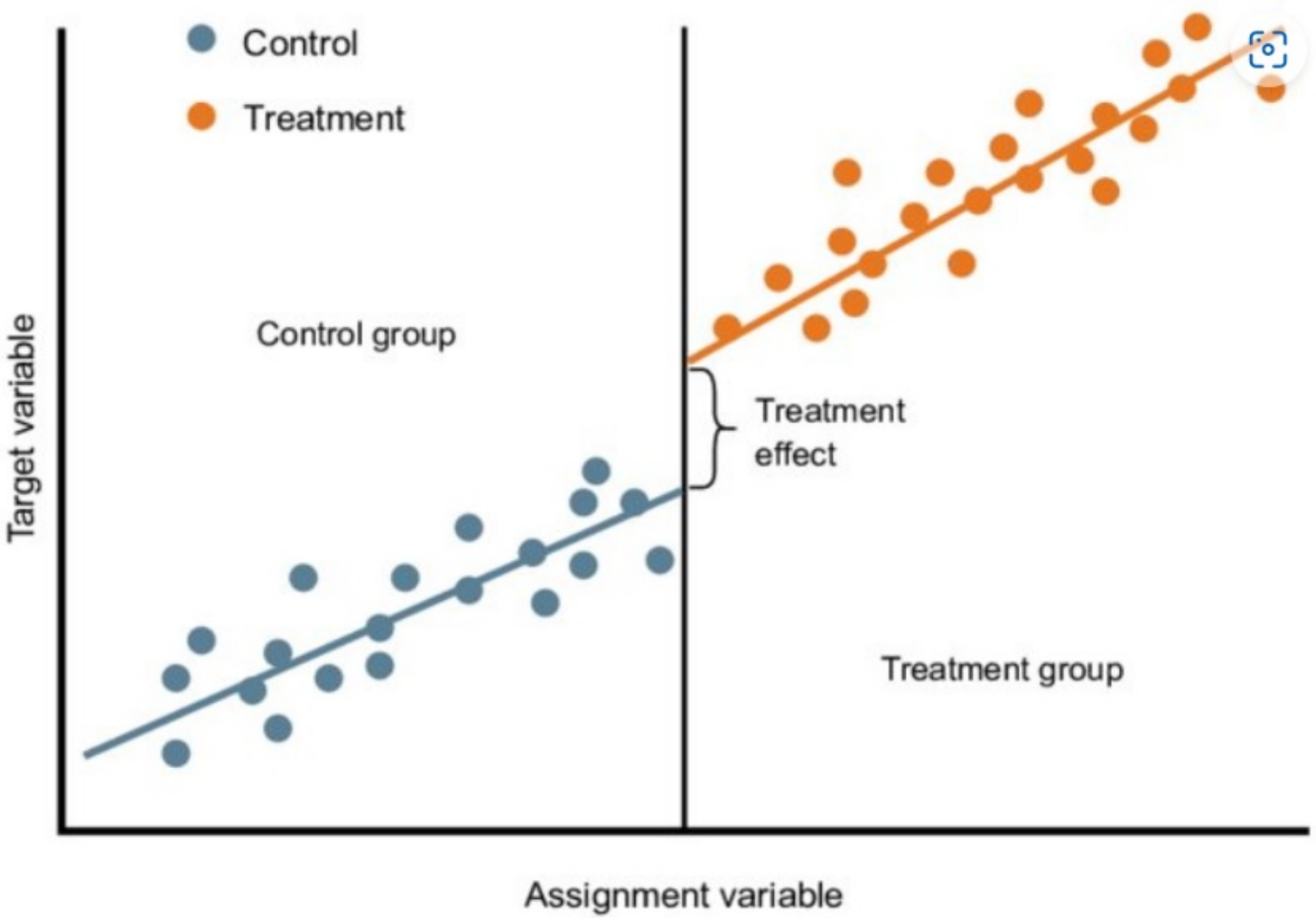


Immanuel Kant, 1724-1804

David Hume, 1711-1776
(portrait by Allan Ramsay)

Regression Discontinuity Design (RDD)

RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomisation is unfeasible. However, it remains impossible to make true causal inference with this method alone, as it does not automatically reject causal effects by any potential confounding variable.


Control
Treatment
Control group
Treatment
effect
Treatment group
Target variable
Assignment variable

Bayesian causal inference in automotive software engineering and online evaluation

Yuchu Liu, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, and Jonn Lantz

Patric Purcell: *Engineering Student Attendance at Lectures: Effect on Examination Performance*

One could argue that the purpose of engineering education is not to produce students capable of scoring well in examinations, but rather to produce students that leave the third level institution equipped to practice as excellent engineers, capable of undertaking continuous professional development. The sparse data relating examination mark to later career performance suggests little correlation between them [10]. Rather than making lecture attendance compulsory, perhaps a more productive approach might be to

Another Study

in a university in which attendance to classes is not mandatory. The methods used are cluster analysis and regression analysis. We find that students form three distinct groups: 1) those who drop out before the final exam, 2) those who attend classes as well as the exam, and 3) those who study independently and attend the exam. Most importantly, we find that in group 2, attendance is positively and significantly related to performance, after controlling for the effect of other variables potentially related to performance. We also find that students in group 3 are characterized by compelling reasons for absenteeism and a good ability to proactively search for information and study independently. The results are relevant for teachers and students alike. First and

David Hume, An Enquiry Concerning Human Understanding.

We may define a cause to be an object, followed by another, and where all objects, similar to the first, are followed by objects similar to the second. Or in other words, where, if the first object had not been, the second never had existed ..." —

If x had not happened, then y would not have happened. Simple? Consider the example

- Person A is about to make a traverse across the Sahara in his Land Rover.

If x had not happened, then y would not have happened. Simple? Consider the example

- Person A is about to make a traverse across the Sahara in his Land Rover.
- The night before A leaves, B puts poison in A's water supply.

If x had not happened, then y would not have happened. Simple? Consider the example

- Person A is about to make a traverse across the Sahara in his Land Rover.
- The night before A leaves, B puts poison in A's water supply.
- Not knowing about B, C empties A's water supply.

If x had not happened, then y would not have happened. Simple? Consider the example

- Person A is about to make a traverse across the Sahara in his Land Rover.
- The night before A leaves, B puts poison in A's water supply.
- Not knowing about B, C empties A's water supply.
- The next day A sets off into the desert and dies. **Who killed A?**


That's all Folks!