The Role of Causal Inference in the Scientific Method

Prof. Marcos López de Prado
Cornell University

Seminar’s Objective

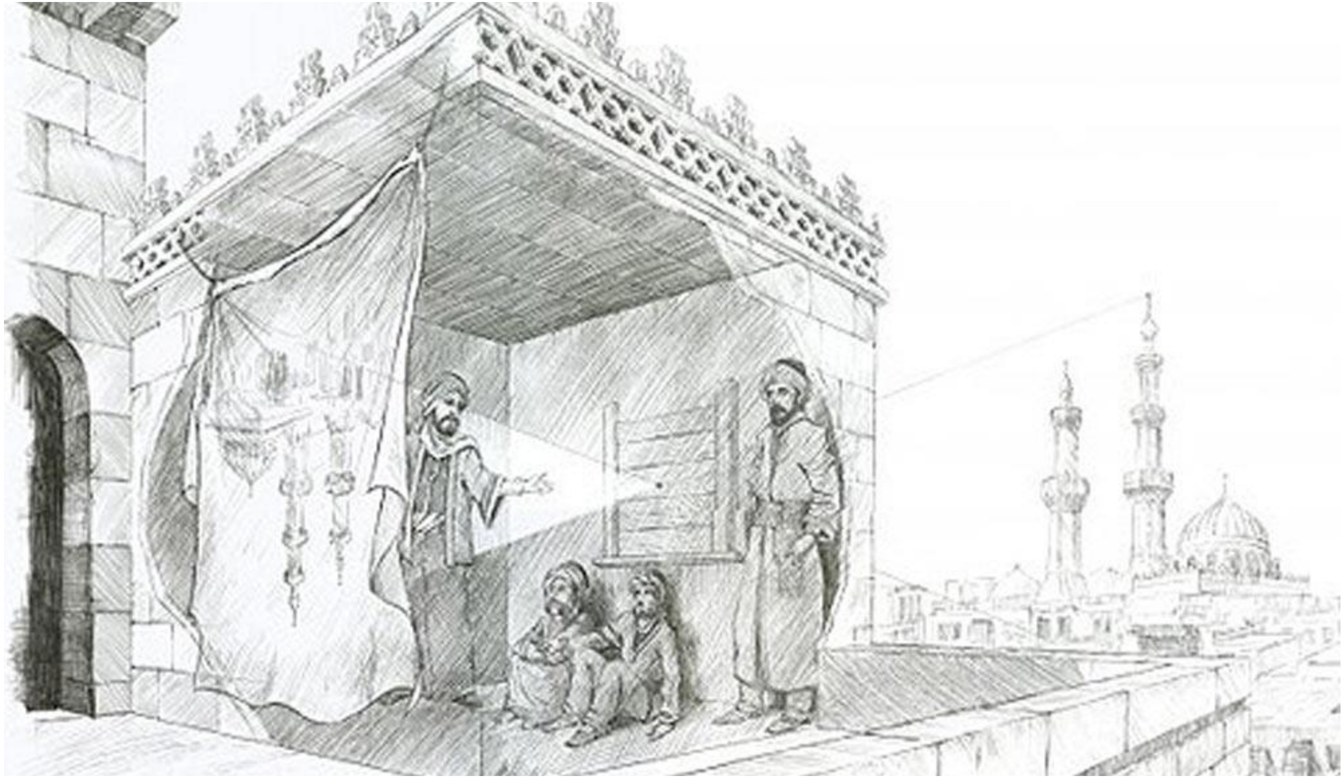
- Every student of statistics learns that correlation does not imply causation
  - Association is an observational property
  - Causation is an interventional concept
- Causality plays a fundamental role in the scientific method
  - Scientific theories are falsifiable statements of the form “X causes Y through mechanism M”
- Much of scientific research is false due to:
  - **Type-A Spuriosity: Statistical flukes**
  - **Type-B Spuriosity: Non-causal association**
- This seminar proposes ways to solve the **replication crisis** that afflicts scientific research

Why Study Cause and Effect?

"Happy the man, who,
studying Nature's laws,
thro' known effects can
trace the secret cause"

The Second Book of the Georgics
Publius Vergilius Maro, "Virgil" (70 – 19 BC)

Around the year 1011 that Arab mathematician Hasan Ibn Al Haytham (965 - 1040) proposed a scientific method for deducing causal mechanisms.



David Hume defined a cause-effect (causal) relation as that “where, if the first object had not been, the second never had existed.”

An Enquiry concerning Human Understanding. Sec. VII. (1748)
David Hume (1711 - 1776)

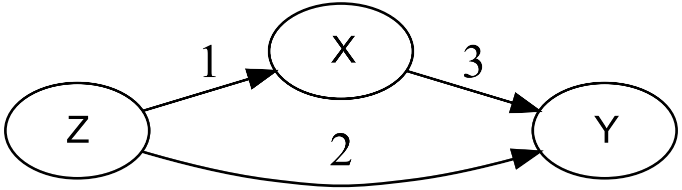
The Three Stages of the Scientific Method

Stage	Statement	Example
Phenomenological (induction)	X is associated with Y	Smoking is associated with lung cancer
Theoretical (abduction)	X causes Y through mechanism M	Smoking causes lung cancer through chemicals that mutate the DNA of lung cells, inducing uncontrolled cell growth
Falsification Refutation (deduction)	Refutation attempts: • X does not cause Y • X causes Y, but not through mechanism M	• Controlled experiments, e.g. animal lab studies • Natural experiments, e.g. regression discontinuity

Association vs. Causation

Paths

- A **data-generating process** can be represented as a directed acyclic graph (DAG)
  - Nodes are variables
  - Arrows indicate the direction of dependence
- A **path** is a sequence of arrows and nodes that connect two variables X and Y, regardless of the direction of causation
- A **directed path** is a path where all arrows point in the same direction
- In a directed path that starts in X and ends in Y
  - X is an **ancestor** of Y, and
  - Y is a **descendant** of X



A DAG is a directed graph with no directed cycles.

In the DAG above, there are two paths between X and Y:

a) A directed path: $X \rightarrow Y$
b) A non-directed path: $X \leftarrow Z \rightarrow Y$

In the DAG above, X is a descendant of Z, and Y a descendant of X and Z.

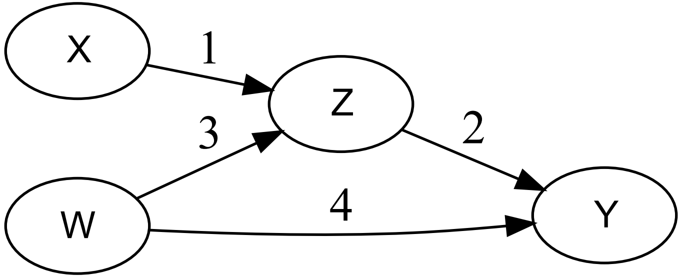
Blocked Paths

- In a DAG with three variables $\{X, Y, Z\}$, a variable Z is
  - a **confounder** when the causal relationships include a structure $X \leftarrow Z \rightarrow Y$
  - a **collider** when the causal relationships are reversed, i.e. $X \rightarrow Z \leftarrow Y$
  - a **mediator** when the causal relationships include a structure $X \rightarrow Z \rightarrow Y$

A path between X and Y is **blocked** if either:

a) the path traverses a collider, and the researcher has not conditioned on that collider or its descendants; or

b) the researcher conditions on a variable in the path between X and Y, where the conditioned variable is not a collider



In the above DAG:

- W is a confounder to Z and Y
- Z is a collider to X and W
- Z is a mediator between X and Y

The path $X \rightarrow Z \leftarrow W \rightarrow Y$ is blocked by Z.

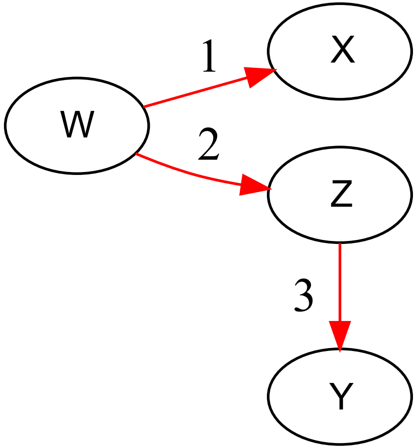
The only unblocked path between X and Y is the causal path, $X \rightarrow Z \rightarrow Y$.

What is Association?

- **Association** flows along an unblocked path between X and Y
  - Association is *symmetric* because paths do not follow the direction of causation
- Probabilistically, two variables X and Y are associated when knowing the value of one conveys information about the value of the other

$$\exists \boldsymbol{x}, \boldsymbol{y} | \boldsymbol{P}[\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}] \neq \boldsymbol{P}[\boldsymbol{Y} = \boldsymbol{y}]$$

- Statistical association is merely an observational statement on the joint distribution of probability
  - $P[Y = y | X = x]$ does not measure the *effect* of X on Y

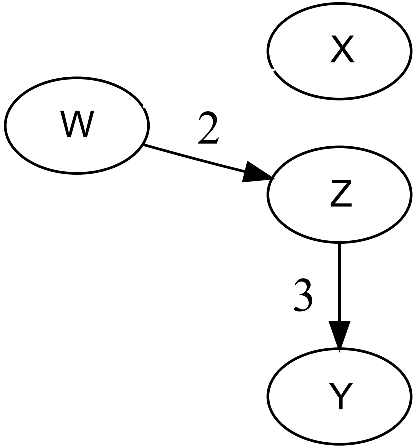


Weather (W) influences ice cream sales (X) and the number of swimmers (Z), hence the number of drownings (Y).

There is no directed path between X and Y, however X and Y are associated (red-dashed undirected edge), because of the unblocked path $X \leftarrow W \rightarrow Z \rightarrow Y$.

What is Causation?

- **Causal association** flows along an unblocked *directed* path that starts in treatment X and ends in outcome Y, denoted the causal path
- Let $do[X = x]$ represent the do-operator on X
  - This is an intervention that sets the value of X to x, hence X is not influenced by any other variable
- <u>**Definition**</u>**: $\boldsymbol{X}$ causes $\boldsymbol{Y}$ iff $\boldsymbol{P\big[Y|do[X]\big] > P[Y]}$**
- Association implies causation only if all non-causal paths are blocked
- Causality is
  - an interventional (beyond observational) concept
  - asymmetric (directional)
  - sequential: X happens first, and then Y adapts



A do-operation on X removes arrow (1), because X is no longer a function of W, while keeping all other things equal ("ceteris paribus"). As a result, there is no unblocked path between X and Y, and $P\big[Y|do[X]\big] = P[Y]$.

Blocking Non-Causal Paths

Interventional Studies

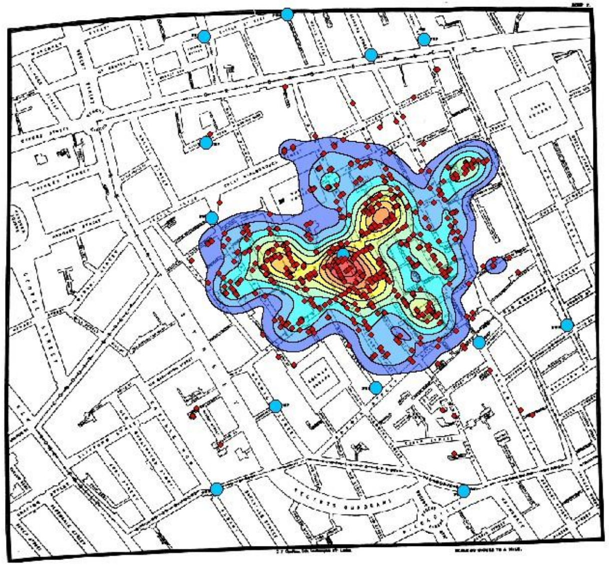
- In a **controlled experiment**, scientists assess causality by observing the effect on Y of changing the values of X, while keeping constant all other variables
  - E.g., Ohm's law of current, Newton's law of gravitation, etc.
- When some of the variables are not under direct experimental control, scientists may execute a **randomized controlled trial** (RCT)
  - E.g., Effectiveness of Pfizer's COVID-19 vaccine
- Under random assignment, subjects in the treatment group ($X = x_1$) are assumed to be indistinguishable from subjects in the control group ($X = x_0$)
  - Thanks to this assumption, the difference in outcomes can be attributed to the treatment

In the 1930s, Ronald Fisher popularized randomized experiments as a way to de-confound variables.

The first published RCT appeared in 1948. Today, well-blinded RCTs are considered the gold standard in experimental research.

Natural Experiments

- Sometimes interventional studies are not possible, because they are unfeasible, unethical, or prohibitively expensive
- In a natural experiment, subjects are assigned to the treatment and control groups determined randomly by Nature or by other factors not controlled by scientists
- Examples of natural experiments include
  - **Regression discontinuity design (RDD)**: When treatment and control groups are comparable in everything but the slight difference in the assignment variable, attributed to noise
  - **Crossover studies (COS)**: When the effect of confounders does not change per subject over time
  - **Difference-in-differences studies (DID)**: When factors other than the treatment influence the outcome over time



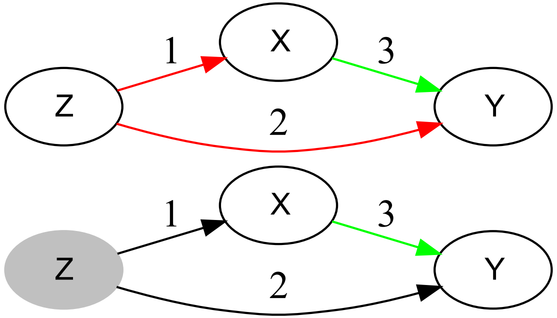
In 1854, Dr. John Snow found that exposure to contaminated water causes cholera. Sick and healthy neighbors of London’s Soho district were comparable in all respects, except by their use of different water pumps.

Simulated Intervention: Backdoor Adjustment

- Under some conditions, we can simulate an intervention
- A **backdoor path** between X and Y is an unblocked non-causal path that connects those two variables
- A set of variables S satisfies the **backdoor criterion** if the following two conditions are true:
  - conditioning on S blocks all backdoor paths between X and Y
  - S does not contain any descendants of X
- Then, S is a sufficient adjustment set, and the causal effect of X on Y can be estimated as:

$$P\left[Y = y \middle| do[X = x]\right] = \sum_{s} P[Y = y|X = x, S = s]P[S = s]$$

- Examples of other adjustments: Front-door, IV, etc.



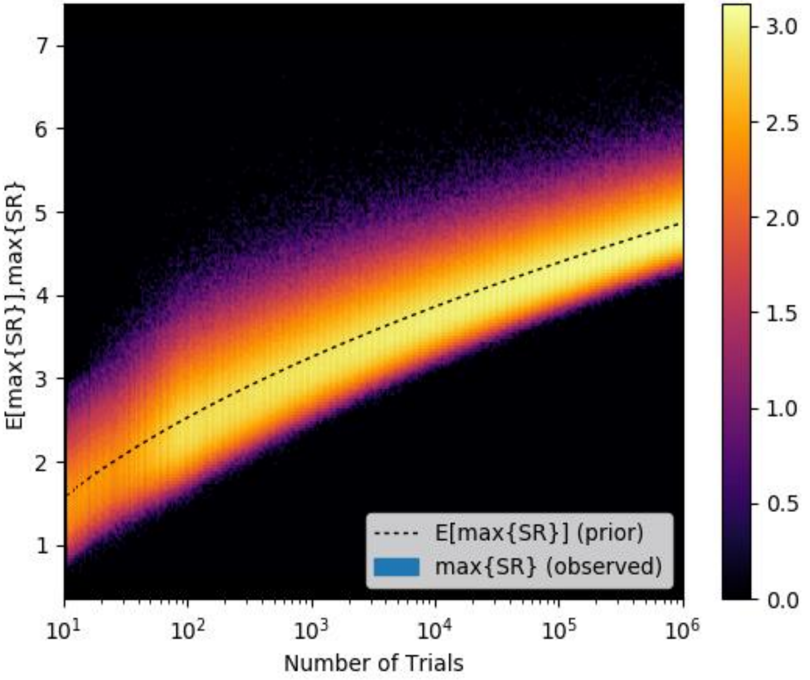
Conditioning on confounder Z (shaded node) blocks the path $X \leftarrow Z \rightarrow Y$, leaving the causal path $X \rightarrow Y$ as the only unblocked path.

Under those circumstances, association does imply causation, and we can simulate the outcome of a do-operation through conditional probabilities.

Spurious Results

Type-A Spuriosity

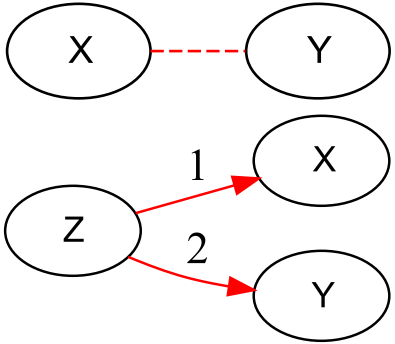
- Type-A spuriosity occurs when a researcher mistakes random variability (noise) for signal, resulting in a *false association*
- Type-A spuriosity has several attributes:
  a) it results in type-1 errors (false positives)
  b) for the same number of trials, it has a lower probability to take place as the sample size grows
  c) it can be corrected through multiple-testing adjustments
- Two main reasons for Type-A spuriosity
  - *p*-hacking, e.g., Hochberg [1988]
  - Backtest overfitting, Bailey and López de Prado [2014]
- “Why most published research findings are false” (John Ioannidis [2005])



Distribution of the maximum t-value as a function of the number of trials, where the true t-value is zero. See “The False Strategy Theorem.”

Type-B Spuriosity

- Type-B spuriosity occurs when a researcher mistakes association for causation (e.g., due to misspecification)
- Type-B spuriosity has several attributes:
  a) it results in type-1 errors and type-2 errors (false positives and false negatives);
  b) it can occur with a single trial;
  c) it has a greater probability to take place as the sample size grows, because the non-causal association can be estimated with lower error; and
  d) it cannot be corrected through multiple-testing adjustments. Its correction requires the injection of extra-statistical information, in the form of a causal theory
- Type-B spurious factors exhibit
  – misattributed causes, due to confounder bias, collider bias, etc.
  – time-varying causes, under non-causal parameter shift

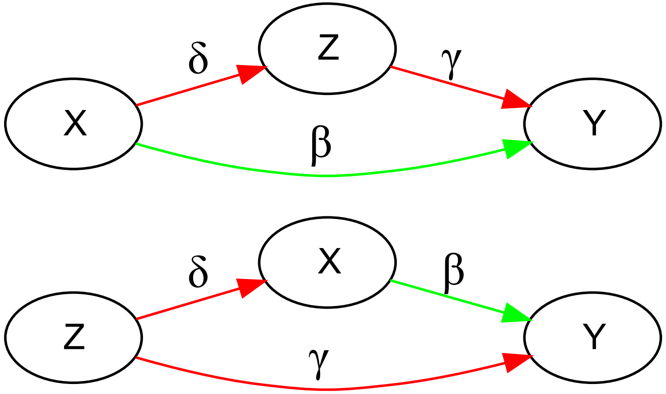


The top graph is an example of false association (type-A spuriosity). The bottom graph is an example of association mistaken for causation (type-B spuriosity).

Type-A and type-B spuriosity are mutually exclusive. For type-B spuriosity to take place, the association must be non-causal but true, which precludes that association from being type-A spurious.

Type-B(1) Spuriosity: Under-Controlling

- Consider a researcher who fits $Y = X\beta + \varepsilon$ on data generated by $Y := X\beta + Z\gamma + u$, where $\gamma \neq 0$ and u is white noise
  - As a consequence, $E[\varepsilon|X] = \gamma E[Z|X]$
  - $E[Z|X] \neq 0 \Rightarrow E[\varepsilon|X] \neq 0$ (exogeneity is not satisfied)
- Case 1: Z is a mediator ($Z \coloneqq X\delta + v$, with $\delta \neq 0$)
  - the chosen specification biases $\hat{\beta}$
  - however $\hat{\beta}$ can still be interpreted as a total causal effect
- Case 2: Z is a confounder ($X \coloneqq Z\delta + v$, with $\delta \neq 0$)
  - the chosen specification biases $\hat{\beta}$
  - $\hat{\beta}$ cannot be interpreted as a causal effect (direct or total)

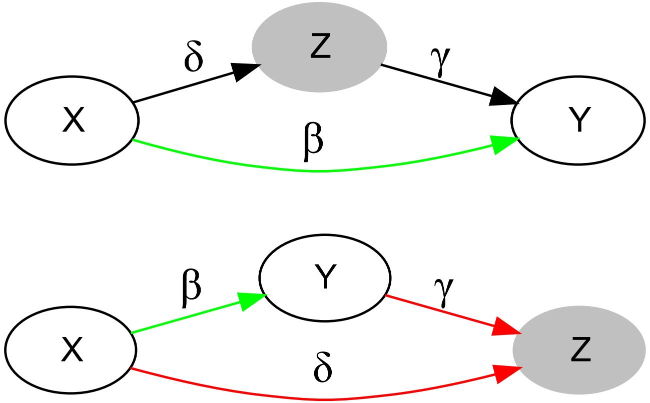


Statistics textbooks often treat all missing variables as equal. **This is a mistake.**

In the top graph, Z is a mediator, and missing Z has mild consequences. In the bottom graph, Z is a confounder, and missing Z will likely lead to false positives or false negatives.

Type-B(2) Spuriosity: Over-Controlling

- Statisticians have been trained for decades to control for any variable Z associated with Y that is not X
  - Statistics textbooks dismiss as a harmless error the inclusion of an irrelevant variable, regardless of the variable's role in the causal graph
- Case 1: Z is a mediator
  - Controlling for a mediator interferes with the mediated effect and the total effect, which the researcher may wish to assess
  - $\hat{\beta}$ measures only the direct effect
- Case 2: Z is a collider
  - Controlling for a collider opens a backdoor path, $X \rightarrow Z \leftarrow Y$ (**Berkson's fallacy**)

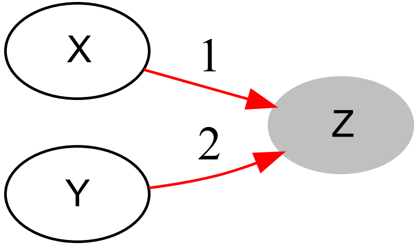


Greene [2012, section 4.3.3] states that the only downside to adding superfluous variables is a reduction in the precision of the estimates. **This is a mistake.**

Over-controlling for a collider has the same consequences as under-controlling for a confounder: **an open backdoor**.

Type-B(3) Spuriosity: Specification-Searching

- The use of explanatory power (an associational, non-causal concept) for selecting the specification of a factor model is inconsistent with that model's causal content
- Specification-searching commingles two separate and sequential stages of the causal analysis:
  1) **Causal discovery**: Finding the causal graph
  2) **Control**: Use the graph to determine the correct specification
- Stage (2) should be informed by stage (1), not the other way around
  - A researcher may achieve higher explanatory power by combining multiple causes of Y, at the expense of biasing the multiple parameters' estimates due to multicollinearity or over-controlling for a collider



Empirical studies often justify the chosen specification in terms of explanatory power. This comingles causal discovery with controlling, and all but ensures that the regressors will include colliders (Berkson's fallacy).

Monte Carlo Experiments

Type-B(1) Spuriosity :
Forks

Confounders

- Consider the fork structure in the right graph
- Applying Bayesian network factorization

$$P[X,Y,Z] = P[Z]P[X|Z]P[Y|Z]$$

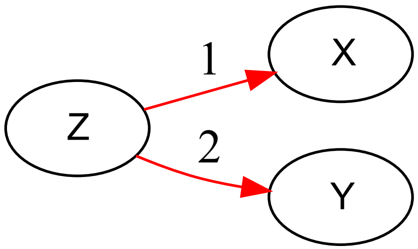
- X and Y are associated, since

$$P[X,Y] = \sum_{Z} P[Z]P[X|Z]P[Y|Z] \neq P[X]P[Y]$$

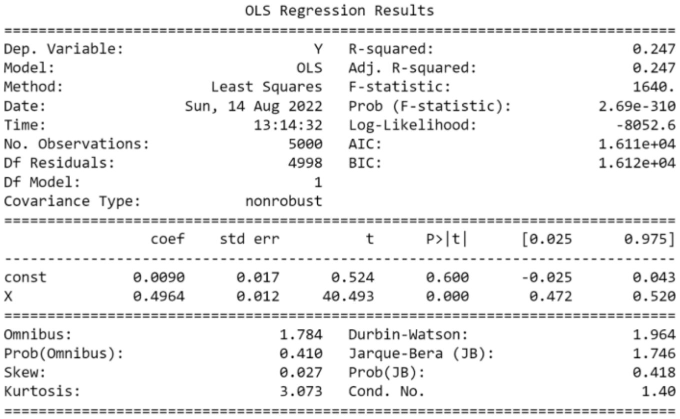
- This is an example of non-causal association
  - X and Y are associated through the backdoor path

$$Y \leftarrow Z \rightarrow X$$

- Given the causal content of the factor model, a statistically significant $\hat{\beta}$ implies that X causes Y
  - This claim of statistical significance is type-B spurious



$$Y_t = \alpha + \beta X_t + \varepsilon_t$$



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.247
Model:                            OLS   Adj. R-squared:                  0.247
Method:                 Least Squares   F-statistic:                     1640.
Date:                Sun, 14 Aug 2022   Prob (F-statistic):          2.69e-310
Time:                        13:14:32   Log-Likelihood:                -8052.6
No. Observations:                5000   AIC:                         1.611e+04
Df Residuals:                    4998   BIC:                         1.612e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0090      0.017      0.524      0.600      -0.025       0.043
X              0.4964      0.012     40.493      0.000       0.472       0.520
==============================================================================
Omnibus:                        1.784   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.410   Jarque-Bera (JB):                1.746
Skew:                           0.027   Prob(JB):                        0.418
Kurtosis:                       3.073   Cond. No.                         1.40
==============================================================================
```

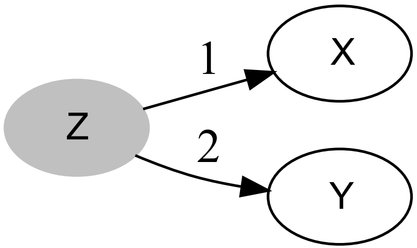
Classification: Public

The Backdoor Adjustment

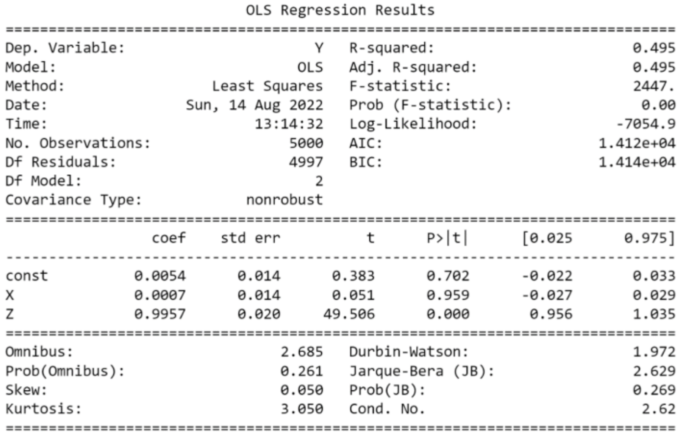
- The effect of conditioning by Z is equivalent to simulating a do-operation (an intervention)
  - It blocks the backdoor path, resulting in the conditional independence of X and Y,

$$P[X,Y|Z]=\frac{P[X,Y,Z]}{P[Z]}=P[X|Z]P[Y|Z]$$

- It is possible to remove the confounder-induced bias by adding Z as a regressor (the partial correlations method)
- With the correct model specification, the researcher concludes that X does not cause Y



$$Y_t=\alpha+\beta X_t+\gamma Z_t+\varepsilon_t$$

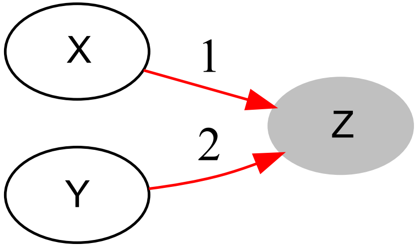


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.495
Model:                            OLS   Adj. R-squared:                  0.495
Method:                 Least Squares   F-statistic:                     2447.
Date:                Sun, 14 Aug 2022   Prob (F-statistic):               0.00
Time:                        13:14:32   Log-Likelihood:                -7054.9
No. Observations:                5000   AIC:                         1.412e+04
Df Residuals:                    4997   BIC:                         1.414e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0054      0.014      0.383      0.702      -0.022       0.033
X              0.0007      0.014      0.051      0.959      -0.027       0.029
Z              0.9957      0.020     49.506      0.000       0.956       1.035
==============================================================================
Omnibus:                        2.685   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.261   Jarque-Bera (JB):                2.629
Skew:                           0.050   Prob(JB):                        0.269
Kurtosis:                       3.050   Cond. No.                         2.62
==============================================================================
```

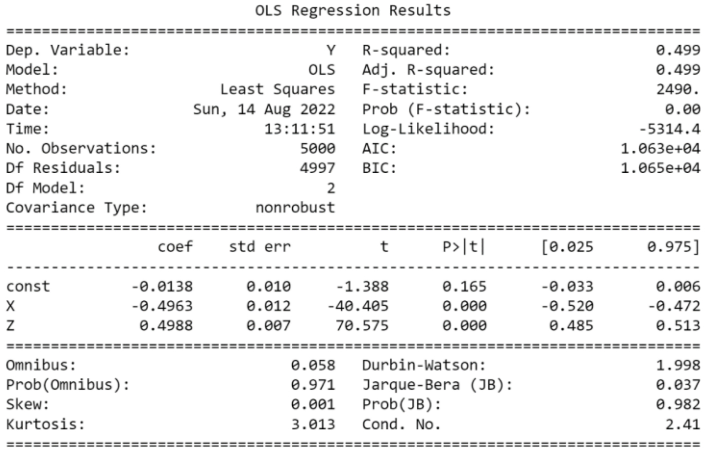
Type-B(2) Spuriosity : Immoralities

Colliders

- This causal graph shows a collider:
  - Variable Z is influenced by both, the treatment X and the outcome Y
- If a researcher controls for Z, the result is a false positive (bottom table)
- Compare the fork structure with the immorality structure
  - When the direction of causality is reversed, a confounder becomes a collider
  - The direction of causality is critical for specification
- One problem is, the direction of causality cannot always be determined from data
  - Causal graphs incorporates extra-statistical (beyond observational) information



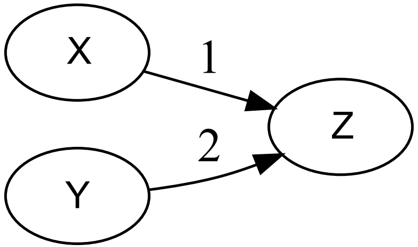
$$Y_t = \alpha + \beta X_t + \gamma Z_t + \varepsilon_t$$



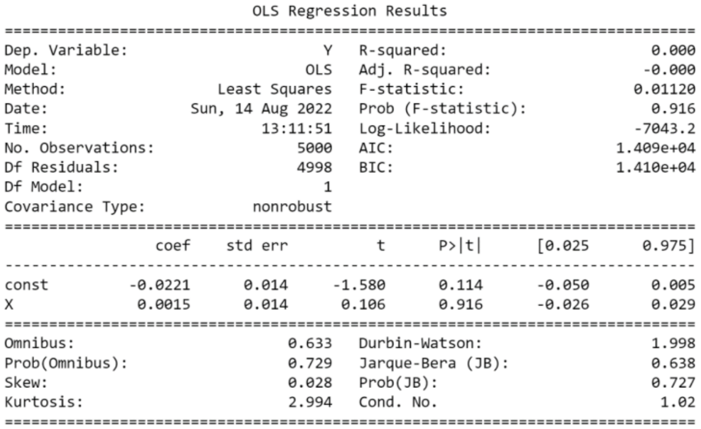
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.499
Model:                            OLS   Adj. R-squared:                  0.499
Method:                 Least Squares   F-statistic:                     2490.
Date:                Sun, 14 Aug 2022   Prob (F-statistic):               0.00
Time:                        13:11:51   Log-Likelihood:                -5314.4
No. Observations:                5000   AIC:                         1.063e+04
Df Residuals:                    4997   BIC:                         1.065e+04
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0138      0.010     -1.388      0.165      -0.033       0.006
X             -0.4963      0.012    -40.405      0.000      -0.520      -0.472
Z              0.4988      0.007     70.575      0.000       0.485       0.513
==============================================================================
Omnibus:                        0.058   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.971   Jarque-Bera (JB):                0.037
Skew:                           0.001   Prob(JB):                        0.982
Kurtosis:                       3.013   Cond. No.                         2.41
==============================================================================
```

Berkson’s Fallacy

- Berkson’s fallacy occurs when a spurious association is observed between two independent variables, as a result of conditioning on a collider
- With a careful selection of colliders, a researcher can present evidence in support of any spurious investment factor
- The correct causal treatment of a collider is to indicate its presence, and justify why researchers should not control for it
- Over-controlling leads to
  - false positives, in the presence of colliders
  - false negatives, in the presence of mediators
    - E.g., controlling for Z in $X \rightarrow Z \rightarrow Y$



$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

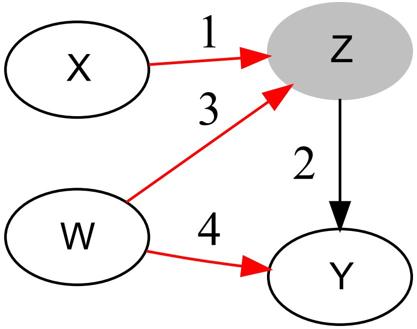


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                   0.01120
Date:                Sun, 14 Aug 2022   Prob (F-statistic):              0.916
Time:                        13:11:51   Log-Likelihood:                -7043.2
No. Observations:                5000   AIC:                         1.409e+04
Df Residuals:                    4998   BIC:                         1.410e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0221      0.014     -1.580      0.114      -0.050       0.005
X              0.0015      0.014      0.106      0.916      -0.026       0.029
==============================================================================
Omnibus:                        0.633   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.729   Jarque-Bera (JB):                0.638
Skew:                           0.028   Prob(JB):                        0.727
Kurtosis:                       2.994   Cond. No.                         1.02
==============================================================================
```

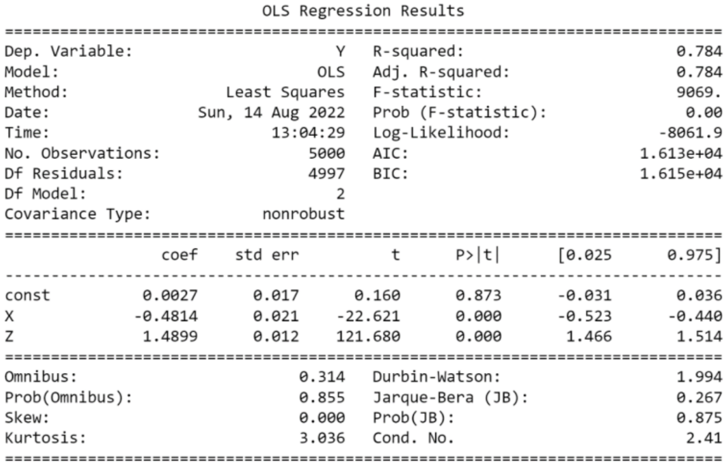
Type-B(3) Spuriosity :
Chains

Confounded Mediators

- This causal graph shows a mediator and a confounder:
  - Variable Z mediates the causal flow from the treatment X to the outcome Y
  - Variable W confounds Z and Y
- If a researcher controls for Z, the outcome is a false positive (bottom table)
  - The reason is that Z also operates as a collider to X and W
    - Controlling for Z opens a backdoor path $X \rightarrow Z \leftarrow W \rightarrow Y$
  - While it is true that X causes Y (through Z), the collider's bias is so strong that the sign of the relationship is reversed ($\hat{\beta} \ll 0$)
- In the absence of link 3, controlling for Z would have led to a false negative



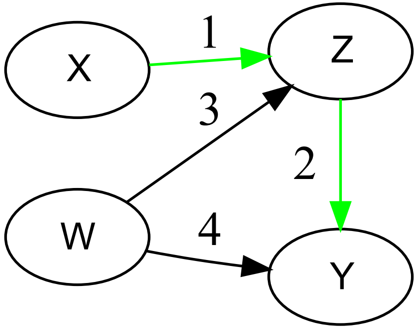
$$Y_t = \alpha + \beta X_t + \gamma Z_t + \varepsilon_t$$



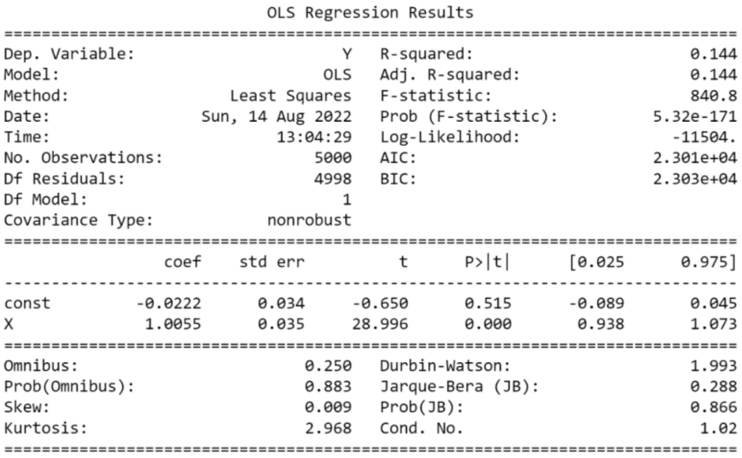
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.784
Model:                            OLS   Adj. R-squared:                  0.784
Method:                 Least Squares   F-statistic:                     9069.
Date:                Sun, 14 Aug 2022   Prob (F-statistic):               0.00
Time:                        13:04:29   Log-Likelihood:                -8061.9
No. Observations:                5000   AIC:                         1.613e+04
Df Residuals:                    4997   BIC:                         1.615e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0027      0.017      0.160      0.873      -0.031       0.036
X             -0.4814      0.021    -22.621      0.000      -0.523      -0.440
Z              1.4899      0.012    121.680      0.000       1.466       1.514
==============================================================================
Omnibus:                        0.314   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.855   Jarque-Bera (JB):                0.267
Skew:                           0.000   Prob(JB):                        0.875
Kurtosis:                       3.036   Cond. No.                         2.41
==============================================================================
```

Mediation Fallacy & Simpson’s Paradox

- The Mediation Fallacy involves conditioning on the mediator when the mediator and the outcome are confounded
- Simpson’s paradox occurs when there is an association in several groups, but it disappears or reverses when the groups are combined
- The solution to Simpson’s paradox is to inject extra-statistical information in the form of a causal graph
- We can estimate the unbiased effect ($\hat{\beta} \gg 0$)
  - Specification-searching would have returned a misspecified model (R^2 drops from 0.78 to 0.14!)
  - Adding W increases R^2 to 0.71 (still below 0.78)



$$Y_t = \alpha + \beta X_t + \varepsilon_t$$



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.144
Model:                            OLS   Adj. R-squared:                  0.144
Method:                 Least Squares   F-statistic:                     840.8
Date:                Sun, 14 Aug 2022   Prob (F-statistic):          5.32e-171
Time:                        13:04:29   Log-Likelihood:                -11504.
No. Observations:                5000   AIC:                         2.301e+04
Df Residuals:                    4998   BIC:                         2.303e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0222      0.034     -0.650      0.515      -0.089       0.045
X              1.0055      0.035     28.996      0.000       0.938       1.073
==============================================================================
Omnibus:                        0.250   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.883   Jarque-Bera (JB):                0.288
Skew:                           0.009   Prob(JB):                        0.866
Kurtosis:                       2.968   Cond. No.                         1.02
==============================================================================
```

Conclusions

Ranking Empirical Evidence

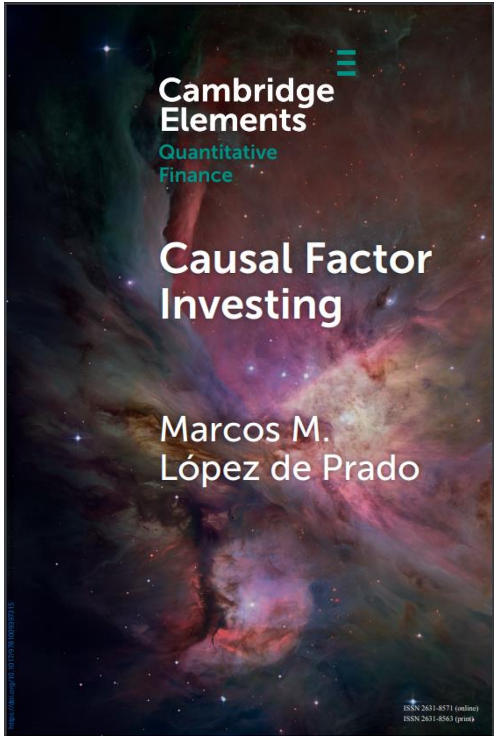
- A scientific theory is a falsifiable statement of the form "X causes Y through mechanism M"
- Not all empirical evidence is equal:
  - Associational evidence does not suffice to formulate a scientific theory
  - Causality allows counter-factual reasoning

- **The adoption of causal inference will have a major positive impact on scientific disciplines**

Type	Rigor	Example
Randomized controlled trials	Very high	Experiments on non-human subjects
Natural experiments	High	Study of genetically-identical twins
Simulated interventions	Medium	After controlling for other known causes of cancer, smoking appears to be a cause
Associational studies	Low	Strong correlation between cancer and smoking
Case studies	Very low	Disproportionate number of smokers among cancer patients
Expert opinion	Anecdotal	People who smoke tend to die earlier

Hierarchy of evidence: Phenomenological vs. scientific evidence.

For More Information



Download for free

Causal Factor Investing

(Cambridge University Press, 2023)

Available at:

https://www.cambridge.org/core/elements/causal-factor-investing/9AFE270D7099B787B8FD4F4CBADE0C6E

Disclaimer

- The views expressed in this presentation are my own, and do not necessarily reflect the views of Cornell University, the Abu Dhabi Investment Authority, or ADIA Lab
- No investment decision or particular course of action is recommended by this presentation
-

www.adialab.ae