Causal Inference for Data Scientists

From Correlation to Causation

Diogo Ribeiro
ESMAD – Escola Superior de Média Arte e Design
Lead Data Scientist, Mysense.ai

October 29, 2025

Outline

Correlation vs Causation: The Fundamental Distinction

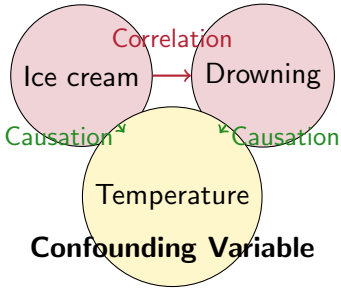
"Correlation does not imply causation, but it sure as hell provides a hint."

– Edward Tufte

The Problem with Pure Prediction:

- Association $\neq$ Causation: High correlation doesn't mean one causes the other
- Confounding: Hidden variables affect both cause and effect
- Spurious correlations: Random chance or third variables
- Simpson's paradox: Correlations reverse when conditioning

Why Causation Matters:



The Causal Hierarchy

1. **Association**: $P(Y|X)$

Famous Examples of Causal Confusion

Spurious Correlation	What People Concluded	Real Explanation
Chocolate consumption vs Nobel prizes	Chocolate makes you smarter	Wealth enables both luxuries and research
Ice cream sales vs crime rates	Ice cream causes crime	Hot weather increases both
Facebook friends vs longevity	Social media extends life	Social people live longer and use Facebook
Stork populations vs birth rates	Storks deliver babies	Rural areas have both more storks and higher birth rates
Shoe size vs reading ability	Big feet help reading	Age affects both shoe size and reading skills

Business Implications:

- Wrong investments: Correlation-based

The Data Science Challenge:

- Most ML focuses on prediction ($P(Y|X)$)

The Potential Outcomes Framework

Rubin Causal Model: Formalize causation through potential outcomes.

Setup:

- Units: $i = 1, 2, \ldots, n$
- Treatment: $T_i \in \{0, 1\}$
- Potential outcomes:
  - $Y_i(1)$: Outcome if treated
  - $Y_i(0)$: Outcome if not treated
- Observed outcome:
  $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$

Individual Treatment Effect:

Average Treatment Effect (ATE):

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

What We Can Observe:

$$\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \mathbb{E}[Y(1)|T=1] - \mathbb{E}[$$

(1)

$$= \underbrace{\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]}_{\text{ATE}}$$

(2)

$$+ \underbrace{\mathbb{E}[Y(0)|T=1] -}_{\text{Selection}}$$

(2)

Directed Acyclic Graphs (DAGs)

Tool: Represent causal assumptions using directed graphs.

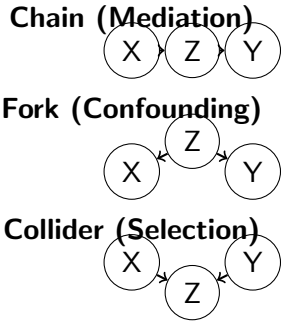
DAG Components:

- **Nodes**: Variables (observed and unobserved)
- **Directed edges**: Direct causal relationships
- **Paths**: Sequences of edges
- **No cycles**: Acyclic assumption

Three Fundamental Structures:

1. **Chain**: $X \to Z \to Y$ (mediation)
2. **Fork**: $X \leftarrow Z \to Y$ (confounding)
3. **Collider**: $X \to Z \leftarrow Y$ (selection bias)

d-separation: Paths are blocked by conditioning on



d-separation Rules

- **Chain/Fork**: Blocked by conditioning on middle node

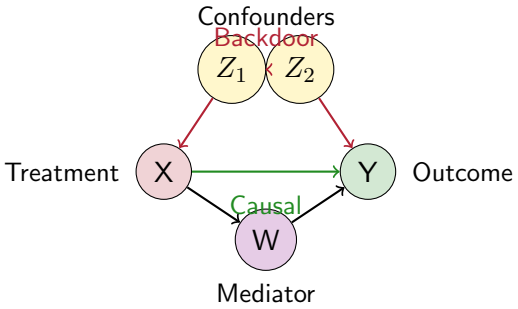
The Backdoor Criterion

Goal: Identify when we can estimate causal effects from observational data.

Definition (Backdoor Criterion)

A set of variables Z satisfies the backdoor criterion relative to (X, Y) if:

1. No node in Z is a descendant of X
2. Z blocks every backdoor path from X to Y



Backdoor path: Any path from X to Y that starts with an arrow into X.

Backdoor Adjustment Formula:

$$P(Y = y|\,\mathrm{do}(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

To estimate causal effect:

- Include: Z_1, Z_2 (block backdoor)

Building and Analyzing DAGs in Python

```python
import networkx as nx
import matplotlib.pyplot as plt
from itertools import combinations
import pandas as pd
import numpy as np

# Create a DAG for education -> income example
def create_education_dag():
    """Create DAG for education's effect on income"""
    G = nx.DiGraph()

    # Add nodes
    nodes = ['Education', 'Income', 'Family_Background',
             'Ability', 'Experience', 'Location']
    G.add_nodes_from(nodes)

    # Add edges (causal relationships)
    edges = [
        ('Family_Background', 'Education'),
        ('Family_Background', 'Income'),
        ('Ability', 'Education'),
        ('Ability', 'Income'),
        ('Education', 'Income'),
        ('Education', 'Experience'),
        ('Experience', 'Income'),
        ('Location', 'Education'),
        ('Location', 'Income')
    ]
    G.add_edges_from(edges)
```

DAG Analysis Steps:

1. **Draw the DAG:**

Randomized Controlled Trials: The Gold Standard

Why Randomization Works: Breaks the link between treatment and confounders.

Randomization Ensures:

$$Y_i(1), Y_i(0) \perp\!\!\!\perp T_i \tag{4}$$

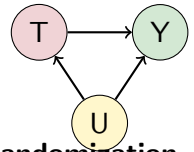
This means:

$$\mathbb{E}[Y(1)|T=1] = \mathbb{E}[Y(1)] \tag{5}$$
$$\mathbb{E}[Y(0)|T=0] = \mathbb{E}[Y(0)] \tag{6}$$

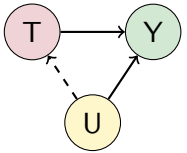
Therefore:

$$\text{ATE} = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \tag{7}$$

Before Randomization



After Randomization



Broken link

When RCTs Are Limited:

- Ethical constraints: Harmful treatments

Instrumental Variables: Finding Natural Experiments

Idea: Use a variable that affects treatment but not outcome directly.

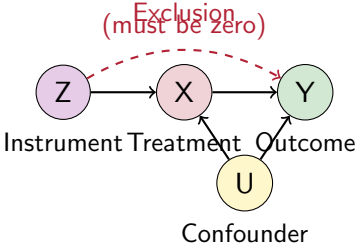
Definition (Instrumental Variable)

Z is an instrument for $X \to Y$ if:

1. **Relevance**: Z affects X (strong first stage)
2. **Exclusion**: Z affects Y only through X
3. **Independence**: Z is as-good-as-randomly assigned

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from scipy import stats
import statsmodels.api as sm
from statsmodels.sandbox.regression.gmm import IV2SLS

# Simulate data with endogeneity
np.random.seed(42)
n = 1000
```



Regression Discontinuity Design

Idea: Exploit arbitrary cutoff rules for treatment assignment.

Setup:

- Running variable: X (e.g., test score, age, income)
- Cutoff: c (treatment threshold)
- Treatment: $T = \mathbb{I}[X \geq c]$
- Outcome: Y

Key Assumption: All other factors vary smoothly around cutoff.

Identification:

$$\tau = \lim_{x \downarrow c} \mathbb{E}[Y|X = x] - \lim_{x \uparrow c} \mathbb{E}[Y|X = x]$$

```
import numpy as np
import matplotlib.pyplot as plt
```

Real-World Examples:

- **Education**: Grade cutoffs for remedial programs
- **Health**: Age thresholds for medical

Difference-in-Differences (DID)

Idea: Compare changes over time between treatment and control groups.

Setup:

- Groups: Treatment ($i = 1$) and Control ($i = 0$)
- Time: Before ($t = 0$) and After ($t = 1$)
- Outcome: Y_{it}

DID Estimator:

$$\hat{\tau}_{DID} = (Y_{11} - Y_{10}) - (Y_{01} - Y_{00}) \quad (8)$$

$$= \Delta Y_{\text{treatment}} - \Delta Y_{\text{control}} \quad (9)$$

Key Assumption: Parallel trends - groups would have followed similar trends without treatment.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
```

Double Machine Learning (DML)

Problem: High-dimensional confounding in observational data.

Traditional Approach Issues:

- Linear models too restrictive
- ML models introduce regularization bias
- Overfitting affects causal estimates
- No honest uncertainty quantification

DML Solution:

1. Use ML for nuisance parameters
2. Cross-fitting to avoid overfitting bias
3. Focus ML on prediction, not causation
4. Get asymptotically normal estimates

DML Algorithm:

Algorithm Double Machine Learning

1: Split data into K folds
2: **for** $k = 1$ to K **do**
3: Train $\hat{g}^{(-k)}(X)$ on all folds except k
4: Train $\hat{m}^{(-k)}(X)$ on all folds except k
5: Predict on fold k:
6: $\tilde{Y}_k = Y_k - \hat{g}^{(-k)}(X_k)$
7: $\tilde{D}_k = D_k - \hat{m}^{(-k)}(X_k)$
8: **end for**
9: Estimate $\hat{\theta} = (\sum \tilde{D}_i \tilde{Y}_i)/(\sum \tilde{D}_i^2)$

Key Properties:

Causal Forests: Heterogeneous Treatment Effects

Goal: Estimate treatment effects that vary across individuals.

```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt

# Simulate heterogeneous treatment effect data
np.random.seed(42)
n = 2000

# Covariates
X1 = np.random.randn(n)
X2 = np.random.randn(n)
X3 = np.random.choice([0, 1], n)

X = np.column_stack([X1, X2, X3])

# Heterogeneous treatment effect
# Effect depends on X1: strong for X1 > 0, weak for X1 <= 0
tau_true = 2 * (X1 > 0) + 0.5 * (X1 <= 0)

# Treatment assignment (confounded)
propensity = 1 / (1 + np.exp(-0.5 * X1 - 0.3 * X2))
T = np.random.binomial(1, propensity, n)

# Outcome
Y0 = 1 + X1 + 0.5 * X2 + 0.3 * X3 + np.random.randn(n) * 0.5
```

Why Heterogeneity Matters:

- **Personalization**: Tailor treatments to individuals

Case Study: Online Marketing Campaign Evaluation

Business Problem: Did our email marketing campaign increase sales?

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
import matplotlib.pyplot as plt
import seaborn as sns

# Simulate marketing campaign data
np.random.seed(42)
n_customers = 5000

# Customer characteristics
age = np.random.normal(40, 15, n_customers)
income = np.random.lognormal(10, 0.5, n_customers)
previous_purchases = np.random.poisson(3, n_customers)
email_engagement = np.random.beta(2, 5, n_customers)

# Selection bias: more engaged customers more likely to
    receive campaign
campaign_prob = 0.3 + 0.4 * email_engagement
received_campaign = np.random.binomial(1, campaign_prob,
    n_customers)

# Outcome: purchase in next month
# Confounded by engagement
base_purchase_prob = (0.1 + 0.2 * email_engagement +
                      0.1 * (previous_purchases > 2) +
```

Case Study: Policy Evaluation with DID

Policy Question: Did minimum wage increase affect employment?

Natural Experiment:

- Some states increased minimum wage in 2019
- Others did not (control group)
- Compare employment changes
- Key assumption: Parallel trends

Data Structure:

- Unit: State
- Time: Monthly, 2017-2021
- Treatment: Minimum wage increase

DID Results:

	Before	After
Treatment States	65.2%	63.8%
Control States	67.1%	66.3%
Difference	-1.9pp	-2.5pp
DID Estimate: -0.6pp (SE: 0.4pp, p=0.12)		

Interpretation:

- Small negative effect on youth employment
- Not statistically significant
- Economically small magnitude

Common Pitfalls and How to Avoid Them

Common Mistakes:

1. **Confusing correlation with causation**
   - Over-interpreting regression coefficients
   - Ignoring selection bias
   - Not considering confounders
2. **Wrong DAG specification**
   - Missing important confounders
   - Including mediators in controls
   - Conditioning on colliders
3. **Weak instruments**
   - First stage F ¡ 10
   - Ignoring LATE interpretation
   - Assuming exclusion restriction
4. **Violated assumptions**
   - Non-parallel trends in DID
   - Manipulation around RDD cutoff

Best Practices:

1. **Start with theory and DAGs**
   - Draw causal assumptions explicitly
   - Use domain knowledge
   - Test implications where possible
2. **Multiple identification strategies**
   - Triangulate with different methods
   - Check robustness across approaches
   - Report sensitivity analyses
3. **Validate assumptions**
   - Pre-trend tests for DID
   - Density tests for RDD
   - Balance tests for matching
   - Placebo and falsification tests
4. **Report honestly**
   - Acknowledge limitations
   - Discuss threats to identification

When Causal Inference Is and Isn't Possible

Good Candidates for Causal Inference:

- **Policy interventions**: Clear before/after
- **Randomized experiments**: Natural or designed
- **Institutional rules**: Arbitrary cutoffs
- **Natural experiments**: Exogenous variation
- **Business experiments**: A/B tests

Requirements:

- Clear treatment definition
- Credible identification strategy
- Testable assumptions

Challenging Cases:

- **Historical questions**: No counterfactual
- **Macro phenomena**: No control group
- **Complex interactions**: Multiple treatments
- **Ethical constraints**: Harmful interventions
- **Long-term effects**: Time horizon issues

Alternative Approaches:

- Structural modeling
- Simulation and calibration
- Cross-country comparisons
- Historical analogies

Key Takeaways

Core Principles:

- **Causation $\neq$ Correlation**: Association doesn't imply causation
- **Design matters**: Identification strategy is crucial
- **Assumptions matter**: Be explicit and test when possible
- **Context matters**: External validity is key

Technical Skills Learned:

- DAG construction and analysis
- Instrumental variables
- Regression discontinuity
- Difference-in-differences

Business Impact:

- Better decisions: Understanding what works
- Resource allocation: Target effective interventions
- Policy evaluation: Evidence-based choices
- Strategic planning: Anticipate intervention effects

Common Applications:

- Marketing campaign effectiveness
- Product feature impact
- Pricing strategy evaluation

Next Steps in Your Causal Journey

Immediate Next Topics:

1. **Experimental Design & A/B Testing**
   - Power analysis and sample sizes
   - Online experimentation platforms
   - Sequential testing methods
2. **Explainable AI & Model Interpretability**
   - SHAP and LIME
   - Global vs local explanations
   - Causal vs predictive explanations
3. **Time Series Analysis**

Advanced Topics to Explore:

- Mediation analysis
- Interference and spillovers
- Machine learning + causality
- Structural equation modeling
- Bayesian causal inference

Practice Projects:

- Analyze marketing campaign data
- Evaluate policy using public data
- Design A/B test for product feature
- Build causal ML pipeline

Essential Reading:

Thank You

Questions & Discussion

Diogo Ribeiro
ESMAD – Escola Superior de Média Arte e Design
Lead Data Scientist, Mysense.ai

dfr@esmad.ipp.pt
https://orcid.org/0009-0001-2022-7072

Slides and code available at:
github.com/diogoribeiro7/academic-presentations

Next: Experimental Design & A/B Testing